MATH TOLLBOX

DIEGO HERRERA

Contents

Date: September 2, 2013.
Key words and phrases. math notation.

1. Basic Tools

1.1. **Iverson Bracket.**

1.1.1. *Definition.* The *Iverson bracket*, named after Kenneth E. Iverson, is a notation that denotes a number that is 1 if the condition in square brackets is satisfied, and 0 otherwise. More exactly,

$$[P]_{\text{iv}} = \begin{cases} 1 & \text{if } P \text{ is true;} \\ 0 & \text{otherwise.} \end{cases}$$

where P is a logical statement; *i.e.*, a statement that can be either true or false.

1.1.2. *Examples.* The notation allows moving boundary conditions of summations (or integrals) as a separate factor into the summand, freeing up space around the summation operator, but more importantly allowing it to be manipulated algebraically. For example

$$\sum_{1 \le i \le 10} i^2 = \sum_i i^2 \, [1 \le i \le 10]_{\text{iv}} \, .$$

In the first sum, the index i is limited to be in the range 1 to 10. The second sum is allowed to range *over all* integers, but where i is strictly less than 1 or strictly greater than 10, the summand is 0, contributing nothing to the sum. Such use of the Iverson bracket can permit easier manipulation of these expressions.

Another use of the Iverson bracket is to simplify equations with special cases. For example, the formula

$$\sum_{\substack{1\le k\le n\\ \gcd(k,n)=1}} k = \frac{1}{2} n\varphi[n] \ ,$$

which is valid for $n > 1$ but which is off by $1/2$ for $n = 1$. To get an identity valid for all positive integers n (*i.e.*, all values for which is defined), a correction term involving the Iverson bracket may be added:

$$\sum_{\substack{1\le k\le n\\ \gcd(k,n)=1}} k = \frac{1}{2} n \left(\varphi[n] + [n=1]_{\text{iv}}\right) \ ,$$

The Kronecker delta notation is a specific case of Iverson notation when the condition is equality. That is,

$$\delta_{ij} = [i=j]_{\text{iv}} \ .$$

The trichotomy of the reals can be expressed:

$$[a<b]_{\text{iv}} + [a=b]_{\text{iv}} + [a>b]_{\text{iv}} = 1 \, .$$

1.2. **Tau Number.** Define a *circle* as the set of points a fixed distance from a given point. Refer to the fixed distance as the *radius of the circle*, denoted r, and to the given point as the *center of the circle*, denoted $\mathcal{O}$.

Then, define the *diameter of the circle*, denoted d, by $d = 2r$ and define the *circumference of the circle*, denoted c, as the linear distance around the outside of the circle.

Next, define the *circle constant*, denoted π, as the ratio of the circumference of the circle to its diameter:

$$\pi \coloneqq \frac{c}{d} \, .$$

The numerical value [1] of π is

$$\pi = 3.141\,592\,653\,589\,793\,238\,462\,643\,3\ldots \, .$$

Finally, define the tau number:
Definition. define the tau number*, denoted τ, as the ratio of the circumference of the circle to its radius:*

$$\tau \coloneqq \frac{c}{r} \, .$$

The numerical value of τ is

$$\tau = 6.283\,185\,307\,179\,586\,476\,925\,286\,6\ldots \, .$$

Note. The transformations between π and τ are given by $\tau = 2\pi$ and $\pi = \tau/2$.

1.3. **Approximation to Earth's gravity.** Gravitational acceleration can be approximated by

$$g \sim \pi^2 = 9.86960\ldots \, .$$

This value is a useful approximation when working with circular or periodic motion.

1.4. **Stirling's Approximation.** *Stirling's approximation* (or Stirling's formula) is an approximation for large factorials.

[1] Mnemonic: "How I want a drink, alcoholic of course, after the heavy lectures involving quantum mechanics." — James Jeans.

1.4.1. *Formula.* The formula as typically used in applications is

$$\ln[n!] = n\ln[n] - n + O[\ln[n]] \ ,$$

a more precise variant of the formula is often written

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \ .$$

Sometimes, bounds for $n!$ rather than asymptotics are required: one has, for all $x \in \mathcal{N}^+$

$$\sqrt{2\pi}\, n^{n+1/2}\, e^{-n} \leq n! \leq e\, n^{n+1/2}\, e^{-n} \ ,$$

so for all $n \geq 1$ the ratio $n!/(n^{n+1/2}\, e^{-n})$ is always, *e.g.*, between 2.5 and 2.8.

1.5. **Transformation of Graphs – Shifts, Scales and Reflections.** Shifting, Scaling and Reflecting graphs can be represented algebraically via the following transformations:

- $g[x] = f[x + a]$: The g-graph is determined by a *horizontal shift* of the f-graph $|a|$ units to the *left* if $a > 0$, or $|a|$ units to the *right* if $a < 0$.
- $h[x] = f[x] + a$: The h-graph is determined by a *vertical shift* of the f-graph $|a|$ units *up* if $a > 0$, or $|a|$ units *down* if $a < 0$.
- $k[x] = f[ax]$: The k-graph is determined by a *horizontal compression* of the f-graph if $a > 1$, or *horizontal stretch* of the f-graph if $0 < a < 1$.
- $j[x] = af[x]$: The j-graph is determined by a *vertical stretch* of the f-graph if $a > 1$, or *vertical compression* of the f-graph if $0 < a < 1$.
- $r[x] = f[-x]$: The r-graph is determined by *reflecting* the f-graph across the y-axis.
- $s[x] = -f[x]$: The s-graph is determined by *reflecting* the f-graph across the x-axis.

Remark. If $f[-x] = f[x]$ for all x in the domain of f, then f is said to be *even* and its graph is *symmetric with respect to the y-axis.* If $g[-x] = -g[x]$ for all x in the domain of g, then g is said to be *odd* and its graph is *symmetric with respect to the origin.*

1.6. **Cartesian Product.** A *Cartesian product* is a mathematical operation which returns a set (or product set) from multiple sets. The Cartesian product is the result of crossing members of each set with one another.

The simplest case of a Cartesian product is the *Cartesian square*, which returns a set from two sets. A table can be created by taking the Cartesian product of a set of rows and a set of columns. If the Cartesian product rows times columns is taken, the cells of the table contain ordered pairs of the form (row value, column value). If the Cartesian product is columns times rows is taken, the cells of the table contain the ordered pairs of the form (column value, row value).

A Cartesian product of n sets can be represented by an array of n dimensions, where each element is an n-tuple. An ordered pair is a 2-tuple.

1.6.1. *Cartesian square and Cartesian power.* The Cartesian square (or binary Cartesian product) of a set $\mathcal{X}$ is the Cartesian product $\mathcal{X}^2 = \mathcal{X} \otimes \mathcal{X}$. An example is the 2-dimensional plane $\mathcal{R}^n = \mathcal{R} \otimes \mathcal{R}$, where $\mathcal{R}$ is the set of real numbers – all points $[x, y]$, where x and y are real numbers.

1.6.2. *Higher powers of a set.* The *Cartesian Power* of a set $\mathcal{X}$ can be defined as:

$$\mathcal{X}^n = \mathcal{X} \otimes \mathcal{X} \cdots \mathcal{X} = \{[x_1, \ldots, x_n] : x_i \in \mathcal{X}, \text{for all } 1 \leq i \leq n\} \ ,$$

i.e., $\mathcal{X}^n$ is the collection of all n-tuples $[x_1, \ldots, x_n]$. Call the $\{x_i\}$ the *components of the tuple.*

An example of this is $\mathcal{R}^3 = \mathcal{R} \otimes \mathcal{R} \otimes \mathcal{R}$, with $\mathcal{R}$ again the set of real numbers and more generally $\mathcal{R}^n$.

1.7. **Function.** In mathematics, a *function* is a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output. An example is the function that relates each real number x to its square x^2. The output of a function f corresponding to an input x is denoted by $f[x]$ (read "f of x" or "f value at the point x). In this example, if the input is -3, then the output is 9, and we may write $f[-3] = 9$. The input variable(s) are sometimes referred to as the *argument(s) of the function*.

Functions are "the central objects of investigation" in most fields of modern mathematics. There are many ways to describe or represent a function. Some functions may be defined by a formula or algorithm that tells how to compute the output for a given input. Others are given by a picture, called the graph of the function. In science, functions are sometimes defined by a table that gives the outputs for selected inputs. A function can be described through its relationship with other functions, for example as an inverse function or as a solution of a differential equation.

The input and output of a function can be expressed as an *ordered pair*, ordered so that the first element is the input (or tuple of inputs, if the function takes more than one input), and the second is the output. In the example above, $f[x] = x^2$, we have the ordered pair $[-3, 9]$. If both input and output are real numbers, this ordered pair can be viewed as the Cartesian coordinates of a point on the graph of the function. But no picture can exactly define every point in an infinite set. In modern mathematics, a function is defined by its set of inputs, called the *domain*, a set containing the outputs, called its *codomain*, and the set of all paired input and outputs, called the *graph*. For example, we could define a function using the rule $f[x] = x^2$ by saying that the domain and codomain are the real numbers, and that the ordered pairs are all pairs of real numbers $\left[x, x^2\right]$. Collections of functions with the same domain and the same codomain are called *function spaces*, the properties of which are studied in such mathematical disciplines as *real analysis* and complex analysis.

In analogy with arithmetic, it is possible to define addition, subtraction, multiplication, and division of functions, in those cases where the output is a number. Another important operation defined on functions is *function composition*, where the output from one function becomes the input to another function.

1.7.1. *Definition.* In order to avoid the use of the not rigorously defined words "rule" and "associates", the above intuitive explanation of functions is completed with a formal definition. This definition relies on the notion of the Cartesian product. The Cartesian product of two sets $\mathcal{X}$ and $\mathcal{Y}$ is the set of all ordered pairs, written $[x, y]$, where x is an element of $\mathcal{X}$ and y is an element of $\mathcal{Y}$. The x and the y are called the *components of the ordered pair*. The Cartesian product of $\mathcal{X}$ and $\mathcal{Y}$ is denoted by $\mathcal{X} \otimes \mathcal{Y}$.

1.7.2. *Definition.*

> A function f from $\mathcal{X}$ to $\mathcal{Y}$ is a subset of the Cartesian product $\mathcal{X} \otimes \mathcal{Y}$ subject to the following condition: every element of $\mathcal{X}$ is the first component of one and only one ordered pair in the subset. In other words, for every x in $\mathcal{X}$ there is exactly one element y such that the ordered pair $[x, y]$ is contained in the subset defining the function f.

This formal definition is a precise rendition of the idea that to each x is associated an element y of $\mathcal{Y}$, namely the uniquely specified element y with the property just mentioned.

1.7.3. *Notation.* A function f with domain $\mathcal{X}$ and codomain $\mathcal{Y}$ is commonly denoted by $f : \mathcal{X} \to \mathcal{Y}$. In this context, the elements of $\mathcal{X}$ are called *arguments of* f. For each argument x, the corresponding unique y in the codomain is called the *function value at* x or the *image of* x *under* f. It is written as $f[x]$. One says that f associates y with x or maps x to y. This is abbreviated by $y = f[x]$.

In order to specify a concrete function, the notation $\mapsto$ (an arrow with a bar at its tail) is used. For example,

$$f : \mathcal{N} \to \mathcal{Z}, x \mapsto 4 - x\,.$$

The first part is read "f is a function from $\mathcal{N}$ (the set of natural numbers) to $\mathcal{Z}$ (the set of integers)" or "f is an $\mathcal{Z}$-valued function of an $\mathcal{N}$-valued variable".

The second part is read "x maps to $4 - x$". In other words, this function has the natural numbers as domain, the integers as codomain. A function is properly defined only when the domain and codomain are specified. For example, the formula $f[x] = 4 - x$ alone (without specifying the codomain and domain) is not a properly defined function. Moreover, the function $g : \mathcal{Z} \to \mathcal{Z}$, such that $x \mapsto 4 - x$ (with different domain) is *not* considered the same function, even though the formulas defining f and g agree, and similarly with a different codomain. Despite that, many authors drop the specification of the domain and codomain, especially if these are clear from the context. So in this example many just write $f[x] = 4 - x$. Sometimes, the maximal possible domain is also understood implicitly: a formula such as $f[x] = \sqrt{x^2 - 5x + 6}$ may mean that the domain of f is the set of real numbers x where the square root is defined (in this case $x \leq 2$ or $x \geq 3$).

1.7.4. *Specifying a Function.* A function can be defined by any mathematical condition relating each argument (input value) to the corresponding output value. If the domain is finite, a function f may be defined by simply tabulating all the arguments x and their corresponding function values $f[x]$. More commonly, a function is defined by a formula, or (more generally) an algorithm – a recipe that tells how to compute the value of $f[x]$ given any x in the domain.

There are many other ways of defining functions. Examples include piecewise definitions, induction or recursion, algebraic or analytic closure, limits, analytic continuation, infinite series, and as solutions to integral and differential equations. The lambda calculus provides a powerful and flexible syntax for defining and combining functions of several variables. In advanced mathematics, some functions exist because of an axiom, such as the Axiom of Choice.

1.7.5. *Basic Properties.* Image and preimage: If $\mathcal{A}$ is any subset of the domain $\mathcal{X}$, then $f[\mathcal{A}]$ is the subset of the codomain $\mathcal{Y}$ consisting of all images of elements of $\mathcal{A}$. We say the $f[\mathcal{A}]$ is the image of $\mathcal{A}$ under f. The *image* of f is given by $f[\mathcal{X}]$. On the other hand, the *inverse* image (or preimage, complete inverse image) of a subset $\mathcal{B}$ of the codomain $\mathcal{Y}$ under a function f is the subset of the domain $\mathcal{X}$ defined by

$$f^{-1}[\mathcal{B}] = \{x \in \mathcal{X} : f[x] \in \mathcal{B}\} \ .$$

So, for example, the preimage of $\{4, 9\}$ under the squaring function is the set $\{-3, -2, 2, 3\}$. The term *range* usually refers to the image, but sometimes it refers to the codomain.

By definition of a function, the image of an element x of the domain is always a single element y of the codomain. Conversely, though, the preimage of a singleton set (a set with exactly one element) may in general contain any number of elements. For example, if $f[x] = 7$ (the constant function taking value 7), then the preimage of $\{5\}$ is the empty set but the preimage of $\{7\}$ is the entire domain. It is customary to write $f^{-1}[b]$ instead of $f^{-1}[\{b\}]$, *i.e.*,

$$f^{-1}[b] = \{x \in \mathcal{X} : f[x] = b\} \ .$$

This set is sometimes called the *fiber* of b under f.

Use of $f[\mathcal{A}]$ to denote the image of a subset $\mathcal{A} \subset \mathcal{X}$ is consistent so long as no subset of the domain is also an element of the domain.

Injective and surjective functions: A function is called *injective* (or one-to-one or an injection)

> if $f[a] \neq f[b]$ for any two different elements a and b of the domain.

It is called *surjective* (or onto) if $f[\mathcal{X}] = \mathcal{Y}$. That is, it is surjective

> if for every element y in the codomain there is an x in the domain such that $f[x] = y$.

Finally f is called *bijective* if it is *both* injective and surjective.

Function composition: The *function composition* of two functions takes the output of one function as the input of a second one. More specifically, the composition of f with a function $g : \mathcal{Y} \to \mathcal{Z}$ is the function $g \circ f : \mathcal{X} \to \mathcal{Z}$ defined by

$$(g \circ f)[x] = g[f[x]] \ .$$

That is, the value of x is obtained by first applying f to x to obtain $y = f[x]$ and then applying g to y to obtain $z = g[y]$. In the notation $g \circ f$, the function on the right, f,

acts first and the function on the left, g acts second, reversing English reading order. The notation can be memorized by reading the notation as "g of f" or "g after f". The composition $g \circ f$ is only defined when the codomain of f is the domain of g. Assuming that, the composition in the opposite order $f \circ g$ need not be defined. Even if it is, *i.e.*, if the codomain of f is the codomain of g, it is not in general true that $g \circ f = f \circ g$. That is, *the order of the composition is important.* For example, suppose $f[x] = x^2$ and $g[x] = x + 1$. Then, $g[f[x]] = x^2 + 1$, while $f[g[x]] = (x + 1)^2$, which is $x^2 + 2x + 1$, a different function.

Identity function: The unique function over a set $\mathcal{X}$ that maps each element to itself is called the *identity function for* $\mathcal{X}$, and typically denoted by $\mathrm{id}_\mathcal{X}$. Each set has its own identity function, so the subscript cannot be omitted unless the set can be inferred from context. Under composition, an identity function is "neutral": if f is any function from $\mathcal{X}$ to $\mathcal{Y}$, then

$$f \circ \mathrm{id}_\mathcal{X} = f\,, \qquad \mathrm{id}_\mathcal{Y} \circ f = f\,.$$

Restrictions and extensions: Informally, a *restriction* of a function f is the result of trimming its domain. More precisely, if $\mathcal{S}$ is any subset of $\mathcal{X}$, the restriction of f to $\mathcal{S}$ is the function $f|_\mathcal{S}$ from $\mathcal{S}$ to $\mathcal{Y}$ such that $f|_\mathcal{S}[s] = f[s]$ for all s in $\mathcal{S}$. If g is a restriction of f, then it is said that f is an *extension* of g.

The *overriding* of $f : \mathcal{X} \to \mathcal{Y}$ by $g : \mathcal{W} \to \mathcal{Y}$ (also called overriding union) is an extension of g denoted as $(f \oplus g) : \mathcal{X} \cup \mathcal{W} \to \mathcal{Y}$. Its graph is the set-theoretical union of the graphs of g and $f|_{\mathcal{X}\setminus\mathcal{W}}$. Thus, it relates any element of the domain of g to its image under g, and any other element of the domain of f to its image under f. Overriding is an associative operation; it has the empty function as an identity element. If $f|_{\mathcal{X}\cap\mathcal{W}}$ and $g|_{\mathcal{X}\cap\mathcal{W}}$ are pointwise equal (*e.g.*, the domains of f and g are disjoint), then the union of f and g is defined and is equal to their overriding union. This definition agrees with the definition of union for binary relations.

Inverse function: An inverse function for f, denoted by f^{-1}, is a function in the opposite direction, from $\mathcal{Y}$ to $\mathcal{X}$, satisfying

$$f \circ f^{-1} = \mathrm{id}_\mathcal{Y}\,, \qquad f^{-1} \circ f = \mathrm{id}_\mathcal{X}\,.$$

That is, the two possible compositions of f and f^{-1} need to be the respective identity maps of $\mathcal{X}$ and $\mathcal{Y}$.

As a simple example, if f converts a temperature in degrees Celsius C to degrees Fahrenheit F, then the function converting degrees Fahrenheit to degrees Celsius would be a suitable f^{-1}:

$$f[C] = \frac{9}{5}C + 32\,, \qquad f^{-1}[F] = \frac{5}{9}(F - 32)\,.$$

Such an inverse function exists if and only if f is *bijective*. In this case, f is called *invertible*.

1.7.6. *Types of Functions.* Real-valued functions: A *real-valued function* f is one whose codomain is the set of real numbers or a subset thereof. If, in addition, the domain is also a subset of the reals, f is a real valued function of a real variable. The study of such functions is called *real analysis*.

Real-valued functions enjoy so-called *pointwise operations*. That is, given two functions $f, g : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a subset of the reals (and $\mathcal{X}$ is an arbitrary set), their (pointwise) sum $f + g$ and product $f \odot g$ are functions with the same domain and codomain. They are defined by the formulas:

$$(f + g)[x] = f[x] + g[x] \qquad \text{and} \qquad (fg)[x] = f[x]\, g[x]\ .$$

Partial and multi-valued functions: In some parts of mathematics, including recursion theory and functional analysis, it is convenient to study partial functions in which some values of the domain have no association in the graph; *i.e.*, single-valued relations. For example, the function f such that $f[x] = 1/x$ does not define a value for $x = 0$, since division by zero is not defined. Hence f is only a *partial function* from the real line to the real line. The *term total* function can be used to stress the fact that every element of the domain does appear as the first element of an ordered pair in the graph. In other parts of mathematics, non-single-valued relations are similarly conflated with functions: these are called *multivalued functions*, with the corresponding term single-valued function for

ordinary functions. For instance, $f[x] = \pm\sqrt{x}$ is not a function in the proper sense, but a multi-valued function: it assigns to each positive real number x two values: the (positive) square root of x, and $-\sqrt{x}$.

Functions with multiple inputs and outputs: The concept of function can be extended to an object that takes a combination of two (or more) argument values to a single result. This intuitive concept is formalized by a function whose domain is the Cartesian product of two or more sets.

For example, consider the function that associates two integers to their product: $f[x, y] = xy$. This function can be defined formally as having domain $\mathcal{Z} \otimes \mathcal{Z}$, the set of all integer pairs; codomain $\mathcal{Z}$; and, for graph, the set of all pairs $[[x, y], xy]$. Note that the first component of any such pair is itself a pair (of integers), while the second component is a single integer.

The function value of the pair $[x, y]$ is $f[[x, y]]$. However, it is customary to drop one set of parentheses and consider $f[x, y]$ a function of two variables, x and y. Functions of two variables may be plotted on the three-dimensional Cartesian as ordered triples of the form $[x, y, f[x, y]]$.

The concept can still further be extended by considering a function that also produces output that is expressed as several variables. For example, consider the integer divide function, with domain $\mathcal{Z} \otimes \mathcal{N}$ and codomain $\mathcal{Z} \otimes \mathcal{N}$. The resultant (quotient, remainder) pair is a single value in the codomain seen as a Cartesian product.

1.7.7. *Functional Equation.* In mathematics, and particularly in functional analysis, a *functional* is a map from a vector space into its underlying scalar field. In other words, it is a function that takes a vector as its input argument, and returns a scalar. Commonly the vector space is a space of functions, thus

> the functional takes a function for its input argument, then it is sometimes considered a function of a function.

Its use originates in the calculus of variations where one searches for a function that minimizes a certain functional. A particularly important application in physics is searching for a state of a system that minimizes the energy functional.

Functional equation: The traditional usage also applies when one talks about a functional equation, meaning an equation between functionals: an equation between functionals can be read as an "equation to solve", with solutions being themselves functions. In such equations there may be several sets of variable unknowns, like when it is said that an additive function f is one *satisfying the functional equation*

$$f[x + y] = f[x] + f[y] \ .$$

1.8. **Periodicity.**

1.8.1. *Definition.* In mathematics, a *periodic function* is a function that repeats its values in regular intervals or periods. The most important examples are the *trigonometric functions*, which repeat over intervals of length τ rad. Periodic functions are used throughout science to describe oscillations, waves and other phenomena that exhibit periodicity. Any function which is not periodic is called a*periodic*.
Definition. Consider a function $f[x]$ and a nonzero, constant real number p. Then, refer to f as a periodic function with period p if f satisfies

$$f[x + p] = f[x] \ ,$$

for all values of x in the domain of f. Call a function that is not periodic a*periodic.*

If there exists a least positive constant p with this property, it is called the *prime period*. A function with period p will repeat on intervals of length p and these intervals are referred to as *periods*.

Geometrically, a periodic function can be defined as a function whose graph exhibits *translational symmetry*. Specifically, a function f is periodic with period p if the graph of f is invariant under translation in the x-direction by a distance of p.

1.8.2. *Properties.* If a function f is periodic with period p, then for all x in the domain of f and all integers n, we have

$$f[x + np] = f[x] \, .$$

If $f[x]$ is a function with period p, then $f[ax + b]$, where a is a positive constant, is periodic with period p/a. For example, $f[x] = \sin[x]$ has period τ, therefore $\sin[5x]$ will have period $\tau/5$.

1.9. **Parity.**

1.9.1. *Definition. Even functions* and *odd functions* are functions which satisfy particular symmetry relations, with respect to taking additive inverses. They are important in many areas of mathematical analysis, especially the theory of power series and Fourier series. They are named for the parity of the powers of the power functions which satisfy each condition: the function $f[x] = x^n$ is an even function if n is an even integer and it is an odd function if n is an odd integer.
Definition. Consider $f[x]$ to be a real-valued function of a real variable. Then, refer to f as even *if it satisfies*

$$f[x] = f[-x] \, ,$$

for all x in the domain of f.

Geometrically speaking, the graph face of an even function is *symmetric with respect to the y-axis*, meaning that its graph remains *unchanged after reflection about the y-axis.*
Definition. Let $g[x]$ be a real-valued function of a real variable. Then, refer to g as odd if it satisfies

$$-g[x] = g[-x] \, ,$$

for all x in the domain of g.

Equivalently, an odd function g satisfies

$$g[x] + g[-x] = 0 \, .$$

Geometrically, the graph of an odd function has *rotational symmetry with respect to the origin*, meaning that its graph remains *unchanged after rotation of τ* rad *about the origin.*

1.9.2. *Properties.* A function's being odd or even does *not* imply differentiability, or even continuity. For example, the Dirichlet function is even, but is nowhere continuous. Properties involving Fourier series, Taylor series, derivatives and so on may only be used when they can be assumed to exist.

- The only function whose domain is all real numbers which is both even and odd is the constant function which is identically zero; *i.e.*, $f[x] = 0$, for all x.
- The sum of two even functions is even, and any constant multiple of an even function is even.
- The sum of two odd functions is odd, and any constant multiple of an odd function is odd.
- The difference between two odd functions is odd.
- The difference between two even functions is even.
- The product of two even functions is an even function.
- The product of two odd functions is an even function.
- The product of an even function and an odd function is an odd function.
- The quotient of two even functions is an even function.
- The quotient of two odd functions is an even function.
- The quotient of an even function and an odd function is an odd function.
- The derivative of an even function is odd.
- The derivative of an odd function is even.
- The composition of two even functions is even, and the composition of two odd functions is odd.
- The composition of an even function and an odd function is even.
- The composition of any function with an even function is even, but not *vice versa*.

- The integral of an odd function from $-A$ to $+A$ is zero, where A is finite and the function has no vertical asymptotes between $-A$ and A.
- The integral of an even function from $-A$ to $+A$ is twice the integral from 0 to $+A$, where A is finite and the function has no vertical asymptotes between $-A$ and A. This also holds true when A is infinite, but only if the integral converges.
- Every function can be expressed as the sum of an even and an odd function.
- The sum of an even and odd function is neither even nor odd, unless one of the functions is equal to zero over the given domain.
- The Maclaurin series of an even function includes only even powers.
- The Maclaurin series of an odd function includes only odd powers.
- The Fourier series of a periodic even function includes only cosine terms.
- The Fourier series of a periodic odd function includes only sine terms.

1.9.3. *The Sum of Odd and Even Functions.* Two theorems: Every function can be expressed as the sum of an even and an odd function. The sum of an even and odd function is neither even nor odd, unless one of the functions is equal to zero over the given domain.

Uniquely write every function $f[x]$ as the *sum of an even function and an odd function*:

$$f[x] = f_{\text{even}}[x] + f_{\text{odd}}[x] \ ,$$

where

$$f_{\text{even}}[x] = \tfrac{1}{2}\left(f[x] + f[-x]\right)$$

and

$$f_{\text{odd}}[x] = \tfrac{1}{2}\left(f[x] - f[-x]\right) \ . \qquad \square$$

For example, if $f[x]$ is $\exp[x]$, then $f_{\text{even}}[x] = \exp[x] = \cosh[x]$ and $f_{\text{odd}}[x] = \exp[x] = \sinh[x]$.

1.10. **Linear Map.** In mathematics, a *linear map*, *aka* linear mapping, linear transformation, or linear operator (in some contexts also called linear function), is a function between two modules (including vector spaces) that preserves the operations of module (or vector) addition and scalar multiplication.

As a result, it always maps linear subspaces in linear subspaces, like straight lines to straight lines or to a single point. The expression "linear operator" is commonly used for linear maps from a vector space to itself (*i.e.*, endomorphisms). Sometimes the definition of a linear function coincides with that of a linear map, while in analytic geometry it does not.

In the language of abstract algebra, a linear map is a homomorphism of modules. In the language of category theory it is a morphism in the category of modules over a given ring.

1.10.1. *Definition.* Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over the same field $\mathcal{K}$. A function $f : \mathcal{V} \to \mathcal{W}$ is said to be a *linear map* if for any two vectors x and y in $\mathcal{V}$ and any scalar α in $\mathcal{K}$, the following two conditions are satisfied:

(1) additivity: $f[x + y] = f[x] + f[y]$;
(2) homogeneity of degree 1: $f[\alpha x] = \alpha f[x]$.

This is equivalent to requiring the same for any linear combination of vectors, *i.e.*, that for any vectors $x1, \ldots, x_m \in \mathcal{V}$ and scalars $a_1, \ldots, a_m \in \mathcal{K}$, the following equality holds:

$$f[a_1 x_1 + \cdots + a_m x_m] = a_1 f[x_1] + \cdots + a_m f[x_m] \ .$$

A linear map from $\mathcal{V}$ to $\mathcal{K}$ (with $\mathcal{K}$ viewed as a vector space over itself) is called a *linear functional*.

In linear algebra, a *linear functional* or *linear form* (also called a *one-form* or *covector*) is a linear map from a vector space to its field of scalars. In $\mathcal{R}^n$, if vectors are represented as column vectors, then linear functionals are represented as row vectors, and their action on vectors is given by the dot product, or the matrix product with the row vector on the left and the column vector on the right.

1.10.2. *Example.* Suppose that vectors in the real coordinate space $\mathcal{R}^n$ are represented as column vectors

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} .$$

Then any linear functional can be written in these coordinates as a sum of the form:

$$f[x] = a_1 x_1 + \cdots + a_n x_n .$$

This is just the matrix product of the row vector $[a_1 \ \ldots \ a_n]$ and the column vector x:

$$f[x] = \begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} .$$

1.11. **Curve Sketching.**

1.11.1. *Increasing and Decreasing Functions.* Let f be a function defined on an interval and let x_1 and x_2 be points in that interval.

- f is *increasing* on the interval if $f[x_1] < f[x_2]$ whenever $x_1 < x_2$ for all points x_1 and x_2.
- f is *decreasing* on the interval if $f[x_1] > f[x_2]$ whenever $x_1 < x_2$ for all points x_1 and x_2.
- f is *constant* on the interval if $f[x_1] = f[x_2]$ for all points x_1 and x_2.

If tangents were drawn to the graph above you would notice that when f, the function, is *increasing* its tangent has a *positive slope* and when f is *decreasing* its tangent has a *negative slope*. When f is *constant* its tangent has *zero slope*. From this it is possible to arrive at the following result.

Note. Let f be a function that is continuous on an interval $[a, b]$ and differentiable on the open interval $]a, b[$.

- If $f'[x] > 0$ for every value of x in $]a, b[$, then f is increasing on $]a, b[$.
- If $f'[x] < 0$ for every value of x in $]a, b[$, then f is decreasing on $]a, b[$.
- If $f'[x] = 0$ for every value of x in $]a, b[$, then f is constant on $]a, b[$.

1.11.2. *Concavity.* Although the sign of the first derivative of f reveals where the graph of f is increasing or decreasing, it does *not* reveal the direction of curvature. The direction of curvature can be either *concave up* (upward curvature) or *concave down* (downward curvature). The following are two suggested ways to characterize the concavity of a differentiable function f on an open interval:

- f is *concave up* on an open interval if its *tangent lines* have *increasing* slopes on that interval and is *concave down* if they have *decreasing* slopes.
- f is concave up on an open interval if its graph lies *above* its tangent line on that interval and is concave down if its graph lies *below* its tangent lines.

Since the slope of the tangent lines to the graph of a differential function f are the values of its derivative f', the above requirements are the same as saying that f' will be increasing on intervals where f'' is positive and f' will be decreasing on intervals where f'' is negative.

Note. Let f be twice differentiable on an open interval $]a, b[$,

- If $f''[x] > 0$ for every value x in $]a, b[$, then f is concave up on $]a, b[$,
- If $f''[x] < 0$ for every value x in $]a, b[$, then f is concave down on $]a, b[$.

1.11.3. *Inflection Points.* Points where a curve changes from concave up to concave down or *vice versa* are of special interest. These points are called *points of inflection* and the following is a more formal definition of what they are.

Definition: If f is continuous on an open interval containing a value x and if f changes the direction of its concavity at the point $[x, f[x]]$, then we say that f has an inflection point at x.

Note. To find the points of inflection of a function f, simply solve $f'' = 0$.

1.11.4. *Relative Maxima and Minima.* Imagine the graph of a function f to be a two-dimensional mountain range with hills and valleys, then the tops of the hills are called *relative maxima* and the bottoms of the valleys are called *relative minima*. A relative maximum need not be the highest point in the entire mountain range, and a relative minimum need not be the lowest – they are just high and low points relative to the nearby terrain.

The relative maxima or minima for all functions occur at points where the graphs of the functions have horizontal tangent lines (slopes equal to zero). A *critical point* of a function f can be defined as a point in the domain of f at which the graph of f has a horizontal tangent line.

Note. To find the critical points of a function f, simply solve $f' = 0$.

1.11.5. *First Derivative Test.* A function f has a relative maximum or minimum at those critical points where f' changes sign.

Suppose that f is continuous at the critical point x_0.

- If $f'[x] > 0$ on an open interval extending left from x_0 and $f'[x] < 0$ on an open interval extending right from x_0, then f has a *relative maximum* at x_0.
- If $f'[x] < 0$ on an open interval extending left from x_0 and $f'[x] > 0$ on an open interval extending right from x_0, then f has a *relative minimum* at x_0.
- If $f'[x]$ has the same sign on an open interval extending left from x_0 as it does on an open interval extending right from x_0, then f does *not* have a relative maximum or minimum at x_0.

1.11.6. *Second Derivative Test.* This is another way (and perhaps an easier way) of classifying critical points that relies on the second derivative of the function f.

Suppose that f is twice differentiable at x_0.

- If $f'[x_0] = 0$ and $f''[x_0] > 0$, then f has a relative minimum at x_0.
- If $f'[x_0] = 0$ and $f''[x_0] < 0$ then f has a relative maximum at x_0.
- If $f'[x_0] = 0$ and $f''[x_0]$, then the test is inconclusive.

1.11.7. *Guidelines.* Remember you don't always need all these steps – calculate as much as you need to get an idea of the shape. If a step is very difficult or laborious, leave it out.

- Domain: use sign charts for polynomials, rational functions, inequalities and so on.
- Roots: find the roots of the function, using num. analysis if necessary.
- Intercepts: x- and y-intercepts, using num. analysis if necessary.
- Symmetry: even ($f[-x] = f[x]$) or odd ($f[-x] = -f[x]$) function or neither, periodic function ($f[x+p] = f[x]$ with period p and $f[x+np] = f[x]$ for all integers n).
- Asymptotes: horizontal ($\lim_{x\to\pm\infty} f[x] = L$) and vertical ($\lim_{x\to a^{\pm}} f[x] = \pm\infty$)
- Intervals of Increase or Decrease: Use the I/D Test (first derivative).
- Local Maximum and Minimum Values: Use critical numbers.
- Concavity and Points of Inflection: Compute $f''[x]$ and use the Concavity Test.
- Alternative representation: change the coordinates of the function to see if the tests are easier to perform using polar, cylindrical, *etc.* coordinates instead of Cartesians.
- Sketch the Curve: Using the information in previous items, draw the graph.

1.11.8. *Domain.* Consider the function $p[x] = (x+4)(x+2)^2(x-2)(x-4)^2$. Find its zeroes (using num. analysis if necessary). Note that $p[x]$ is already in *factored form*, so the zeros of a polynomial in factored form can be read off without trouble. We have $\{-4, -2, 2, 4\}$. The multiplicities of -2 and 4 are two. Thus we have four branch points as shown on the chart below.

Note that the four branch points divide the number line into five test intervals, $]-\infty, -4[$, $]-4, -2[$, $]-2, 2[$, $]2, 4[$, $]4, \infty$. Select a test point from each interval. Let's take $\{-5, -3, 0, 3, 5\}$.

To determine the sign of the function at each test point, build a matrix with test points listed down the side and factors listed along the top. In the current case...

The power of the technique shows up here. It does not matter which point in the interval is selected as the test point. The sign of the function does not change over a test interval. You can see from the sign chart that $p[x]$ changes sign at -4 from positive to negative and at 2 from negative to positive. If the problem we are given is to solve the inequality $p[x] \geq 0$, we could do this easily at this stage. The solution to $p[x] > 0$ is just $]-\infty, -4[\cup]2, 4[\cup]4, \infty[$. There are four zeros of p to add to this set, so we get $]-\infty, -4] \cup \{2\} \cup [2, \infty[$.

1.12. **Examples.**

1.12.1. *Transformations.* Calculate the values of the cosine function in terms of the sine function.

Solution. Graph transformation can be used to calculate the values of functions. For instance, the graph of the cosine function is the same as the graph of the sine function but shifted horizontally $\tau/4$ units to the left; *i.e.*,

$$\cos[x] = \sin[x + \tau/4] \ .$$

To numerically verify this, set $x = 0$ to find, by direct calculation, $\cos[0] = 1$ or, indirectly,

$$\cos[0] = \sin[0 + \tau/4] = \sin[\pi/2] = 1 \, ,$$

which agrees with the direct calculation.

With this little trick, we need only memorize the values of the sine function and apply the shift to find those for the cosine function.

2. Foundations of Dimensional Analysis

[Grigory Isaakovich Barenblatt. Scaling, Self-similarity, and Intermediate Asymptotics: Dimensional Analysis and Intermediate Asymptotics]

2.1. **Measurement of Physical Quantities, Units of Measurement. System of Units.** In general, we express all physical quantities in terms of numbers. These numbers are attained by *measuring* the physical quantities.

The process of *measurement* is the direct or indirect comparison of a certain quantity with an appropriate standard – a *unit of measurement*; *e.g.*, if the length of a ruler is 0.25 m, it means that the length has been compared with a unit of measurement of length – the meter.

The units for measuring physical quantities can be classified into *fundamental* and *derived*.

When a class of phenomena (mechanics, heat transfer, *etc*) is singled out for study, certain quantities are listed and standard references reduces for these quantities – natural or artificial – are adopted as fundamental. Once the fundamental units have been chosen, derived units are obtained from the fundamental units *using the definitions of the quantities involved*. These definitions always involve describing at least a conceptual method for measuring the physical quantity in question. For instance, density is by definition the ratio of some mass to the volume occupied by that mass. Thus, the density of an homogeneous body that contains one unit of mass per unit volume – a cube with a side equal to one unit of length – can be adopted as a unit of density.

It can be seen, then, that it is precisely the class of phenomena under discussion (the complete set of physical quantities in which we are interested) that ultimately determines whether or not a given set of fundamental units is sufficient for its measurement. A set of fundamental units that is *sufficient* for measuring the properties of the class of phenomena under consideration is called a *system of units*. The common one in use in science and technology is the SI.

Systems of units depend on the phenomena under consideration. For instance,

- properties of geometric objects: length (that's why the metric is an important concept!);
- kinematic phenomena require one more unit, besides length: time;
- dynamic phenomena require one more unit, besides kinematic units: force or mass;
- heat and mass transfer require one more unit, besides dynamic units: temperature.

However, the system of units need not be *minimal*. For instance, length could be measured in cm, m or even in.

2.2. **Classes of System of Units.** Consider a system of units, say MKS, and consider a second system, say cgs. These two systems of units share the same property: *standard quantities of the same physical nature (mass, length and time) are used as fundamental units*. To generalize, a system of units that differs only in the magnitude (but not in the physical nature) of the fundamental units is called a *class of systems of units*. Then, for instance, consider the MKS system, the corresponding units for an arbitrary system in this class are

$$\begin{aligned} \text{unit of length} &= \text{m}/L\,; \\ \text{unit of mass} &= \text{kg}/M\,; \\ \text{unit of time} &= \text{s}/T\,, \end{aligned}$$

where L, M and T are *abstract positive numbers* that indicate the factors by which the fundamental units of length, mass and time decrease in passing from the original system (in this case, MKS) to another system in the same class. The call is called the LMT class. Other class frequently used is the FLT class, where force, length and time are chosen as fundamental units.

As an example, consider a class where the units of length, mass and time are chosen as fundamental were given by

$$\text{m}/L\,, \qquad \text{kg}/M \qquad \text{and} \qquad \text{hr}/T\,.$$

This set is the same as the MKS. The only difference is the representation of the LMT class: in the second representation, we have $L = 1$, $M = 1$ and $T = 3600$.

2.3. **Dimensions.** Upon decreasing the units of mass by a factor M and the unit of length by a factor L, we find that the new density is a factor M/L^3 smaller than the original unit, so that the numerical values of all densities are thus decreased by a factor of M/L^3. The changes in the numerical values of physical quantities upon passage of one system of units to another one within the *same class* are determined by their *dimension*.

> The function that determines the factor by which the numerical value of a physical quantity changes upon passage from the original system of units to another system of units within a given class is called the *dimension function* or *dimension* of that quantity.

We denote the dimension of a given physical quantity ϕ by $\dim \phi$. We emphasize that the dimension of a given physical quantity is different in different classes of systems of units. For instance, the dimension of density ρ in the MLT class is $\dim \rho = M/L^3$; in the FLT class, it is $\dim \rho = FT^2/L^4$.

> Quantities whose numerical values are identical in all systems of units within a given class are called *dimensionless*.

Therefore, always mention the system of units when expressing the dimensions of a physical quantity. Say,

> The dimensions of pressure p in the MLT class is $\dim p = [M/LT]$.

Analogously,

> the dimensions of pressure p in the FLT class is $\dim_{FLT} = \left[F/L^2\right]$.

The dimension function has two important properties:

(1) The dimension function is always a power-law monomial.
(2) All systems with in a given class are equivalent; *i.e.*, there are no distinguished, somehow preferred, systems among them.

The first property follows from the second one: [demonstration in the text].

In practice, convenient systems of units have been proposed for use with some special classes of problems. For instance, in classical electrodynamics, Kapitza proposed a natural system of units based on the classical radius of the electron as the unit of length, the rest-mass energy of the electron as the units of energy and the mass of the electron as the unit of mass; *i.e.*, a LEM class. This system is convenient, since it allows one to avoid very large or very small numeric values for all quantities of practical interest. But, it does not mean that this system is preferred over the SI system, for instance.

Example: at the beginning of the 20th century, the physico-chemists E. Bose, D. Rauert and M. Bose published a series of experimental studies on the internal turbulent friction of various fluids. The experiments were carried out in the following way: various fluids (water, chloroform, bromoform, *etc*) were allowed to flow through a pipe in a regime of steady turbulence. The time t required to fill a vessel with a certain fixed volume v and the pressure drop p between the ends of the pipe were measured. As was customary, the results of the measurements were presented in the form of a series of tables and curves showing the pressure drop as a function of the filling time.

T. von Karman was attracted by the work of Bose and Rauert and he subjected their results to a processing procedure using dimensional analysis. von Karman analysis can be presented as this: the pressure drop between the ends of the pipe p depends on the time required for the vessel to be filled t and its volume v, as well as on the properties of the fluid, its dynamic viscosity μ and mass density ρ. The dimensions of the quantities are as follows:

$$\dim p = [M/LT]\ ,\ \dim t = [T]\ ,\ \dim v = \left[L^2\right]\ ,\ \dim \mu = [M/LT] \text{ and } \dim \rho = \left[M/L^3\right]\ ,$$

where the MLT system of units was chosen.

According to the Pi-theorem, the number of required dimensionless parameters is $5 - 3 = 2$. Thus,

$$\Pi_1 = \frac{pt}{\mu} \qquad \text{and} \qquad \Pi_2 = \frac{\rho v^{2/3}}{\mu t}\ .$$

The parameters were found in Wolfram Alpha with the inputs:

- for Π_1: pressure, time, volume, dynamic viscosity and mass density and
- for Π_2: time, volume, dynamic viscosity and mass density. However, in this case, some processing was needed. The raw output was $\Pi_2 = \mu^2 t^3/\rho^3 v^2$. Then, the cubic root was taken and the final expression was inversed, since p depends directly on ρ (denser fluid, then greater pressure lost), so ρ must live "upstairs".

After finding the dimensionless parameters, the model can be found by applying the principle of dimensionally homogeneity of physical laws:

$$f[\Pi_1, \Pi_2] = 0 \implies \Pi_1 = f[\Pi_2] \implies \frac{pt}{\mu} = f\left[\frac{\rho v^{2/3}}{\mu t}\right] \implies \frac{pt}{\mu} = \Pi \frac{\rho v^{2/3}}{\mu t},$$

where *kdim* is a dimensionless parameter that must be obtained by experimentation.

2.4. **Approach to Problem Solving.** [dim analysis with case studies in mechanics]

A series of physical quantities can describe natural phenomena and engineering problems such that the physical laws governing those phenomena and problems can be understood. Revealing those physical laws involves three steps:

(1) Classifying physical quantities of a given phenomenon or problem according to the natures of these physical quantities.
(2) Finding correlations that connect the physical quantities.
(3) Finding causality that connects the physical quantities.

To determine causality, it is necessary to understand physical links and relations in a phenomenon or problem. Fundamental principles of physics may then be used to find parameters of cause and effect governing the phenomenon or problem. Parameters must be ranked according to importance, and only parameters in the same class can be compared in terms of magnitude. Deeper analysis means better results, so the analyst needs rich experience and resourcefulness in order to succeed. Trial and error is the usual way to achieve satisfactory results.

3. Dimensional Analysis

> The idea of dimensional analysis is that units [...] are artificial. The universe cares not for our choice of units. Valid physical laws must have the same form in any system of units. Only dimensionless quantities – pure numbers – are the same in every unit system, so we write equations in a universe-friendly, dimensionless form. Often, there is only one such form. Then, without doing any work, we have solved the problem.
>
> — order-of-magnitude physics: understanding the world with dimensional analysis, educated guesswork, and white lies, Peter Goldreich, Sanjoy Mahajan and Sterl Phinney, 1999.

[dim. analysis, bridgman] [the endeavor of dimensional analysis] is to find, without going trough a detailed solution of the problem, certain relations which must be satisfied by the various measurable quantities in which we are interested. The usual procedure is as follows. We first make a list of all the quantities on which the answer may be supposed to depend; we then write down the dimensions of these quantities, and then we demand that these quantities be combined into a functional relation in such a way that the relation remains true no matter what the size of the units in terms of which the quantities are measured.

[Some Nomenclature:

physical property: A physical property is any property that is measurable whose value describes a physical system's state. The changes in the physical properties of a system can be used to describe its transformations (or evolutions between its momentary states).

Physical properties can be intensive or extensive or neither. An intensive property does not depend on the size or amount of matter in the object, while an extensive property does in an additive manner. In addition to extensiveness, properties can also be either isotropic if their values do not depend on the direction of observation or anisotropic otherwise. Physical properties are referred to as observables. They are not modal properties.

Often, it is difficult to determine whether a given property is physical or not. Color, for example, can be "seen"; however, what we perceive as color is really an interpretation of the reflective properties of a surface. In this sense, many ostensibly physical properties are termed as supervenient. A supervenient property is one which is actual (for dependence on the reflective properties of a surface is not simply imagined), but is secondary to some underlying reality. This is similar to the way in which objects are supervenient on atomic structure. A "cup" might have the physical properties of mass, shape, color, temperature, *etc.*, but these properties are supervenient on the underlying atomic structure, which may in turn be supervenient on an underlying quantum structure.

Physical properties are contrasted with chemical properties which determine the way a material behaves in a chemical reaction. Properties that do not change the chemical nature of matter.

physical quantity: Property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a number and a reference. Hence the value of a physical quantity q is expressed as the product of a numerical value N_q and a unit of measurement u_q: $q = N_q \times u_q$.

dimensionless quantity: Property of a phenomenon, body, or substance, where the property has a magnitude that can be expressed as a pure number; *i.e.*, with no units!]

In physics and all science, *dimensional analysis* is the practice of checking relations among physical quantities by identifying their dimensions. The dimension of any physical quantity is the combination of the *basic physical dimensions* that compose it. Some fundamental physical dimensions are length, mass, time, and electric charge. Speed has the dimension length (or distance) per unit of time, and may be measured in meters per second, miles per hour, or other units. Similarly electrical current is electrical charge per unit time (flow rate of charge) and is measured in coulombs (a unit of electrical charge) per second, or equivalently, amperes. Dimensional analysis is based on the fact that

> a physical law must be independent of the units used to measure the physical variables.

A straightforward practical consequence is that any meaningful equation (and any inequality and inequation) must have the same dimensions on the left and right sides. Checking this is the basic way of performing dimensional analysis.

Dimensional analysis is routinely used to check the plausibility of derived equations and computations. It is also used to form reasonable hypotheses about complex physical situations that can be tested by experiment or by more developed theories of the phenomena, and to categorize types of physical quantities and units based on their relations to or dependence on other units, or their dimensions if any.

3.1. **Great Principle of Similitude.** The basic principle of dimensional analysis was known to Isaac Newton (1686) who referred to it as the *Great Principle of Similitude*. James Clerk Maxwell played a major role in establishing modern use of dimensional analysis by distinguishing mass, length, and time as *fundamental dimensions*, while referring to other units as *derived*. The 19th-century French mathematician Joseph Fourier made important contributions based on the idea that physical laws like $f = ma$ should be *independent* of the units employed to measure the physical variables. This led to the conclusion that

> meaningful laws must be *homogeneous* equations in their various units of measurement,

a result which was eventually formalized in the Buckingham π theorem. This theorem describes how every physically meaningful equation involving n variables can be equivalently rewritten as an equation of $n - m$ dimensionless parameters, where m is the number of fundamental dimensions used. Furthermore, and most importantly, it provides a method for computing these dimensionless parameters from the given variables.

A dimensional equation can have the dimensions reduced or eliminated through nondimensionalization, which begins with dimensional analysis, and involves scaling quantities by characteristic units of a system or natural units of nature. This gives insight into the fundamental properties of the system, as illustrated in the examples below.

3.1.1. *Definition.* The *dimension* of a physical quantity can be expressed as a product of the basic physical dimensions mass, length, time, electric charge, and absolute temperature, represented by $[[M], [L], [T], [Q], [\Theta]]$, each raised to a rational power.

The term dimension is more abstract than scale unit: mass is a dimension, while kilograms are a scale unit (choice of standard) in the mass dimension.

As examples, the dimension of the physical quantity speed is length/time ($[L/T]$, and the dimension of the physical quantity force is "mass times acceleration" or "mass times (length/time)/time" ($\left[ML/T^2\right]$). In principle, other dimensions of physical quantity could be defined as "fundamental" (such as momentum or energy or electric current) in lieu of some of those shown above. Most physicists do not recognize temperature, $[\Theta]$, as a fundamental dimension of physical quantity since it essentially expresses the energy per particle per degree of freedom, which can be expressed in terms of energy (or mass, length, and time). Still others do not recognize electric charge, Q, as a separate fundamental dimension of physical quantity, since it has been expressed in terms of mass, length, and time in unit systems such as the cgs system. There are also physicists that have cast doubt on the very existence of incompatible fundamental dimensions of physical quantity.

The unit of a physical quantity and its dimension are related, but not identical concepts. The units of a physical quantity are defined by convention and related to some standard; *e.g.*, length may have units of meters, feet, inches, miles or micrometres; but any length always has a dimension of $[L]$, independent of what units are arbitrarily chosen to measure it. Two different units of the same physical quantity have conversion factors that relate them. For example: $1\,\text{in} = 2.54\,\text{cm}$; then $(2.54\,\text{cm/in})$ is the conversion factor, and is itself dimensionless and equal to one. Therefore, multiplying by that conversion factor does not change a quantity. Dimensional symbols do not have conversion factors.

3.1.2. *Mathematical properties.* The dimensions that can be formed from a given collection of basic physical dimensions, such as $[M]$, $[L]$, and $[T]$, form a *group*: The identity is written as 1; $[L_0] = 1$, and the inverse to $[L]$ is $1/[L]$ or $\left[L^{-1}\right]$. $[L]$ raised to any rational power p is a member of the group, having an inverse of $[1/L^p]$. The operation of the group is multiplication, having the usual rules for handling exponents ($[L^n][L^m] = \left[L^{n+m}\right]$).

This group can be described as a vector space over the rational numbers, with for example dimensional symbol $\left[M^i L^j T^k\right]$ corresponding to the vector $[i, j, k]$. When physical measured quantities (be they like-dimensioned or unlike-dimensioned) are multiplied or divided by one other, their dimensional units are likewise multiplied or divided; this corresponds to addition or subtraction in the vector space. When measurable quantities are raised to a rational power, the same is done to the dimensional symbols attached to those quantities; this corresponds to scalar multiplication in the vector space.

A basis for a given vector space of dimensional symbols is called a *set of fundamental dimensions*, and all other vectors are called *derived units*. As in any vector space, one may choose different bases, which yields different systems of units (*e.g.*, choosing whether the unit for charge is derived from the unit for current, or *vice versa*).

The group identity 1, the dimension of dimensionless quantities, corresponds to the origin in this vector space. The set of units of the physical quantities involved in a problem correspond to a set of vectors (or a matrix). The kernel describes some number (*e.g.*, m) of ways in which these vectors can be combined to produce a zero vector. These correspond to producing (from the measurements) a number of dimensionless quantities, $\{\Pi_1, \ldots, \Pi_m\}$. (In fact these ways completely span the null subspace of another different space, of powers of the measurements.) Every possible way of multiplying (and exponating) together the measured quantities to produce something with the same units as some derived quantity X can be expressed in the general form

$$X = \prod_{i=1}^{m} (\Pi_i)^{k_i} \ .$$

Consequently,

> every possible commensurate equation for the physics of the system can be rewritten in the form
>
> $$f[\Pi_1, \Pi_2, \ldots, \Pi_m] = 0 \, .$$

Knowing this restriction can be a powerful tool for obtaining new insight into the system.

3.1.3. *Polynomials and transcendental functions.* Scalar arguments to transcendental functions such as exponential, trigonometric and logarithmic functions, or to inhomogeneous polynomials *must* be dimensionless quantities.

While most mathematical identities about dimensionless numbers translate in a straightforward manner to dimensional quantities, care must be taken with logarithms of ratios: the identity $\log[a/b] = \log[a] - \log[b]$, where the logarithm is taken in any base, holds for dimensionless numbers a and b, but it does not hold if a and b are dimensional, because in this case the left-hand side is well-defined but the right-hand side is not.

Similarly, while one can evaluate monomials (x^n) of dimensional quantities, one cannot evaluate polynomials of mixed degree with dimensionless coefficients on dimensional quantities.

3.1.4. *Examples.* Period of a harmonic oscillator: What is the period of oscillation T of a mass m attached to an ideal linear spring with spring constant k suspended in gravity of strength g? That period is the solution for T of some dimensionless equation in the variables T, m, k and g with dimensions $\left[[T], [M], \left[M/T^2\right], \left[L/T^2\right]\right]$. From these we can form only one dimensionless product of powers of our chosen variables, $\Pi_1 = T^2 k/m$, where $\dim \Pi_1 = \left[T^2.M/T^2/M = 1\right]$, and putting for some dimensionless constant $\Pi = \Pi_1$ gives the dimensionless equation sought. The dimensionless product of powers of variables is sometimes referred to as a *dimensionless group* of variables; here the term "group" means "collection" rather than mathematical group. They are often called *dimensionless numbers* as well.

Note that the variable g does not occur in the group. It is easy to see that it is impossible to form a dimensionless product of powers that combines g with k, m and T, because g is the only quantity that involves the dimension $[L]$. This implies that in this problem g is irrelevant. Dimensional analysis can sometimes yield strong statements about the *irrelevance* of some quantities in a problem or the need for additional parameters. If we have chosen enough variables to properly describe the problem, then from this argument

we can conclude that the period of the mass on the spring is independent of g: it is the same on the earth or the moon. The equation demonstrating the existence of a product of powers for our problem can be written in an entirely equivalent way: $T = \Pi\sqrt{m/k}$, for some dimensionless constant Π (equal to Π from the original dimensionless equation).

When faced with a case where dimensional analysis rejects a variable (g, here) that one intuitively expects to belong in a physical description of the situation, another possibility is that the rejected variable is in fact relevant, but that some other relevant variable has been omitted, which might combine with the rejected variable to form a dimensionless quantity. That is, however, not the case here.

When dimensional analysis yields only one dimensionless group, as here, there are no unknown functions, and the solution is said to be "complete" – although it still may involve unknown dimensionless constants, such as Π.

Energy of a vibrating wire: Consider the case of a vibrating wire of length l ($[L]$) vibrating with an amplitude A ($[L]$). The wire has a linear density ρ ($[M/L]$) and is under tension s ($\left[ML/T^2\right]$), and we want to know the energy E ($\left[ML^2/T^2\right]$) in the wire. Let Π_1 and Π_2 be two dimensionless products of powers of the variables chosen, given by

$$\Pi_1 = E/As \qquad \text{and} \qquad \Pi_2 = l/A\,.$$

The linear density of the wire is not involved. The two groups found can be combined into an equivalent form as an equation

$$g[\Pi_1, \Pi_2] = g[E/As, l/A] = 0\,.$$

where g is some unknown function, or, equivalently as

$$E = As f[l/A]\ ,$$

where f is some other unknown function. Here the unknown function implies that our solution is now incomplete, but dimensional analysis has given us something that may not have been obvious: *the energy is proportional to the first power of the tension.* Barring further analytical analysis, we might proceed to experiments to discover the form for the unknown function f. But our experiments are simpler than in the absence of dimensional analysis. We'd perform none to verify that the energy is proportional to the tension. Or perhaps we might guess that the energy is proportional to l, and so infer that $E = ls$. The power of dimensional analysis as an aid to experiment and forming hypotheses becomes evident.

The power of dimensional analysis really becomes apparent when it is applied to situations, unlike those given above, that are more complicated, the set of variables involved are not apparent, and the underlying equations hopelessly complex. Consider, for example, a small pebble sitting on the bed of a river. If the river flows fast enough, it will actually raise the pebble and cause it to flow along with the water. At what critical velocity will this occur? Sorting out the guessed variables is not so easy as before. But dimensional analysis can be a powerful aid in understanding problems like this and is usually the *very first tool* to be applied to complex problems where the underlying equations and constraints are poorly understood. In such cases, the answer may depend on a dimensionless number such as the Reynolds number, which may be interpreted by dimensional analysis.

3.1.5. *Percentages and derivatives.* Percentages are dimensionless quantities, since they are ratios of two quantities with the same dimensions.

Derivatives with respect to a quantity add the dimensions of the variable one is differentiating with respect to on the denominator. Thus:

- position (x) has dimension of $[L]$ (Length);
- derivative of position with respect to time ($\mathrm{d}x/\mathrm{d}t$, velocity) has units of $[L/T]$ – Length from position, Time from the derivative;
- the second derivative ($\mathrm{d}^2x/\mathrm{d}t^2$, acceleration) has units of $\left[L/T^2\right]$.

3.1.6. *Dimensionless concepts.* Constants: The dimensionless constants that arise in the results obtained come from a more detailed analysis of the underlying physics and *often* arises from integrating some differential equation. Dimensional analysis itself has little to say about these constants, but it is useful to know that they *very often have a magnitude of order unity.* This observation can allow one to sometimes make "back of the envelope"

calculations about the phenomenon of interest, and therefore be able to more efficiently design experiments to measure it, to judge whether it is important, *etc.*

Formalisms: Paradoxically, dimensional analysis can be a useful tool even if all the parameters in the underlying theory are dimensionless.

It has been argued by some physicists, *e.g.*, Michael Duff, that the laws of physics are inherently dimensionless. The fact that we have assigned incompatible dimensions to Length, Time and Mass is, according to this point of view, just a matter of convention, borne out of the fact that before the advent of modern physics, there was no way to relate mass, length, and time to each other. The three independent dimensionful constants: c, $\hbar$ and G, in the fundamental equations of physics must then be seen as mere *conversion factors* to convert Mass, Time and Length into each other.

One can recover the results of dimensional analysis in the appropriate scaling limit; *e.g.*, dimensional analysis in mechanics can be derived by reinserting the constants $\hbar$, c, and G (but we can now consider them to be *dimensionless*) and demanding that a nonsingular relation between quantities exists in the limit $c \to \infty$, $\hbar \to 0$ and $G \to 0$. In problems involving a gravitational field the latter limit should be taken such that the field stays finite.

3.1.7. *Dimensional Equivalences.* Following are tables of commonly occurring expressions in physics, related to the dimensions of energy, momentum, and force.

[`https://en.wikipedia.org/wiki/Dimensional_analysis#Dimensional_equivalences`.]

[`https://en.wikipedia.org/wiki/List_of_physical_quantities`.]

One use of the table: estimate the internal energy of an ideal gas using dim. analysis.

In the table, in the thermal section, energy e is the equivalent of the product between pressure p and volume v: pv. Thus, $e \sim pv$. On the other hand, we know that for ideal gases the ideal gas law holds: $pv = nk_{\mathrm{b}}t$, where n is the number of molecules, k_{b} Boltzmann constant and t thermodynamic temperature. Therefore, energy can be computed as

$$e \sim nk_{\mathrm{b}}t = \Pi nk_{\mathrm{b}}t\,,$$

where Π is a dimensionless quantity. Physically, e represents the gas internal energy and Π the gas *specific heat capacity at constant volume*, denoted c_v. c_v takes the values:

$$c_v \sim \begin{cases} 3/2 & \text{for mono-atomic gases}\,, \\ 5/2 & \text{for di-atomic gases}\,, \\ 3 & \text{for complex molecules}\,. \end{cases}$$

3.2. **Buckingham Pi Theorem.** The Buckingham Π theorem is a key theorem in dimensional analysis. It is a formalization of Rayleigh's method of dimensional analysis. The theorem loosely states that

> if we have a physically meaningful equation involving a certain number, n, of physical quantity, and these quantities are expressible in terms of k independent fundamental physical dimensions, then the original expression is equivalent to an equation involving a set of $p = n - k$ dimensionless quantities constructed from the original quantities: it is a scheme for nondimensionalization.

This provides a method for computing sets of dimensionless quantities from the given physical quantities, even if the *form* of the equation is still unknown. However, the choice of dimensionless quantities is not unique: Buckingham's theorem only provides a way of generating sets of dimensionless quantities, and will not choose the most 'physically meaningful'.

3.2.1. *Statement.* More formally, the number of dimensionless terms that can be formed, p, is equal to the nullity of the dimensional matrix, and k is the rank. For the purposes of the experimenter, different systems which share the same description in terms of these dimensionless quantities are equivalent.

In mathematical terms, if we have a physically meaningful equation such as

$$f[q_1, q_2, \ldots, q_n] = 0\,,$$

where the $\{q_i\}$ are the n physical quantity, and they are expressed in terms of k independent physical dimensions, then the above equation can be restated as

$$F[\Pi_1, \Pi_2, \ldots, \Pi_p] = 0\,,$$

where the $\{\Pi_i\}$ are dimensionless quantities constructed from the q_i by $p = n - k$ equations of the form

$$\Pi_i = q_1^{a_1} q_2^{a_2} \cdots q_n^{a_n} = \prod_{i=1}^{n} q_i^{a_i}\,,$$

where the exponents a_i are rational numbers (they can always be taken to be integers: just raise it to a power to clear denominators).

The use of the $\{\Pi_i\}$ as the dimensionless quantities was introduced by Edgar Buckingham in his original 1914 paper on the subject from which the theorem draws its name.

3.2.2. *Significance.* The Vaschy-Buckingham Π theorem provides a method for computing sets of dimensionless parameters from the given variables, even if the form of the equation is still unknown. However, the choice of dimensionless parameters is not unique: Buckingham's theorem only provides a way of generating sets of dimensionless parameters, and will not choose the most 'physically meaningful'.

Two systems for which these parameters coincide are called *similar* (as with similar triangles, they differ only in scale); they are equivalent for the purposes of the equation, and the experimentalist who want to determine the form of the equation can choose the most convenient one.

3.2.3. *Rayleigh's Method of Dimensional Analysis. Rayleigh's method of dimensional analysis* is a conceptual tool used in physics, chemistry and engineering. This form of dimensional analysis expresses a functional relationship of some variables in the form of an exponential equation. It was named after Lord Rayleigh.

The method involves the following steps:

- Gather all the independent variables that are likely to influence the dependent variable.
- If R is a variable that depends upon independent variables $\{R_1, R_2, \ldots, R_n\}$, then the functional equation can be written as $R = f[R_1, R_2, \ldots, R_n]$.
- Write the above equation in the form $X = \Pi X_1^a X_2^b \cdots X_n^m$ where Π is a dimensionless constant and $a, b, c, \ldots, m$ are arbitrary exponents.
- Express each of the quantities in the equation in some fundamental units in which the solution is required.
- By using dimensional homogeneity, obtain a set of simultaneous equations involving the exponents $a, b, c, \ldots, m$.
- Solve these equations to obtain the value of exponents $a, b, c, \ldots, m$.
- Substitute the values of exponents in the main equation, and form the non-dimensional parameters by grouping the variables with like exponents.

3.2.4. *Examples.* Pipe flow: We consider the problem of determining the pressure drop of a fluid flowing through a pipe. If the pipe is long compared to its diameter, we shall assume that the pressure drop is proportional to the length of the pipe, all other factors being equal. Thus we really look for the (average) pressure gradient ∇p, and presume the length of the pipe to be irrelevant.

Variables that are relevant clearly include other properties of the *pipe*: Its diameter d, and its roughness e. To a first approximation, we just let e be the average size of the unevennesses of the inside surface of the pipe; thus it is a length.

Also relevant are *fluid* properties. We shall use the kinematic viscosity $\nu = \mu/\rho$ with the density ρ. In a Newtonian fluid in shear motion, the shear tension (a force per unit area) is proportional to a velocity gradient, and the dynamic viscosity μ is the required constant of proportionality: Thus the dimensions of μ are $[M/LT]$, and therefore those of ν are $\left[L^2/T\right]$.

Finally, the average fluid velocity v is most definitely needed.

The dimension matrix can be written as follows: ... [matrix here!]

We can find the null space of this, and hence use it to find the dimensionless combinations. However, it is in fact easier to find dimensionless combinations by inspection. It is easy to see that the matrix has rank 3, so with 6 variables, we must find $6 - 3 = 3$ independent dimensionless combinations. There are, of course, an infinite number of possibilities, since the choice corresponds to choosing a basis for the nullspace of A (matrix). In this case we may be guided by common practice, however, and pick dimensionless quantities as follows:

$$\Pi_{\mathrm{re}} = \frac{vd}{\nu}\,, \qquad \text{[Reynold's number]}$$
$$\varepsilon = \frac{e}{d}\,, \qquad \text{[Relative roughness]}$$
$$\frac{\nabla p d}{\rho v^2}\,. \qquad \text{[No name]}$$

Since we expect the ∇p to be a function of the other variables, we should have a relationship between the dimensionless constant that yields a unique solution for ∇p:

$$\frac{\nabla p}{\rho \nu^2} = f[\Pi_{\mathrm{re}}, \varepsilon]\ ,$$

or

$$\nabla p = \frac{2\rho\nu^2}{d} f[\Pi_{\mathrm{re}}, \varepsilon]\ .$$

The factor of 2 was introduced, because then f is known as *Fanning's friction factor*.

One final remark. We did not really need to write down the dimension matrix. It is quite clear that the three dimensionless quantities we found are independent, since each of them contains at least one variable which is not present in the two others. Since there were only three fundamental units involved, the dimension matrix could not possibly have rank greater than 3, and therefore there could not exist more than three independent dimensionless combinations. Still, the dimension matrix provides a convenient way to summarize the dimensions and to reduce everything to a problem in linear algebra.

3.2.5. *Guidelines to Perform Dimensional Analysis.* Every valid physical equation can be written in a form without units. To find such forms, you follow these steps (see also, section 3.2.5):

- Write down – by magic, intuition, or luck – the physically relevant quantities. For illustration, let's say that there are n of them.
- Determine the dimensions of each quantity. Count how many *independent* dimensions these quantities comprise. Call this number r. Usually, length, mass, and time are all that you need, so $r = 3$.
- By playing around, or by guessing with inspiration, find $n - r$ independent dimensionless combinations of the physical quantities. These combinations are the dimensionless quantities, or Pi groups – named for the Buckingham Pi theorem.
- To verify the dimless groups, head to Wolfram Alpha website: `https://www.wolframalpha.com`, and then type the names of the physical quantities separated by commas; *i.e.*, if "force, density, velocity, radius" are typed, then the results show the dimensions of the quantities, their units and a dimless combination of them. No more playing around or guessing. Awesome!
- Write down the most general relation using these groups. Then try to eliminate dimensionless groups, or to restrict the form of the relation, using physical arguments.
- Using physical arguments, simplify the dimensionless relation by eliminating dimensionless groups, or by otherwise constraining the form of the relation.
- When you solve a difficult problem – such as computing the drag force – simplify by assuming an extreme case: Assume that one or more of the dimensionless variables are nearly zero or nearly infinity.
- Using your solution, check your assumption!

Dimensional analysis is to be complemented, iterated and refined by plugging in some numbers, perhaps known from a given problem or from thought experiments, using bibliographic information, see section 3.2.5). The procedure to solve such problems is:

FIGURE 1. The Pi machine. After lucky guessing, we decide what variables to feed the grouper. We feed them in, along with their units. The *grouper* hunts for combinations that have no units (for dimensionless groups), and spits them out (here, there is only one group). That group enters the *relation finder*, which produces the most general dimensionless relation from its inputs. The *simplifier* simplifies this equation, presenting the answer $F = \Pi ma$.

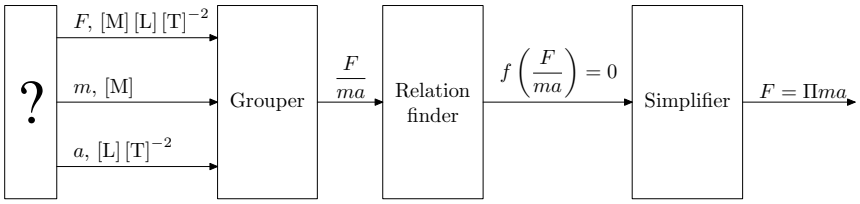
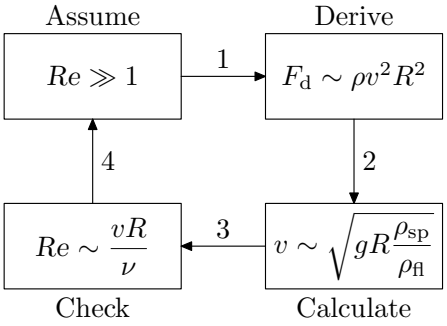


FIGURE 2. As an example: the correct solution order for terminal velocity in turbulent flow. By starting in the *assume box*, we simplified the solution procedure. Step 1: On that assumption, we estimated the drag force (the *derive box*). Step 2: From the drag force, we estimated the terminal velocity (the *calculate box*). Step 3: From the terminal velocity, we estimated the Reynolds' number (the *check box*). Step 4: To close the loop, we verified the starting condition. Note: For compactness, the terminal-velocity formula ignores the normally small effect of buoyancy.



- Use the least amount of assumptions to solve problems. Only then do you build up adding more complexity, more effects.
- Estimate intermediate physical quantities with given information or with bibliographical numbers.
- Estimate the desired physical quantity with the intermediate physical quantities.
- Verify the starting condition: the assumptions: are they satisfied? Does the model give useful information? Is the model still applicable? More parameters are needed?

3.3. **Nondimensionalization.** *Nondimensionalization* is the partial or full removal of physical dimensions from an equation involving physical quantities by a suitable substitution of variables. This technique can simplify and parameterize problems where measured units are involved. It is closely related to dimensional analysis. In some physical systems, the term scaling is used interchangeably with nondimensionalization, in order to suggest that certain quantities are better measured relative to some appropriate unit. These units refer to quantities intrinsic to the system, rather than units such as SI units. Nondimensionalization is not the same as converting extensive quantities in an equation to intensive quantities, since the latter procedure results in variables that still carry units.

Nondimensionalization can also recover characteristic properties of a system. For example, if a system has an intrinsic resonance frequency, length, or time constant, nondimensionalization can recover these values. The technique is especially useful for systems that can be described by differential equations. One important use is in the analysis of control systems. One of the simplest characteristic units is the doubling time of a system experiencing exponential growth, or conversely the half-life of a system experiencing exponential decay; a more natural pair of characteristic units is mean age/mean lifetime, which correspond to base e rather than base 2.

Many illustrative examples of nondimensionalization originate from simplifying differential equations. This is because a large body of physical problems can be formulated in terms of differential equations. Consider the following:

- List of dynamical systems and differential equations topics
- List of partial differential equation topics
- Differential equations of mathematical physics

Although nondimensionalization is well adapted for these problems, it is not restricted to them. An example of a non-differential-equation application is dimensional analysis, while another is normalization in statistics.

Measuring devices are practical examples of nondimensionalization occurring in everyday life. Measuring devices are calibrated relative to some known unit. Subsequent measurements are made relative to this standard. Then, the absolute value of the measurement is recovered by scaling with respect to the standard.

3.3.1. *Rationale.* Suppose a pendulum is swinging with a particular period T. For such a system, it is advantageous to perform calculations relating to the swinging relative to T. In some sense, this is normalizing the measurement with respect to the period.

Measurements made relative to an intrinsic property of a system will apply to other systems which also have the same intrinsic property. It also allows one to *compare a common property of different implementations of the same system.* Nondimensionalization determines in a systematic manner the *characteristic units* of a system to use, without relying heavily on prior knowledge of the system's intrinsic properties (one should not confuse characteristic units of a system with natural units of nature). In fact, *nondimensionalization can suggest the parameters which should be used for analyzing a system.* However, it is necessary to start with an equation that describes the system appropriately.

3.3.2. *Nondimensionalization Steps.* To nondimensionalize a system of equations, one must do the following:

(1) Identify all the independent and dependent variables;
(2) Replace each of them with a quantity scaled relative to a characteristic unit of measure to be determined;
(3) Divide through by the coefficient of the highest order polynomial or derivative term;
(4) Choose judiciously the definition of the characteristic unit for each variable so that the coefficients of as many terms as possible become 1;
(5) Rewrite the system of equations in terms of their new dimensionless quantities.

The last three steps are usually specific to the problem where nondimensionalization is applied. However, almost all systems require the first two steps to be performed.

Note. Additionally, and more importantly, give a physical interpretation of the nondim. equation, of the nondim. quantities and of the characteristic quantities.

Example. Non-dimensionalize the following first order differential equation with constant coefficients:

$$a\frac{\mathrm{d}x}{\mathrm{d}t} + bx = Af[t] \ .$$

Solution.

(1) In this equation the independent variable here is t, and the dependent variable is x.
(2) Set $x = \bar{x}x_c$ and $t = \bar{t}t_c$. This results in the equation

$$a\frac{x_c}{t_c}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + bx_c\bar{x} = Af[\bar{t}t_c] \coloneqq AF[\bar{t}] \ .$$

(3) The coefficient of the highest ordered term is in front of the first derivative term. Dividing by this gives
$$\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \frac{bt_c}{a}\bar{x} = \frac{At_c}{ax_c}F[\bar{t}]\ .$$
(4) The coefficient in front of $\bar{x}$ only contains one characteristic variable t_c, hence it is easiest to choose to set this to unity first:
$$\frac{bt_c}{a} = 1 \implies t_c = \frac{a}{b}\,.$$
Subsequently,
$$\frac{At_c}{ax_c} = \frac{A}{bx_c} = 1 \implies x_c = \frac{A}{b}\,.$$
(5) The final dimensionless equation in this case becomes completely independent of any parameters with units:
$$\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \bar{x} = F[\bar{t}]\ .$$

3.3.3. *Substitutions.* Suppose for simplicity that a certain system is characterized by two variables – a dependent variable x and an independent variable t, where x is a function of t, $x[t]$. Both x and t represent quantities with units. To scale these two variables, assume there are two intrinsic units of measurement x_c and t_c with the same units as x and t respectively, such that these conditions hold:
$$\bar{t} = \frac{t}{t_c} \implies t = \bar{t}t_c \qquad \text{and} \qquad \bar{x} = \frac{x}{x_c} \implies x = \bar{x}x_c\,.$$
These equations are used to replace x and t when nondimensionalizing. If differential operators are needed to describe the original system, their scaled counterparts become dimensionless differential operators.

3.3.4. *Conventions.* There are no restrictions on the variable names used to replace x and t. However, they are generally chosen so that it is convenient and intuitive to use for the problem at hand. For example, if x represented mass, the letter m might be an appropriate symbol to represent the dimensionless mass quantity.

In this article, the following conventions have been used:

- t represents the independent variable – usually a time quantity. Its nondimensionalized counterpart is $\bar{t}$.
- x represents the dependent variable – can be mass, voltage, or any measurable quantity. Its nondimensionalized counterpart is $\bar{x}$.

The subscripted c added to a quantity's variable-name is used to denote the characteristic unit used to scale that quantity. For example, if x is a quantity, then x_c is the characteristic unit used to scale it.

3.3.5. *Differential Operators.* Consider the relationship:
$$t = \bar{t}t_c \implies \mathrm{d}t = t_c\mathrm{d}\bar{t} \implies \frac{\mathrm{d}\bar{t}}{\mathrm{d}t} = \frac{1}{t_c}\,.$$
The dimensionless differential operators with respect to the independent variable becomes (the chain rule used)
$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\mathrm{d}\bar{t}}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\bar{t}} = \frac{1}{t_c}\frac{\mathrm{d}}{\mathrm{d}\bar{t}} \implies \frac{\mathrm{d}^n}{\mathrm{d}t^n} = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n = \left(\frac{1}{t_c}\frac{\mathrm{d}}{\mathrm{d}\bar{t}}\right)^n = \frac{1}{t_c^n}\frac{\mathrm{d}^n}{\mathrm{d}\bar{t}^n}\,.$$

3.3.6. *Forcing Function.* If a system has a forcing function $f[t]$, then
$$f[t] = f[\bar{t}t_c] = f[t[\bar{t}]] = F[\bar{t}]\ .$$
Hence, the new forcing function F is made to be dependent on the dimensionless quantity $\bar{t}$.

3.3.7. *Linear Differential Equations with Constant Coefficients.*

3.3.8. *First order system.* Let us consider the differential equation for a first order system:

$$a\frac{\mathrm{d}x}{\mathrm{d}t} + bx = Af[t] \ .$$

The derivation of the characteristic units for this system gives

$$t_c = \frac{a}{b} \qquad \text{and} \qquad x_c = \frac{A}{b} \ .$$

3.3.9. *Second order system.* The second order system has the form

$$a\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + b\frac{\mathrm{d}x}{\mathrm{d}t} + cx = Af[t] \ .$$

Substitution step: Replace the variables x and t with their scaled quantities. The equation becomes

$$a\frac{x_c}{t_c^2}\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + b\frac{x_c}{t_c}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + cx_c\bar{t} = Af[\bar{t}t_c] = AF[\bar{t}] \ .$$

This new equation is not dimensionless, although all the variables with units are isolated in the coefficients. Dividing by the coefficient of the highest ordered term, the equation becomes

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + t_c\frac{b}{a}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + t_c^2\frac{c}{a}\bar{x} = \frac{At_c^2}{ax_c}F[\bar{t}] \ .$$

Now it is necessary to determine the quantities of x_c and t_c so that the coefficients become normalized. Since there are two free parameters, at most only two coefficients can be made to equal unity.

Determination of characteristic units: Consider the variable t_c:

- If $t_c = a/b$, then the first order term is normalized.
- If $t_c^2 = a/b$, then the zeroth order term is normalized.

Both substitutions are valid. However, for pedagogical reasons, the latter substitution is used for second order systems. Choosing this substitution allows x_c to be determined by normalizing the coefficient of the forcing function:

$$1 = \frac{At_c^2}{ax_c} = \frac{A}{cx_c} \implies x_c = \frac{A}{c}$$

The differential equation becomes

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + \frac{b}{\sqrt{ac}}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \bar{x} = F[\bar{t}] \ .$$

The coefficient of the first order term is dimentionless. Define

$$2\zeta := \frac{b}{\sqrt{ac}} \ .$$

The factor 2 is present so that the solutions can be parameterized in terms of ζ. In the context of mechanical or electrical systems, ζ is known as the *damping ratio* and is an important parameter required in the analysis of control systems. 2ζ is also known as the *linewidth of the system.* The result of the definition is the *universal oscillator equation.*

3.3.10. *Mechanical Oscillations.* Suppose we have a mass attached to a spring and a damper, which in turn are attached to a wall, and a force acting on the mass along the same line. Define

- x = displacement from equilibrium, $[L]$;
- t = time, $[T]$;
- F = external force or "disturbance" applied to system, , $\left[M.L/T^2\right]$;
- m = mass of the block, $[M]$;
- B = damping constant of dashpot, $[M/H]$;
- k = force constant of spring, $\left[M/T^2\right]$.

Suppose the applied force is a sinusoid $F = F_0 \cos[\omega t]$, the differential equation that describes the motion of the block is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + B\frac{\mathrm{d}x}{\mathrm{d}t} + kx = F_0 \cos[\omega t] \ .$$

Nondimensionalizing this equation the same way as described under second order system yields several characteristics of the system.

The intrinsic unit (characteristic quantity) x_c corresponds to the *distance* the block moves per unit force

$$x_c = \frac{F_0}{k} \ ,$$

the characteristic variable t_c is equal to the *period* of the oscillations

$$t_c = \sqrt{\frac{m}{k}} \ ,$$

and the dimensionless variable 2ζ corresponds to the *linewidth* of the system. ζ itself is the damping ratio

$$2\zeta = \frac{B}{\sqrt{mk}} \ .$$

3.3.11. *Nonlinear Differential Equation.* Since there are no general methods of solving nonlinear differential equations, each case has to be considered on an individual basis when nondimensionalizing.

Quantum harmonic oscillator: The Schrödinger equation for the one dimensional time independent quantum harmonic oscillator is

$$\left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{1}{2}m\omega^2x^2\right)\psi[x] = E\psi[x] \ .$$

The modulus square of the wavefunction $|\psi|^2$ represents probability, which is in a sense already dimensionless and normalized. Therefore, there is no need to nondimensionalize the wavefunction. However, it should be rewritten as a function of a dimensionless variable. Furthermore, the variable x has dimensions of length. Hence substitute

$$\bar{x} = \frac{x}{x_c} \qquad \text{and} \qquad \psi[x] = \psi[x[\bar{x}]] = \psi[\bar{x}] \ .$$

The differential equation becomes

$$\begin{aligned} &\left(-\frac{\hbar^2}{2mx_c}\frac{\mathrm{d}^2}{\mathrm{d}\bar{x}^2} + \frac{m\omega^2x_c^2}{2}\bar{x}^2\right)\psi[\bar{x}] = E\psi[\bar{x}] \\ \implies &\left(-\frac{\mathrm{d}^2}{\mathrm{d}\bar{x}^2} + \frac{m^2\omega^2x_c^4}{\hbar^2}\bar{x}^2\right)\psi[\bar{x}] = \frac{2Emx_c^2}{\hbar^2}\psi[\bar{x}] \ . \end{aligned}$$

To make the term in front of $\bar{x}^2$ dimensionless, set

$$\frac{m^2\omega^2x_c^4}{\hbar^2} = 1 \implies x_c = \sqrt{\frac{\hbar}{m\omega}} \ .$$

Hence, the fully nondimensionalized equation is

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}\bar{x}^2} + \bar{x}^2\right)\psi[\bar{x}] = \frac{2E}{\hbar\omega}\psi[\bar{x}] \coloneqq \bar{E}\psi[\bar{x}] \ .$$

The nondimensionalization factor for the energy is the same as the ground state of the harmonic oscillator. Usually, the energy term is not made dimensionless because a primary emphasis of quantum mechanics is determining the energies of the states of a system. Rearranging the first equation, the familiar equation for the harmonic oscillator is

$$\frac{\hbar\omega}{2}\left(-\frac{\mathrm{d}^2}{\mathrm{d}\bar{x}^2} + \bar{x}^2\right)\psi[\bar{x}] = E\psi[\bar{x}] \ .$$

3.4. **Similitude.** *Similitude* is a concept applicable to the testing of engineering models. A model is said to have similitude with the real application if the two share geometric similarity, kinematic similarity and dynamic similarity. Similarity and similitude are interchangeable in this context.

The term dynamic similitude is often used as a catch-all because it implies that geometric and kinematic similitude have already been met.

Similitude's main application is in hydraulic and aerospace engineering to test fluid flow conditions with scaled models. It is also the primary theory behind many textbook formulas in fluid mechanics.

3.4.1. *Overview.* Engineering models are used to study complex fluid dynamics problems where calculations and computer simulations aren't reliable. Models are usually smaller than the final design, but not always. Scale models allow testing of a design prior to building, and in many cases are a critical step in the development process.

Construction of a scale model, however, must be accompanied by an analysis to determine what conditions it is tested under. While the geometry may be simply scaled, other parameters, such as pressure, temperature or the velocity and type of fluid may need to be altered. Similitude is achieved when testing conditions are created such that the test results are applicable to the real design.

The following criteria are required to achieve similitude:

- *Geometric similarity* – The model is the same shape as the application, usually scaled.
- *Kinematic similarity* – Fluid flow of both the model and real application must undergo similar time rates of change motions. (fluid streamlines are similar)
- *Dynamic similarity* – Ratios of all forces acting on corresponding fluid particles and boundary surfaces in the two systems are constant.

To satisfy the above conditions the application is analyzed:

(1) All parameters required to describe the system are identified using principles from continuum mechanics.
(2) Dimensional analysis is used to express the system with as few independent variables and as many dimensionless parameters as possible.
(3) The values of the dimensionless parameters are held to be the same for both the scale model and application. This can be done because they are dimensionless and will ensure dynamic similitude between the model and the application. The resulting equations are used to derive scaling laws which dictate model testing conditions.

It is often impossible to achieve strict similitude during a model test. The greater the departure from the application's operating conditions, the more difficult achieving similitude is. In these cases some aspects of similitude may be neglected, focusing on only the most important parameters.

The design of marine vessels remains more of an art than a science in large part because dynamic similitude is especially difficult to attain for a vessel that is partially submerged: a ship is affected by wind forces in the air above it, by hydrodynamic forces within the water under it, and especially by wave motions at the interface between the water and the air. The scaling requirements for each of these phenomena differ, so models cannot replicate what happens to a full sized vessel nearly so well as can be done for an aircraft or submarine – each of which operates entirely within one medium.

Similitude is a term used widely in fracture mechanics relating to the strain life approach. Under given loading conditions the fatigue damage in an un-notched specimen is comparable to that of a notched specimen. Similitude suggests that the component fatigue life of the two objects will also be similar.

3.4.2. *An Example.* Consider a submarine modeled at 1/40th scale. The application operates in sea water at 0.5 °C, moving at 5 m/s. The model will be tested in fresh water at 20 °C. Find the power required for the submarine to operate at the stated speed.

A free body diagram is constructed and the relevant relationships of force and velocity are formulated using techniques from continuum mechanics. The variables which describe the system are:

- Variable – Application – Scaled mode – Units
- L (diameter of submarine) – 1 – 1/40 – m
- V (speed) – 5 – calculate – m/s
- ρ (density) – 1028 – 998 – kg/m^3
- μ (dynamic viscosity) – 1.88×10^{-3} – 1.00×10^{-3} – Pa s or N s/m^2
- F (force) – calculate – to be measured – N or kg m/s^2

This example has five independent variables and three fundamental units. The fundamental units are meter, kilogram, second.

Invoking the Buckingham π theorem shows that the system can be described with two dimensionless numbers and one independent variable.

Dimensional analysis is used to re-arrange the units to form the Reynolds number Π_{re} and pressure coefficient C_{p}. These dimensionless numbers account for all the variables listed above except F, which will be the test measurement. Since the dimensionless parameters will stay constant for both the test and the real application, they will be used to formulate scaling laws for the test.

Scaling laws:

$$\Pi_{\mathrm{re}} = \left(\frac{\rho V L}{\mu}\right) \qquad \Longrightarrow\ V_{\mathrm{m}} = V_{\mathrm{a}} \left(\frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{m}}}\right) \left(\frac{L_{\mathrm{a}}}{L_{\mathrm{m}}}\right) \left(\frac{\mu_{\mathrm{m}}}{\mu_{\mathrm{a}}}\right) ,$$

$$C_{\mathrm{p}} = \left(\frac{2\Delta p}{\rho V^2}\right) F = \Delta p L^2 \qquad \Longrightarrow\ F_{\mathrm{m}} = F_{\mathrm{a}} \left(\frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{m}}}\right) \left(\frac{V_{\mathrm{a}}}{V_{\mathrm{m}}}\right)^2 \left(\frac{L_{\mathrm{a}}}{L_{\mathrm{m}}}\right)^2 .$$

where the subscript a stands for *application* and m for *model*.

This gives a required test velocity of:

$$V_{\mathrm{m}} = 21.9 V_{\mathrm{a}} .$$

The force measured from the model at that velocity is then scaled to find the force that can be expected for the real application:

$$F_{\mathrm{a}} = 3.44 F_{\mathrm{m}} .$$

The power P in watts required by the submarine is then:

$$P = F_{\mathrm{a}} V_{\mathrm{a}} = 17.2 F_{\mathrm{m}} .$$

Note that even though the model is scaled smaller, the water velocity needs to be increased for testing. This result shows how similitude in nature is often counter-intuitive.

3.4.3. *Typical applications.* Similitude has been well documented for a large number of engineering problems and is the basis of many textbook formulas and dimensionless quantities. These formulas and quantities are easy to use without having to repeat the laborious task of dimensional analysis and formula derivation. Simplification of the formulas (by neglecting some aspects of similitude) is common, and needs to be reviewed by the engineer for each application.

Similitude can be used to predict the performance of a new design based on data from an existing, similar design. In this case, the model is the existing design. Another use of similitude and models is in validation of computer simulations with the ultimate goal of eliminating the need for physical models altogether.

Another application of similitude is to replace the operating fluid with a different test fluid. Wind tunnels, for example, have trouble with air liquefying in certain conditions so helium is sometimes used. Other applications may operate in dangerous or expensive fluids so the testing is carried out in a more convenient substitute.

Some common applications of similitude and associated dimensionless numbers:

- Incompressible flow: Reynolds number, Pressure coefficient, Froude number and Weber number for open channel hydraulics;
- Compressible flows: Reynolds number, Mach number, Prandtl number, Specific heat ratio;
- Flow-excited vibration: Strouhal number;

- Centrifugal compressors: Reynolds number, Mach number, Pressure coefficient, Velocity ratio;
- Boundary layer thickness: Reynolds number, Womersley number, Dynamic similarity.

3.5. **Economy of Graphical Representation.**

> A good table of functions of one variable may require a page; that of a function of two variables a volume; that of a function of three variables a bookcase; and that of a function of four variables a library.
>
> — HAROLD JEFFREYS, quoted in Street-Fighting Mathematics

One of the most significant benefits of using dimensionless versus physical variables is the very substantial saving of space and effort in the logistic in presenting the relations graphically. To illustrate, suppose we wish to find from a chart the dependent variable for k distinct values of the independent variable and the p number of parameters, each of which can also assume k distinct values. To present such a function graphically, how many curves do we need? The answer is found by the following simple argument:

- if there are zero parameters, then we need $k^0 = 1$ curves;
- if there is one parameter, then we need $k^1 = k$ curves;
- if there are two parameters, then we need k^2 curves;
- so, in general, if there are p parameters, then we need k^p curves.

The last generalization can be written as

$$N_{\text{curves}} = k^p \,,$$

where N_{curves} represents the number of curves needed to present the given relationship, k the number of distinct values for the dependent variable and each parameter and p the number of parameters.

If N_{var} is the number of variables, then the number of parameters is generally

$$p = N_{\text{var}} - 2 \,,$$

since we consider *one* variable to be independent and *one* dependent. Thus, by plugging the last equation into the number of curves one has

$$N_{\text{curves}} = k^{N_{\text{var}}-2} \,,$$

A chart can have one *family of curves* characterized by a single *parameter*. Thus, each curve has a particular value of this parameter assigned to it, and therefore there are k curves in a chart. It follows that the number of charts required to present a relation of N_{var} variables is

$$\begin{cases} N_{\text{chart}} = k^{N_{\text{var}}-3} & \text{if } N_{\text{var}} > 2 \,, \\ N_{\text{chart}} = 1 & \text{if } N_{\text{var}} = 2 \,. \end{cases}$$

One can see from these equations that the number of curves and charts to be plotted grows rather vehemently with the number of variables – so much so that the situation soon becomes unmanageable. Is there a way out of this predicament? Yes, there is, use dimensionless variables!

As seen from the Buckingham Pi Theorem:

$$N_{\text{dimless}} = N_{\text{var}} - N_{\text{dims}} \,,$$

where N_{dimless} is the number of dimensionless variables which can be formed from (and by) N_{var} number of physical variables and N_{dims} is the number of dimensions.

Here we assumed that the rank of the dimensional matrix is not less than the number of dimensions N_{dims}. If this condition is not fulfilled, then we must delete one or more dimensions to gain the equality of N_{dims} with the rank of dimensional matrix.

So, if we have N_{dims} dimensions in a relation, then the number of dimensionless variables that can be formed is always less than the number of variables in that relation. This

results in a very significant improvement in the logistics of graphical presentation. If we substitute N_{dimless} for N_{var}, we obtain

$$\begin{cases} N'_{\text{curve}} = k^{N_{\text{var}}-N_{\text{dim}}-2}\,, \\ N'_{\text{chart}} = k^{N_{\text{var}}-N_{\text{dim}}-3}\,. \end{cases}$$

Hence,

$$z = \frac{N_{\text{curve}}}{N'_{\text{curve}}} = \frac{N_{\text{chart}}}{N'_{\text{chart}}}\,.$$

Note that this ratio is *independent* of the number of variables and is dependent only upon the number of dimensions N_{dim} and the number of distinct values k for each parameter. For example, if we have a relation of three dimensions and six distinct values for each parameter (variable), then by *not* employing dimensionless variables, we would need $z = k^{N_{\text{dim}}} = 63 = 216$ times as many curves and charts as we would if we used only these variables.

Example: Gravitational Acceleration on a Celestial Body as a Function of Altitude (Positive or Negative):

Solution. To find the altitude-dependent gravitational acceleration on a celestial body, we deal with the following variables:

- gravitational acceleration, g, $\mathrm{m/s^2}$;
- distance from center of the celestial body, R, m;
- universal gravitational const., G, $\mathrm{m^3/(kg\,s^2)}$;
- mass of celestial body, M, kg;
- radius of celestial body, R_0, m.

We have five variables and hence, if the number of distinct values of variables is $k = 6$, then to represent the relation $g = \Phi[R, G, M, R_0]$ graphically, we need 216 curves drawn on 36 charts. What do we have if we use dimensionless variables? There are five variables and three dimensions, therefore we have $5 - 3 = 2$ dimensionless variables, from which we can find the dimensionless quantities

$$\Pi_1 = \frac{gR_0^2}{GM} \qquad \text{and} \qquad \Pi_2 = \frac{R}{R_0}\,.$$

So now we have only *two* (dimensionless) variables and, therefore, the relation can be plotted by a *single* curve – a 216-fold improvement!

To continue now, we can write

$$\Pi_1 = \Phi[\Pi_2]\,,$$

which, by assuming monomial form, can be written

$$\Pi_1 = c\Pi_2^n\,,$$

where c and n are pure numbers still to be determined. As a simple analytic derivation (not presented here) can show, $c = 1$ and

$$\begin{cases} n = -2 & \text{if } R > R_0\,, \\ n = 1 & \text{if } R < R_0\,. \end{cases}$$

Fig. [...] (graph: in x-axis Π_2 and in y-axis Π_1, because the relation is $\Pi_1 = \Phi[\Pi_2]$) presents this "one-curve" plot. Note the interesting feature that n is not constant but varies – rather abruptly – with the sign of the altitude. The plot shows that g is maximum at "sea level" on any homogeneous celestial body.

3.6. **Steps of Dimensional Analysis.** [Qing-Ming Tan, dim analysis with case studies in mechanics]

The principles of dimensional analysis provide the only way to solve complex problems when there is no available mathematical model. To deal with such problems, it is natural for me to apply the steps of dimensional analysis:

- Analyze physical effects involved in the problem.
- Select corresponding governing parameters.
- Design experiments.
- Analyze and synthesize experimental data.
- Establish dimensionless relationships between cause and effect parameters.

3.7. **Steps of Dimensional Analysis – Again.** Use the following procedure to perform dim. analysis:

- Identify the problem domain: geometry, mechanics, thermal energy transfer, mass transfer and so on.
- Choose a dim-independent set of physical quantities accordingly to the problem domain: $\{L\}$, $\{L,T\}$, $\{M,L,T\}$, $\{F,L,T\}$, $\{E,L,T,\Theta\}$ and so forth. Count the elements as m.
- Identify the transport properties (fluid velocity, thermal flux, mass flux, ...) and the material properties (viscosity, thermal conduction, density, specific thermal capacity, ...). Count all the properties as n. (Do *not* forget to include the quantity being sought!)
- Define the dimensions of all the properties in the dim-ind system.
- Calculate the number of dimensionless quantities: $r = n - m$.
- Form the dimless quantities: Π_i. (The first one, Π_1, should be the one containing the quantity being sought!)
- Apply the principle of dimensional homogeneity of physical laws:

$$f[\Pi_i] = 0\,.$$

- Classify the different dimensionless quantities and give them physical interpretation.
- When possible, use physical information, intuition or approximate methods (simplification, extreme cases, ...) to reduce the number of dimless quantities.
- If required make experiments to find the functional form of f.
- Make calculations with the final form of f in order to obtain numerical values that can be used to review assumptions, simplifications, approximations, *etc.*

3.8. **Examples.**

3.8.1. *Nondimensionalized the equation of motion for a simple pendulum.* The motion of a simple pendulum is modeled using the following (ordinary) differential equation and boundary and initial conditions:

$$\begin{cases} \dfrac{\mathrm{d}^2\phi}{\mathrm{d}t^2} = -\dfrac{g}{l}\sin[\phi]\ , \\ \phi_0 = \phi[0]\ , \\ \dfrac{\mathrm{d}\phi}{\mathrm{d}t} = 0\,. \end{cases}$$

To non-dim. the equation of motion, follow the five-step method:

(1) the independent variable is t, whereas the dependent one is ϕ.
(2) ϕ is already dimensionless, so nothing to do here. For t, find the characteristic time t_c as $t = t_c\bar{t}$, where $\bar{t}$ is the scaled (dimensionless) time. With this replacement, find the derivatives

$$t = t_c\bar{t} \implies \mathrm{d}t = t_c\mathrm{d}\bar{t} \implies \mathrm{d}t^2 = t_c^2\mathrm{d}\bar{t}^2\,.$$

(3) Replace the scaled variables and their derivatives in the original differential equation and boundary and initial conditions

$$\begin{cases} \dfrac{1}{t_c^2}\dfrac{\mathrm{d}^2\phi}{\mathrm{d}\bar{t}^2} + \dfrac{g}{l}\sin[\phi] = 0 \implies \dfrac{\mathrm{d}^2\phi}{\mathrm{d}\bar{t}^2} + \dfrac{gt_c^2}{l}\sin[\phi] = 0\,, \\ \phi_0 = \phi[0]\ , \\ \dfrac{1}{t_c}\dfrac{\mathrm{d}\phi}{\mathrm{d}\bar{t}} = 0 \implies \dfrac{\mathrm{d}\phi}{\mathrm{d}\bar{t}} = 0\,. \end{cases}$$

(4) In the non-dim. model differential equation, the coefficient in front of $\sin[\phi]$ has only one characteristic quantity: t_c. So set it to unity to find

$$\frac{gt_c^2}{l} = 1 \implies t_c = \sqrt{\frac{l}{g}}\,.$$

Note that t_c equals $\sqrt{l/g}$, which, in turn, equals the pendulum period. In other words, the pendulum period is a characteristic quantity: the period is the pendulum's own clock!

(5) With these changes, the model becomes finally

$$\begin{cases} \dfrac{\mathrm{d}^2\phi}{\mathrm{d}\bar{t}^2} + \sin[\phi] = 0\,, \\ \phi_0 = \phi[0]\ , \\ \dfrac{\mathrm{d}\phi}{\mathrm{d}\bar{t}} = 0\,. \end{cases}$$

Not only is this final equation, and its boundary and initial conditions, dimensionless, but also parameter-free!

An analytic solution to the pendulum equation can be found by considering small swings, where the approximation $\sin[\phi] \sim \phi$ is valid. The equation for the model thus becomes

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}t^2} + \frac{g}{l}\phi = 0\,.$$

After integration and replacement of boundary and initial conditions, one finds

$$\phi = \phi_0 \cos\left[t/\sqrt{l/g}\right]\ .$$

Note that the argument of the cosine function is dimensionless and equals the pendulum period; *i.e.*, as expected, the model was dimensionless all along!

3.8.2. *Nondimensionalized the equation of motion for a vertical projectile.* Suppose that a projectile of mass m is thrown vertically into the air with an initial velocity v_0.

Note by x the position of the projectile at any time t. Then, the equations of motion for the projectile are

$$\begin{cases} m\ddot{x} = -mg\,, \\ x[0] = 0\,, \\ \dot{x}[0] = v_0\,. \end{cases}$$

In this set of equations, the independent quantity is t, the dependent one x and the parameters g and v_0.

Choose the characteristic quantities for the phenomenon by

$$\begin{aligned} &\bar{x} = x/x_c \implies x = x_c\bar{x}\,, \\ &\bar{t} = x/t_c \implies t = t_c\bar{x} \implies \mathrm{d}t = t_c\mathrm{d}\bar{t} \implies \mathrm{d}t^2 = t_c^2\mathrm{d}\bar{t}^2\,. \end{aligned}$$

With these choices, the equations of motion become

$$\begin{cases} \dfrac{\bar{x}}{\bar{t}^2}\dfrac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} = -g\,, \\ x[0] = 0 \implies \bar{x}[0] = 0\,, \\ \dfrac{\bar{x}}{\bar{t}}\dfrac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] = v_0\,. \end{cases}$$

Noting that $\dim v_0^2/g = [L]$ and that $\dim v_0/g = [T]$, find x_c and t_c by

$$x_c = \frac{v_0^2}{g} \qquad \text{and} \qquad t_c = \frac{v_0}{g}\,.$$

Replace x_c and t_c in the equations of motion to have

$$\begin{cases} \dfrac{v_0^2}{g}\dfrac{g^2}{v_0^2}\dfrac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} = -g \implies \dfrac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} = -1\,, \\ \bar{x}[0] = 0\,, \\ \dfrac{v_0^2}{g}\dfrac{g}{v_0}\dfrac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] = v_0 \implies \dfrac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] = 1\,. \end{cases}$$

Finally, write the equations of motion as

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} = -1\,, \qquad \bar{x}[0] = 0 \qquad \text{and} \qquad \frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] = 1\,.$$

Note that the last set of equations is dimensionless and parameter free!

3.8.3. *Nondimensionalized the equation of motion for a vertical projectile – again.* Suppose that a projectile of mass m is thrown vertically into the air with an initial velocity v_0.

The modeling equations for the particle motion for this system are:

$$\begin{cases} \dfrac{\mathrm{d}^2x}{\mathrm{d}t^2} = -g \\ x[0] = 0 \\ \dfrac{\mathrm{d}x}{\mathrm{d}t} = v_0 \,. \end{cases}$$

The independent variable is t, the dependent one is x and the parameters are g and v_0.

Choose the characteristic and scaled quantities by

$$x = x_c\bar{x} \implies \mathrm{d}x = x_c\mathrm{d}\bar{x} \implies \mathrm{d}^2x = x_c\mathrm{d}^2\bar{x}\,,$$
$$t = t_c\bar{t} \implies \mathrm{d}t = t_c\mathrm{d}\bar{t} \implies \mathrm{d}t^2 = t_c^2\mathrm{d}\bar{t}^2\,.$$

Replace the chosen quantities in the physical model:

$$\begin{aligned} \frac{x_c}{t_c^2}\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} &= -g\,, \\ x_c\bar{x}[0] = 0 &\implies \bar{x}[0] = 0\,, \\ \frac{x_c}{t_c}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] &= v_0\,. \end{aligned}$$

Manipulate the last set of equations to have

$$\begin{aligned} \frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} &= -\frac{gt_c^2}{x_c}\,, \\ x_c\bar{x}[0] = 0 &\implies \bar{x}[0] = 0\,, \\ \frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] &= \frac{t_cv_0}{x_c}\,. \end{aligned}$$

Equate the coefficients of the RHSs of the last equations to unity

$$\frac{gt_c^2}{x_c} = 1 \qquad \text{and} \qquad \frac{t_cv_0}{x_c} = 1\,,$$

to find that

$$x_c = \frac{v_0^2}{g} \qquad \text{and} \qquad t_c = \frac{v_0}{g}\,.$$

Replace the characteristic quantities into the set of equations for the model to have

$$\begin{aligned} \frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} &= -\frac{v_0^2/g^2}{v_0^2/g}g = -1\,, \\ \bar{x}[0] &= 0\,, \\ \frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] &= \frac{v_0/g}{v_0^2/g}v_0 = 1\,. \end{aligned}$$

This replacements leave the final set of equations for the model:

$$\begin{aligned} \frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} &= -1\,, \\ \bar{x}[0] &= 0\,, \\ \frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}}[0] &= 1\,. \end{aligned}$$

To go back to the original, dimensional, physical quantities, the transformation equations are

$$x = x_c\bar{x} \implies \bar{x} = x/x_c = x/(v_0^2/g) = gx/v_0^2\,,$$
$$t = t_c\bar{t} \implies \bar{t} = t/t_c = t/(v_0/g) = gt/v_0\,.$$

4. Calculus

> When guys at MIT or Princeton had trouble doing a certain integral, it was because they couldn't do it with the standard methods they had learned in school. [..] I come along and try differentiating under the integral sign, and often it worked. So I got a great reputation for doing integrals, only because my box of tools was different from everybody else's, and they had tried all their tools on it before giving the problem to me.
>
> — RICHARD FEYNMAN, Surely You're Joking, Mr. Feynman!

4.1. **Derivative.** In calculus, the *derivative* is a measure of how a function changes as its input changes. Loosely speaking, a derivative can be thought of as how much one quantity is changing in response to changes in some other quantity; *e.g.*, the derivative of the position of a moving object with respect to time is the object's instantaneous velocity.

The derivative of a function at a chosen input value describes *the best linear approximation of the function* near that input value. Informally, the derivative is the ratio of the infinitesimal change of the output over the infinitesimal change of the input producing that change of output. For a real-valued function of a single real variable, the derivative at a point equals the slope of the tangent line to the graph of the function at that point. In higher dimensions, the derivative of a function at a point is a linear transformation called the linearization. A closely related notion is the *differential of a function*.

The process of finding a derivative is called *differentiation*. The reverse process is called *antidifferentiation*. The fundamental theorem of calculus states that antidifferentiation is the same as integration. Differentiation and integration constitute the two fundamental operations in single-variable calculus.

4.1.1. *Definition via difference quotients.* Let f be a real valued function. In classical geometry, the *tangent line to the graph of the function f at a real number a* was the unique line through the point $[a, f[a]]$ that did not meet the graph of f transversally, meaning that the line did not pass straight through the graph. The derivative of y with respect to x at a is, geometrically, the slope of the tangent line to the graph of f at a. The slope of the tangent line is very close to the slope of the line through $[a, f[a]]$ and a nearby point on the graph, *e.g.*, $[a + h, f[a + h]]$. These lines are called *secant lines*. A value of h close to zero gives a good approximation to the slope of the tangent line and smaller values (in absolute value) of h will, in general, give better approximations. The slope m of the secant line is the difference between the y values of these points divided by the difference between the x values, that is,

$$m = \frac{\Delta f[a]}{\Delta a} = \frac{f[a+h] - f[a]}{(a+h) - a} = \frac{f[a+h] - f[x]}{h}\,.$$

This expression is *Newton's difference quotient*. The derivative is the value of the difference quotient as the secant lines approach the tangent line. Formally, the *derivative of the function f at a* is the limit

$$f'[a] = \lim_{h \to 0} \frac{f[a+h] - f[x]}{h}\,.$$

of the difference quotient as h approaches zero, if this limit exists. If the limit exists, then f is differentiable at a. Here $f'[a]$ is one of several common notations for the derivative.

Equivalently, the derivative satisfies the property that

$$\lim_{h \to 0} \frac{f[a+h] - f[a] - f'[a]\,h}{h}\,,$$

which has the intuitive interpretation that the tangent line to f at a gives *the best linear approximation* [2]

$$f[a+h] \sim f[a] + f'[a]\,h$$

[2] Example: calculate $\sin[0.3]$. Solution: let f be sin, $a = 0$ and $h = 0.3$, then find $\sin[0.3] \sim \sin[0 + 0.3] \sim \sin[0] + \cos[0]\,0.3 \sim 0 + (1)(0.3) \sim 0.3$. Verification: the exact solution is $\sin[0.3] = 0.2955202\ldots$.

to f near a (*i.e.*, for small h). This interpretation is the easiest to generalize to other settings.

Substituting 0 for h in the difference quotient causes division by zero, so the slope of the tangent line cannot be found directly using this method. Instead, define $Q[h]$ to be the difference quotient as a function of h:

$$Q[h] = \frac{f[a+h] - f[a]}{h} .$$

$Q[h]$ is the slope of the secant line between $[a, f[a]]$ and $[a+h, f[a+h]]$. If f is a *continuous function*, meaning that its graph is an unbroken curve with no gaps, then Q is a continuous function away from $h = 0$. If the limit $\lim_{h\to 0} Q[h]$ exists, meaning that there is a way of choosing a value for $Q[0]$ that makes the graph of Q a continuous function, then the function f is differentiable at a, and its derivative at a equals $Q[0]$.

In practice, the existence of a continuous extension of the difference quotient $Q[h]$ to $h = 0$ is shown by modifying the numerator to cancel h in the denominator. Such manipulations can make the limiting value of Q for small h clear even though Q is still not defined at $h = 0$. This process can be long and tedious for complicated functions and many shortcuts are commonly used to simplify the process.

4.1.2. *Continuity and Differentiability.* If $y = f[x]$ is differentiable at a, then f must also be continuous at a. However, even if a function is continuous at a point, it may not be differentiable there. Even a function with a smooth graph is not differentiable at a point where its tangent is vertical. In summary: for a function f to have a derivative it is necessary for the function f to be continuous, but continuity alone is not sufficient.

4.1.3. *The Derivative as a Function.* Let f be a function that has a derivative at every point a in the domain of f. Because every point a has a derivative, there is a function that sends the point a to the derivative of f at a. This function is written $f'[x]$ and is called the *derivative function* or the *derivative of* f. The derivative of f collects all the derivatives of f at all the points in the domain of f.

Sometimes f has a derivative at most, but not all, points of its domain. The function whose value at a equals $f'[a]$ whenever $f'[a]$ is defined and elsewhere is undefined is *also called the derivative of* f. It is still a function, but its domain is strictly smaller than the domain of f.

Using this idea, *differentiation becomes a function of functions*: The derivative is an *operator* whose domain is the set of all functions that have derivatives at every point of their domain and whose range is a set of functions. If we denote this operator by D, then $\mathrm{D}f$ is the function f'. Since $\mathrm{D}f$ is a function, it can be evaluated at a point a. By the definition of the derivative function, then

$$\mathrm{D}f[a] = f'[a] .$$

4.1.4. *Higher derivatives.* Let f be a differentiable function and let $f'[x]$ be its derivative. The derivative of $f'[x]$ (if it has one) is written $f''[x]$ and is called the *second derivative of* f. Similarly, the derivative of a second derivative, if it exists, is written $f'''[x]$ and is called the *third derivative of* f. These repeated derivatives are called *higher-order derivatives*.

If $x[t]$ represents the *position* of an object at time t, then the higher-order derivatives of x have physical interpretations. The second derivative of x is the derivative of $x'[t]$, the *velocity*, and by definition this is the object's *acceleration*. The third derivative of x is defined to be the *jerk* and the fourth derivative is defined to be the *jounce*.

A function that has k successive derivatives is called *k times differentiable*. If in addition the k-th derivative is continuous, then the function is said to be of differentiability class C^k. (This is a stronger condition than having k derivatives. A function that has infinitely many derivatives is called *infinitely differentiable* or *smooth*.

On the real line, every polynomial function is infinitely differentiable. By standard differentiation rules, if a polynomial of degree n is differentiated n times, then it becomes a *constant function*. All of its subsequent derivatives are identically zero. In particular, they exist, so polynomials are smooth functions.

The derivatives of a function f at a point x provide *polynomial approximations to that function near x*. For example, if f is twice differentiable, then

$$f[x+h] \sim f[x] + f'[x]\,h + \frac{1}{2} f''[x]\,h^2\,,$$

in the sense that

$$\lim_{h\to 0} \frac{f[x+h] - f[x] - f'[x]\,h - \frac{1}{2} f''[x]\,h^2}{h^2}\,.$$

If f is infinitely differentiable, then this is the beginning of the *Taylor series for f*.

4.1.5. *Inflection point.* A point where the second derivative of a function changes sign is called an *inflection point*. At an inflection point, the second derivative may be zero, as in the case of the inflection point $x = 0$ of the function $y = x^3$, or it may fail to exist, as in the case of the inflection point $x = 0$ of the function $y = x^{1/3}$. *At an inflection point, a function switches from being a convex function to being a concave function or vice versa.*

4.2. **Chain Rule.** In calculus, the *chain rule* is a formula for computing the derivative of the composition of two or more functions. That is, if f is a function and g is a function, then the chain rule expresses the derivative of the composite function $f \circ g$ in terms of the derivatives of f and g. For example, the chain rule for $(f \circ g)[x] \coloneqq f[g[x]]$ is

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x}\,.$$

In other notation: $(f \circ g)'[x] = f'[g[x]]\,g[x]$.
Example. Analyze the physics of a skydiver who jumps from an aircraft. Assume that t seconds after his jump, his height above sea level in meters is given by $g[t] = 4000 - 4.9t^2$. One model for the atmospheric pressure at a height h is $f[h] = 101325e^{-0.0001h}$. These two equations can be differentiated and combined in various ways to produce the following data:

- $g'[t] = -9.8t$ is the velocity of the skydiver at time t
- $f'[h] = -10.1325e^{-0.0001h}$ is the rate of change in atmospheric pressure with respect to height at the height h and is proportional to the buoyant force on the skydiver at h meters above sea level. (The true buoyant force depends on the volume of the skydiver.)
- $(f \circ g)[t]$ is the atmospheric pressure the skydiver experiences t seconds after his jump.
- $(f \circ g)'[t]$ is the rate of change in atmospheric pressure with respect to time at t seconds after the skydiver's jump and is proportional to the buoyant force on the skydiver at t seconds after his jump.

The chain rule gives a method for computing $(f \circ g)'[t]$ in terms of f' and g'. While it is always possible to directly apply the definition of the derivative to compute the derivative of a composite function, this is usually very difficult. The utility of the chain rule is that it turns a complicated derivative into several easy derivatives.

The chain rule states that, under appropriate conditions, $(f \circ g)'[x] = f'[g[x]]\,g[x]$. In this example, this equals

$$(f \circ g)'[t] = (-10.1325\,e^{-0.0001(4000-4.9t^2)})(-9.8t)\,.$$

In the statement of the chain rule, f and g play slightly different roles because f' is evaluated at $g[t]$, whereas g' is evaluated at t. This is necessary to make the units work out correctly. For example, suppose that we want to compute the rate of change in atmospheric pressure ten seconds after the skydiver jumps. This is $(f \circ g)'[10]$ and has units of Pascals per second. The factor $g'[10]$ in the chain rule is the velocity of the skydiver ten seconds after his jump, and it is expressed in meters per second. $f'[g[10]]$ is the change in pressure with respect to height at the height $g[10]$ and is expressed in Pascals per meter. The product of $f'[g[10]]$ and $g'[10]$ therefore has the correct units of Pascals per second. It is not possible to evaluate f anywhere else. For instance, because the 10 in the problem represents ten seconds, the expression $f'[10]$ represents the change in pressure at a height of ten seconds, which is nonsense. Similarly, because $g'[10 = -98\,\mathrm{m/s}]$, the expression $f'[g'[10]]$ represents the change in pressure at a height of -98 meters per second, which is

also nonsense. However, $g[10]$ is 3020 meters above sea level, the height of the skydiver ten seconds after his jump. This has the correct units for an input to f.

4.3. **Riemann Sums and Definite Integrals.**

4.3.1. *Riemann Sums.* For a function f defined on $[a,b]$, a *partition* P of $[a,b]$ into a collection of subintervals

$$[a = x_0, x_1],\ [x_1, x_2],\ \ldots,\ [x_{n-1}, x_n = b]\,,$$

for each $i = 1, 2, \ldots, n$, a point x_i^* in $[x_{i-1}, x_i]$, the sum

$$\sum_{i=1}^{n} f[x_i^*]\,(x_i - x_{i-1}) = \sum_{i=1}^{n} f[x_i^*]\ \Delta x$$

is called a *Riemann sum* for f determined by the partition P. Let

$$|P| = \max\{x_i - x_{i-1} \text{ for all } i = 1, 2, \ldots, n\}$$

denote the longest length of all the subintervals.

4.3.2. *The Definite Integral.* The *definite integral* of f from a to b is the number

$$\int_a^b f[x]\ \mathrm{d}x = \lim_{|P|\to 0} \sum_{i=1}^{n} f[x_i^*]\ \Delta x$$

provided the limit exists. (We in this case say f is *integrable* on $[a,b]$).

4.3.3. *Computing Riemann Sums.* For a continuous function f on $[a,b]$, then $\int_a^b f[x]\ \mathrm{d}x$ always exists and can be computed by

$$\int_a^b f[x]\ \mathrm{d}x = \lim_{n\to\infty} \sum_{i=1}^{n} f[x_i^*]\ \Delta x$$

for any choice of the x_i^* in $[x_{i-1} - x_i]$ with $\Delta x = (b-a)/n$ and $x_i = a + i\Delta x$. This is, P partitions $[a,b]$ into equal length subintervals, called a *regular partition*.
Example. Compute the Riemann sum $\sum_{i=1}^{n} f[x_i^*]\ \Delta x$ for the function $f[x] = 1/x$ on $[1,6]$ with a regular partition into $n = 5$ subintervals and width $x_i^* = x_i$.
Solution. Note that $a = 1$, $b = 6$ and $n = 5$. Compute then the following

$$\Delta x = \frac{b-a}{n} = \frac{6-1}{5} = 1\,,$$
$$x_i = a + i\Delta x = 1 + i\,,\ \text{for each } i\,,$$
$$f[x_i^*] = f[x_i] = \frac{1}{1+i}\,,\ \text{for each } i\,.$$

Therefore, we have

$$\sum_{i=1}^{5} f[x_i^*]\ \Delta x = \sum_{i=1}^{5} \frac{1}{1+i} = \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}\,.$$

4.3.4. *Properties of the Integral.*

- Linearity with respect to the integrand: if $f_1, f_2, \ldots, f_n$ are integrable in $[a,b]$, then so is $c_1 f_1, c_2 f_2, \ldots, c_n f_n$ for all reals $c_1, c_2, \ldots, c_n$, and
$$\int_a^b \sum_{k=1}^{n} c_k f_k = \sum_{k=1}^{n} c_k \int_a^b f_k\,.$$
- Additivity with respect to the interval of integration: if two of the following three integrals exist, then the third also exists, and we have
$$\int_a^b f + \int_b^c f = \int_a^c f\,.$$
- Invariance under translation: if f is integrable on $[a,b]$, then for every real c we have
$$\int_a^b f[x]\ \mathrm{d}x = \int_{a+c}^{b+c} f[x-c]\ \mathrm{d}x\,.$$

- Expansion and contraction of the interval of integration: if f is integrable on $[a,b]$, then for every real $k \neq 0$ we have
$$\int_a^b f[x]\ \mathrm{d}x = \frac{1}{k}\int_{ka}^{kb} f\left[\frac{1}{k}\right]\ \mathrm{d}x\,.$$
When $k=-1$, the last theorem is called *reflection property*.
- Comparison theorem: if both f and g are integrable on $[a,b]$ and if $g[x] \leq f[x]$ for every x in $[a,b]$, then we have
$$\int_a^b g \leq \int_a^b f\,.$$
In particular, when $g[x]=0$ for every x. In this case, if $f[x] \geq 0$ everywhere on $[a,b]$, then $\int_a^b f[x]\ \mathrm{d}x \geq 0$. In other words, a nonnegative function has a nonnegative integral.

4.4. **Definite Integral.** If f is a continuous function defined on $[a,b]$, if $[a,b]$ is divided into n equal subintervals of width $\Delta x = (b-a)/n$ and if $x_k = a + k\Delta x$ is the right endpoint of the subinterval k, then the definite integral of f from a to b is the number
$$\int_a^b f[x]\ \mathrm{d}x = \lim_{n\to\infty}\sum_{k=1}^{n} f[x_k]\ \Delta x\,.$$

4.5. **Leibniz Integral Rule.** In Calculus, *Leibniz's rule for differentiation under the integral sign* tells us that if we have an integral of the form
$$\int_{y_0}^{y_1} f[x,y]\ \mathrm{d}y\,,$$
then for x in $[x_0,x_1]$ the derivative of this integral is thus expressible
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{y_0}^{y_1} f[x,y]\ \mathrm{d}y\right) = \int_{y_0}^{y_1} f_{,x}[x,y]\,\mathrm{d}y\,,$$
provided that f and its partial derivative $f_{,x}$ are both continuous over a region in the form $[x_0,x_1]\otimes[y_0,y_1]$.

Thus under certain conditions, one may *interchange the integral and partial differential operators*. This important result is particularly useful in the *differentiation of integral transforms*. An example of such is the moment generating function in probability theory, a variation of the Laplace transform, which can be differentiated to generate the moments of a random variable. Whether Leibniz's Integral Rule applies is essentially a question about the interchange of limits.

A Leibniz integral rule for three dimensions is
$$\frac{\mathrm{d}}{\mathrm{d}t}\iint_{\Sigma[t]} F[x,t]\cdot\mathrm{d}A = \iint_{\Sigma[t]} \left(F_{,t}[x,t] + (\nabla\cdot F[x,t])\,v\right)\cdot\mathrm{d}A - \oint_{\partial\Sigma[t]} (v\times F[x,t])\cdot\mathrm{d}s\,,$$
where $F[x,t]$ is a *vector field* at the *spatial position vector* x at *time* t; Σ is a *moving surface* in three-space bounded by the closed curve $\partial\Sigma$; $\mathrm{d}A$ is a *vector element of the surface* Σ; $\mathrm{d}s$ is a *vector element of the curve* $\partial\Sigma$; v is the *velocity vector* of movement of the region Σ; $\nabla\cdot$ is the *vector divergence* and $\times$ is the vector cross product. The *double integrals* are *surface integrals over the surface* Σ, and the *line integral* is over the bounding curve $\partial\Sigma$.

[another form] *Differentiation under the integral sign* is a useful operation in calculus. Formally it can be stated as follows:
Theorem. Let $f[x,t]$ be a function such that both $f[x,t]$ and its partial derivative $f_{,x}[x,t]$ are continuous in t and x in some region of the (x,t)-plane, including $a[x] \leq t \leq b[x]$, $x_0 \leq x \leq x_1$. Also suppose that the functions $a[x]$ and $b[x]$ are both continuous and both have continuous derivatives for $x_0 \leq x \leq x_1$. Then $x_0 \leq x \leq x_1$:
$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{a[x]}^{b[x]} f[x,t]\ \mathrm{d}t\right) = f[x,b[x]]\,b'[x] - f[x,a[x]]\,a'[x] + \int_{a[x]}^{b[x]} f_{,x}[x,t]\ \mathrm{d}t\,.$$

This formula is the general form of the Leibniz integral rule and can be derived using the fundamental theorem of calculus. The [second] fundamental theorem of calculus is just a particular case of the above formula, for $a[x] = a$, a constant, $b[x] = x$ and $f[x,t] = f[t]$.

The following three basic theorems on the interchange of limits are essentially equivalent:

- the interchange of a derivative and an integral (differentiation under the integral sign; *i.e.*, Leibniz integral rule),
- the change of order of partial derivatives,
- the change of order of integration (integration under the integral sign; *i.e.*, Fubini's theorem).

The Leibniz integral rule can be extended to multidimensional integrals. In two and three dimensions, this rule is better known from the field of fluid dynamics as the *Reynolds transport theorem*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{D[t]} \phi[x,t]\ \mathrm{d}V = \int_{D[t]} \phi_{,t}[x,t]\ \mathrm{d}V + \int_{\partial D[t]} \phi[x,t]\, v_b \cdot \mathrm{d}\Sigma\,,$$

where $\phi[x,t]$ is a *scalar function*, $D[t]$ and $\partial D[t]$ denote a *time-varying connected region of* $\mathcal{R}^3$ *and its boundary*, v_b is the *Eulerian velocity of the boundary* and $\mathrm{d}\Sigma = n\mathrm{d}S$ is the *unit normal component of the surface element*.

4.6. **Lagrangian and Eulerian Specification of the Flow Field.** In fluid dynamics and finite-deformation plasticity the *Lagrangian specification of the flow field* is a way of looking at fluid motion where the

> observer follows an individual fluid parcel as it moves through space and time.

Plotting the position of an individual parcel through time gives the *pathline* of the parcel. This can be visualized as sitting in a boat and drifting down a river.

The *Eulerian specification of the flow field* is a way of looking at fluid motion that focuses on

> specific locations in the space through which the fluid flows as time passes.

This can be visualized by sitting on the bank of a river and watching the water pass the fixed location.

The Lagrangian and Eulerian specifications of the flow field are sometimes loosely denoted as the Lagrangian and Eulerian frame of reference. However, in general both the Lagrangian and Eulerian specification of the flow field can be applied in any observer's frame of reference, and in any coordinate system used within the chosen frame of reference.

4.6.1. *Description.* In the Eulerian specification of the flow field, the flow quantities are depicted as a function of position x and time t. Specifically, the flow is described by a function

$$v[x,t]$$

giving the flow *velocity at position* x *at time* t.

On the other hand, in the Lagrangian specification, individual *fluid parcels* are followed through time. The fluid parcels are labelled by some (time-independent) *vector field* a. (Often, a is chosen to be the *center of mass of the parcels at some initial time* t_0. It is chosen in this particular manner to account for the possible changes of the shape over time. Therefore, the center of mass is a good parametrization of the velocity v of the parcel.) In the Lagrangian description, the flow is described by a function

$$X[a,t]$$

giving the position of the parcel labeled a at time t.

The two specification are related as follows:

$$v[X[a,t]] = X_{,t}[a,t]\ .$$

because both sides describe the velocity of the parcel labeled a at time t.

Within a chosen coordinate system, a and x are referred to as the Lagrangian coordinates and Eulerian coordinates of the flow.

4.6.2. *Substantial Derivative.* The Lagrangian and Eulerian specifications of the kinematics and dynamics of the flow field are related by the *substantial derivative* (*aka*, the Lagrangian derivative, convective derivative, material derivative, particle derivative and so on).

Suppose we have a flow field with Eulerian specification v, and we are also given some function (vector field or scalar field) $f[x,t]$ defined for every position x and every time t. (For instance, f could be an external force field – vector field – or temperature – scalar field.) Now one might ask about the total rate of change of f experienced by a specific flow parcel. This can be computed as

$$\frac{\mathrm{D}f}{\mathrm{D}t} = \frac{\partial f}{\partial t} + (v\cdot\nabla)f \implies \dot{f} = f_{,t} + (v\cdot\nabla)f\,,$$

(where ∇ denotes the gradient with respect to x and the operator $v\cdot\nabla$ is to be applied to each component of f.) This tells us that

> the total rate of change of the function f as the fluid parcels moves through a flow field described by its Eulerian specification v is equal to the sum of the local rate of change and the convective rate of change of f.

This is a consequence of the *chain rule* since we are differentiating the function $f[X[a,t],t]$ with respect to t; *viz.*, let $\phi[x,t]$ be a scalar function where x in turns depend on t: $x = [x[t], y[t], z[t]]$. Then, by applying the chain rule, we have

$$\frac{\mathrm{D}\phi}{\mathrm{D}t} = \frac{\partial\phi}{\partial t} + \frac{\partial\phi}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial\phi}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial\phi}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t} = \phi_{,t} + \dot{x}\phi_{,x} + \dot{y}\phi_{,y} + \dot{z}\phi_{,z}\,.$$

Noting that $v = [\dot{x},\dot{y},\dot{z}]$ is the velocity vector, then, we find

$$\mathrm{D}_t\phi = \dot{\phi} = \phi_{,t} + v\cdot\nabla\phi\,. \qquad \square$$

Physically, the last equation describes the change rate of a scalar quantity with a parcel of fluid moving along the fluid, Lagrange description of fluid motion. The term $f_{,t}$ is the Eulerian description of fluid motion, so the relationship between the two descriptions is

$$\phi_{,t} = \mathrm{D}_t\phi - v\cdot\nabla\phi\,. \qquad \square$$

4.6.3. *Material Derivative.* In continuum mechanics, the *material derivative* describes the time rate of change of some physical quantity (like heat or momentum) for a material element subjected to a space-and-time-dependent velocity field. The material derivative can serve as a link between Eulerian and Lagrangian descriptions of continuum deformation.

For example, in fluid dynamics, take the case that the velocity field under consideration is the flow velocity itself, and the quantity of interest is the temperature of the fluid. Then the material derivative describes the temperature evolution of a certain fluid parcel in time, as it is being *moved along its pathline* (trajectory) while following the fluid flow.

There are many other names for this operator, including:

- convective derivative,
- advective derivative,
- substantive derivative,
- substantial derivative,
- Lagrangian derivative,
- Stokes derivative,
- particle derivative,
- hydrodynamic derivative,
- derivative following the motion,
- *total derivative.*

The material derivatives of a *scalar field* $\phi[x[t],t]$ and a *vector field* $u[x[t],t]$ are defined as:

$$\begin{cases}\mathrm{D}_t\phi = \phi_{,t} + v\cdot\nabla\phi\,,\\ \mathrm{D}_t u = u_{,t} + v\cdot\nabla u\,.\end{cases}$$

where the distinction is that $\nabla\phi$ is the *gradient of a scalar*, while ∇u is the *covariant derivative of a vector*. In case of the material derivative of a vector field, the term $v\cdot\nabla u$ can both be interpreted as $v\cdot(\nabla u)$ involving the tensor derivative of u, or as $(v\cdot\nabla)u$, leading to the same result.

Confusingly, the term convective derivative is both used for the whole material derivative $\mathrm{D}_t\phi$ or $\mathrm{D}_t u$, and for only the spatial rate of change part, $v\cdot\nabla\phi$ or $v\cdot\nabla u$ respectively. For that case, the convective derivative only equals D/t for time independent flows.

These derivatives are physical in nature and describe the transport of a scalar or vector quantity in a velocity field $v[x,t]$. The effect of the time independent terms in the definitions are for the scalar and vector case respectively known as *advection* and *convection*.

4.7. **Partial Derivative.** In mathematics, a *partial derivative* of a function of several variables is its derivative with respect to one of those variables, with the others held constant (that is the rate of change is taken along one of the coordinate curves, all other coordinates being constant, as opposed to the total derivative, in which all variables are allowed to vary). Partial derivatives are used in vector calculus and differential geometry.

The partial derivative of a function f with respect to the variable x is variously denoted by

$$f'_x, f_x, f_{,x}, \partial_x f \text{ or } \frac{\partial f}{\partial x}\,.$$

4.8. **Directional Derivative.** In mathematics, the *directional derivative* of a multivariate differentiable function along a given vector v at a given point x intuitively represents the instantaneous rate of change of the function, moving through x with a velocity specified by v. It therefore generalizes the notion of a partial derivative, in which the rate of change is taken along one of the coordinate curves, all other coordinates being constant.

The directional derivative is a special case of the Gâteaux derivative.

4.8.1. *Definition.* The directional derivative of a scalar function $f[x] = f[x_1, x_2, \ldots, x_n]$ along a vector $v = [v_1, v_2, \ldots, v_n]$ is the function defined by the limit

$$\nabla_v f[x] = \lim_{h\to 0} \frac{f[x+hv] - f[x]}{h}\,,$$

where h is a scalar.

If the function f is differentiable at x, then the directional derivative exists along any vector v, and one has

$$\nabla_v f[x] = \nabla f[x] \cdot v\,,$$

where the ∇ on the right denotes the gradient (scalar part of the geometric derivative) and $\cdot$ is the dot product. At any point x, the directional derivative of f intuitively represents the rate of change in f moving at a rate and direction given by v at the point x.

Some authors define the directional derivative to be with respect to the vector v after normalization, thus ignoring its magnitude. In this case, one has

$$\nabla_v f[x] = \lim_{h\to 0} \frac{f[x+hv] - f[x]}{h|v|}\,,$$

or in case f is differentiable at x,

$$\nabla_v f[x] = \nabla f[x] \cdot \frac{v}{|v|}\,,$$

This definition has several disadvantages: it only applies when the norm of a vector is defined and the vector is not null. It is also incompatible with notation used elsewhere in mathematics and physics and engineering, and so should *not* be used.

4.8.2. *Notation.* Directional derivatives can be also denoted by:

$$\nabla_v f[x] \sim \frac{\partial f[x]}{\partial |v|} \sim f'_v[x] \sim \mathrm{D}_v f[x] \sim v \cdot \nabla f[x]\ .$$

4.8.3. *Properties.* Many of the familiar properties of the ordinary derivative hold for the directional derivative. These include, for any functions (vector or scalar) f and g defined in a neighborhood of, and differentiable at, p:

(1) The sum rule: $\nabla_v fg = \nabla_v f + \nabla_v g$.
(2) The sum rule: for any constant c, $\nabla_v cf = c\nabla_v f$.
(3) The product rule (or Leibniz rule): $\nabla_v fg = \nabla_v f\, g + f\nabla_v g = g\nabla_v f + f\nabla_v g$. (The rearrangement of $\nabla_v f\, g$ for $g\nabla_v f$ is possible when *not* using geometric calculus (*i.e.*, when using vector calculus), for the geometric product is not generally commutative; *viz.*, $\nabla_v f\, g \neq g\nabla_v f$. In any case, geometric calculus or vector calculus, the rearrangement is possible if both functions are scalar functions.)

(4) The chain rule: If g is differentiable at p and h is differentiable at $g[p]$, then $\nabla_v h \circ g[p] = h'[g[p]] \nabla_v g[p]$.

4.9. **Vector Area.** In geometry, for a finite planar surface of scalar area $|S|$, the *vector area* S is defined as a vector whose magnitude is $|S|$ and whose direction is perpendicular to the plane, as determined by the right hand rule on the rim:

$$S = n|S| \, .$$

For an orientable surface S of a set $|S_i|$ of flat facet areas, the vector area of the surface is given by

$$S = \sum_i n_i |S_i| \, ,$$

where n_i is the unit normal vector to the area $|S_i|$.

For bounded, oriented curved surfaces that are sufficiently well-behaved, we can still define vector area. First, we split the surface into infinitesimal elements, each of which is effectively flat. For each infinitesimal element of are, we have an area vector, also infinitesimal:

$$\mathrm{d}S = n \, \mathrm{d}|S| \, ,$$

where n is the local unit vector perpendicular to $\mathrm{d}S$. Integrating gives the vector area for the surface:

$$S = \int \mathrm{d}S \, .$$

For a curved or faceted surface, the vetor area is smaller in magnitude than the area. As an extreme example, a closed surface can possess arbitrarily large area, but its vector area is necessarily zero. Surfaces that share a boundary may have very different areas, but they must have the same vector area – the vector area is entirely determined by the boundary. These are consequences of Stokes theorem.

The concept of an area vector simplifies the equation for determining the flux through the surface. Consider a planar surface in a uniform field. The flux can be written as the dot product of the field and area vector. This is much simpler than multiplying the field strength by the surface area and the cosine of the angle between the field and the surface normal.

Projection of area onto planes: the projected area onto (for example) the $x - y$ plane is equivalent to the z-component of the vector area and is given by

$$S_z = |S| \, \cos[\theta] \, ,$$

where θ is the angle between the plane normal and the z-axis.

4.10. **Line Integral.** In mathematics, a *line integral* (sometimes called a *path integral*, contour integral or curve integral); not to be confused with calculating arc length using integration) is an integral where the function to be integrated is evaluated along a curve.

The function to be integrated may be a scalar field or a vector field. The value of the line integral is the sum of values of the field at all points on the curve, weighted by some scalar function on the curve (commonly arc length or, for a vector field, the scalar product of the vector field with a differential vector in the curve). This weighting distinguishes the line integral form simpler integrals defined on intervals. Many simple formulae in physics (for example, $w = f \cdot s$) have natural continuous analogs in terms of line integrals ($w = \oint_C f \cdot \mathrm{d}S$). The line integral finds the work done on an object moving through an electric or gravitational field, for example.

4.10.1. *Vector Calculus.* In qualitative terms, a line integral in vector calculus can be thought of as a measure of the total effect of a given field along a given curve. More specifically, the line integral over a scalar field can be interpreted as the area under the field carved out by a particular curve. This can be visualized as the surface created by $z = f[x, y]$ and a curve $\mathcal{C}$ in the $x - y$ plane. The line integral of f would be the area of the "curtain" created when the points of the surface that are directly over $\mathcal{C}$ are carved out.

4.10.2. *Line Integral of a Scalar Field.* Definition: for some scalar field $f : \mathcal{U} \subset \mathcal{R}^n \to \mathcal{R}$, the line integral along a piecewise smooth curve $\mathcal{C} \subset \mathcal{U}$ is deea

$$\int_C f \, \mathrm{d}s = \int_a^b f[r[t]] \, |r'[t]| \, \mathrm{d}t \,,$$

where $r : [a, b] \to \mathcal{C}$ is an arbitrary bijective parametrization of $\mathcal{C}$ such that $r[a]$ and $r[b]$ give the endpoints of $\mathcal{C}$ and $a < b$.

The function f is called the integrand, the curve $\mathcal{C}$ is the domain of integration and the symbol $\mathrm{d}s$ may be intuitively interpreted as an elementary arc length. Line integrals of scalar fields over a curve $\mathcal{C}$ do not depend on the chosen parametrization r of $\mathcal{C}$.

Geometrically, when the scalar field f is defined over a plane, its graph is a surface $z = f[x, y]$ in space and the line integral gives the (signed) cross-sectional area bounded by the curve $\mathcal{C}$ and the graph of f.

4.10.3. *Line Integral of a Vector Field.* Definition: for a vector field $f : \mathcal{U} \subset \mathcal{R}^n \to \mathcal{R}^n$, the line integral along a piecewise smooth curve $\mathcal{C} \subset \mathcal{U}$, in the direction of r, is defined is

$$\int_C f[r] \cdot \mathrm{d}r = \int_a^b f[r[t]] \cdot r'[t] \; \mathrm{d}t \,,$$

where $\cdot$ is the dot product and $r : [a, b] \to \mathcal{C}$ is an arbitrary bijective parametrization of $\mathcal{C}$ such that $r[a]$ and $r[b]$ give the endpoints of $\mathcal{C}$ and $a < b$.

A line integral of a scalar field is thus a line integral of a vector field were the vectors are always tangential to the line.

Line integrals of vector fields are independent of the parametrization r in absolute value, but they do depend on its orientation. Specifically, a reversal in the orientation of the parametrization changes the sign of the line integral.

4.10.4. *Path Independence.* If a vector field f is the gradient of a scalar field g (*i.e.*, f is conservative); *i.e.*, $\operatorname{grad} g = f$, then the derivative of the composition of g and $r[t]$ is

$$\frac{\mathrm{d}g[r[t]]}{\mathrm{d}t} = \operatorname{grad} g[r[t]] \cdot r'[t] = f[r[t]] \cdot r'[t] \;,$$

which happens to be the integrand for the line integral of f on $r[t]$. It follows that, given a path C, then

$$\int_C f[r] \cdot \mathrm{d}r = \int_a^b f[r[t]] \cdot r'[t] \; \mathrm{d}t = \int_a^b \frac{\mathrm{d}g[r[t]]}{\mathrm{d}t} \, \mathrm{d}t = g[r[b]] - g[r[a]] \;.$$

In other words, the integral of f over C depends solely on the values of g in the points $r[b]$ and $r[a]$ and is thus independent of the path between them. For this reason, a line integral of a conservative vector field is called *path independent*.

4.11. **Surface Integral.** A *surface integral* is a definite integral taken over a surface. It can be thought of as the double integral analog of the line integral. Given a surface, one may integrate over its scalar fields (that is, functions which return scalars as values), and vector field (that is, functions which return vectors as values).

Surface integrals have applications in physics, particularly with the classical theory of electromagnetism.

4.11.1. *Surface Integrals of Scalar Fields.* To find an explicit formula for the surface integral, we need to parametrize the surface of interest, S, by considering a system of curvilinear coordinates on S, like the latitude and longitude on a sphere. Let such a parametrization be $x[s, t]$, where $[s, t]$ varies in some region T in the plane. Then, the surface integral is given by

$$\int_S f \, \mathrm{d}S = \iint_T f[x[s,t]] \left| \frac{\partial x}{\partial s} \times \frac{\partial x}{\partial t} \right| \mathrm{d}s\mathrm{d}t \,,$$

where the expression between bars on the right-hand side is the magnitude of the cross product of the partial derivatives of $x[s, t]$ and is known as the surface element.

4.11.2. *Surface Integrals of Vector Fields.* Consider a vector field v on S, that is, for each x in S, $v[x]$ is a vector.

The surface integral can be defined component-wise according to the definition of te surface integral of a scalar field; the result is a vector. This applies for instance in the expression of the electric field at some fixed point due to an electrically charged surface, or the gravity at some fixed point due to a sheet of material.

Alternatively, if we integrate the normal component of the vector field, the result as a scalar. Imagine that we have a fluid flowing through S, such that $v[x]$ determines the velocity of the fluid at x. The flux is defined as the quantity of fluid flowing through S in unit amount of time.

This illustration implies that if the vector field is tangent to S at each point, then the flux is zero, because the fluid just flows in parallel to S and neither in nor out. This also implies that if v does not just flow along S, that is, if v has both a tangential and a normal component, then only the normal component contributes to the flux. Based on this reasoning, to find the flux, we need to take the dot product of v with the unit surface normal to S at each point, which will give us a scalar field and integrate the obtained field as above. We find the formula

$$\int_S v \cdot \mathrm{d}S = \int_S v \cdot n \, \mathrm{d}S = \iint_T v[x[s,t]] \cdot \left(\frac{\partial x}{\partial s} \times \frac{\partial x}{\partial t} \right) \mathrm{d}s\mathrm{d}t \, .$$

The cross product on the right-hand side of this expression is a surface normal determined by the parametrization. This formula *defines* the integral on the left (note the dot and the vector notation for the surface element).

4.12. **Volume Integral.** A *volume integral* refers to an integral over a 3-dimensional domain.

A volume integral is a triple integral of the constant function 1, which gives the volume of the region $\mathcal{D}$. That is, the integral

$$V[\mathcal{D}] = \iiint_{\mathcal{D}} \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \, .$$

It can also mean a triple integral within a region $\mathcal{D}$ in $\mathcal{R}^3$ of a function $f[x, y, z]$ and is usually written as

$$\iiint_D f[x, y, z] \; \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \, .$$

4.13. **Note on Notation: Surface integrals in terms of double-integrals.** Surface integrals can be calculated in Cartesian coordinates. This is the case usually when the surface S does not have any symmetries, such as rotational symmetry. For this the surface S is projected onto Cartesian plane $x_1 \wedge x_2$X1X2 which gives a domain like S (Fig.). Then,

$$\int_S A \cdot n \, \mathrm{d}S = \iint_S A \cdot n \, \frac{\mathrm{d}x_1 \mathrm{d}x_2}{n \cdot \gamma_3} \, .$$

The transformation between the area element $\mathrm{d}S$ and its projection $\mathrm{d}x_1\mathrm{d}x_2$ is given by

$$\mathrm{d}x_1\mathrm{d}x_2 = (n\mathrm{d}S) \cdot \gamma_3 = (n \cdot \gamma_3) \, dS \, ,$$

where n is the unit normal vector to $\mathrm{d}S$ and γ_3 is the unit normal vector to $\mathrm{d}x_1\mathrm{d}x_2$.

4.14. **Arc Length.** Determining the *arc length of an irregular arc segment* is also called rectification of a curve. Historically, many methods were used for specific curves. The advent of calculus led to a general formula that provides closed-form solutions in some cases.

4.14.1. *General Approach.* a curve in the plane can be approximated by connecting a finite number of points on the curve using line segments to create a polygonal path. Since it is straightforward to calculate the length of each linear segment (using the Pythagorean theorem in Euclidean space, for instance), the total length of the approximation can be found by summing the lengths of each linear segment.

Polygonal approximations are linearly dependent on the curve in a few select cases. One of these cases is when the curve is simply a point function as is its polygonal approximation. Another case where the polygonal approximation is linearly dependent on the curve is

when the curve is linear. This would mean the approximation is also linear and the curve and its approximation overlap. Both of these two circumstances result in an eigenvalue equal to one. There are also a set of circumstances where the polygonal approximation is still linearly dependent but the eigenvalue is equal to zero. This case is a function with petals where all points for the polygonal approximation are at the origin.

If the curve is not already a polygonal path, better approximations to the curve can be obtained by following the shape of the curve increasingly more closely. The approach is to use an increasingly larger number of segments of smaller lengths. The lengths of the successive approximations do not decrease and will eventually keep increasing – possibly indefinitely, but for smooth curves this will tend to a limit as the lengths of the segments get arbitrarily small.

For some curves there is a smallest number L that is an upper bound on the length of any polygonal approximation. If such a number exists, then the curve is said to be *rectifiable* and the curve is defined to have *arc length* L.

4.14.2. *Definition.* Let C be a curve in Euclidean (or, more generally, a metric) space $\mathcal{X} = \mathcal{R}^n$, so C is the image of a continuous function $f : [a, b] \to \mathcal{X}$ of the interval $[a, b]$ into $\mathcal{X}$.

From a partition $a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$ of the interval $[a, b]$, we obtain a finite collection of points $f[t_0], f[t_1], \ldots, f[t_{n-1}], f[t_n]$ on the curve C. Denote the distance from $f[t_i]$ to $f[t_{i+1}]$ by $d[f[t_i], f[t_{i+1}]]$, which is the length of the line segment connecting the two points.

The *arc length* L of C is then defined to be

$$L[C] = \sup_{a=t_0,\ldots,t_n=b} \sum_{i=0}^{n-1} d[f[t_i], f[t_{i+1}]] \ ,$$

where the supremum is taken over all possible partitions of $[a, b]$ and n is unbounded.

The arc length L is either finite or infinite. If $L < \infty$, then we say that C is *rectifiable*, and is *non-rectifiable* otherwise. This definition of arc length does not require that C be defined by a differentiable function. In fact, in general, the notion of differentiabibity is not defined on a metric space.

A curve may be parametrized in many ways. Suppose C also has the parametrization $g : [c, b] \to \mathcal{X}$. Provided that f and g are injective, there is a continuous monotone function S from $[a, b]$ to $[c, d]$ so that $g[S[t]] = f[t]$ and an inverse function S^{-1} from $[c, d]$ to $[a, b]$. It is clear that any sum of the form

$$\sum_{i=0}^{n-1} d[f[t_i], f[t_{i+1}]]$$

can be made equal to a sum of the form

$$\sum_{i=0}^{n-1} d[g[u_i], g[u_{i+1}]]$$

by taking $u_i = S[t_i]$ and similarly a sum involving g can be made equal to a sum involving f. So the arc length is an intrinsic property of the curve, meaning that it does not depend on the choice of parametrization. The definition of arc length for the curve is analogous to the definition of the total variation of a real-valued function.

4.14.3. *Finding Arc Lengths by Integrating.* Consider a real function f such that $y = f[x]$ and $f'[x] = \mathrm{d}y/\mathrm{d}x$ (its derivative with respect to x) are continuous on $[a, b]$. The length s of the part of the graph of f between $x = a$ and $x = b$ can be found as follows:

Consider an infinitesimal part of the curve $\mathrm{d}s$ (or consider this as a limit in which the change in s approaches $\mathrm{d}s$). According to Pythagoras' theorem $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2$, from which

$$\frac{\mathrm{d}s^2}{\mathrm{d}x^2} = 1 + \frac{\mathrm{d}y^2}{\mathrm{d}x^2} \implies \mathrm{d}s = \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x \,, \implies s = \int_a^b \sqrt{1 + (f'[x])^2}\,\mathrm{d}x \,.$$

If a curve is defined parametrically by $x = X[t]$ and $y = Y[t]$, then its arc length between $t = a$ and $t = b$ is

$$s = \int_a^b \sqrt{(X'[t])^2 + (Y'[x])^2}\,\mathrm{d}t\,.$$

A useful mnemonic is

$$s = \lim \sum_a^b \sqrt{\Delta x^2 + \Delta y^2} = \int_a^b \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2} = \int_a^b \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t\,.$$

4.15. **Flow and Flux.** In the various subfields of physics, there exist two common usages of the term flux, both with rigorous mathematical frameworks. A simple and ubiquitous concept throughout physics and applied mathematics is the flow of a physical property in space, frequently also with time variation. It is the basis of the field concept in physics and mathematics, with two principle applications: in transport phenomena and surface integrals. The terms "flux", "current", "flux density", "current density", can sometimes be used interchangeably and ambiguously, though the terms used below match those of the contexts in the literature.

In transport phenomena (heat transfer, mass transfer and fluid dynamics), *flux*, j, is defined as the rate of flow of a property per unit area, which has the dimensions $\dim j = [Q]/([T]\,[A])$, where $[Q]$ refers to the dimensions of the property, $[T]$ to the dimension of time and $[A]$ to the dimension of area. For example, the magnitude of a river's current, *i.e.* the amount of water that flows through a cross-section of the river each second, or the amount of sunlight that lands on a patch of ground each second is also a kind of flux.

4.15.1. *General Mathematical Definition (Transport).* In this definition, flux is generally a *vector* due to the widespread and useful definition of *vector area*, although there are some cases where only the magnitude is important (like in number fluxes). The frequent symbol is j (or J), and a definition for *scalar flux of physical quantity* q is the limit:

$$j = \lim_{\Delta A \to 0} \frac{\Delta I}{\Delta A} = \frac{\mathrm{d}I}{\mathrm{d}A}\,,$$

where

$$I = \lim_{\Delta t \to 0} \frac{\Delta q}{\Delta t} = \frac{\mathrm{d}q}{\mathrm{d}t}$$

is the *flow of quantity* q per unit time t and A is the area through which the quantity flows.

For vector flux, the surface integral of j over a surface S, followed by an integral over the time duration t_1 to t_2, gives the total amount of the property flowing through the surface in that time $(t_2 - t_1)$:

$$q = \int_{t_1}^{t_2} \iint_S j \cdot n\,\mathrm{d}A\mathrm{d}t\,.$$

The area required to calculate the flux is real or imaginary, flat or curved, either as a cross-sectional area or a surface. The *vector area* A is a combination of the magnitude of the area through which the mass passes through, $|A|$, and a unit vector normal to the area, n. The relation is $A = |A|n$.

If the flux j passes through the area at an angle θ to the area normal n, then

$$j \cdot n = j\cos\theta\,,$$

where $\cdot$ is the dot product of the unit vectors. This is, the component of flux passing *through* the surface (*i.e.*, *normal* to it) is $j\cos\theta$, while the component of flux passing *tangential* to the area is $j\sin\theta$, but there is *no* flux actually passing through the area in the *tangential* direction. The *only* component of flux passing normal to the area is the cosine component.

4.15.2. *Transport fluxes.* Eight of the most common forms of flux from the transport phenomena literature are defined as follows:

- *Momentum flux*: the rate of transfer of momentum across a unit area: $\mathrm{N\,s/m^2\,s}$. (Newton's law of viscosity)
- *Heat flux*: the rate of heat flow across a unit area: $\mathrm{J/m^2\,s}$. (Fourier's law of conduction) (This definition of heat flux fits Maxwell's original definition.)
- *Diffusion flux*: the rate of movement of molecules across a unit area: $\mathrm{mol/m^2\,s}$. (Fick's law of diffusion)
- *Volumetric flux*: the rate of volume flow across a unit area: $\mathrm{m^3/m^2\,s}$. (Darcy's law of groundwater flow)
- *Mass flux*: the rate of mass flow across a unit area: $\mathrm{kg/m^2\,s}$. (Either an alternate form of Fick's law that includes the molecular mass or an alternate form of Darcy's law that includes the density.)
- *Radiative flux*: the amount of energy transferred in the form of photons at a certain distance from the source per steradian per second $\mathrm{J/m^2\,s}$. (Used in astronomy to determine the magnitude and spectral class of a star. Also acts as a generalization of heat flux, which is equal to the radiative flux when restricted to the infrared spectrum.)
- *Energy flux*: the rate of transfer of energy through a unit area: $\mathrm{J/m^2\,s}$. (The radiative flux and heat flux are specific cases of energy flux.)
- *Particle flux*: the rate of transfer of particles through a unit area: $\mathrm{N_p/m^2\,s}$, where $\mathrm{N_p}$ represents the number of particles.

These fluxes are vectors at each point in space, and have a definite magnitude and direction. Also, one can take the divergence of any of these fluxes to determine the accumulation rate of the quantity in a control volume around a given point in space. For incompressible flow, the divergence of the volume flux is zero.

4.15.3. *Chemical diffusion.* As mentioned above, chemical molar flux of a component A in an isothermal, isobaric system is defined in Fick's law of diffusion as:

$$j_{\mathrm{A}} = -D_{\mathrm{AB}} \nabla c_{\mathrm{A}} \,,$$

where D_{AB} is the diffusion coefficient ($\mathrm{m^2/s}$) of component A diffusing through component B, c_{A} is the concentration ($\mathrm{mol/m^3}$) of component A. This flux has units of $\mathrm{mol/(m^2\,s)}$ and fits Maxwell's original definition of flux.

For dilute gases, kinetic molecular theory relates the diffusion coefficient D to the particle density $n = N/V$, the molecular mass M, the collision cross section σ and the absolute temperature T by

$$D = \frac{3}{2n\sigma} \sqrt{\frac{k_{\mathrm{b}} T}{\pi M}} \,,$$

where the second factor is the mean free path and the square root (with Boltzmann's constant k_{b}) is the mean velocity of the particles.

In turbulent flows, the transport by eddy motion can be expressed as a grossly increased diffusion coefficient.

5. Grad, Div, Curl and All That

5.1. **Scalar and Vector Fields.** A scalar field is a scalar-valued function of the position vector; *i.e.*, a scalar field assigns a scalar to every point of a region in space. It's called scalar field because it returns a scalar when a vector is plugged into it. For instance, the temperature distribution of a body.

A vector field is a vector-valued function of the position vector; *i.e.*, a vector field assigns a vector to every point of a region in space. It's called vector field because it returns a vector when a vector is plugged into it. For instance, the velocity vector of a fluid.

5.2. **Gradient.** The *gradient of a scalar field* is a *vector field* that points in the direction of the greatest rate of increase of the scalar field and whose magnitude is that rate of increase. In simple terms,

> the variation in space of any quantity can be represented (*e.g.*, graphically) by a slope. The gradient represents the steepness and direction of that slope.

A generalization of the gradient for functions on a Euclidean space that have values in another Euclidean space is the Jacobian. A further generalization for a function from one Banach space to another is the Fréchet derivative.

5.2.1. *Interpretation.* Consider a room in which the temperature is given by a scalar field, T, so at each point $[x, y, z]$ the temperature is $T[[x, y, z]]$. (We will assume that the temperature does not change over time.) At each point in the room, the gradient of T at that point will show the direction the temperature rises most quickly. The magnitude of the gradient will determine how fast the temperature rises in that direction.

Consider a surface whose height above sea level at a point $[x, y]$ is $H[[x, y]]$. The gradient of H at a point is a vector pointing in the direction of the steepest slope or grade at that point. The steepness of the slope at that point is given by the magnitude of the gradient vector.

The gradient can also be used to measure how a scalar field changes in other directions, rather than just the direction of greatest change, by taking a dot product. Suppose that the steepest slope on a hill is 40%. If a road goes directly up the hill, then the steepest slope on the road will also be 40%. If, instead, the road goes around the hill at an angle, then it will have a shallower slope. For example, if the angle between the road and the uphill direction, projected onto the horizontal plane, is 60°, then the steepest slope along the road will be 20%, which is 40% times the cosine of 60°.

This observation can be mathematically stated as follows. If the hill height function H is differentiable, then the gradient of H "dotted" with a unit vector gives the slope of the hill in the direction of the vector. More precisely,

> when H is differentiable, the dot product of the gradient of H with a given unit vector is equal to the *directional derivative of H in the direction of that unit vector*.

5.2.2. *Definition.* The *gradient (or gradient vector field) of a scalar function* $f\left[x^1, x^2, \ldots, x^n\right]$ is denoted ∇f, where ∇ (the nabla symbol) denotes the vector differential operator, *del*. The notation grad f is also commonly used for the gradient. The gradient of f is defined as the unique vector field whose dot product with any vector v at each point x is the directional derivative of f along v. That is,

$$(\nabla f[x]) \cdot v = \nabla_v f[x]$$

In a rectangular coordinate system, the gradient is the vector field whose components are the partial derivatives of f:

$$\nabla f = \operatorname{grad} f = \gamma^k \partial_k f = \frac{\partial f}{\partial x^1}\gamma^1 + \cdots + \frac{\partial f}{\partial x^n}\gamma^n \,,$$

where the $\left\{\gamma^i\right\}$ are the *orthogonal unit vectors* pointing in the coordinate directions. When a function also depends on a parameter such as time, the gradient often refers simply to the vector of its spatial derivatives only.

5.2.3. *Linear Approximation to a Function.* The gradient of a function f from the Euclidean space $\mathcal{E}^n$ to $\mathcal{R}$ at any particular point x_0 in $\mathcal{E}^n$ characterizes *the best linear approximation to f at x_0*. The approximation is as follows

$$f[x] \sim f[x_0] + (\nabla f)_{x_0} \cdot (x - x_0) \ ,$$

for x close to x_0, where $(\nabla f)_{x_0}$ is the gradient of f computed at x_0, and the dot denotes the dot product on $\mathcal{E}^n$. This equation is equivalent to the first two terms in the multi-variable Taylor Series expansion of f at x_0.

5.2.4. *Properties of the Gradient.*

- Linearity: The gradient is linear in the sense that if f and g are two real-valued functions differentiable at the point $a \in \mathcal{E}^n$ and α and β are two constants, then $\alpha f + \beta g$ is differentiable at a. Moreover

$$\nabla (\alpha f + \beta g)[a] = \alpha \nabla f[a] + \beta \nabla g[a] \ .$$

- Product rule: If f and g are real-valued functions differentiable at a point $a \in \mathcal{E}^n$, then the product rule asserts that the product $(fg)[x] = f[x]\, g[x]$ of the functions f and g is differentiable at a and

$$\nabla (fg)[a] = f[a]\, \nabla g[a] + g[a]\, \nabla f[a] \ .$$

- Chain rule: Suppose that $f : \mathcal{A} \to \mathcal{B}$ is a real-valued function defined on a subset $\mathcal{A}$ of $\mathcal{E}^n$, and that f is differentiable at a point a. There are two forms of the chain rule applying to the gradient. First, suppose that the function g is a parametric curve; that is, a function $g : \mathcal{I} \to \mathcal{E}^n$ maps a subset $\mathcal{I} \subset \mathcal{R}$ into $\mathcal{E}^n$. If g is differentiable at a point $c \in \mathcal{I}$ such that 1, then

$$(f \circ g)'[c] = \nabla f[a] \cdot g'[c] \ ,$$

  where $\circ$ is the composition operator. More generally, if instead $\mathcal{I} \subset \mathcal{E}^k$, then the following holds:

$$\nabla (f \circ g)[c] = (\mathrm{D}g[c])^T (\nabla f[a]) \, ,$$

  where (Dg)T denotes the transpose *Jacobian matrix*.

  For the second form of the chain rule, suppose that $h : \mathcal{I} \to \mathcal{R}$ is a real valued function on a subset $\mathcal{I}$ of $\mathcal{R}$, and that h is differentiable at the point $f[a] \in \mathcal{I}$. Then,

$$\nabla (h \circ f)[a] = h'[f[a]]\, \nabla f[a] \ .$$

5.2.5. *Conservative Vector Fields and the Gradient Theorem.* The gradient of a function is called a *gradient field*. A (continuous) gradient field is *always* a conservative vector field: its line integral along any path depends *only* on the endpoints of the path and can be evaluated by the gradient theorem (the fundamental theorem of calculus for line integrals). Conversely,

> a (continuous) conservative vector field is always the gradient of a function.

This is the starting point of Lagrange's formulation of mechanics.

5.2.6. *Integral Definition of the Gradient of a Scalar Field.* The gradient of a scalar, ϕ, at a point xx, is determined by considering a small volume ΔV, which is bounded by the surface S, as shown in figure 4.3. The outward unit normal to the surface is n. Then, the gradient is defined as

$$\operatorname{grad} \phi = \lim_{\Delta V \to 0} \frac{1}{\Delta V} \int_S \phi n \mathrm{d}S \, .$$

It is evident that the gradient is a vector, since it is proportional to the integral of the unit normal and the scalar function ϕ.

The value of the gradient in Cartesian coordinates is determined by considering a cubic volume $\Delta x^1 \Delta x^2 \Delta x^3$ about the point $x = \left[x^1, x^2, x^3\right]$, as shown in figure 4.4. This volume has six faces, and the surface integral in the last equation is the sum of the contributions due to these six faces. The directions of the outward unit normals should be noted; the outward unit normal to the face A at $(x^2 + \Delta x^2/2)$ is $+\gamma_1$, while the outward unit normal

to the face B at $(x^2 - \Delta x^2/2)$ is $-\gamma_2$. Similarly, for the rest of the faces. With these, the surface integral in the definition of the gradient becomes

$$\int_S \phi n \mathrm{d}S = \gamma_1 \Delta x^2 \Delta x^3 \left(\phi\big[x^1 + \Delta x^1/2, x^2, x^3\big] - \phi\big[x^1 - \Delta x^1/2, x^2, x^3\big] \right) + \cdots .$$

When this surface integral is divided by the volume $\Delta x^1 \Delta x^2 \Delta x^3$, and the limit Δx^1, Δx^2 and $\Delta x^3 \to 0$ is taken, we get the definition of the gradient,

$$\operatorname{grad} \phi = \gamma_1 \frac{\partial \phi}{\partial x^1} + \gamma_2 \frac{\partial \phi}{\partial x^2} + \gamma_3 \frac{\partial \phi}{\partial x^3} .$$

The physical significance of the gradient is as follows. The variation in the scalar ϕ, due to a small variation in the position vector Δx a point, can be written as the dot product of the gradient of ϕ and the vector displacement:

$$\begin{aligned} \phi[x + \Delta x] - \phi[x] &= \Delta x^1 \frac{\partial \phi}{\partial x^1} + \Delta x^2 \frac{\partial \phi}{\partial x^2} + \Delta x^3 \frac{\partial \phi}{\partial x^3} , \\ &= \left(\gamma_1 \Delta x^1 + \gamma_2 \Delta x^2 + \gamma_3 \Delta x^3 \right) \cdot \left(\gamma_1 \frac{\partial \phi}{\partial x^1} + \gamma_2 \frac{\partial \phi}{\partial x^2} + \gamma_3 \frac{\partial \phi}{\partial x^3} \right) , \\ &= \Delta x \cdot \nabla \phi . \end{aligned}$$

Two important results arise from the above relation:

(1) The gradient provides the direction of maximum variation of the scalar ϕ. In the last equation, if we keep the magnitude of the displacement $|\Delta x|$ a constant, and vary the direction, then the dot product $(\Delta x \cdot \nabla \phi)$ is a maximum when Δx and $\nabla \phi$ are in the same direction. Thus,
   the direction of the gradient is the direction of the maximum variation of the function.

(2) If the displacement Δx is perpendicular to $\nabla \phi$, then there is no variation in the function ϕ due to the displacement. Thus,
   the gradient vector is perpendicular to the surface of constant ϕ.

The variation in the scalar ϕ in the last equation was defined for a differential displacement Δx. This can be used to obtain the variation in ϕ for two points separated by a macroscopic distance, by connecting the two points by a path and summing the variation in ϕ over the differential displacements $x^{(i)}$ along this path, as shown in figure 4.5:

$$\phi[x_{\mathcal{B}}] - \phi[x_{\mathcal{A}}] = \sum_i \Delta x^{(i)} \cdot \nabla \phi = \int_{x_{\mathcal{A}}}^{x_{\mathcal{B}}} \mathrm{d}x \cdot \nabla \phi .$$

This is the equivalent of the integral relation 4.26 for the gradient. A consequence of this is that since the difference in ϕ between $x_{\mathcal{A}}$ and $x_{\mathcal{B}}$ is independent of the path used to reach $\mathcal{B}$ from $\mathcal{A}$, the integral $\int \mathrm{d}x \cdot \nabla \phi$ is equal for all paths between the two points $\mathcal{A}$ and $\mathcal{B}$. Another consequence is that the integral of the gradient over a closed path is always equal to zero.

5.3. **Divergence.** *Divergence* is a vector operator that measures the magnitude of a vector field's source or sink at a given point, in terms of a signed scalar. More technically,

> the divergence represents the volume density of the outward flux of a vector field from an infinitesimal volume around a given point.

For example, consider air as it is heated or cooled. The relevant vector field for this example is the velocity of the moving air at a point. If air is heated in a region it will expand in all directions such that the velocity field points outward from that region. Therefore, the divergence of the velocity field in that region would have a *positive value*, as the region is a *source*. If the air cools and contracts, the divergence is *negative* and the region is called a *sink*.

5.3.1. *Definition of Divergence.* In physical terms, the divergence of a three dimensional vector field is the extent to which the vector field flow behaves like a source or a sink at a given point. It is a local measure of its "outgoingness" – the extent to which there is more exiting an infinitesimal region of space than entering it. If the divergence is *nonzero* at some point then there must be a source or sink at that position. (Note that we are

imagining the vector field to be like the velocity vector field of a fluid – in motion – when we use the terms flow, sink and so on.)

More rigorously, the divergence of a vector field f at a point $\mathcal{P}$ is defined as the limit of the net flow of f across the smooth boundary of a three dimensional region $\mathcal{V}$ divided by the volume of $\mathcal{V}$ as $\mathcal{V}$ shrinks to $\mathcal{P}$. Formally,

$$\operatorname{div} f[\mathcal{P}] = \lim_{\mathcal{V}\to\mathcal{P}} \iint_{\mathcal{S}[\mathcal{V}]} \frac{f\cdot n}{|\mathcal{V}|}\,\mathrm{d}S\,,$$

where $|\mathcal{V}|$ is the volume of $\mathcal{V}$, $\mathcal{S}[\mathcal{V}]$ is the boundary of $\mathcal{V}$ and the integral is a *surface integral* with n being the outward unit normal to that surface. The result, $\operatorname{div} f$, is a function of $\mathcal{P}$. From this definition it also becomes explicitly visible that $\operatorname{div} f$ can be seen as the *source density of the flux of* f.

In light of the physical interpretation, a vector field with constant zero divergence is called *incompressible* or *solenoidal* – in this case, no net flow can occur across any closed surface.

The intuition that the sum of all sources minus the sum of all sinks should give the net flow outwards of a region is made precise by the *divergence theorem*.

Note the equivalent definition

$$\operatorname{div} f = \nabla\cdot f\,.$$

5.3.2. *Application in Cartesian Coordinates.* Let $\{x, y, z\}$ be a system of Cartesian coordinates in 3-dimensional Euclidean space, and let $\{\gamma_k\}$ be the corresponding basis of unit vectors.

The divergence of a continuously differentiable vector field $f = u\gamma_x + v\gamma_y + w\gamma_z$ is equal to the scalar-valued function:

$$\operatorname{div} f = \nabla\cdot f = u_{,x} + v_{,y} + w_{,z} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\,.$$

Although expressed in terms of coordinates,

> $\operatorname{div} f$ is *invariant* under orthogonal transformations, as the physical interpretation suggests.

The common notation for the divergence $\nabla\cdot f$ is a convenient mnemonic, where the dot denotes an operation reminiscent of the dot product: take the components of ∇, apply them to the components of f and sum the results. Because applying an operator is different from multiplying the components, this is considered an abuse of notation.

5.3.3. *Decomposition Theorem.* It can be shown that any stationary flux $v[r]$ which is at least two times continuously differentiable in $\mathcal{E}^3$ and vanishes sufficiently fast for $|r|\to\infty$ can be decomposed into an *irrotational part* $e[r]$ and a *source-free* part $b[r]$. Moreover, these parts are explicitly determined by the respective source-densities (divergence) and circulation densities (curl):

For the irrotational part one has

$$e = -\nabla\phi[r]\ ,$$

with

$$\phi[r] = \int_{\mathcal{E}^3} \mathrm{d}^3r'\,\frac{\operatorname{div} v[r']}{4\pi|r-r'|}$$

The source-free part, b, can be similarly written: one only has to replace the scalar potential $\phi[r]$ by a vector potential $a[r]$ and the terms $-\nabla\phi$ by $+\nabla\times a$ and the source-density $\operatorname{div} v$ by the circulation-density $\nabla\times v$.

This "decomposition theorem" is in fact a by-product of the stationary case of electrodynamics. It is a special case of the more general Helmholtz decomposition which works in dimensions greater than three as well.

5.3.4. *Properties.* The following properties can all be derived from the ordinary differentiation rules of calculus.

- the divergence is a linear operator, *i.e.*,

  $$\operatorname{div}(af + bg) = a\operatorname{div} f + b\operatorname{div} g\,,$$

  for all *vector* fields f and g and all real numbers a and b.

- There is a product rule of the following type: if ϕ is a scalar valued function and g is a vector field, then
$$\operatorname{div}(\phi f) = \operatorname{grad}\phi \cdot f + \phi \operatorname{div} f$$
or in more suggestive notation
$$\nabla \cdot (\phi f) = (\nabla \phi) \cdot f + \phi(\nabla \cdot f)\,.$$
- Another product rule for the cross product of two vector fields f and g in three dimensions involves the curl and reads as follows:
$$\operatorname{div}(f \times g) = \operatorname{curl} f \cdot g - f \cdot \operatorname{curl} g\,.$$
- The Laplacian of a scalar field is the divergence of the field's gradient:
$$\nabla^2 \phi = \operatorname{div}\operatorname{grad}\phi\,.$$
- The divergence of the curl of any vector field (in three dimensions) is equal to zero:
$$\nabla \cdot (\nabla \times f) = 0\,.$$

5.4. **Curl.** *Curl* is a vector operator that describes the infinitesimal rotation of a 3-dimensional vector field. At every point in the field, the curl of that field is represented by a vector. The attributes of this vector (length and direction) characterize the rotation at that point.

The direction of the curl is the axis of rotation, as determined by the right-hand rule, and the magnitude of the curl is the magnitude of rotation. If the vector field represents the flow velocity of a moving fluid, then

> the curl is the circulation density of the fluid.

A vector field whose curl is zero is called *irrotational*. The curl is a form of differentiation for vector fields. The corresponding form of the fundamental theorem of calculus is Stokes' theorem, which relates the surface integral of the curl of a vector field to the line integral of the vector field around the boundary curve.

The alternative terminology *rotor* or *rotational* and alternative notations $\operatorname{rot} f$ and $\nabla \times f$ are often used (the former especially in many European countries, the latter, using the del operator and the cross product, is more used in other countries) for curl and $\operatorname{curl} f$.

Unlike the gradient and divergence, *curl does not generalize as simply to other dimensions*; some generalizations are possible, but only in three dimensions is the geometrically defined curl of a vector field again a vector field. This is a similar phenomenon as in the 3 dimensional cross product, and the connection is reflected in the notation $\nabla \times$ for the curl.

5.4.1. *Definition.* The curl of a vector field f, denoted by $\operatorname{curl} f$, $\operatorname{rot} f$ or $\nabla \times f$, at a point is defined in terms of its projection onto various lines through the point. If n is any unit vector, the projection of the curl of f onto n is defined to be the limiting value of a *closed line integral* in a plane orthogonal to n as the path used in the integral becomes infinitesimally close to the point, divided by the area enclosed.

As such, the curl operator maps C^1 functions from $\mathcal{E}^3$ to $\mathcal{E}^3$ to C^0 functions from $\mathcal{E}^3$ to $\mathcal{E}^3$.

Implicitly, curl is defined by:
$$(\nabla \times f) \cdot n = \lim_{\mathcal{A}\to 0} \left(\frac{1}{|\mathcal{A}|} \oint_{\mathcal{C}} f \cdot \mathrm{d}r \right)\,,$$
where $\oint$ is a line integral along the boundary of the area in question, and $|\mathcal{A}|$ is the magnitude of the area. If v is an outward pointing in-plane normal, whereas n is the unit vector perpendicular to the plane, then the orientation of $\mathcal{C}$ is chosen so that a tangent vector ω to $\mathcal{C}$ is positively oriented if and only if $\{n, v, \omega\}$ forms a positively oriented basis for $\mathcal{E}^3$ (right-hand rule).

The above formula means that the curl of a vector field is defined as the infinitesimal area density of the circulation of that field. To this definition fit naturally

- the Kelvin-Stokes theorem, as a global formula corresponding to the definition, and

- the following "easy to memorize" definition of the curl in curvilinear orthogonal coordinates, *e.g.*, in cartesian coordinates, spherical, cylindrical, or even elliptical or parabolical coordinates:

$$(\operatorname{curl} f)_3 = \frac{1}{a_1 a_2}\left(\frac{\partial(a_2 f_2)}{\partial u_1} - \frac{\partial(a_1 f_1)}{\partial u_2}\right).$$

If $\left[x^1, x^2, x^3\right]$ are the Cartesian coordinates and $\left[u^1, u^2, u^3\right]$ are the orthogonal coordinates, then

$$a_i = \sqrt{\sum_{j=1}^{3}\left(\frac{\partial x^j}{\partial u^i}\right)^2}$$

is the length of the coordinate vector corresponding to u^i. The remaining two components of curl result from cyclic permutation of indices: $3,1,2 \to 1,2,3 \to 2,3,1$.

5.4.2. *Intuitive Interpretation.* Suppose the vector field describes the velocity field of a fluid flow (such as a large tank of liquid or gas) and a small ball is located within the fluid or gas (the centre of the ball being fixed at a certain point). If the ball has a rough surface, the fluid flowing past it will make it rotate. The rotation axis (oriented according to the right hand rule) points in the direction of the curl of the field at the centre of the ball, and the angular speed of the rotation is half the magnitude of the curl at this point.

5.5. **Laplace Operator.** The *Laplace operator* or Laplacian is a differential operator given by the divergence of the gradient of a function on Euclidean space. It is usually denoted by the symbols $\nabla \cdot \nabla$, ∇^2 or ∇^2. The Laplacian $\nabla^2 f[p]$ of a function f at a point p, up to a constant depending on the dimension, is the rate at which the average value of f over spheres centered at p, deviates from $f[p]$ as the radius of the sphere grows. In a Cartesian coordinate system, the Laplacian is given by sum of second partial derivatives of the function with respect to each independent variable. In other coordinate systems such as cylindrical and spherical coordinates, the Laplacian also has a useful form.

The Laplace operator is named after the French mathematician Pierre-Simon de Laplace (1749–1827), who first applied the operator to the study of celestial mechanics, where the operator gives a constant multiple of the mass density when it is applied to a given gravitational potential. Solutions of the equation $\nabla^2 f = 0$, now called Laplace's equation, are the so-called harmonic functions and represent the possible gravitational fields in free space.

The Laplacian occurs in differential equations that describe many physical phenomena, such as electric and gravitational potentials, the diffusion equation for heat and fluid flow, wave propagation and quantum mechanics. The Laplacian represents the flux density of the gradient flow of a function. For instance, the net rate at which a chemical dissolved in a fluid moves toward or away from some point is proportional to the Laplacian of the chemical concentration at that point; expressed symbolically, the resulting equation is the diffusion equation. For these reasons, it is extensively used in the sciences for modelling all kinds of physical phenomena. The Laplacian is the simplest elliptic operator and is at the core of Hodge theory as well as the results of de Rham cohomology. In image processing and computer vision, the Laplacian operator has been used for various tasks such as blob and edge detection.

5.5.1. *Definition.* The Laplace operator is a second order differential operator in the n-dimensional Euclidean space, defined as the divergence $\nabla \cdot$ of the gradient $\operatorname{grad} f$. Thus if f is a twice-differentiable real-valued function, then the Laplacian of f is defined by

$$\nabla^2 f = \nabla^2 f = \nabla \cdot \nabla f = \operatorname{div} \operatorname{grad} f.$$

Equivalently, the Laplacian of f is the sum of all the unmixed second partial derivatives in the Cartesian coordinates

$$\nabla^2 f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

As a second-order differential operator, the Laplace operator maps C^k-functions to C^{k-2}-functions for $k \geq 2$. The last expressions define an operator $\nabla^2 : C^k[\mathcal{R}^n] \to C^{k-2}[\mathcal{R}^n]$ or more generally an operator $\nabla^2 : C^k[\Omega] \to C^{k-2}[\Omega]$ for any open set Ω.

5.6. **Motivation.**

5.6.1. *Diffusion.* In the physical theory of diffusion, the Laplace operator (via Laplace's equation) arises naturally in the mathematical description of equilibrium. Specifically, if u is the density at equilibrium of some quantity such as a chemical concentration, then the net flux of u through the boundary of any smooth region V is zero, provided there is no source or sink within V:

$$\int_{\partial V} \nabla u \cdot n \, \mathrm{d}S = 0 \,,$$

where n is the outward unit normal to the boundary of V. By the divergence theorem,

$$\int_V \operatorname{div} \nabla u \, \mathrm{d}V = \int_{\partial V} \nabla u \cdot n \, \mathrm{d}S \,.$$

Since this holds for all smooth regions V, it can be shown that this implies

$$\operatorname{div} \nabla u = \nabla^2 u = 0 \,.$$

The left-hand side of this equation is the Laplace operator. The Laplace operator itself has a physical interpretation for non-equilibrium diffusion as the extent to which a point represents a source or sink of chemical concentration, in a sense made precise by the diffusion equation.

5.6.2. *Density Associated to a Potential.* If ϕ denotes the electrostatic potential associated to a charge distribution q, then the charge distribution itself is given by the Laplacian of ϕ:

$$q = \nabla^2 \phi \,.$$

This is a consequence of Gauss's law. Indeed, if V is any smooth region, then by Gauss's law the flux of the electrostatic field E is equal to the charge enclosed (in appropriate units):

$$\int_{\partial V} E \cdot n \, \mathrm{d}S = \int_{\partial V} \operatorname{grad} \phi \cdot n \, \mathrm{d}S = \int_V q \, \mathrm{d}V \,,$$

where the first equality uses the fact that the electrostatic field is the gradient of the electrostatic potential. The divergence theorem now gives

$$\int_V \nabla^2 \phi \, \mathrm{d}V = \int_V q \, \mathrm{d}V$$

and since this holds for all regions V, $q = \nabla^2 \phi$ follows.

The same approach implies that the Laplacian of the gravitational potential is the mass distribution. Often the charge (or mass) distribution are given, and the associated potential is unknown. Finding the potential function subject to suitable boundary conditions is equivalent to solving Poisson's equation.

5.7. **Divergence Theorem.** The *divergence theorem*, *aka* Gauss's theorem, Green's theorem or Ostrogradsky's theorem, is a result that relates the flow (that is, *flux*) of a vector field through a surface to the behavior of the vector field inside the surface.

More precisely, the divergence theorem states that

> the outward flux of a vector field through a closed surface is equal to the volume integral of the divergence over the region inside the surface.

Intuitively, it states that

> the sum of all sources minus the sum of all sinks gives the net flow out of a region.

The divergence theorem is an important result for the mathematics of engineering, in particular in electrostatics and fluid dynamics.

In physics and engineering, the divergence theorem is usually applied in three dimensions. However, it generalizes to any number of dimensions. In one dimension, it is equivalent to the fundamental theorem of calculus.

The theorem is a special case of the more general Stokes' theorem.

5.7.1. *Intuition.* If a fluid is flowing in some area, and we wish to know how much fluid flows out of a certain region within that area, then we need to add up the sources inside the region and subtract the sinks. The fluid flow is represented by a vector field, and the vector field's divergence at a given point describes the strength of the source or sink there. So, integrating the field's divergence over the interior of the region should equal the integral of the vector field over the region's boundary. The divergence theorem says that this is true.

The divergence theorem is thus a *conservation law* which states that the volume total of all sinks and sources, the volume integral of the divergence, is equal to the net flow across the volume's boundary.

5.7.2. *Mathematical Statement.* Suppose $\mathcal{V}$ is a subset of $\mathcal{E}^n$ (in the case of $n = 3$, $\mathcal{V}$ represents a volume in 3D space) which is compact and has a piecewise smooth boundary $\mathcal{S}$. If f is a continuously differentiable *vector field* defined on a neighborhood of $\mathcal{V}$, then we have

$$\int_{\mathcal{V}} (\nabla \cdot f)\, \mathrm{d}\mathcal{V} = \oint_{\mathcal{S}} (f \cdot n)\, \mathrm{d}\mathcal{S}\,.$$

The left side is a *volume integral* over the volume $\mathcal{V}$, the right side is the *surface integral* over the boundary of the volume $\mathcal{V}$. The closed manifold $\partial\mathcal{V}$ is quite generally the boundary of $\mathcal{V}$ oriented by outward-pointing normals, and n is the outward pointing unit normal field of the boundary $\partial\mathcal{V}$. ($\mathrm{d}\mathcal{S}$ may be used as a shorthand for $n\mathrm{d}S$.) By the symbol within the two integrals it is stressed once more that $\partial\mathcal{V}$ is a closed surface. In terms of the intuitive description above, the left-hand side of the equation represents the total of the sources in the volume $\partial\mathcal{V}$, and the right-hand side represents the total flow across the boundary $\partial\mathcal{V}$.

5.7.3. *Applications.* Differential form and integral form of physical laws: As a result of the divergence theorem, a host of physical laws can be written in both a differential form (where one quantity is the divergence of another) and an integral form (where the flux of one quantity through a closed surface is equal to another quantity). Three examples are Gauss's law (in electrostatics), Gauss's law for magnetism, and Gauss's law for gravity.

Continuity equations: Continuity equations offer more examples of laws with both differential and integral forms, related to each other by the divergence theorem. In fluid dynamics, electromagnetism, quantum mechanics, relativity theory, and a number of other fields, there are continuity equations that describe the conservation of mass, momentum, energy, probability, or other quantities. Generically, these equations state that the divergence of the flow of the conserved quantity is equal to the distribution of sources or sinks of that quantity. The divergence theorem states that any such continuity equation can be written in a differential form (in terms of a divergence) and an integral form (in terms of a flux).

Inverse-square laws: Any inverse-square law can instead be written in a Gauss' law-type form (with a differential and integral form, as described above). Two examples are Gauss' law (in electrostatics), which follows from the inverse-square Coulomb's law, and Gauss' law for gravity, which follows from the inverse-square Newton's law of universal gravitation. The derivation of the Gauss' law-type equation from the inverse-square formulation (or *vice versa*) is exactly the same in both cases.

5.8. **Gradient Theorem.** The *gradient theorem*, *aka* the fundamental theorem of calculus for line integrals, says that a line integral through a gradient field can be evaluated by evaluating the original scalar field at the endpoints of the curve:

$$\phi[q] - \phi[p] = \int_{\gamma[p,q]} \operatorname{grad} \phi[r] \cdot \mathrm{d}r\,.$$

It is a generalization of the fundamental theorem of calculus to any curve in a plane or space (generally n-dimensional) rather that just the real line.

The gradient theorem implies that line integrals through irrotational vector fields are path independent. In physics, this theorem is one of the ways of defining a "conservative" force. By placing ϕ as potential, $\operatorname{grad}\phi$ is a conservative field. Work done by conservative forces does not depend on the path followed by the object, but only the end points, as the above equation shows.

The gradient theorem also has an interesting converse:

> any conservative vector field can be expressed as the gradient of a scalar field.

Just like the gradient theorem itself, this converse has many striking consequences and applications in both pure and applied mathematics.

5.9. **Continuity Equation.** A *continuity equation* in physics is an equation that describes the transport of a conserved quantity. Since mass, energy, momentum, electric charge and other natural quantities are conserved under their respective appropriate conditions, a variety of physical phenomena may be described using continuity equations.

Continuity equations are a stronger, *local* form of conservation laws. For example, it is true that "the total energy in the universe is conserved". But this statement does not immediately rule out the possibility that energy could disappear from Earth while simultaneously appearing in another galaxy. A stronger statement is that energy is locally conserved: Energy can neither be created nor destroyed, nor can it "teleport" from one place to another – it can only move by a *continuous flow*. A continuity equation is the mathematical way to express this kind of statement.

Continuity equations more generally can include "source" and "sink" terms, which allow them to describe quantities which are often but not always conserved, such as the density of a molecular species which can be created or destroyed by chemical reactions. In an everyday example, there is a continuity equation for the number of living humans; it has a "source term" to account for people giving birth, and a "sink term" to account for people dying.

Any continuity equation can be expressed in an "integral form" (in terms of a *flux integral*), which applies to any finite region, or in a "differential form" (in terms of the *divergence operator*) which applies at a point. Continuity equations underlie more specific transport equations such as the convection-diffusion equation, Boltzmann transport equation, and Navier-Stokes equations.

5.9.1. *Preliminary Description.* As stated above, the idea behind the continuity equation is the flow of some property, such as mass, energy, electric charge, momentum, and even probability, through surfaces from one region of space to another. The surfaces, in general, may either be open or closed, real or imaginary, and have an arbitrary shape, but are *fixed* for the calculation (*i.e.*, *not* time-varying, which is appropriate since this complicates the maths for no advantage). Let this property be represented by just one scalar variable, q, and let the volume density of this property (the amount of q per unit volume V) be ρ, and the union of all surfaces be denoted by S. Mathematically, ρ is a ratio of two infinitesimal quantities:

$$\rho = \frac{\mathrm{d}q}{\mathrm{d}V}\,,$$

which has the dimension $[quantity] \,/\, \left[L^3\right]$ (where $[L]$ is length).

There are different ways to conceive the continuity equation:

- either the flow of particles carrying the quantity q, described by a velocity field v, which is also equivalent to a flux j of q (a vector function describing the flow per unit area per unit time of q), or,
- in the cases where a velocity field is not useful or applicable, the flux j of the quantity q only (no association with velocity).

In each of these cases, the transfer of q occurs as it passes through two surfaces, the first S_1 and the second S_2.

The flux j should represent some flow or transport, which has dimensions $[quantity] \,/\, [T] \left[L^2\right]$. In cases where particles/carriers of quantity q are moving with velocity v, such as particles of mass in a fluid or charge carriers in a conductor, j can be related to v by:

$$j = \rho v\,.$$

This relation is only true in situations where there are particles moving and carrying q – it can't always be applied. To illustrate this: if j is electric current density (electric current per unit area) and ρ is the charge density (charge per unit volume), then the velocity of the charge carriers is v. However, if j is heat flux density (heat energy per unit time per

unit area), then even if we let ρ be the heat energy density (heat energy per unit volume) it does not imply the "velocity of heat" is v (this makes no sense and is not practically applicable). In the latter case only j (with ρ) may be used in the continuity equation.

5.9.2. *Elementary Vector Form.* Consider the case when the surfaces are flat and planar cross-sections. For the case where a velocity field can be applied, *dimensional analysis* leads to this form of the continuity equation:

$$\rho_1 v_1 \cdot S_1 = \rho_2 v_2 \cdot S_2 \,,$$

where the left hand side is the initial amount of q flowing per unit time through surface S_1, the right hand side is the final amount through surface S_2 and S_1 and S_2 are the vector areas for the surfaces S_1 and S_2.

Notice the dot products $v_i \cdot S_i$ are *volumetric flow rates* of q. The dimension of each side of the equation is $[quantity] / \left[L^3\right] \times [L] / [T] \times \left[L^2\right] = [quantity] / [T]$. For the more general cases, independent of whether a velocity field can be used or not, the continuity equation becomes:

$$j_1 \cdot S_1 = j_2 \cdot S_2 \,.$$

This has exactly the same dimensions as the previous version. The relation between j and v allows us to pass from the velocity version to this *flux equation*, but not always the other way round (as explained above – velocity fields are not always applicable). These results can be generalized further to curved surfaces by reducing the vector surfaces into infinitely many differential surface elements (that is $S \to \mathrm{d}S$), then integrating over the surface:

$$\int_{S_1} \rho_1 v_1 \cdot \mathrm{d}S_1 = \int_{S_2} \rho_2 v_2 \cdot \mathrm{d}S_2$$

or, more generally still,

$$\int_{S_1} j_1 \cdot \mathrm{d}S_1 = \int_{S_2} j_2 \cdot \mathrm{d}S_2 \,,$$

in which $\int_S \mathrm{d}S = \int_S n \mathrm{d}S$ denotes a surface integral over the surface S, n is the outward-pointing unit normal to the surface S. Note that the scalar area S and vector area S are related by $\mathrm{d}S = n\mathrm{d}S$. Either notations may be used interchangeably.

5.9.3. *Differential form.* The differential form for a general continuity equation is (using the same q, ρ and j as above):

$$\frac{\partial \rho}{\partial t} + \nabla \cdot j = \sigma \,,$$

where $\nabla \cdot$ is divergence, t is time, σ is the *generation of q per unit volume per unit time*. Terms that generate ($\sigma > 0$) or remove ($\sigma < 0$) q are referred to as a "sources" and "sinks".

This general equation may be used to derive *any* continuity equation, ranging from as simple as the volume continuity equation to as complicated as the Navier-Stokes equations. This equation also generalizes the *advection equation*. Other equations in physics, such as Gauss's law of the electric field and Gauss's law for gravity, have a similar mathematical form to the continuity equation, but are not usually called by the term "continuity equation", because j in those cases does *not* represent the flow of a real physical quantity.

In the case that q is a conserved quantity that cannot be created or destroyed (such as energy), this translates to $\sigma = 0$ and the continuity equation is:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot j = 0 \,.$$

5.9.4. *Integral Form.* By the divergence theorem (see below), the continuity equation can be rewritten in an equivalent way, called the "integral form":

$$\frac{\mathrm{d}q}{\mathrm{d}t} + \oint_S j \cdot \mathrm{d}S = \Sigma \,,$$

where S is a surface as described above – except this time it has to be a *closed surface* that encloses a volume V, $\oint_S \mathrm{d}S$ denotes a surface integral over a closed surface, $\int_V \mathrm{d}V$ denotes a volume integral over V. $q = \int_V \rho \mathrm{d}V$ is the total amount of ρ in the volume V; $\Sigma = \int_V \sigma \mathrm{d}V$ is the total generation (negative in the case of removal) per unit time by the sources and sinks in the volume V.

In a simple example, V could be a building and q could be the number of people in the building. The surface S would consist of the walls, doors, roof and foundation of the building. Then, the continuity equation states that the number of people in the building increases when people enter the building (an inward flux through the surface), decreases when people exit the building (an outward flux through the surface), increases when someone in the building gives birth (a "source" where $\sigma > 0$) and decreases when someone in the building dies (a "sink" where $\sigma < 0$).

5.9.5. *Derivation of the Differential Form.* Suppose first an amount of quantity q is contained in a region of volume V, bounded by a closed surface S, as described above. This can be written as

$$q[t] = \int_V \rho[r,t]\ \mathrm{d}V\,.$$

The rate of change of q is simply the time derivative:

$$\frac{\mathrm{d}q}{\mathrm{d}t}[t] = \frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho[r,t]\ \mathrm{d}V = \int_V \frac{\partial\rho}{\partial t}[r,t]\ \mathrm{d}V\,.$$

The derivative is changed from the total to partial as it enters the integral because the integrand is not only a function of time, but also of coordinates due to the density nature of ρ. The rate of change of q can also be expressed as a sum of the flow through the surface S taken with the minus sign (the flow is from inside to outside) and the rate of production of q:

$$\frac{\mathrm{d}q}{\mathrm{d}t}[t] = -\int_S j[r,t]\cdot\mathrm{d}S + \Sigma[t]\ .$$

Now equating these expressions:

$$-\int_S j[r,t]\cdot\mathrm{d}S + \Sigma[t] = \int_V \frac{\partial\rho}{\partial t}[r,t]\ \mathrm{d}V\,.$$

Using the divergence theorem on the left-hand side:

$$-\int_V \nabla\cdot j[r,t]\ \mathrm{d}V + \int_V \sigma[r,t]\ \mathrm{d}V = \int_V \frac{\partial\rho}{\partial t}[r,t]\ \mathrm{d}V\,.$$

Since the volume V is arbitrary chosen this is only true if the integrands are equal, which directly leads to the differential continuity equation:

$$\nabla\cdot j[r,t] = -\frac{\partial\rho}{\partial t}[r,t] + \sigma[r,t]\ ,$$
$$\nabla\cdot j + \frac{\partial\rho}{\partial t} = \sigma \implies \nabla\cdot\rho v + \frac{\partial\rho}{\partial t} = \sigma\,.$$

Either form may be useful and quoted, both can appear in hydrodynamics and electromagnetism, but for quantum mechanics and energy conservation, only the first is used. Therefore the first is more general.

5.9.6. *Electromagnetism.* In electromagnetic theory, the continuity equation is an empirical law expressing (local) charge conservation. Mathematically it is an automatic consequence of Maxwell's equations, although charge conservation is more fundamental than Maxwell's equations. It states that the divergence of the current density J (in amperes per square meter) is equal to the negative rate of change of the charge density ρ (in coulombs per cubic metre),

$$\nabla\cdot J = -\frac{\partial\rho}{\partial t}\,.$$

Current is the movement of charge. The continuity equation says that if charge is moving out of a differential volume (*i.e.*, divergence of current density is positive) then the amount of charge within that volume is going to decrease, so the rate of change of charge density is negative. Therefore the continuity equation amounts to a conservation of charge.

5.9.7. *Fluid Dynamics.* In fluid dynamics, the continuity equation states that, in any steady state process, the rate at which mass enters a system is equal to the rate at which mass leaves the system.

The differential form of the continuity equation is:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0 \,,$$

where ρ is fluid density, t is time, u is the flow velocity vector field.

In this context, this equation is also one of Euler equations. The Navier-Stokes equations form a vector continuity equation describing the conservation of linear momentum.

If ρ is a constant, as in the case of incompressible flow, the mass continuity equation simplifies to a volume continuity equation:

$$\nabla \cdot u = 0 \,,$$

which means that the divergence of velocity field is zero everywhere. Physically, this is equivalent to saying that the local volume dilation rate is zero.

5.9.8. *Heat Transfer.* Conservation of energy (which, in non-relativistic situations, can only be transferred, and not created or destroyed) leads to a continuity equation, an alternative mathematical statement of energy conservation to the thermodynamic laws. Letting u to be the local energy density (energy per unit volume), q the energy flux (transfer of energy per unit cross-sectional area per unit time) as a vector, then the continuity equation is:

$$\nabla \cdot q + \frac{\partial u}{\partial t} = 0 \,.$$

5.10. **Green's Identities.** In mathematics, Green's identities are a set of three identities in vector calculus. They are named after the mathematician George Green, who discovered Green's theorem.

5.10.1. *Green's first identity.* This identity is derived from the divergence theorem applied to the vector field : Let φ and ϕ be scalar functions defined on some region $\mathcal{U}$ in $\mathcal{R}^3$, and suppose that φ is twice continuously differentiable, and ϕ is once continuously differentiable. Then,

$$\int_{\mathcal{U}} \left(\phi \nabla^2 \varphi + \operatorname{grad} \varphi \cdot \operatorname{grad} \phi \right) \, \mathrm{d}\mathcal{V} = \oint_{\partial \mathcal{U}} \phi (\operatorname{grad} \varphi \cdot n) \, \mathrm{d}S \,.$$

where ∇^2 is the Laplace operator, $\partial \mathcal{U}$ is the boundary of region $\mathcal{U}$ and n is the outward pointing unit normal of surface element $\mathrm{d}S$. This theorem is essentially the higher dimensional equivalent of integration by parts with ϕ and the gradient of φ replacing u and v.

Note that Green's first identity above is a special case of the more general identity derived from the divergence theorem by substituting $F = \phi\gamma$:

$$\int_{\mathcal{U}} \left(\phi \operatorname{grad} \cdot \gamma + \gamma \cdot \operatorname{grad} \phi \right) \, \mathrm{d}\mathcal{V} = \oint_{\partial \mathcal{U}} \phi (\gamma \cdot n) \, \mathrm{d}S \,.$$

5.10.2. *Green's second identity.* If φ and ϕ are both twice continuously differentiable on $\mathcal{U}$ in $\mathcal{R}^3$, and ϵ is once continuously differentiable, we can choose $F = \phi\epsilon \operatorname{grad} \varphi - \varphi\epsilon \operatorname{grad} \phi$ and obtain:

$$\int_{\mathcal{U}} \left(\phi \operatorname{grad} \cdot (\epsilon \operatorname{grad} \varphi) - \varphi \operatorname{grad} \cdot (\epsilon \operatorname{grad} \phi) \right) \, \mathrm{d}\mathcal{V} = \oint_{\partial \mathcal{U}} \epsilon \left(\phi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \phi}{\partial n} \right) \, \mathrm{d}S \,.$$

For the special case of $\epsilon = 1$ all across $\mathcal{U}$ in $\mathcal{R}^3$ then:

$$\int_{\mathcal{U}} \left(\phi \nabla^2 \varphi - \varphi \nabla^2 \phi \right) \, \mathrm{d}\mathcal{V} = \oint_{\partial \mathcal{U}} \left(\phi \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \phi}{\partial n} \right) \, \mathrm{d}S \,.$$

In the equation above $\partial\phi/\partial n$ is the directional derivative of ϕ in the direction of the outward pointing normal n to the surface element $\mathrm{d}S$:

$$\frac{\partial \phi}{\partial n} = \operatorname{grad} \phi \cdot n \,.$$

5.11. **Stokes' Theorem.** In differential geometry, *Stokes' theorem* (also called the generalized Stokes' theorem) is a statement about the integration of differential forms on manifolds, which both simplifies and generalizes several theorems from vector calculus. Stokes' theorem says that the integral of a differential form ω over the boundary of some orientable manifold Ω is equal to the integral of its exterior derivative $\mathrm{d}\omega$ over the whole of Ω, *i.e.*,

$$\int_{\partial\Omega} \omega = \int_{\Omega} \mathrm{d}\omega \,.$$

This modern form of Stokes' theorem is a vast generalization of a classical result first discovered by Lord Kelvin, who communicated it to George Stokes in a letter dated July 2, 1850. Stokes set the theorem as a question on the 1854 Smith's Prize exam, which led to the result bearing his name. This classical Kelvin-Stokes theorem relates the surface integral of the curl of a vector field F over a surface Σ in Euclidean three-space to the line integral of the vector field over its boundary $\partial\Sigma$:

$$\iint_{\Sigma} \nabla \times F \cdot \mathrm{d}\Sigma = \oint_{\partial\Sigma} F \cdot \mathrm{d}x \,,$$

where x is the position vector.

This classical statement, as well as the classical Divergence theorem, fundamental theorem of calculus, and Green's Theorem are simply special cases of the general formulation stated above.

6. Partial Differential Equations

[The Princeton Companion to Maths – Timothy Gowers, June Barrow-Green, Imre Leader]

Partial differential equations are of immense importance in physics, and have inspired a vast amount of mathematical research. Three basic examples will be discussed here, as an introduction to more advanced articles later in the volume.

6.1. **Heat Equation.** The first is the *heat equation*, which, as its name suggests, describes the way the distribution of heat in a physical medium changes with time:

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) = \kappa \nabla^2 T \,.$$

Here, $T[x, y, z, t]$ is a function that specifies the temperature at the point $[x, y, z]$ at time t.

It is one thing to read an equation like this and understand the symbols that make it up, but quite another to see what it really means. However, it is important to do so, since of the many expressions one could write down that involve partial derivatives, only a minority are of much significance, and these tend to be the ones that have interesting interpretations. So let us try to interpret the expressions involved in the heat equation.

The left-hand side, $T_{,t}$, is quite simple. It is the rate of change of the temperature $T[x, y, z, t]$, when the spatial coordinates x, y and z are *kept fixed* and t *varies*. In other words, it tells us how fast the point $[x, y, z]$ is heating up or cooling down at time t. What would we expect this to depend on? Well, heat takes time to travel through a medium, so although the temperature at some distant point $[x', y', z']$ will eventually affect the temperature at $[x, y, z]$, the way the temperature is changing right now (that is, at time t) will be affected only by the temperatures of points very close to $[x, y, z]$: if points in the immediate neighborhood of $[x, y, z]$ are hotter, on average, than $[x, y, z]$ itself, then we expect the temperature at $[x, y, z]$ to be increasing, and if they are colder then we expect it to be decreasing.

The expression in brackets on the right-hand side appears so often that it has its own shorthand. The symbol ∇^2, defined by

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

is known as the *Laplacian*. What information does $\nabla^2 f$ give us about a function f? The answer is that it captures the idea in the last paragraph: it tells us how the value of f at $[x, y, z]$ compares with the average value of f in a small neighborhood of $[x, y, z]$, or, more precisely, with the limit of the average value in a neighborhood of $[x, y, z]$ as the size of that neighborhood shrinks to zero.

(Note: In Euclidean space $\mathcal{E}^3$ in Cartesian coordinates $[x, y, z]$, using index notation and Einstein summation convention, Laplacian can be written as

$$\nabla^2 = \partial^i \partial_j = g^{ij} \partial_i \partial_j = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \,.)$$

This is not immediately obvious from the formula, but the following (not wholly rigorous) argument in one dimension gives a clue about why second derivatives should be involved. Let f be a function that takes real numbers to real numbers. Then to obtain a good approximation to the second derivative of f at a point x, one can look at the expression $(f'[x] - f'[x - h])/h$ for some small h. (If one substitutes $-h$ for h in the above expression, one obtains the more usual formula, but this one is more convenient here.) The derivatives $f'[x]$ and $f'[x - h]$ can themselves be approximated by $(f[x + h] - f[x])/h$ and $(f[x] - f[x - h])/h$, respectively, and if we substitute these approximations into the earlier expression, then we obtain

$$\frac{1}{h} \left(\frac{f[x + h] - f[x]}{h} - \frac{f[x] - f[x - h]}{h} \right) ,$$

which equals $(f[x + h] - 2f[x] + f[x - h])/h^2$. Dividing the top of this last fraction by 2, we obtain $(1/2)(f[x + h] + f[x - h]) - f[x]$: that is, the difference between the value of f at x and the average value of f at the two surrounding points $x + h$ and $x - h$.

In other words, the second derivative conveys just the idea we want – a comparison between the value at x and the average value near x. It is worth noting that if f is linear, then the average of $f[x-h]$ and $f[x+h]$ will be equal to $f[x]$, which fits with the familiar fact that the second derivative of a linear function f is zero.

Just as, when defining the first derivative, we have to divide the difference $f[x+h]-f[x]$ by h so that it is not automatically tiny, so with the second derivative it is appropriate to divide by h^2. (This is appropriate, since, whereas the first derivative concerns linear approximations, the second derivative concerns quadratic ones: the best quadratic approximation for a function f near a value x is $f[x+h] = f[x] + hf'[x] + (1/2)h^2 f''[x]$, an approximation that one can check is exact if f was a quadratic function to start with.)

It is possible to pursue thoughts of this kind and show that if f is a function of three variables then the value of $\nabla^2 f$ at $[x, y, z]$ does indeed tell us how the value of f at $[x, y, z]$ compares with the average values of f at points nearby. (There is nothing special about the number 3 here – the ideas can easily be generalized to functions of any number of variables.) All that is left to discuss in the heat equation is the parameter κ. This measures the conductivity of the medium. If κ is small, then the medium does not conduct heat very well and $\nabla^2 T$ has less of an effect on the rate of change of the temperature; if it is large then heat is conducted better and the effect is greater.

6.2. **Laplace Equation.** A second equation of great importance is the Laplace equation [3], $\nabla^2 f = 0$. Intuitively speaking, this says of a function f that its value at a point $[x, y, z]$ is always equal to the average value at the immediately surrounding points. If f is a function of just one variable x, this says that the second derivative of f is zero, which implies that f is of the form ax + b. However, for two or more variables, a function has more flexibility – it can lie above the tangent lines in some directions and below it in others. As a result, one can impose a variety of boundary conditions on f (that is, specifications of the values f takes on the boundaries of certain regions), and there is a much wider and more interesting class of solutions.

6.3. **Wave Equation.** A third fundamental equation is the wave equation. In its one-dimensional formulation it describes the motion of a vibrating string that connects two points A and B. Suppose that the height of the string at distance x from A and at time t is written $h[x, t]$. Then the wave equation says that

$$\frac{1}{v^2}\frac{\partial^2 h}{\partial t^2} = \frac{\partial^2 h}{\partial x^2}\,.$$

Ignoring the constant $1/v^2$ for a moment, the left-hand side of this equation represents the acceleration (in a vertical direction) of the piece of string at distance x from A. This should be proportional to the force acting on it. What will govern this force? Well, suppose for a moment that the portion of string containing x were absolutely straight. Then the pull of the string on the left of x would exactly cancel out the pull on the right and the net force would be zero. So, once again, what matters is how the height at x compares with the average height on either side: if the string lies above the tangent line at x, then there will be an upwards force, and if it lies below, then there will be a downwards one. This is why the second derivative appears on the right-hand side once again. How much force results from this second derivative depends on factors such as the density and tautness of the string, which is where the constant comes in. Since h and x are both distances, v^2 has dimensions of $\left[L^2/T^2\right]$, which means that v represents a speed, which is, in fact, the speed of propagation of the wave.

Similar considerations yield the three-dimensional wave equation, which is, as one might now expect,

$$\frac{1}{v^2}\frac{\partial^2 h}{\partial t^2} = \frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} + \frac{\partial^2 h}{\partial z^2}\,,$$

or, more concisely,

$$\frac{1}{v^2}\frac{\partial^2 h}{\partial t^2} = \nabla^2 h\,.$$

[3] The quantity $\nabla^2 f$ has been termed the concentration of f and its value at any point indicates the "excess" of the value of f there over its mean value in the neighbourhood of the point.

One can be more concise still and write this equation as $\Box h = 0$, where $\Box h$ is shorthand for

$$\nabla^2 h - \frac{1}{v^2}\frac{\partial^2 h}{\partial t^2}\,.$$

The operation $\Box$ is called the *d'Alembertian*.

In Minkowski spacetime in standard coordinates $[t, x, y, z]$ with signature $\operatorname{sig}[+---]$, using index notation and Einstein summation convention, d'Alembertian can be written as

$$\Box = \partial^\mu \partial_\nu = g^{\mu\nu}\partial_\mu\partial_\nu = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2}\,.$$

6.4. **Diffusion Equation.** In the physical theory of diffusion, the Laplace operator (via Laplace's equation) arises naturally in the mathematical description of equilibrium. Specifically, if u is the density at equilibrium of some quantity such as a chemical concentration, then the net flux of u through the boundary of any smooth region $\mathcal{V}$ is zero, provided there is no source or sink within $\mathcal{V}$:

$$\int_{\partial\mathcal{V}} \operatorname{grad} u \cdot n \,\mathrm{d}S = 0\,,$$

where n is the outward unit normal (vector) to the boundary of $\mathcal{V}$. By the divergence theorem,

$$\int_{\mathcal{V}} \operatorname{div}\operatorname{grad} u \,\mathrm{d}\mathcal{V} = \int_{\partial\mathcal{V}} \operatorname{grad} u \cdot n \,\mathrm{d}S = 0\,.$$

Since this holds for all smooth regions $\mathcal{V}$, it can be shown that this implies

$$\operatorname{div}\operatorname{grad} u = \nabla^2 u = 0\,.$$

The right-hand side of this equation is the Laplace operator. The Laplace operator itself has a physical interpretation for non-equilibrium diffusion as the extent to which a point represents a source or sink of chemical concentration, in a sense made precise by the diffusion equation.

6.5. **Density associated to a potential.** If ϕ denotes the electrostatic potential associated to a charge distribution q, then the charge distribution itself is given by the Laplacian of ϕ:

$$q = \nabla^2\phi\,.$$

This is a consequence of Gauss's law. Indeed, if $\mathcal{V}$ is any smooth region, then by Gauss's law the flux of the electrostatic field E is equal to the charge enclosed (in appropriate units):

$$\int_{\partial\mathcal{V}} E \cdot n \,\mathrm{d}S = \int_{\partial\mathcal{V}} \operatorname{grad}\phi \cdot n \,\mathrm{d}S = \int_{\mathcal{V}} q \,\mathrm{d}\mathcal{V}\,,$$

where the first equality uses the fact that the electrostatic field is the gradient of the electrostatic potential. The divergence theorem [4] now gives

$$\int_{\mathcal{V}} \nabla^2\phi \,\mathrm{d}\mathcal{V} = \int_{\mathcal{V}} q \,\mathrm{d}\mathcal{V}$$

and, since this holds for all regions $\mathcal{V}$, then $q = \nabla^2\phi$ follows.

The same approach implies that the Laplacian of the gravitational potential is the mass distribution. Often the charge (or mass) distribution are given, and the associated potential is unknown. Finding the potential function subject to suitable boundary conditions is equivalent to solving Poisson's equation.

[4] The divergence theorem can be put in alternate notation: $\int_{\partial\mathcal{V}} \operatorname{grad} u \cdot n \,\mathrm{d}S = \int_{\mathcal{V}} \operatorname{div}\operatorname{grad} u \,\mathrm{d}\mathcal{V}$.

6.6. **Energy minimization.** Another motivation for the Laplacian appearing in physics is that solutions to $\nabla^2 f = 0$ in a region $\mathcal{U}$ are functions that make the Dirichlet energy functional stationary:

$$E[f] = \frac{1}{2}\int_{\mathcal{U}} |\text{grad}\, f|^2 \,\mathrm{d}x\,.$$

To see this, suppose $f : \mathcal{U} \to \mathcal{R}$ is a function, and $u : \mathcal{U} \to \mathcal{R}$ is a function that vanishes on the boundary of $\mathcal{U}$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} E[f + \epsilon u] = \int_{\mathcal{U}} \text{grad}\, f \cdot \text{grad}\, u \,\mathrm{d}x = -\int_{\mathcal{U}} u\nabla^2 f \,\mathrm{d}x\,,$$

where the last equality follows using Green's first identity. This calculation shows that if $\nabla^2 f = 0$, then E is stationary around f. Conversely, if E is stationary around f, then $\nabla^2 f = 0$ by the fundamental lemma of calculus of variations.

7. Sequences and Series

7.1. **Sequence.** In mathematics, informally speaking, a *sequence* is an ordered list of objects (or events). Like a set, it contains members (also called elements, or terms). The number of ordered elements (possibly infinite) is called the length of the sequence. Unlike a set, order matters, and exactly the same elements can appear multiple times at different positions in the sequence. Most precisely, a sequence can be defined as a function whose domain is a countable totally ordered set, such as the natural numbers.

7.1.1. *Indexing.* Indexing notation is used to refer to a sequence in the abstract. It is also a natural notation for sequences whose elements are related to the index n (the element's position) in a simple way. For instance, the sequence of the first 10 square numbers could be written as $\{a_1, a_2, \ldots, a_n\}$ such that $a_k = k^2$ for all k. This represents the sequence $\{1, 4, 9, \ldots, 100\}$. This notation is often simplified further as $\{a_k\}_{k=1}^{10}$ with $a_k = k^2$. Here the subscript ($k = 1$) and superscript 10 together tell us that the elements of this sequence are the a_k such that $k = 1, 2, \ldots, 10$.

Sequences can be indexed beginning and ending from any integer. The infinity symbol ∞ is often used as the superscript to indicate the sequence including all integer k-values starting with a certain one. The sequence of all positive squares is then denoted $\{a_k\}_{k=1}^{\infty}$ with $a_k = k^2$.

In some cases the elements of the sequence are related naturally to a sequence of integers whose pattern can be easily inferred. In these cases the index set may be implied by a listing of the first few abstract elements. For instance, the sequence of squares of odd numbers could be denoted $\left\{(2k-1)^2\right\}_{k=1}^{\infty}$.

7.1.2. *Specifying a Sequence by Recursion.* Sequences whose elements are related to the previous elements in a straightforward way are often specified using *recursion*. This is in contrast to the specification of sequence elements in terms of their position.

To specify a sequence by recursion requires a *rule to construct each consecutive element in terms of the ones before it*. In addition, *enough initial elements must be specified* so that new elements of the sequence can be specified by the rule. The *principle of mathematical induction* can be used to prove that a sequence is well-defined, which is to say that that every element of the sequence is specified at least once and has a single, unambiguous value. Induction can also be used to prove properties about a sequence, especially for sequences whose most natural specification is by recursion.

The Fibonacci sequence can be defined using a recursive rule along with two initial elements. The rule is that each element is the sum of the previous two elements, and the first two elements are 0 and 1.

$$a_n = a_{n-1} + a_{n-2}\,, \qquad \text{with} \qquad a_0 = 0 \text{ and } a_1 = 1\,.$$

The first ten terms of this sequence are 0, 1, 1, 2, 3, 5, 8, 13, 21 and 34.

Not all sequences can be specified by a rule in the form of an equation, recursive or not, and some can be quite complicated. For example, the sequence of prime numbers is the set of prime numbers in their natural order. This gives the sequence $\{2, 3, 5, 7, 11, 13, 17, \ldots\}$.

7.1.3. *Definition.* A sequence is usually defined as a function whose domain is a countable totally ordered set, although in many disciplines the domain is restricted, such as to the natural numbers. In real analysis a sequence is a function from a subset of the natural numbers to the real numbers. In other words, a sequence is a map $f[n] : \mathcal{N} \mapsto \mathcal{R}$. To recover our earlier notation we might identify $a_k = f[n]\ \forall n$ or just write $a_n : \mathcal{N} \mapsto \mathcal{R}$.

7.1.4. *Finite and Infinite.* The *length* of a sequence is defined as the number of terms in the sequence.

A sequence of a finite length n is also called an n-tuple. Finite sequences include the empty sequence $\{\}$ that has no elements.

Normally, the term *infinite sequence* refers to a sequence which is infinite in one direction, and finite in the other – the sequence has a first element, but no final element (a *singly infinite sequence*). A sequence that is infinite in both directions – it has neither a first nor a final element-is called a *bi-infinite sequence*, two-way infinite sequence, or doubly infinite sequence. For instance, a function from all integers into a set, such as the sequence

of all even integers $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$ is bi-infinite. This sequence could be denoted $\{2n\}_{n=\infty}^{\infty}$.

7.1.5. *Increasing and Decreasing.* A sequence is said to be *monotonically increasing* if each term is greater than or equal to the one before it. For a sequence this can be written as $a_k \leq a \ \forall n \geq 1$. If each consecutive term is strictly greater than the previous term then the sequence is called *strictly monotonically increasing*. A sequence is *monotonically decreasing* if each consecutive term is less than or equal to the previous one and *strictly monotonically decreasing* if each is strictly less than the previous. If a sequence is either increasing or decreasing it is called a *monotone sequence*. This is a special case of the more general notion of a monotonic function.

7.1.6. *Bounded.* If the sequence of real numbers $\{a_n\}$ is such that all the terms, after a certain one, are less than some real number M, then the sequence is said to be *bounded from above*. In less words, this means $a_n \leq M \ \forall n > N$ for some pair M and N. Any such M is called an *upper bound*. Likewise, if, for some real m $a_n \geq m$ for all n greater than some N, then the sequence is *bounded from below* and any such m is called a *lower bound*. If a sequence is both bounded from above and bounded from below then the sequence is said to be *bounded*.

7.1.7. *Convergence.* One of the most important properties of a sequence is *convergence*. Informally, a sequence converges if it has a limit. Continuing informally, a (singly-infinite) sequence has a limit if it approaches some value L, called the limit, as n becomes very large. That is, for an abstract sequence $\{a_k\}_{k=1}^{\infty}$ (with n running from 1 to infinity understood) the value of the a_n's approaches L as n approaches infinity, denoted

$$\lim_{n\to\infty} a_n = L\,.$$

More precisely, the

> sequence converges if there exists a limit, L, such that the remaining a_n's are arbitrarily close to L for some n large enough.

If a sequence converges to some limit, then it is *convergent*; otherwise it is *divergent*.

If the a_n's get arbitrarily large as n approaches infinity then we write $\lim_{n\to\infty} a_n = \infty$. In this case the sequence diverges, or that it converges to infinity. If the a_n's become arbitrarily "small" negative numbers (large in magnitude) as n goes to positive infinity then we write $\lim_{n\to\infty} a_n = -\infty$ and say that the sequence diverges or converges to minus infinity.

7.1.8. *Applications and Important Results.* Important results for convergence and limits of (one-sided) sequences of real numbers include the following. These equalities are all true at least when both sides exist. For a discussion of when the existence of the limit on one side implies the existence of the other see a real analysis text such as can be found in the references.

- The limit of a sequence is unique.
- $\lim_{n\to\infty}(a_n \pm b_n) = \lim_{n\to\infty} a_n \pm \lim_{n\to\infty} b_n$.
- $\lim_{n\to\infty} ca_n = c \lim_{n\to\infty} a_n$.
- $\lim_{n\to\infty} a_n b_n = (\lim_{n\to\infty} a_n)(\lim_{n\to\infty} b_n)$.
- $\lim_{n\to\infty} (a_n/b_n) = \lim_{n\to\infty} a_n / \lim_{n\to\infty} b_n$ provided $\lim_{n\to\infty} b_n \neq 0$.
- $\lim_{n\to\infty} a_n^p = (\lim_{n\to\infty} a_n)^p$.
- If $a_n \leq b_n$ for all n greater than some N, then $\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n$.
- (Squeeze Theorem) If $a_n \leq c_n \leq b_n$ for all $n > N$ and $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$, then $\lim_{n\to\infty} c_n = L$.
- If a sequence is bounded and monotonic then it is convergent.
- A sequence is convergent if and only if every subsequence is convergent.

7.2. **Series.** A *series* is, informally speaking, the sum of the terms of a sequence. Finite sequences and series have defined first and last terms, whereas infinite sequences and series continue indefinitely.

In mathematics, given an infinite sequence of numbers $\{a_n\}$, a series is informally the result of adding all those terms together: $a_1 + a_2 + a_3 + \cdots$. These can be written more

compactly using the summation symbol $\sum$. An example is the famous series from Zeno's dichotomy and its mathematical representation:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots .$$

The *terms of the series* are often produced according to a certain rule, such as by a formula, or by an algorithm. As there are an infinite number of terms, this notion is often called an *infinite series*. Unlike finite summations, infinite series need tools from mathematical analysis, and specifically the notion of limits, to be fully understood and manipulated. In addition to their ubiquity in mathematics, infinite series are also widely used in other quantitative disciplines such as physics, computer science, and finance.

7.2.1. *Definition.* For any sequence of rational numbers, real numbers, complex numbers, functions thereof, *etc.*, the *associated series* is defined as the ordered formal sum

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \dots .$$

The *sequence of partial sums* $\{S_k\}$ associated to a series $\sum_{n=0}^{\infty} a_n$ is defined for each k as the sum of the sequence $\{a_k\}$ from a_0 to a_k

$$S_k = \sum_{n=0}^{k} a_n = a_0 + a_1 + \cdots + a_k .$$

By definition the series $\sum_{n=0}^{\infty} a_n$ *converges* to a limit L if and only if the associated sequence of partial sums $\{S_k\}$ converges to L. This definition is usually written as

$$L = \sum_{n=0}^{\infty} a_n \iff L = \lim_{k\to\infty} S_k .$$

7.2.2. *Convergent Series.* A series $\sum_{a_n}$ is said to *converge* or to *be convergent* when the sequence S_N of partial sums has a finite limit. If the limit of S_N is infinite or does not exist, the series is said to *diverge*. When the limit of partial sums exists, it is called the *sum of the series*

$$\sum_{n=0}^{\infty} a_n = \lim_{N\to\infty} S_N = \lim_{N\to\infty} \sum_{n=0}^{N} a_n .$$

An easy way that an infinite series can converge is if all the a_n are zero for n sufficiently large. Such a series can be identified with a finite sum, so it is only infinite in a trivial sense.

7.3. **Recursion.** *Recursion* is the process of repeating items in a self-similar way. For instance, when the surfaces of two mirrors are exactly parallel with each other the nested images that occur are a form of infinite recursion. The term has a variety of meanings specific to a variety of disciplines ranging from linguistics to logic. The most common application of recursion is in mathematics and computer science, in which it refers to a method of defining functions in which the function being defined is applied within its own definition. Specifically this defines an infinite number of instances (function values), using a finite expression that for some instances may refer to other instances, but in such a way that no loop or infinite chain of references can occur. The term is also used more generally to describe a process of repeating objects in a self-similar way.

7.3.1. *Definition.* A class of objects or methods exhibit recursive behavior when they can be defined by two properties:

(1) A simple base case (or cases).
(2) A set of rules that reduce all other cases toward the base case.

The Fibonacci sequence is a classic example of recursion:

- fib[0] is 0 [base case];
- fib[1] is 1 [base case];
- For all integers $n > 1$: fib[n] is (fib[$n-1$] + fib[$n-2$]) [recursive definition].

Many mathematical axioms are based upon recursive rules. For example, the formal definition of the natural numbers by the Peano axioms can be described as: 0 is a natural number, and each natural number has a successor, which is also a natural number. By this base case and recursive rule, one can generate the set of all natural numbers.

Recursively defined mathematical objects include functions, sets and especially fractals.

7.3.2. *Recursive definition.* A *recursive definition* (or inductive definition) is used to define an object in terms of itself.

A recursive definition of a function defines values of the functions for some inputs in terms of the values of the same function for other inputs. For example, the factorial function $n!$ is defined by the rules

$$0! = 1 \qquad \text{and} \qquad (n+1)! = (n+1)n!\,.$$

This definition is valid for all n, because the recursion eventually reaches the *base case* of 0. The definition may also be thought of as giving a procedure describing how to construct the function $n!$, starting from $n = 0$ and proceeding onward with $n = 1$, $n = 2$, $n = 3$, *etc.* That such a definition indeed defines a function can be proved by induction.

An inductive definition of a set describes the elements in a set in terms of other elements in the set. For example, one definition of the set $\mathcal{N}$ of natural numbers is:

- 1 is in $\mathcal{N}$.
- If an element n is in $\mathcal{N}$ then $n + 1$ is in $\mathcal{N}$.
- $\mathcal{N}$ is the smallest set satisfying the previous conditions.

There are many sets that satisfy the two first conditions; *e.g.*, the set $\{1, 1.649, 2, 2.649, 3, 3.649, \ldots\}$ satisfies the definition. However, the last condition specifies the set of natural numbers by removing the sets with extraneous members.

Properties of recursively defined functions and sets can often be proved by an induction principle that follows the recursive definition. For example, the definition of the natural numbers presented here directly implies the *principle of mathematical induction* for natural numbers: if a property holds of the natural number 0, and the property holds of $n + 1$ whenever it holds of n, then the property holds of all natural numbers.

7.3.3. *Form of recursive definitions.* Most recursive definition have three foundations: a base case (basis), an inductive clause, and an extremal clause.

The difference between a circular definition and a recursive definition is that a recursive definition must always have base cases, cases that satisfy the definition *without* being defined in terms of the definition itself, and all other cases comprising the definition must be "smaller" (closer to those base cases that terminate the recursion) in some sense. In contrast, a circular definition may have no base case, and define the value of a function in terms of that value itself, rather than on other values of the function. Such a situation would lead to an infinite regress.

8. Taylor Series

In mathematics, a *Taylor series* is a representation of a function as an infinite sum of terms calculated from the values of the function's derivatives at a single point.

If the Taylor series is centered at zero, then that series is also called a *Maclaurin series*.

It is common practice to approximate a function by using a finite number of terms of its Taylor series. Taylor's theorem gives quantitative estimates on the error in this approximation. Any *finite number of initial terms of the Taylor series generated by a function* is called a *Taylor polynomial*. The Taylor series of a function is the limit of that function's Taylor polynomials, provided that the limit exists. A function may not be equal to its Taylor series, even if its Taylor series converges at every point. A function that is equal to its Taylor series in an open interval (or a disc in the complex plane) is known as an *analytic function*.

Definition. Consider $f[x]$ to be a real-valued function infinitely differentiable in a neighborhood of a real number $a \in \mathcal{R}$. Then, define the Taylor series generated by f at the point a, denoted $T_\infty f[x\,;a]$, by the power series

$$T_\infty f[x\,;a] = \sum_{k=0}^{\infty} \frac{f^{(k)}[a]}{k!}(x-a)^k\,, \tag{8.1}$$

where $k!$ denotes the factorial of k and $f^{(k)}[a]$ the kth derivative of f evaluated at the point a. The derivative of order zero f is defined to be f itself and $(x-a)^0$ and $0!$ are both defined to be 1.

Call any finite number of initial terms, say n, of the Taylor series of the function f a *Taylor polynomial of degree n generated by f at the point a*, denoted $T_n f[x\,;a]$, by

$$T_n f[x\,;a] = \sum_{k=0}^{n} \frac{f^{(k)}[a]}{k!}(x-a)^k\,, \tag{8.2}$$

Also refer to the Taylor polynomial of degree n generated by f at the point a as the *n-degree Taylor polynomial generated by f at a*.

Finally, when $a = 0$, refer to Taylor series as *Maclaurin series*.

8.1. **Properties.** The Taylor operator $T_n[\,;]$ has the following properties:

- Linearity property: if c_1 and c_2 are constants, then
$$T_n[\,;]\,(c_1 f + c_2 g) = c_1 T_n f[\,;] + c_2 T_n g[\,;]\,.$$
- Differentiation property: the derivative of a Taylor polynomial of f is a Taylor polynomial of f':
$$(T_n f[\,;])' = T_{n-1} f'[\,;]\,.$$
- Integration property: an indefinite integral of a Taylor polynomial of f is a Taylor polynomial of and indefinite integral of f. More precisely, if $g[x] = \int_a^x f[t]\,\mathrm{d}t$, then
$$T_{n+1} g[x\,;a] = \int_a^x T_n f[t\,;]\,\mathrm{d}t\,.$$
- The Maclaurin series of an even function includes only even powers.
- The Maclaurin series of an odd function includes only odd powers.

8.2. **Geometry.** The 1-degree Taylor polynomial is the tangent line to $f[x]$ at $x = a$:

$$T_1 f[x\,;a] = f[a] + f'[a]\;(x-a)\,.$$

This is often called the *linear approximation to $f[x]$ near $x = a$*; *i.e.*, the *tangent line to the graph*. Therefore, view Taylor polynomials as a generalization of *linear approximations*. In particular, the 2-degree Taylor polynomial is sometimes called the *quadratic approximation*, the 3-degree Taylor polynomial is the *cubic approximation* and so forth.

8.3. **Applications.** Use Taylor series and Taylor polynomials for three important applications:

(1) to find the sum of a series;
(2) to evaluate limits;
(3) to approximate functions.

8.4. **Taylor's Theorem.** Let $k \geq 1$ be an integer and let the function $f : \mathcal{R} \to \mathcal{R}$ be k times differentiable at the point $a \in \mathcal{R}$. Then, there exists a function $h_k : \mathcal{R} \to \mathcal{R}$ such that

$$f[x] = f[a] + f'[a]\,(x-a) + \frac{1}{2!}f''[a]\,(x-a)^2 + \cdots + \frac{1}{k!}f^{(k)}[a]\,(x-a)^k + h_k[x]\,(x-a)^k$$

and $\lim_{x\to a} h_k[x] = 0$. This is called the *Peano form of the remainder*.

The polynomial appearing in Taylor's theorem is the k-th order Taylor polynomial

$$P_k[x] = f[a] + f'[a]\,(x-a) + \frac{1}{2!}f''[a]\,(x-a)^2 + \cdots + \frac{1}{k!}f^{(k)}[a]\,(x-a)^k$$

of the function f at the point a. The Taylor polynomial is the unique "asymptotic best fit" polynomial in the sense that if there exists a function $h_k : \mathcal{R} \to \mathcal{R}$ and a k-th order polynomial p such that

$$f[x] = p[x] + h_k[x]\,(x-a)^k\,, \qquad \lim_{x\to a} h_k[x] = 0\,,$$

then $p = P_k$. Taylor's theorem describes the asymptotic behavior of the remainder term

$$R_k[x] = f[x] - P_k[x]\ ,$$

which is the approximation error when approximating f with its Taylor polynomial.

8.5. **Formulae for the Remainder.**

8.5.1. *Mean-value forms of the remainder.* Let $f : \mathcal{R} \to \mathcal{R}$ be $k+1$ times differentiable on the open interval and continuous on the closed interval between a and x. Then,

$$R_k[x] = \frac{f^{(k+1)}(\xi_L)}{(k+1)!}(x-a)^{k+1}\,,$$

for some real number ξ_L between a and x. This is the *Lagrange form of the remainder*. Similarly,

$$R_k[x] = \frac{f^{(k+1)}(\xi_C)}{k!}(x-\xi_C)^k(x-a)\,,$$

for some real number ξ_C between a and x. This is the *Cauchy form of the remainder*.

8.5.2. *Integral Form of the Remainder.* Let $f[k]$ be absolutely continuous on the closed interval between a and x. Then,

$$R_k[x] = \int_a^x \frac{f^{(k+1)}[t]}{k!}(x-t)^k\,\mathrm{d}t\,.$$

8.5.3. *Estimates for the Remainder.* It is often useful in practice to be able to estimate the remainder term appearing in the Taylor approximation, rather than having a specific form of it. Suppose that f is $(k+1)$-times continuously differentiable in an interval I containing a. Suppose that there are real constants q and Q such that

$$q \leq f^{(k+1)}[x] \leq Q\,,$$

throughout I. Then, the remainder term satisfies the inequality

$$q\frac{(x-a)^{k+1}}{(k+1)!} \leq R_k[x] \leq Q\frac{(x-a)^{k+1}}{(k+1)!}$$

if $x > a$, and a similar estimate if $x < a$. This is a simple consequence of the Lagrange form of the remainder. In particular, if

$$|f^{(k+1)}[x]| \leq M$$

on an interval $I =]a-r, a+r[$ with some $r > 0$, then

$$|R_k[x]| \leq M\frac{|x-a|^{k+1}}{(k+1)!} \leq M\frac{r^{k+1}}{(k+1)!}$$

for all $x \in]a-r, a+r[$. The second inequality is called a *uniform estimate*, because it holds uniformly for all x on the interval $x \in]a-r, a+r[$.

8.6. **Examples.**

8.6.1. *Taylor Series and Maclaurin series.* Calculate the Taylor series of $\sin x$ at $x = 0$.

Solution. Let $f[x] = \sin x$. Then,

$$\begin{aligned}
&f[x] = \sin x \implies f[0] = \sin 0 = 0\,,\\
&f'[x] = \cos x \implies f'[0] = \cos 0 = 1\,,\\
&f''[x] = -\sin x \implies f''[0] = -\sin 0 = 0\,,\\
&f'''[x] = -\cos x \implies f'''[0] = -\cos 0 = -1\,,\\
&f^{iv}[x] = \sin x \implies f^{iv}[0] = \sin 0 = 0\,,\\
&\cdots
\end{aligned}$$

Note that the fourth derivative takes us back to the start point, so these values repeat in a cycle of four as $0, 1, 0, -1$; $0, 1, 0, -1$ and so on, with only the odd powers of x appearing in the polynomials. Besides, note that the function sin is infinitely differentiable and that all of its derivatives exist at $x = 0$. Therefore, plug these results into eq. (8.2) to find

$$T_{2n+1}\sin[x\,;0] = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty}(-1)^n\,\frac{x^{2n+1}}{(2n+1)!}\,.$$

Working on a similar fashion, find the Maclaurin series for the cosine function

$$T_{2n}\cos[x\,;0] = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots = \sum_{n=0}^{\infty}(-1)^n\,\frac{x^{2n}}{(2n)!}\,. \qquad \square$$

Note. The calculation of Maclaurin series is eased by noting the parity of functions. For instance, the sine function is odd; that is, $-\sin x = \sin -x$. Thus, Maclaurin series generated by the sine function will have only odd powers. In this way, calculate only the odd powers of the series, not all.

Similarly, the cosine function is even: $\cos x = \cos -x$. Thus, Maclaurin series generated by the cosine function will have only even powers. Therefore, calculate only the even powers of the series.

Calculate the Maclaurin series of e^x.

Solution. Let $f[x] = e^x$. Then,

$$\begin{aligned}
&f[x] = e^x \implies f[0] = e^0 = 1\,,\\
&f'[x] = e^x \implies f'[0] = e^0 = 1\,,\\
&f''[x] = e^x \implies f''[0] = e^0 = 1\,,\\
&\cdots
\end{aligned}$$

Since e^x is infinitely differentiable and that all of its derivatives exist at 0, then, plug these results into eq. (8.1) to find

$$T_{\infty}e[x\,;0] = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty}\frac{x^n}{n!}\,. \qquad \square$$

8.6.2. *Sum of Series.* Find the sum of the following series:

$$\sum_{n=0}^{\infty}\frac{1}{n!} = 1 + \frac{1}{1!} + \frac{2}{2!} + \frac{3}{3!} + \cdots\,.$$

Solution. Substitute $x = 1$ in the Taylor series generated by e^x to find the sum of the given series:

$$T_{\infty}e[1\,;0] = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots = e\,.$$

If $x = -1$ is substituted, then

$$T_{\infty}e[-1\,;0] = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!} = \frac{1}{e}\,.$$

8.6.3. *Limits.* Evaluate $\lim_{x\to 0} \dfrac{\sin x - x}{x^3}$.

Solution. Plug in the Taylor series generated by $\sin x$:

$$\begin{aligned}
\lim_{x\to 0} \frac{\sin x - x}{x^3} &= \lim_{x\to 0} \frac{\left(x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \cdots\right) - x}{x^3}\,, \\
&= \lim_{x\to 0} \frac{-\dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \cdots}{x^3}\,, \\
&= \lim_{x\to 0} \left(-\frac{1}{3!} + \frac{1}{5!}x^2 - \frac{1}{7!}x^4 + \cdots\right)\,, \\
&= -\frac{1}{6}\,.
\end{aligned}$$

□

8.6.4. *Approximations.* Find the 5-degree Taylor polynomial generated by $\sin\theta$ at $\theta = 0$. Use this result to approximate $\sin 0.3$.

Solution. The 5-degree Taylor polynomial generated by $\sin\theta$ is given by

$$T_{2n+1}\sin[\theta\,;0] = \sum_{k=0}^{5} (-1)^n \frac{\theta^{2n+1}}{(2n+1)!} = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} = \theta - \frac{\theta^3}{6} + \frac{\theta^5}{120}\,.$$

Approximate $\sin 0.3$ by setting $\theta = 0.3$ in the last equation

$$T_{2n+1}\sin[\theta\,;0] = 0.3 - \frac{0.3^3}{6} + \frac{0.3^5}{120} = 0.295\,520\,25\,.$$

□

9. Fourier Series

In mathematics, a *Fourier series* decomposes periodic functions or periodic signals into the sum of a (possibly infinite) set of simple oscillating functions, namely sines and cosines (or complex exponentials). The study of Fourier series is a branch of Fourier analysis.

The heat equation is a partial differential equation. Prior to Fourier's work, no solution to the heat equation was known in the general case, although particular solutions were known if the heat source behaved in a simple way, in particular, if the heat source was a sine or cosine wave. These simple solutions are now sometimes called eigensolutions. Fourier's idea was

> to model a complicated heat source as a superposition (or linear combination) of simple sine and cosine waves and to write the solution as a superposition of the corresponding eigensolutions. This superposition or linear combination is called the Fourier series.

Although the original motivation was to solve the heat equation, it later became obvious that the same techniques could be applied to a wide array of mathematical and physical problems and especially those involving linear differential equations with constant coefficients, for which the eigensolutions are sinusoids. The Fourier series has many such applications in electrical engineering, vibration analysis, acoustics, optics, signal processing, image processing, quantum mechanics, econometrics, thin-walled shell theory, *etc.*

9.1. **Definition.** In this section, $f[x]$ denotes a function of the real variable x. This function is usually taken to be periodic, of period 2π, which is to say that $f[x+2\pi] = f[x]$, for all real numbers x. We will attempt to write such a function as an infinite sum, or series, of simpler 2π-periodic functions. We will start by using an infinite sum of sine and cosine functions on the interval $[-\pi, \pi]$, as Fourier did and we will then discuss different formulations and generalizations.

9.2. **Fourier's formula for 2π-periodic functions using sines and cosines.** For a periodic function $f[x]$ that is integrable on $[-\pi, \pi]$, the numbers

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f[x]\,\cos[nx]\,\mathrm{d}x\,, \qquad n \geq 0$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f[x]\,\sin[nx]\,\mathrm{d}x\,. \qquad n \geq 1$$

are called the *Fourier coefficients of* f. One introduces the *Fourier partial sums of degree* n *generated by* f *at the point* x, often denoted by

$$F_N f[x] = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos[nx] + b_n \sin[nx]\right)\,, \qquad N \geq 0\,.$$

The partial sums for f are *trigonometric polynomials*. One expects that the functions $F_N f[x]$ approximate the function f and that the approximation improves as $N \to \infty$. The infinite sum

$$F_\infty f[x] = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos[nx] + b_n \sin[nx]\right)\,, \qquad N \geq 0\,.$$

is called the *Fourier series of generated by* f *at the point* x. These trigonometric functions can themselves be expanded, using multiple angle formulae.

The Fourier series does *not* always converge and even when it does converge for a specific value x_0 of x, the sum of the series at x_0 may differ from the value $f[x_0]$ of the function. It is one of the main questions in harmonic analysis to decide when Fourier series converge and when the sum is equal to the original function. If a function is square-integrable on the interval $[-\pi, \pi]$, then the Fourier series converges to the function at almost every point. In engineering applications, the Fourier series is generally presumed to converge everywhere except at discontinuities, since the functions encountered in engineering are more well behaved than the ones that mathematicians can provide as counter-examples to this presumption. In particular, the Fourier series converges absolutely and uniformly to $f[x]$ whenever the derivative of f (which may not exist everywhere) is square integrable.

It is possible to define Fourier coefficients for more general functions or distributions, in such cases convergence in norm or weak convergence is usually of interest.

9.3. **Properties.** We say that f belongs to $C^k[T]$ if f is a 2π-periodic function on $\mathcal{R}$ which is k times differentiable and its kth derivative is continuous.

- If f is a 2π-periodic *odd* function, then $a_n = 0$ for all n.
- If f is a 2π-periodic *even* function, then $b_n = 0$ for all n.

9.4. **Examples.** We now give a Fourier series expansion of a very simple function. Consider a sawtooth wave

$$\begin{aligned} f[x] &= \frac{x}{\pi} && \text{for } -\pi < x < \pi\,, \\ f[x + 2\pi k] &= f[x] && \text{for } -\infty < x < \infty \text{ and } k \in \mathcal{Z}\,. \end{aligned}$$

In this case, the Fourier coefficients are given by

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f[x]\ \mathrm{d}x = 0\,, \\ a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f[x]\cos[nx]\ \mathrm{d}x = 0\,, \qquad n > 0 \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f[x]\sin[nx]\ \mathrm{d}x = -\frac{2}{n}\cos[n\pi] + \frac{2}{\pi n^2}\sin[n\pi] = 2\frac{(-1)^{n+1}}{n}\,, \qquad n \geq 1\,. \end{aligned}$$

It can be proven that the Fourier series converges to $f[x]$ at every point x where f is differentiable and therefore:

$$\begin{aligned} f[x] = F_\infty f[x] &= \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n\cos[nx] + b_n\sin[nx]\right)\,, \\ &= 2\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\ \sin[nx] \qquad \text{for } x - \pi \in 2\pi\mathcal{Z}\,. \end{aligned}$$

When $x = \pi$, the Fourier series converges to 0, which is the half-sum of the left- and right-limit of f at $x = \pi$. This is a particular instance of the Dirichlet theorem for Fourier series.

10. Legendre Transform

The *Legendre transform*, *aka* Legendre transformation, is an involutive transformation on the real-valued convex functions of one real variable. Its generalisation to convex functions of affine spaces is sometimes called the Legendre-Fenchel transformation. It is commonly used in thermodynamics and to derive the Hamiltonian formalism of classical mechanics out of the Lagrangian formulation.

The Legendre transform exploits the equivalence between lines and points.

The Legendre transform is invertible by its own inverse: *i.e.*, apply the Legendre transform to f to find $f_\star$; then, apply the Legendre transform to $f_\star$ to arrive back to f.

[a graphical derivation of the legendre transform, Sam Kennerly]

The Legendre transform is a trick for representing a function in terms of its first derivative.

While mathematically rigorous descriptions are arguably unnecessary for many applications, some caution is necessary to avoid serious errors in practice. A few common sources of confusion are:

- failing to clearly state the necessary existence/uniqueness conditions,
- using notation which confuses numbers with functions and
- misinterpreting the somewhat-ambiguous formula $px - f[x]$.

The second error is especially popular with physicists. The symbol $y[x]$ is often used to represent both "the function $y[]$" and "the value of y at x". This *abuse of notation* is usually harmless, but it can be dangerous when *change-of-variable techniques* are used. Here we will use y to mean a number, $y[]$ to mean a function, and $y[x]$ to mean "the output of $y[]$ when given x as an input".

10.1. **Existence/uniqueness conditions for a Legendre transform.** Suppose all of the following statements are true:

(1) A well-behaved function $f[]$ is defined over some chunk $\mathcal{D}$ of the real line.
(2) For any $x \in \mathcal{D}$, you know how to find $f[x]$ and $f'[x]$.
(3) The graph of $f[x]$ always curves upward: for any $x \in \mathcal{D}$, then $f''[x] > 0$.

Condition 1 is deliberately vague; what do "well-behaved" and "chunk" mean? The point is: when using functions that fail common tests (*e.g.*, continuity, non-singularity, smoothness), then be careful. A more rigorous treatment than the one provided here may be necessary.

Condition 2 simply requires that an explicit formula for $f[x]$, its derivative can be found either by hand or by computer and that derivative is also well-behaved.

Condition 3 is *not always* stated explicitly, but it should be. Legendre transformations behave very badly if the curvature of $f[]$ changes sign as x changes. (If $f''[x]$ fails to exist at some points, see the subsection "Convex functions and convex sets.")

Suppose that instead of using x as a variable, you would prefer a new variable p such that $p[x] = f'[x]$. The Legendre transform produces a formula, in terms of p, for a new function $g[]$. The transform is *invertible*, so knowing $g[p]$ tells you everything about $f[x]$.

10.2. **Geometric interpretation of the Legendre transform.** Plot $f[x]$. At each point, imagine a line tangent to the plot. This line intersects $[x, f[x]]$ and has slope $p = f'[x]$. Any straight line with slope p must look like this for some $g \in \mathcal{R}$:

$$y[x] = px - g\,.$$

Here g means "the *negative* y-intercept of the line tangent to $f[]$ at the point $[x, f[x]]$". (We could have defined g to be the positive y-intercept, but that's not the usual convention.)

Since $f''[x] > 0$ everywhere, there is *only one* tangent line for each possible slope p. Draw pictures to convince yourself that if a function always curves upward, it can't have two tangent lines with the same slope.

For each possible slope p, there is *exactly one* tangent line. That tangent line has its y-intercept at $y = -g[p]$. We want to find the function $g[]$ that maps p's to g's.

The really useful thing about $g[]$ is this:

> each point $[x, f[x]]$ has exactly one "evil twin" point $[p, g[p]]$. Knowing $g[]$ then gives us complete information about the $f[]$ and *vice versa*.

10.3. **Recipe to Find the Legendre Transform.** The recipe for the Legendre transform is:

(1) Check that $f[]$ satisfies the existence/uniqueness conditions.
(2) Define a new function $p[]$ such that $p[x] \coloneqq f'[x]$. Then, invert $p[]$ and call the result $x[]$.
(3) Define g to be the negative of the y-intercept of the line tangent to $f[]$ at x:
$$g = p[x]\; x - f[x]\;.$$
(4) Use the formula for $x[p]$ to write the x's as functions of p. Call the result $g[p]$:
$$g[p] = p\, x[p] - f[x[p]]\;.$$

Just be careful to remember what $x[p]$ means: it is the value of x at which the slope of $f'[]$ is $f'[x] = p$. Otherwise this equation won't make any sense.

Additionally, change the notation [5] and wording to agree with more "standard" sources:

- Notation: denote the transformed function f by $f_\star$ and denote the transformed variable p by $x_\star$.
- Wording: consider a function f with domain $\mathcal{D}$ and consider $x \in \mathcal{D}$. Then, call $f_\star$ the *Legendre transform of* f and call $x_\star$ the *conjugate variable of* x.

10.4. **Examples.** Find the Legendre transform of $f : x \mapsto x^2$ with $x \in \mathcal{R}$.

Solution. Follow the steps in section 10.3 to find the solution:

- f is well behaved, f' is also well behaved and $f'' > 0$.
- Define $p[x] \coloneqq f'[x] = 2x$. Invert this to find $x[p]$: $p[x] = 2x \implies x[p] = p/2$.
- Define $g[x] \coloneqq p[x]\, x - f[x]$: $g[x] = (2x)x - x^2 = x^2$.
- Use $x[p] = p/2$ to write the x's as functions of p: $g[p] = (p/2)^2 = p^2/4$.
- The Legendre transform is finally: $f_\star[p_\star] = p_\star^2/4$.

Notice how we have changed the function f for its equivalent $f_\star$: analytically, the derivative $x_\star$ has replaced the independent variable x. Geometrically, the slope (tangent) $x_\star$ has replaced the x-coordinate; *i.e.*, a line has replaced a point.

Example. Invert the last Legendre transform.

Solution. To invert a Legendre transform means to apply the Legendre transform to an already transformed function. In the present case, transform the function $f : x \mapsto x^2/4$ with $x \in \mathcal{R}$.

- f is well behaved, f' is also well behaved and $f'' > 0$.
- Define $p[x] \coloneqq f'[x] = x/2$. Invert this to find $x[p]$: $p[x] = x/2 \implies x[p] = 2p$.
- Define $g[x] \coloneqq p[x]\, x - f[x]$: $g[x] = (x/2)x - (x^2/4) = x^2/4$.
- Use $x[p] = 2p$ to write the x's as functions of p: $g[p] = (2p)^2/4 = p^2$.
- The Legendre transform is finally: $f_\star[p_\star] = p_\star^2$, which is the original function we started with in the last example!

Example. Imagine a pendulum made of a very light, rigid rod of length r with a dense, point-like blob of mass m on one end. The other end is attached to a ball bearing which allows the pendulum to rotate 360° in a vertical plane. Define θ to be the angle between the rod and a vertical line and set $\theta = 0$, when the blob is at maximum height. Choose positive θ to be clockwise or counter-clockwise. Ignore friction but don't ignore gravity. Find the Lagrangian of the system and, then, its Legrendre transform: the Hamiltonian of the system.

Solution. The (approximate) gravitational potential energy of this object is $V = mgy = mgr\cos\theta$. The (approximate) rotational kinetic energy is $K = i\omega^2/2 = mr^2\omega^2/2$, where ω is the pendulum's angular velocity. The Lagrangian describing the system is $L = K - V$; *i.e.*,
$$L[\theta, \omega] = K - V = \frac{mr^2\omega^2}{2} - mgr\cos\theta\,.$$

The Hamiltonian of this system is found by Legendre-transforming L to remove the variable ω. (The variable θ comes along for the ride. For our purposes, θ can be thought of as a constant during the Legendre-transform process.) First, define $p[\omega] = L'[\omega]$:
$$p[\omega] = L'[\omega] = mr^2\omega\,.$$
Is $L''[\omega] > 0$ for all ω? Since $L''[\omega] = mr^2$ and $m > 0$, then it is. Note that p has a physical interpretation as the pendulum's angular momentum $mr^2\omega = i\omega$.

[5] The usage of g and p hides the transformation of f and x. That's why, we prefer more explicit notation: $f_\star$ and $x_\star$, respectively.

Now invert $p[\omega]$ to find $\omega[p] = p/mr^2$. Define $g = p[\omega] - L[\omega]$ as usual, use $\omega[p]$ to write everything in terms of p's, and call the result $g[p]$:

$$g[p] = \frac{p^2}{2mr^2} + mgr\cos\theta\,.$$

Remembering that θ is not really a constant, we should call it $g[\theta, p]$. Also, traditional notation uses H and L instead of g and p for "Hamiltonian" and "angular momentum".

$$H[\theta, L] = \frac{L^2}{2mr^2} + mgr\cos\theta\,.$$

This is the pendulum's Hamiltonian function. It has a physical interpretation as the total (kinetic plus potential) energy of the pendulum in terms of angular position and momentum.

In most simple physical systems like this one, using a Legendre transform to find the particle's Hamiltonian seems like extra work for no clear benefit; why not just write $H = K + V$ in the first place? For many practical calculations, this is an excellent criticism. The method is primarily important for providing a theoretical motivation for quantum mechanics.

11. Coordinate Systems

11.1. **Polar Coordinates.** In mathematics, the *polar coordinate system* is a two-dimensional coordinate system in which each point on a plane is determined by a distance from a fixed point and an angle from a fixed direction.

The fixed point (analogous to the origin of a Cartesian system) is called the *pole*, and the ray from the pole in the fixed direction is the polar axis. The distance from the pole is called the *radial coordinate* or *radius*, and the angle is the *angular coordinate*, *polar angle*, or azimuth.

11.1.1. *Conventions.* The radial coordinate is often denoted by r, and the angular coordinate by θ or t.

Angles in polar notation are generally expressed in either degrees or radians (2π rad being equal to 360°). Degrees are traditionally used in navigation, surveying, and many applied disciplines, while radians are more common in mathematics and mathematical physics.

In many contexts, a positive angular coordinate means that the angle θ is measured counterclockwise from the axis.

In mathematical literature, the polar axis is often drawn horizontal and pointing to the right.

11.1.2. *Converting between polar and Cartesian coordinates.* The polar coordinates r and θ can be converted to the Cartesian coordinates x and y by using the trigonometric functions sine and cosine:

$$x = r\cos[\theta] \qquad \text{and} \qquad y = r\sin[\theta]\ .$$

The Cartesian coordinates x and y can be converted to polar coordinates r and θ with $r \geq 0$ and θ in the interval $]-\pi, \pi]$ by:

$r^2 = x^2 + y^2 \implies r = \sqrt{x^2 + y^2}$ [as in the Pythagorean theorem or the Euclidean norm]

$\theta = \operatorname{atan2}[y, x]$,

where $\operatorname{atan2}[y, x]$ is defined in `http://en.wikipedia.org/wiki/Polar_coordinate_system`.

11.1.3. *Calculus.* Calculus can be applied to equations expressed in polar coordinates.

The angular coordinate θ is expressed in radians throughout this section, which is the conventional choice when doing calculus.

Differential calculus. Using $x = r\cos[\theta]$ and $y = r\sin[\theta]$, one can derive a relationship between derivatives in Cartesian and polar coordinates.

We have the following formulae:

$$r\frac{\partial}{\partial r} = x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} \qquad \text{and} \qquad \frac{\partial}{\partial \theta} = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}\ .$$

Using the inverse coordinates transformation, an analogous reciprocal relationship can be derived between the derivatives:

$$\frac{\partial}{\partial x} = \cos[\theta]\,\frac{\partial}{\partial r} - \frac{1}{r}\sin[\theta]\,\frac{\partial}{\partial \theta} \qquad \text{and} \qquad \frac{\partial}{\partial y} = \sin[\theta]\,\frac{\partial}{\partial r} + \frac{1}{r}\cos[\theta]\,\frac{\partial}{\partial \theta}\ .$$

Integral calculus (arc length). The arc length (length of a line segment) defined by a polar function is found by the integration over the curve $r[\theta]$. Let L denote this length along the curve starting from points A through to point B, where these points correspond to $\theta = a$ and $\theta = b$ such that $0 < b - a < 2\pi$. The length of L is given by the following integral

$$L = \int_a^b \sqrt{(r[\theta])^2 + \left(\frac{\mathrm{d}r}{\mathrm{d}\theta}[\theta]\right)^2}\ \mathrm{d}\theta\,.$$

Integral calculus (area). Let $\mathcal{R}$ denote the region enclosed by a curve $r[\theta]$ and the rays $\theta = a$ and $\theta = b$, where $0 < b - a \leq 2\pi$. Then, the area of $\mathcal{R}$ is

$$\frac{1}{2}\int_a^b (r[\theta])^2\ \mathrm{d}\theta\,.$$

Using Cartesian coordinates, an infinitesimal area element can be calculated as $\mathrm{d}A = \mathrm{d}x\mathrm{d}y$. The substitution rule for multiple integrals states that, when using other coordinates, the Jacobian determinant of the coordinate conversion formula has to be considered:

$$J = \det \frac{\partial[x,y]}{\partial[r,\theta]} = \begin{bmatrix} x_{,r} & x_{,\theta} \\ y_{,r} & y_{,\theta} \end{bmatrix} = \begin{bmatrix} \cos[\theta] & -r\sin[\theta] \\ \sin[\theta] & r\cos[\theta] \end{bmatrix} = r\cos^2[\theta] + r\sin^2[\theta] = r\,.$$

Hence, an area element in polar coordinates can be written as

$$\mathrm{d}A = \mathrm{d}x\mathrm{d}y = J\mathrm{d}r\mathrm{d}\theta = r\mathrm{d}r\mathrm{d}\theta\,.$$

Now, a function that is given in polar coordinates can be integrated as follows:

$$\iint_{\mathcal{R}} f[x,y]\ \mathrm{d}A = \int_a^b \int_a^{r[\theta]} f[r,\theta]\, r\,\mathrm{d}r\mathrm{d}\theta\,.$$

Here, $\mathcal{R}$ is the same region as above, namely, the region enclosed by a curve $r[\theta]$ and the rays $\theta = a$ and $\theta = b$.

The formula for the area of $\mathcal{R}$ mentioned above is retrieved by taking f identically equal to 1. A more surprising application of this result yields the Gaussian integral

$$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}\,.$$

Vector calculus. Vector calculus can also be applied to polar coordinates. For a planar motion, let x be the position vector $[r\cos[\theta]\,, r\sin[\theta]]$, with r and θ depending on time t.

We define the unit vectors

$$\hat{r} = [\cos[\theta]\,, \sin[\theta]]$$

in the direction of r and

$$\hat{\theta} = [-\sin[\theta]\,, \cos[\theta]] = \hat{k} \times \hat{r}$$

in the plane of the motion perpendicular to the radial direction, where $\hat{k}$ is a unit vector normal to the plane of the motion.

Then,

$$\begin{aligned} r &= [x,y] = r\,[\cos[\theta]\,, \sin[\theta]] = r\hat{r}\,, \\ \dot{r} &= [\dot{x},\dot{y}] = \dot{r}\,[\cos[\theta]\,, \sin[\theta]] + r\dot{\theta}\,[-\sin[\theta]\,, \cos[\theta]] = \dot{r}\hat{r} + r\dot{\theta}\hat{\theta}\,. \\ \ddot{r} &= [\ddot{x},\ddot{y}] = [\ldots] = \left(\ddot{r} - r\dot{\theta}^2\right)\hat{r} + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\hat{\theta}\,. \end{aligned}$$

11.2. **Spherical Coordinate System.** In mathematics, a *spherical coordinate system* is a coordinate system for three-dimensional space where the position of a point is specified by three numbers: the *radial distance* of that point from a fixed origin, its *polar angle* measured from a fixed zenith direction, and the *azimuth angle* of its orthogonal projection on a reference plane that passes through the origin and is orthogonal to the zenith, measured from a fixed reference direction on that plane.

The radial distance is also called the radius or radial coordinate. The polar angle may be called colatitude, zenith angle, normal angle, or inclination angle.

11.2.1. *Coordinate system conversions.* As the spherical coordinate system is only one of many three-dimensional coordinate systems, there exist equations for converting coordinates between the spherical coordinate system and others.

Cartesian coordinates: The spherical coordinates (radius r, inclination θ, azimuth ϕ) of a point can be obtained from its Cartesian coordinates $[x,y,z]$ by the formulae

$$r^2 = x^2 + y^2 + z^2\,, \qquad \theta = \arccos[z/r] \qquad \text{and} \qquad \phi = \arctan[y/x]\ .$$

The inverse tangent denoted in $\phi = \arctan[y/x]$ must be suitably defined, taking into account the correct quadrant of $[x,y]$. See article atan2.

11.2.2. *Kinematics.* In spherical coordinates the position of a point is written,

$$r = r\,\hat{r}\,,$$

its velocity is then,

$$v = \dot{r} = \dot{r}\,\hat{r} + r\dot{\theta}\,\hat{\theta} + r\dot{\phi}\sin[\theta]\;\hat{\phi}$$

and its acceleration is,

$$\begin{aligned} a = \dot{v} = &+ \left(\ddot{r} - r\dot{\theta}^2 - r\dot{\phi}^2\sin^2[\theta]\right)\hat{r} \\ &+ \left(r\ddot{\theta} + 2\dot{r}\dot{\theta} - r\dot{\phi}^2\sin[\theta]\cos[\theta]\right)\hat{\theta} \\ &+ \left(r\ddot{\phi}\sin[\theta] + 2\dot{r}\dot{\phi}\sin[\theta] + 2r\dot{\theta}\dot{\phi}\cos[\theta]\right)\hat{\phi} \end{aligned}$$

In the case of a constant ϕ or $\theta = \pi/2$ this reduces to vector calculus in polar coordinates.

11.3. **Generalization.** [James Foster, David Nightingale, A short course in general relativity]

In Cartesian coordinates, a point's position x in $\mathcal{E}^3$ is determined by three coordinates $[x, y, z]$. These three coordinates are associated with three orthonormal vectors: $\left[\hat{\imath}, \hat{\jmath}, \hat{k}\right]$. Since these vectors form a basis, then the point's position can be expressed as a linear combination of them

$$x = \gamma_k x^k = x\hat{\imath} + y\hat{\jmath} + k\hat{k}\,.$$

Say, we want to express the position of a vector using another coordinate system: $[u, v, w]$ whose relationships with the Cartesian coordinates are given by

$$x = f[u, v, w]\;, \qquad y = g[u, v, w] \qquad \text{and} \qquad z = h[u, v, w]\;.$$

It is possible, now, to express the Cartesian basis in the alternative coordinate system by using the *tangent vectors* $[\gamma_u, \gamma_v, \gamma_w]$ *defined* by

$$\gamma_u = x_{,u} = \frac{\partial x}{\partial u}\,, \qquad \gamma_v = x_{,v} = \frac{\partial x}{\partial v} \qquad \text{and} \qquad \gamma_w = x_{,w} = \frac{\partial x}{\partial w}\,.$$

These tangent vectors need *not* be orthogonal nor have unit length.

The metric coefficients g_{ij} are found by

$$g_{ij} = \gamma_i \cdot \gamma_j\,.$$

Onto this basis, x can be expressed as a liner combination of the basis elements

$$x = x^k \gamma_k\,,$$

where the components x^k are found via

$$x^k = x \cdot \gamma_k\,.$$

Using the inverse transformation, on the other hand,

$$u = f[x, y, z]\;, \qquad v = g[x, y, z] \qquad \text{and} \qquad w = h[x, y, z]\;,$$

it's possible to define a *normal basis* whose elements are defined by

$$\gamma^u = \operatorname{grad} u\,, \qquad \gamma^v = \operatorname{grad} v \qquad \text{and} \qquad \gamma^w = \operatorname{grad} w\,.$$

Onto this basis, the position of a particle can be expanded as

$$x = x_k \gamma^k$$

and the components x_k can be found via

$$x_k = x \cdot \gamma^k\,.$$

The metric coefficients g^{ij} are found by

$$g^{ij} = \gamma^i \cdot \gamma^j\,.$$

Finally, if everything went OK, the following condition must hold

$$g_{ij} g^{ij} = \delta^i_j\,.$$

Example. Consider $\mathcal{E}^2$. Express the position vector x in polar coordinates and then find $\dot{x}$.

Solution. In Cartesian coordinates, the position vector is written as

$$x = [x, y] \ .$$

The transformation from Cartesian coordinates to polar coordinates is given by

$$x = r\cos[\theta] \qquad \text{and} \qquad y = y\sin[\theta] \ .$$

Then, the position vector becomes

$$x = [x, y] = [r\cos[\theta], r\sin[\theta]] = r\,[\cos[\theta], \sin[\theta]] \ .$$

The basis vectors, thus, in polar coordinates can be calculated by their definitions

$$\gamma_r = \frac{\partial x}{\partial r} = [\cos[\theta], \sin[\theta]] \implies |\gamma_r| = 1 \implies \hat{r} = \gamma_r$$

in the r direction and

$$\gamma_\theta = \frac{\partial x}{\partial \theta} = [-r\sin[\theta], r\cos[\theta]] \implies |\gamma_\theta| = r \implies \hat{\theta} = \frac{\gamma_\theta}{r} = [-\sin[\theta], \cos[\theta]]$$

in the θ direction.

Therefore, the position vector can be rewritten as

$$x = [x, y] = r\,[\cos[\theta], \sin[\theta]] = r\gamma_r \ .$$

The velocity vector, next, can be calculated as

$$\dot{x} = \frac{\mathrm{d}(r\gamma_r)}{\mathrm{d}t} = \dot{r}\gamma_r + r\dot{\gamma_r} \ ,$$

where $\dot{\gamma_r}$ is given by

$$\dot{\gamma_r} = \frac{\mathrm{d}}{\mathrm{d}t}\,[\cos[\theta], \sin[\theta]] = \Big[-\sin[\theta]\,\dot{\theta}, \cos[\theta]\,\dot{\theta}\Big] = \dot{\theta}\,[-\sin[\theta], \cos[\theta]] = \dot{\theta}\hat{\theta} \ .$$

Finally, the velocity vector is

$$\dot{x} = \dot{r}\gamma_r + r\dot{\gamma_r} = \dot{r}\hat{r} + r\dot{\theta}\hat{\theta} \ . \qquad \square$$

Solution. An alternative form to represent the position vector is by using the definition of components

$$x^u = x \cdot \gamma_u \ .$$

Using this, the components of the position vector in polar coordinates are

$$x^r = x \cdot \gamma_r = [r\cos[\theta], r\sin[\theta]] \cdot [\cos[\theta], \sin[\theta]] = r\cos^2[\theta] + r\sin^2[\theta] = r$$

and

$$x^\theta = x \cdot \gamma_\theta = [r\cos[\theta], r\sin[\theta]] \cdot [-r\sin[\theta], r\cos[\theta]] = -r^2\sin[\theta]\cos[\theta] + r^2\sin[\theta]\cos[\theta] = 0 \ .$$

With this, the position vector becomes

$$x = x^r\gamma_r + x^\theta\gamma_\theta = r\gamma_r + 0 = r\gamma_r \ .$$

11.4. **Finally Formulas!** Given a frame $\{\gamma_k\}$, any vector, say x, can be expressed as a linear combination of the frame elements

$$x = x^k\gamma_k \ .$$

The components x^k can be found via

$$x^k = x \cdot \gamma_k \ .$$

The reciprocal frame $\left\{\gamma^k\right\}$ is given by

$$\gamma^k = {\gamma_k}^{-1} \ .$$

Thus, x can be expressed as

$$x = x_k\gamma^k \ .$$

The components x_k can be found via

$$x_k = x \cdot \gamma^k \ .$$

11.5. **Polar Coordinates Revisited!** Polar coordinates $[r, \theta]$ are related to Cartesian coordinates via

$$x = r\cos[\theta] \qquad \text{and} \qquad y = r\sin[\theta]\ .$$

Then, the position of a particle x can be written as

$$x = [x, y] = [r\cos[\theta]\,, r\sin[\theta]]\ .$$

The natural basis elements for polar coordinates become

$$\gamma_r = \frac{\partial x}{\partial r} = [\cos[\theta]\,, \sin[\theta]] \qquad \text{and} \qquad \gamma_\theta = \frac{\partial x}{\partial \theta} = [-r\sin[\theta]\,, r\cos[\theta]]\ .$$

Note that $|\gamma_r| = 1$, thus it is a unit vector, whereas $|\gamma_\theta| = r^2 \neq 1$, thus it is not a unit vector.

The components of x onto the natural basis are

$$\begin{cases} x^r = x\cdot\gamma_r = [r\cos[\theta]\,, r\sin[\theta]]\cdot[\cos[\theta]\,, \sin[\theta]] = r\cos^2[\theta] + r\sin^2[\theta] = r\,, \\ x^\theta = x\cdot\gamma_\theta = [r\cos[\theta]\,, r\sin[\theta]]\cdot[-r\sin[\theta]\,, r\cos[\theta]] = -r^2\sin[\theta]\cos[\theta] + r^2\sin[\theta]\cos[\theta] = 0\,. \end{cases}$$

Thus, x becomes

$$x = x^k\gamma_k = x^r\gamma_r + x^\theta\gamma_\theta = r\gamma_r + 0 = r\gamma_r\,.$$

The components of the metric g are

$$\begin{cases} g_{rr} = \gamma_r\cdot\gamma_r = [\cos[\theta]\,, \sin[\theta]]\cdot[\cos[\theta]\,, \sin[\theta]] = \cos^2[\theta] + \sin^2[\theta] = 1\,, \\ g_{\theta\theta} = \gamma_\theta\cdot\gamma_\theta = [-r\sin[\theta]\,, r\cos[\theta]]\cdot[-r\sin[\theta]\,, r\cos[\theta]] = r^2\sin^2[\theta] + r^2\cos^2[\theta] = r^2\,, \\ g_{r\theta} = g_{\theta r} = \gamma_r\cdot\gamma_\theta = [\cos[\theta]\,, \sin[\theta]]\cdot[-r\sin[\theta]\,, r\cos[\theta]] = -r\sin[\theta]\cos[\theta] + r\sin[\theta]\cos[\theta] = 0\,. \end{cases}$$

Note that, since $g_{r\theta} = g_{\theta r} = 0$, then the natural basis vectors are orthogonal.

The matrix representation of the metric is given by

$$g = \begin{bmatrix} 1 & 0 \\ 0 & r^2 \end{bmatrix}\ .$$

With the metric elements g_{ij}, we can find the square of the separation vector between two neighboring points

$$\mathrm{d}s^2 = \mathrm{d}x\mathrm{d}x = \mathrm{d}x\cdot\mathrm{d}x = g_{ij}\mathrm{d}x^i\mathrm{d}x^j = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2\,,$$

and then the Lagrangian L for a free particle as

$$L = \frac{1}{2}m\,\mathrm{d}\dot{x}\mathrm{d}\dot{x} = \frac{1}{2}m\,g_{ij}\mathrm{d}\dot{x^i}\mathrm{d}\dot{x^j} = \frac{1}{2}m\left(\mathrm{d}\dot{r}^2 + r^2\mathrm{d}\dot{\theta}^2\right)\ .$$

11.6. **Procedure.** Consider that the position of a particle can be determined by a position vector $x \in \mathcal{E}^3$. Then, to change the coordinate system follow the next procedure:

(1) Express x in Cartesian coordinates: $x = [x, y, z]$.
(2) Find the transformation between Cartesian coord. to the alternate coordinate system:

$$x = f[u, v, w]\ , \qquad y = g[u, v, w] \qquad \text{and} \qquad z = h[u, v, w]$$

and the inverse transformation

$$u = f[x, y, z]\ , \qquad v = g[x, y, z] \qquad \text{and} \qquad w = h[x, y, z]\ .$$

(3) Calculate the elements of the natural (tangent) basis

$$\gamma_u = \frac{\partial x}{\partial u}\,, \qquad \gamma_v = \frac{\partial x}{\partial v} \qquad \text{and} \qquad \gamma_w = \frac{\partial x}{\partial w}\,.$$

(4) Calculate the metric coefficients for the natural basis

$$g_{ij} = \gamma_i\cdot\gamma_j\,.$$

(5) Express x as a linear combination of the natural basis elements

$$x = x^k\gamma_k\,,$$

where the components x^k are given by

$$x^k = x\cdot\gamma_k\,.$$

(6) Using the inverse transformation, find the elements of the dual (normal) basis

$$\gamma^u = \operatorname{grad} u\,, \qquad \gamma^v = \operatorname{grad} v \qquad \text{and} \qquad \gamma^w = \operatorname{grad} w\,.$$

(7) Calculate the metric coefficients for the dual basis

$$g^{ij} = \gamma^i \cdot \gamma^j\,.$$

(8) Express x as a linear combination of the natural basis elements

$$x = x_k \gamma^k\,,$$

where the components x_k are given by

$$x_k = x \cdot \gamma^k\,.$$

(9) Finally, to verify results, the following condition must hold:

$$g_{ij} g^{ij} = \delta^i_j\,.$$

11.7. **Tangents and gradients.** By dropping the requirement that our coordinate systems be orthogonal, we have found ourselves in the position of having two different, but related, bases at each point of space. Is this one two many? To avoid confusion, should we reject one of them and retain the other? If so, which one? As we shall see, each has its uses, and there are situations where it is appropriate to use the natural basis $\{\gamma_i\}$ defined by the tangents to the coordinate curves, while in other situations it is appropriate to use the dual basis $\{\gamma^i\}$ defined by the normals to the coordinate surfaces. Let us start by looking at the tangent vector to a curve in space.

11.7.1. *Tangents.* Suppose we put

$$u = u[t]\,, \qquad v = v[t] \qquad \text{and} \qquad w = w[t]\,,$$

where $u[t]$, $v[t]$ and $w[t]$ are differentiable functions of t for t belonging to some interval I. Then the points with coordinates given by the last equation will lie on a curve $\mathcal{C}$ parameterized by t. The position vector of these points is

$$x = x[u[t]\,, v[t]\,, w[t]]\,\hat{\imath} + y[u[t]\,, v[t]\,, w[t]]\,\hat{\jmath} + z[u[t]\,, v[t]\,, w[t]]\,\hat{k}\,.$$

and for each t in I the derivative $\dot{x}[t] = \mathrm{d}x/\mathrm{d}t$ gives a tangent vector to the curve (provided $\dot{x}[t] \neq 0$). Using the chain rule we have

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial x}{\partial u}\frac{\mathrm{d}u}{\mathrm{d}t} + \frac{\partial x}{\partial v}\frac{\mathrm{d}v}{\mathrm{d}t} + \frac{\partial x}{\partial u}\frac{\mathrm{d}w}{\mathrm{d}t}\,,$$

which can be written as

$$\dot{x}[t] = \dot{u}[t]\,\gamma_u + \dot{v}[t]\,\gamma_v + \dot{w}[t]\,\gamma_w\,.$$

The suffix notation version of this last equation is

$$\dot{x}[t] = \dot{u^i}[t]\,\gamma_i\,,$$

showing that the derivatives $\dot{u^i}[t]$ are the components of the tangent vector to the curve $\mathcal{C}$ relative to the natural basis γ_i. So

for tangents to curves, it is appropriate to use the natural basis.

The length of the curve $\mathcal{C}$ is obtained by integrating $|\dot{x}|$ with respect to t over the interval I. Now

$$|\dot{x}|^2 = \dot{x}\dot{x} = \dot{x} \cdot \dot{x} = \dot{u^i}\gamma_i \cdot \dot{u^j}\gamma_j = g_{ij}\dot{u^i}\dot{u^j}\,,$$

on using equation the definition of the quantities g_{ij}. So if I is given by $a \leq t \leq b$, then the length of $\mathcal{C}$ is given by

$$L = \int_a^b \left(g_{ij}\dot{u^i}\dot{u^j}\right)^{1/2} \mathrm{d}t\,.$$

The infinitesimal version of the equation $\dot{x}[t] = \dot{u^i}[t]\,\gamma_i$ is $\mathrm{d}x = \mathrm{d}u^i\gamma_i$, which gives

$$\mathrm{d}s^2 = \mathrm{d}x\mathrm{d}x = \mathrm{d}u^i\gamma_i \cdot \mathrm{d}u^j\gamma_j$$

for the distance between points whose coordinates differ by $\mathrm{d}u^i$. We thus arrive at the formula

$$\mathrm{d}s^2 = g_{ij}\mathrm{d}u^i\mathrm{d}u^j\,.$$

11.7.2. *Gradients.* Suppose now that we take a differentiable function $\phi[u, v, w]$ of the coordinates u, v, w. This will give us a function of position and therefore a scalar field. Its gradient is

$$\operatorname{grad}\phi = \nabla\phi = \frac{\partial\phi}{\partial x}\hat{\imath} + \frac{\partial\phi}{\partial y}\hat{\jmath} + \frac{\partial\phi}{\partial z}\hat{k}\,,$$

where, in calculating these partial derivatives, we are regarding ϕ as a function of x, y, z got by substituting the expressions for u, v, w in terms of x, y, z (the inverse transformation):

$$\phi = \phi[u[x, y, z]\,, v[x, y, z]\,, w[x, y, z]]\ .$$

The chain rule gives

$$\frac{\partial\phi}{\partial x} = \frac{\partial\phi}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial\phi}{\partial v}\frac{\partial v}{\partial x} + \frac{\partial\phi}{\partial w}\frac{\partial w}{\partial x}$$

with similar expressions for $\phi_{,y}$ and $\phi_{,z}$. Hence we can say that

$$\operatorname{grad}\phi = \cdots = \frac{\partial\phi}{\partial u}\nabla u + \frac{\partial\phi}{\partial v}\nabla v + \frac{\partial\phi}{\partial w}\nabla w\,.$$

That is,

$$\operatorname{grad}\phi = \frac{\partial\phi}{\partial u}\gamma^{u} + \frac{\partial\phi}{\partial v}\gamma^{v} + \frac{\partial\phi}{\partial w}\gamma^{w}\,,$$

on using the definitions of the dual basis elements. The suffix notation version of this is

$$\operatorname{grad}\phi = \nabla\phi = \frac{\partial\phi}{\partial u^{i}}\gamma^{i}\,,$$

showing that the partial derivatives $\partial\phi/\partial u^{i}$ are the components of $\operatorname{grad}\phi$ relative to the dual basis. Note that, in letting the repeated suffix imply summation in the last equation, we are regarding the suffix i on $\partial\phi/\partial u^{i}$ as a subscript. We can make this point more clearly by shortening the partial differential operator $\partial/\partial u^{i}$ to ∂_i, so that $\partial\phi/\partial u^{i} = \partial_i\phi$. The notation $\phi_{,i}$ is also used to mean the same thing. We can then rewrite the last equation as

$$\nabla\phi = \partial_i\phi\gamma^{i} = \phi_{,i}\gamma^{i}\,,$$

with the suffix correctly occupying the subscript position.

11.7.3. *Conclusion.* Thus, we see that when dealing with tangents to curves it is appropriate to use the natural basis $\{\gamma_i\}$ defined by the coordinate system, but when dealing with gradients of scalar fields it is appropriate to use the dual basis $\{\gamma^{i}\}$. This conclusion is not surprising, given the way in which the two bases are defined.

11.8. **Yet Another Way of Calculating Basis Elements.** A particle position x in Cartesian coordinates is given by $x = [x, y]$.

Transform the position components to polar coordinates by

$$x = r\cos[\theta] \qquad \text{and} \qquad y = r\sin[\theta]\ ,$$

where $\dim r = [L]$ and $\dim\theta = [1]$.

Express the position in polar coordinates:

$$x = [r\cos[\theta]\,, r\sin[\theta]]\ .$$

Then, find the basis elements for polar coordinates

$$\gamma_r = \frac{\partial x}{\partial r} = [\cos[\theta]\,, \sin[\theta]] \quad \text{and} \quad \gamma_\theta = \frac{\partial x}{\partial\theta} = [-r\sin[\theta]\,, r\cos[\theta]] = r\,[-\sin[\theta]\,, \cos[\theta]]\ ,$$

where $\dim\gamma_r = [1]$, but $\dim\gamma_\theta = [L]$.

Next, rewrite x as a linear combination of the basis elements

$$x = [x, y] = [r\cos[\theta]\,, r\sin[\theta]] = r\,[\cos[\theta]\,, \sin[\theta]] = r\gamma_r\,.$$

Consider now that $\theta = \theta[t]$. Then, the differential of the particle position becomes

$$\mathrm{d}x = \mathrm{d}\,(r\cos[\theta]\,, r\sin[\theta]) = (\mathrm{d}r\cos[\theta] - r\sin[\theta]\,\mathrm{d}\theta, \mathrm{d}r\sin[\theta] + r\cos[\theta]\,\mathrm{d}\theta)\ ,$$

where the product and chain rules were used.

Separate the components of the vector in the last equation to

$$\mathrm{d}x = (\mathrm{d}r\cos[\theta]\,, \mathrm{d}r\sin[\theta]) + (-r\sin[\theta]\,\mathrm{d}\theta, r\cos[\theta]\,\mathrm{d}\theta)\ .$$

Factor out the common terms in the vectors components to have

$$\mathrm{d}x = \mathrm{d}r\,(\cos[\theta]\,,\sin[\theta]) + r\mathrm{d}\theta\,(-\sin[\theta]\,,\cos[\theta])\ .$$

Use the definition of the basis elements in the last equation

$$\mathrm{d}x = \mathrm{d}r\,\gamma_r + \mathrm{d}\theta\,\gamma_\theta\ .$$

Note that this expression is dimensionally homogeneous: $\dim \mathrm{d}x = [L]$, $\dim \mathrm{d}r\gamma_r = [L]\,[1]$ and $\dim \mathrm{d}\theta\gamma_\theta = [1]\,[L]$.

To find the particle velocity, divide the last equation by the time differential $\mathrm{d}t$ and use the dot notation to represent time derivatives:

$$\dot{x} = \dot{r}\,\gamma_r + \dot{\theta}\,\gamma_\theta\ .$$

Refer to $\dot{r}$ as the *radial velocity* and to $\dot{\theta}$ as the *angular velocity* whose dimensions are $\dim \dot{r} = [L/T]$ and $\dim \dot{\theta} = [1/T]$.

Normalize the basis elements, so to have unit length and dimensionless basis elements

$$\hat{r} = \gamma_r/|\gamma_r| = \gamma_r \quad \text{and} \quad \hat{\theta} = \gamma_\theta/|\gamma_\theta| = \gamma_\theta/r\,,$$

which implies that $\gamma_\theta = r\hat{\theta}$.

Then, the equation for the particle position becomes

$$x = r\gamma_r = r\hat{r}\,.$$

Therefore, the equation for the particle velocity turns into

$$\dot{x} = \dot{r}\gamma_r + \dot{\theta}\gamma_\theta = \dot{r}\hat{r} + \dot{\theta}(r\hat{\theta}) = \dot{r}\hat{r} + r\dot{\theta}\hat{\theta}\,.$$

Refer to $r\dot{\theta}$ as the *tangencial velocity*. Note that $\dim r\dot{\theta} = [L/T]$; *i.e.*, the last equation is dimensionally homogeneous. Additionally, see that the normalized (normal) basis elements are dimensionless and thus r in $r\dot{\theta}$ acts as a *conversion factor* between polar and Cartesian coordinates – r links linear and circular motions.

Note. Remember that the components of the particle position in Cartesian coordinates all measure lengths. But in polar coordinates one component measure lengths, while the other one angles. Then, when using non-normalized basis elements, the basis elements themselves provide a measure for lengths. On the other hand, when using normalized basis elements, r provides the conversion factor mentioned above, so that the components of the position also measure lengths, leaving, thus, the normal basis elements dimensionless.

As an example of the last note, consider cylindrical coordinates $\{\rho, \phi, z\}$ and the related normal basis $\left\{\hat{\rho}, \hat{\phi}, \hat{z}\right\}$.

According to their definitions, find that

$$\dim \rho = [L]\ , \quad \dim \phi = [1] \quad \text{and} \quad \dim z = [L]\ .$$

Then, consider Laplace operator in such a coordinate system:

$$\nabla = \frac{\partial}{\partial \rho}\hat{\rho} + \frac{1}{\rho}\frac{\partial}{\partial \phi}\hat{\phi} + \frac{\partial}{\partial z}\,.$$

Note that $\dim \nabla = [1/L]$ in the LHS and, since normal, the basis elements are dimensionless in the RHS. Thus, ρ in $(1/\rho)\partial_\phi$ provides the "missing" dimensional factor of length, so to render the operator dimensionally homogeneous.

11.9. **Coordinates and Langrangian.** Consider $\mathcal{E}^n$ and consider a system of curvilinear coordinates $\left\{q^i : i : 1, \ldots, n\right\}$. Then, express the position vector x as a tuple of the curvilinear coordinates

$$x = \left[q^1, \ldots, q^n\right]\ .$$

Find a tangent frame to the system of curvilinear coordinates with elements defined by

$$\gamma_i = \frac{\partial x}{\partial q^i}\,;$$

note that the frame need not be orthonormal.

Then, find the metric coefficients g_{ij} by

$$g_{ij} = \gamma_i \cdot \gamma_j\,.$$

Next, write the position vector as $x = \gamma_i q^i$. Then, the position vector differential becomes

$$\mathrm{d}x = \gamma_i \mathrm{d}q^i \, .$$

Thus, the differential distance turns into

$$\mathrm{d}s^2 = \mathrm{d}x\mathrm{d}x = \mathrm{d}x \cdot \mathrm{d}x = \gamma_i q^i \cdot \gamma_j q^j = \gamma_i \cdot \gamma_j q^i q^j = g_{ij} q^i q^j \, .$$

Calculate the velocity by $\dot{x} = \mathrm{d}x/\mathrm{d}t = \gamma_i \dot{q}^i$. Then, compute the kinetic energy by

$$k = \frac{1}{2} m \dot{x}^2 = \frac{1}{2} m \gamma_i \dot{q}^i \cdot \gamma_j \dot{q}^j = \frac{1}{2} m \gamma_i \cdot \gamma_j \dot{q}^i \dot{q}^j = \frac{1}{2} m g_{ij} \dot{q}^i \dot{q}^j \, .$$

Finally, the Lagrangian becomes

$$L = k = \frac{1}{2} m g_{ij} \dot{q}^i \dot{q}^j \, .$$

12. Lagrangian Mechanics

Nature is thrifty in all its actions.

— pierre-louis moreau de maupertuis, wiki: principle of least action ;)

The term *generalized coordinates* refers to the parameters that describe the configuration of the system relative to some reference configuration. These parameters must *uniquely* define the configuration of the system relative to the reference configuration. The *generalized velocities* are the time derivatives of the generalized coordinates of the system.

An example of a generalized coordinate is the angle (anti-clockwise from some reference point) that locates a point moving on a circle. The adjective "generalized" distinguishes these parameters from the traditional use of the term coordinate to refer to Cartesian coordinates – canonical coordinates; *e.g.*, describing the location of the point on the circle using x and y coordinates.

Although there may be many choices for generalized coordinates for a physical system, parameters are usually selected which are convenient for the specification of the configuration of the system and which make the solution of its equations of motion easier. If these parameters are independent of one another, then number of independent generalized coordinates is defined by the number of degrees of freedom of the system.

Finally, canonical coordinates (Cartesian coordinates) all measure lengths, whereas generalized coordinates measure different dimensions; for instance, in polar coordinates $\{r, \theta\}$, r measures lengths and θ measures angles; *i.e.*, θ is dimensionless. Therefore, generalized velocities may not measure in dimensions of length per unit time, generalized momenta may not measure in dimensions of mass times length, generalized forces may not measure in dimensions of mass times length per squared unit time and so on.

Example. Dynamic model of a simple pendulum. The relationship between the use of generalized coordinates and Cartesian coordinates to characterize the movement of a mechanical system can be illustrated by considering the constrained dynamics of a simple pendulum.

Solution. A simple pendulum consists of a mass m hanging from a pivot point so that it is constrained to move on a circle of radius l. The position of the mass is defined by the coordinate vector $r = [x[t]\,, y[t]]$ measured in the plane of the circle such that y is in the vertical direction. The coordinates $x[t]$ and $y[t]$ are related by the equation of the circle

$$f[x[t]\,, y[t]] = x^2[t] + y^2[t] - l^2 = x^2 + y^2 - l^2 = 0\,,$$

that constrains the movement of m. This equation also provides a constraint on the velocity components,

$$\dot{f}[x[t]\,, y[t]] = f_{,x}\dot{x} + f_{,y}\dot{y} = 2x[t]\,x'[t] + 2y[t]\,y'[t] = 2x\dot{x} + 2y\dot{y}\,.$$

Now introduce the parameter θ, that defines the angular position of m from the vertical direction. It can be used to define the coordinates $x[t]$ and $y[t]$, such that the coordinate vector

$$r = [x[t]\,, y[t]] = [l\sin\theta[t]\,, -l\cos\theta[t]] = l\,[\sin\theta, \cos\theta]$$

and the velocity becomes

$$\dot{r} = \left[l\dot{\theta}[t]\cos\theta[t]\,, l\dot{\theta}[t]\sin\theta[t]\right] = l\dot{\theta}\,[\cos\theta, \sin\theta]\ .$$

The use of $\theta[t]$ to define the configuration of this system avoids the constraint provided by the equation of the circle.

12.1. **Generalized Coordinates.** A set of *generalized coordinates* $\left\{q^1, q^2, \ldots, q^n\right\} = \left\{q^i\right\}$ is a set that completely describes the positions of all particles in a mechanical system. In a system with d_{f} degrees of freedom and k constraints, $n = d_{\mathrm{f}} - k$ *independent* generalized coordinates are needed to completely specify all the positions. A *constraint* is a relation among coordinates, such as $x^2 + y^2 + z^2 = a^2$ for a particle moving on a sphere of radius a. In this case, $d_{\mathrm{f}} = 3$ and $k = 1$, so we could eliminate z in favor of x and y, *i.e.*, by writing $z = \pm\sqrt{a^2 - x^2 - y^2}$, or we could choose as coordinates the polar and azimuthal angles θ and ϕ.

For the moment we will assume that $n = d_{\rm f} - k$, and that the generalized coordinates are *independent*, satisfying no additional constraints among them. Later on we will learn how to deal with any remaining constraints among the generalized coordinates $\{q^i\}$.

The generalized coordinates may have dimensions of length, or angle, or perhaps something totally different. In the theory of small oscillations, the normal coordinates are conventionally chosen to have dimensions of $(\left[M^{1/2}.L\right])$. However, once a choice of generalized coordinate is made, with a concomitant set of dimensions, the dimensions of the *conjugate momentum* p^i and *conjugate force* f^i are determined:

$$\dim p^i = \frac{\left[M.L^2\right]}{[T]} \frac{1}{\dim q^i} \qquad \dim f^i = \frac{\left[M.L^2\right]}{[T^2]} \frac{1}{\dim q^i} \,.$$

These choices are such that, if q^i has dimensions of length, then p^i has dimensions of momentum and f^i has dimensions of force. If q^i is dimensionless, as is the case for an angle, p^i has dimensions of angular momentum $(\left[M.L^2/T\right])$ and f^i has dimensions of torque $(\left[M.L^2/T^2\right])$.

12.2. **Principle of Least Action.** In physics, the *principle of least action* – or, more accurately, the *principle of stationary action* – is a variational principle that, when applied to the action of a mechanical system, can be used to obtain the equations of motion for that system. The principle led to the development of the Lagrangian and Hamiltonian formulations of classical mechanics.

The principle remains *central* in modern physics and mathematics, being applied in the theory of relativity, quantum mechanics and quantum field theory, and a focus of modern mathematical investigation in Morse theory. The chief examples of the principle of stationary action are Maupertuis' principle and Hamilton's principle.

The action principle is preceded by earlier ideas in surveying and optics. The rope stretchers of ancient Egypt stretched corded ropes between two points to measure the path which minimized the distance of separation, and Claudius Ptolemy, in his Geographia (Bk 1, Ch 2), emphasized that one must correct for "deviations from a straight course"; in ancient Greece Euclid states in his Catoptrica that, for the path of light reflecting from a mirror, the angle of incidence equals the angle of reflection; and Hero of Alexandria later showed that this path was the shortest length and least time. But the credit for the formulation of the principle as it applies to the action is often given to Pierre-Louis Moreau de Maupertuis, who wrote about it in 1744 and 1746. However, scholarship indicates that this claim of priority is not so clear; Leonhard Euler discussed the principle in 1744 and there is evidence that Gottfried Leibniz preceded both by 39 years.

Maupertuis: Credit for the formulation of the principle of least action is commonly given to Pierre-Louis Moreau de Maupertuis, who felt that

> Nature is thrifty in all its actions,

and applied the principle broadly:

> The laws of movement and of rest deduced from this principle being precisely the same as those observed in nature, we can admire the application of it to all phenomena. The movement of animals, the vegetative growth of plants [...] are only its consequences; and the spectacle of the universe becomes so much the grander, so much more beautiful, the worthier of its Author, when one knows that a small number of laws, most wisely established, suffice for all movements.

This notion of Maupertuis, although somewhat deterministic today, *does* capture much of the essence of mechanics.

Lagrange and Hamilton: Much of the calculus of variations was stated by Joseph Louis Lagrange in 1760 and he proceeded to apply this to problems in dynamics. In *Méchanique Analytique* (1788) Lagrange derived the general equations of motion of a mechanical body. William Rowan Hamilton in 1834 and 1835 applied the variational principle to the classical Lagrangian function

$$L = k - v \,,$$

to obtain the *Euler-Lagrange equations* in their present form.

12.3. **Hamilton's Principle.** The equations of motion of classical mechanics are embodied in a variational principle, called *Hamilton's principle*. Hamilton's principle states that the motion of a system is such that the *action functional*

$$A[q[t]] = \int_{t_1}^{t_2} \mathrm{d}t\, L[q, \dot{q}, t] \ ,$$

is an extremum, *i.e.*, $\delta A = 0$. Here, $q = \left\{q^i\right\}$ is a complete set of *generalized coordinates* for our mechanical system, and

$$L = k - v$$

is the *Lagrangian*, where k is the *kinetic energy* and v is the *potential energy*. Setting the first variation of the action to zero gives the *Euler-Lagrange equations*,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0\,,$$

where the first term inside the parenthesis $L_{,\dot{q}^i} = p^i$ is the generalized momentum and the second term $L_{,q^i} = f^i$ is the generalized force. Thus, the Euler-Lagrange equations imply the familiar $\dot{p}^i = f^i$, *aka Newton's second law of motion*. Note, however, that the $\left\{q^i\right\}$ are generalized coordinates, so p^i may not have dimensions of momentum, nor f^i of force. For example, if the generalized coordinate in question is an angle ϕ, then the corresponding generalized momentum is the angular momentum about the axis of ϕ's rotation, and the generalized force is the torque.

12.4. **Momentum conservation.** Whenever L is independent of a generalized coordinate q^i, the conjugate force f^i vanishes and therefore the conjugate momentum p^i is conserved. This is an example of a deep result known as *Noether's theorem*. Noether's theorem guarantees that to every *continuous symmetry of* L there corresponds an associated *conserved quantity*.

12.5. **Invariance of the equations of motion.** The equations of motion are *invariant under a shift of* L by a total time derivative of a function of coordinates and time.

12.6. **Remarks on the Choice of Generalized Coordinates.** Any choice of generalized coordinates will yield an equivalent set of equations of motion. However, some choices result in an apparently simpler set than others. This is often true with respect to the form of the potential energy. Additionally, certain constraints that may be present are more amenable to treatment using a particular set of generalized coordinates.

The kinetic energy k is always simple to write in Cartesian coordinates, and it is good practice, at least when one is first learning the method, to write k in Cartesian coordinates and then convert to generalized coordinates. In Cartesian coordinates, the kinetic energy of a single particle of mass m is

$$k = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)\,.$$

If the motion is two-dimensional, and confined to the plane $z = \text{const.}$, one $2k = m(\dot{x}^2 + \dot{y}^2)$.

Two other commonly used coordinate systems are the cylindrical and spherical systems. In cylindrical coordinates $[\rho, \phi, z]$, ρ is the radial coordinate in the $[x, y]$ plane and ϕ is the azimuthal angle:

$$x = \rho\cos\phi \qquad\qquad \dot{x} = \dot{\rho}\cos\phi - \rho\dot{\phi}\sin\phi\,,$$
$$y = \rho\sin\phi \qquad\qquad \dot{y} = \dot{\rho}\sin\phi + \rho\dot{\phi}\cos\phi\,.$$

and the third, orthogonal coordinate is z. The kinetic energy is then

$$k = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) = \tfrac{1}{2}m(\dot{\rho}^2 + \rho^2\dot{\phi}^2 + \dot{z}^2)\,.$$

When the motion is confined to a plane with $z = \text{const.}$, this coordinate system is often referred to as "two-dimensional polar" coordinates.

12.7. **How to Solve Mechanics Problems.** Here are some simple steps you can follow toward obtaining the equations of motion:

(1) Choose a set of generalized coordinates $\{q^1, \dots, q^n\}$.
(2) Find the kinetic energy $k[q, \dot{q}, t]$, the potential energy $v[q, t]$ and the Lagrangian $L[q, \dot{q}, t] = k - v$. It is often helpful to first write the kinetic energy in Cartesian coordinates for each particle before converting to generalized coordinates.
(3) Find the canonical momenta $p^i = L_{,\dot{q}^i}$ and the generalized forces $f^i = L_{,q^i}$.
(4) Evaluate the time derivatives $\dot{q}^i$ and write the equations of motion $\dot{p}^i = f^i$. Be careful to differentiate properly, using the chain rule and the Leibniz rule where appropriate.
(5) Identify any conserved quantities.
(6) Note about wording: when using Cartesian coordinates, *aka*, canonical coordinates, all of the quantities are named *canonical*; *e.g.*, canonical coordinates, canonical momentum, canonical force and so on. When using other coordinates, *aka*, generalized coordinates, all of the quantities are named *generalized*; *e.g.*, generalized coordinates, generalized momentum, generalized force and so on.
(7) Note about physical interpretation: the generalized quantities have physical interpretations depending on the underling coordinates being used.

12.8. **Examples.** Find the equation of one-dimensional motion.

Solution. Use Cartesian coordinates – canonical coordinates. Then, choose the canonical coordinate x to describe the one-dimensional mechanical system. These system has potential energy $v[x]$ and, thus, the Lagrangian is

$$L = k - v = \tfrac{1}{2} m\dot{x}^2 - v[x] \, .$$

The canonical momentum (canonical because we are using Cartesian coordinates) is

$$p = \frac{\partial L}{\partial \dot{x}} = m\dot{x} \, .$$

and the canonical force

$$f = \frac{\partial L}{\partial x} = v'[x] \, .$$

The equation of motion is finally

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0 \implies m\ddot{x} = -v'[x] \, ,$$

which is $f = ma$.

Note that we can multiply the equation of motion by $\dot{x}$ to get

$$0 = m\dot{x}\left(m\ddot{x} + v'[x]\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\tfrac{1}{2} m\dot{x}^2 + v[x]\right) = \frac{\mathrm{d}E}{\mathrm{d}t} = \dot{E} \, ,$$

where $E = k + v$ is the total energy of the system.

Example. Consider next a particle of mass m moving in two dimensions under the influence of a potential $v[\rho]$ which is a function of the distance from the origin $\rho^2 = x^2 + y^2$. Find the equations of motion.

Solution. Using polar coordinates, the Lagrangian becomes

$$L = \tfrac{1}{2} m\left(\dot{\rho}^2 + \rho\dot{\phi}^2\right) + v'[\rho] \, .$$

The equations of motions via the Euler-Lagrange equations are

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\rho}}\right) - \frac{\partial L}{\partial \rho} = m\ddot{\rho} - m\rho\dot{\phi}^2 + v'[\rho] \, , \\ &\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\phi}}\right) - \frac{\partial L}{\partial \phi} = 0 \implies \frac{\mathrm{d}}{\mathrm{d}t}\left(m\rho^2\dot{\phi}\right) = 0 \, . \end{aligned}$$

Note that the canonical momentum conjugate to ϕ, which is to say the angular momentum, is conserved

$$p^\phi = m\rho^2\dot{\phi} = \text{const.} \, .$$

Use this to eliminate $\dot{\phi}$ from the first Euler-Lagrange equation to obtain

$$m\ddot{\rho} = \frac{(p^\phi)^2}{m\rho^3} - v'[\rho] \, .$$

The total energy $E = k + v$ can then be written as

$$E = \tfrac{1}{2} m\dot{\rho}^2 + \frac{(p^\phi)^2}{m\rho^3} + v[\rho] \,,$$

from which it may be shown that E is also a constant:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \left(m\ddot{\rho} - \frac{(p^\phi)^2}{m\rho^3} + v'[\rho] \right) \dot{\rho} = 0 \,. \qquad \square$$

12.9. **Conserved Quantities.** A conserved quantity $\Lambda[q, \dot{q}, t]$ is one which does not vary throughout the motion of the system. This means

$$\left. \frac{\mathrm{d}\Lambda}{\mathrm{d}t} \right|_{q=q[t]} = 0 \,.$$

Momentum conservation: The simplest case of a conserved quantity occurs when the Lagrangian does *not* explicitly depend on one or more of the generalized coordinates, *i.e.*, when

$$f^i = \frac{\partial L}{\partial q^i} = 0 \,.$$

We then say that L is *cyclic in the coordinate* q^i (*i.e.*, when the generalized force in the coordinate q^i vanishes). In this case, the Euler-Lagrange equations $\dot{p}^i = f^i$ say that the conjugate momentum p^i is conserved. (This case is the exact analog to the Newtonian formalism: when there are no forces acting on a particle, the particle linear momentum is preserved: $\dot{p} = f$, but $f = 0$. Thus, $\dot{p} = 0$, or p is constant.)

Example. Consider the motion of a particle of mass m near the surface of the earth. Let $[x, y]$ be coordinates parallel to the surface and z the height. Find the equations of motion.

Solution. The Lagrangian for the particle motion is

$$L = k - v = \tfrac{1}{2} m \left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2 \right) - mgz \,.$$

Since $f^x = L_{,x} = 0$ and $f^y = L_{,y} = 0$, then p^x and p^y are conserved with

$$p^x = \frac{\partial L}{\partial \dot{x}} = m\dot{x} \qquad \text{and} \qquad p^y = \frac{\partial L}{\partial \dot{y}} = m\dot{y} \,.$$

Integrate the first order Euler-Lagrange equations to yield

$$x[t] = x[0] + \tfrac{p^x}{m} t \qquad \text{and} \qquad y[t] = y[0] + \tfrac{p^y}{m} t \,.$$

The z equation is

$$\dot{p}^z = m\ddot{z} = -mg = f^z \,,$$

with solution

$$z[t] = z[0] + \dot{z}[0]\, t - \tfrac{1}{2} g t^2 \,. \qquad \square$$

12.10. **Summary of Lagrangian mechanics.** [advanced mechanics, Eric Poisson]

The methods of Newtonian mechanics, based on the vectorial equation $f = \dot{q}$, are very powerful and they can be applied to *all* mechanical systems. But they lack in efficiency when Cartesian coordinates $[x, y, z]$ do not give the simplest description of a mechanical system.

To increase the efficiency of the theoretical methods of mechanics, a number of scientists in the centuries following Newton endeavored to recast the Newtonian laws into a more flexible formulation. The most famous players include Leonhard Euler (1707-1783), Joseph Lagrange (1736-1813), William Rowan Hamilton (1805-1865), and Carl Gustav Jacobi (1804-1851). Their new techniques proved extremely useful and they allowed them and others to solve increasingly challenging problems, most notably in the context of celestial mechanics. These new powerful techniques are the topic of this chapter on Lagrangian mechanics and the following chapter on Hamiltonian mechanics.

It is important to point out that

> the Lagrangian and Hamiltonian formulations of the laws of mechanics are largely restricted to forces that can be derived from a potential; *i.e.*, conservative forces.

For other problems, such as a particle subjected to air resistance, the new techniques cannot be applied in a very straightforward way and it is usually best to go back to the old Newtonian methods. In this chapter and the next, we shall consider only forces that can be derived from a potential.

The entire content of Lagrangian mechanics is summarized in the following simple recipe:

(1) Verify the applicability of the Lagrangian formalism: forces must be conservative, for only then they can be expressed as a potential: $f_{\text{cons}} = -\nabla v$. (Conservative forces depend only on the position vector $f_{\text{cons}}[x, t]$, whereas non-conservative forces on position and velocity $f_{\text{non-cons}}[x, \dot{x}, t]$.)
(2) Select generalized coordinates $\{q^i : i = 1, 2, \dots\}$ to describe the degrees of freedom of a mechanical system. These coordinates are completely *arbitrary*. They need not be the original Cartesian coordinates associated with an inertial frame. Indeed, there is no need for the coordinates to even be attached to an inertial frame. The index i labels each one of the generalized coordinates; there is one coordinate for each degree of freedom.
(3) In terms of the generalized coordinates, calculate the system's total kinetic energy k and total potential energy v. Then form what is known as the Lagrangian function of the system, denoted $L[q^i, \dot{q}^i]$; this depends on the generalized coordinates q^i and the generalized velocities $\dot{q}^i := \mathrm{d}q/\mathrm{d}t$. The Lagrangian is defined by $L = k - v$; it is the *difference* between the kinetic and potential energies.
(4) Substitute the Lagrangian into the Euler-Lagrange (EL) equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0\,.$$

This returns *an equation of motion for each generalized coordinate* $q^i[t]$; *i.e.*, there is one EL equation for each generalized coordinate.
(5) Identify any conserved quantities.
(6) The rest of the recipe is concerned with solving the equations of motion. The methods for doing this are varied, and they depend on the particular situation, just as they do in the Newtonian formulation.
(7) Note about wording: when using Cartesian coordinates, *aka*, canonical coordinates, all of the quantities are named *canonical*; *e.g.*, canonical coordinates, canonical momentum, canonical force and so on. When using other coordinates, *aka*, generalized coordinates, all of the quantities are named *generalized*; *e.g.*, generalized coordinates, generalized momentum, generalized force and so on.
(8) Note about physical interpretation: the generalized quantities have physical interpretations depending on the underling coordinates being used.

Let us first verify that the recipe is compatible with Newton's laws.

Example. Consider a particle moving in three-dimensional space and subjected to a potential $v[x, y, z]$. Describe the motion of the particle using Cartesian coordinates – canonical coordinates.

Solution. In this case, therefore, the generalized coordinates are chosen as $q^1 = x$, $q^2 = y$ and $q^3 = z$. The particle's kinetic energy is $2k = m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$ and the Lagrangian function is

$$L[x, y, z, \dot{x}, \dot{y}, \dot{z}] = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - v[x, y, z]\ .$$

To substitute this into the EL equation for $q^1 = x$, say, we must first evaluate the generalized momentum $p^x = L_{,\dot{x}}$. This is the partial derivative of L with respect to $\dot{x}$, treating all other variables (including x) as constant parameters. This is given by

$$p^x = \frac{\partial L}{\partial \dot{x}} = m\dot{x}\,.$$

(The generalized momentum coincides with the canonical momentum, since we're using Cartesian coordinates.)

We next differentiate this with respect to t to get the rate change of generalized momentum $\dot{p}^x$:

$$\dot{p}^x = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = m\ddot{x}\,.$$

Finally, we partially differentiate L with respect to x, treating all other variables (including $\dot{x}$) as constant parameters; this gives the generalized force f^x:

$$f^x = \frac{\partial L}{\partial x} = -\frac{\partial v}{\partial x}\,.$$

(The generalized force coincides with the canonical force, since we're using Cartesian coordinates.)

Substituting these results into the EL equation for x, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0 \implies m\ddot{x} + \frac{\partial v}{\partial x} = 0\,.$$

Repeating these calculations for y and z would eventually return the full vectorial equation

$$ma + \nabla v = 0\,.$$

or $ma = f$ if we recall that the force is derived from the potential, so that $f = -\nabla v$. This exercise reveals that indeed,

the Lagrangian recipe is compatible with the Newtonian law.

The true power of the recipe, however, is revealed when the generalized coordinates are not Cartesian, not canonical. Let us see what the recipe produces in the case of a pendulum.

Example. Find the equations of motion for a pendulum.

Solution. The pendulum's single degree of freedom is best represented by the swing angle θ; this will be our generalized coordinate for this problem and we write $q = \theta$. (We do not need a label, index, i in this case, as there is only one generalized coordinate.) The relation between θ and the original Cartesian coordinates is $x = l\sin\theta$ and $z = l\cos\theta$, with l denoting the length of the rod. The pendulum's kinetic energy is $2k = m(\dot{x}^2 + \dot{z}^2)$. Its potential energy is $v = -mgz = -mgl\cos\theta = -ml^2\omega^2\cos\theta$, where we have introduced the quantity $\omega^2 = g/l$. The pendulum's Lagrangian function is

$$L\left[\theta, \dot{\theta}\right] = ml^2\left(\tfrac{1}{2}\dot{\theta}^2 + \omega^2\cos\theta\right)\,.$$

To substitute this into the EL equation we must first evaluate the generalized momentum p^θ – the partial derivative of L with respect to $\dot{\theta}$. This is

$$p^\theta = \frac{\partial L}{\partial \dot{\theta}} = ml^2\dot{\theta}\,.$$

Next we calculate the change in momentum $\dot{p}^\theta$ – we differentiate the last equation with respect to time:

$$\dot{p}^\theta = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{\theta}} = ml^2\ddot{\theta}\,.$$

Finally, we calculate the generalized force f^θ – the partial derivative of L with respect to θ:

$$f^\theta = \frac{\partial L}{\partial \theta} = ml^2\ddot{\theta}\,.$$

Substituting these results into the EL equation produces:

$$\dot{p}^\theta - f^\theta = 0 \implies ml^2\left(\ddot{\theta} + \omega^2\sin\theta\right) = 0 \implies \ddot{\theta} + \omega^2\sin\theta = 0\,.$$

This last equation can also be obtained by Newtonian methods – the calculation is fairly laborious. However, comparing the computations carried out here to those required by Newtonian methods shows the greater efficiency of the Lagrangian recipe.

12.11. **The Classical Lagrangian.** [advanced mechanics and general relativity, joel franklin]

12.11.1. *The Lagrangian.* A Lagrangian is the integrand of an action – while this is not the usual definition, it is, upon definition of action, more broadly applicable than the usual "kinetic minus potential" form. In classical mechanics, the Lagrangian leading to Newton's second law reads, in Cartesian coordinates, with $r[t]$ being the position vector:

$$L \coloneqq \tfrac{1}{2}m\,\dot{r}[t]\cdot\dot{r}[t] - v[r[t]] = \tfrac{1}{2}m\,v[t]\cdot v[t] - v[r[t]]\ ,$$

where we view x, y and z as functions of a parameter t which we normally interpret as "time". The first term is the kinetic energy (denoted k), the second is the potential energy. Remember, the ultimate goal of classical mechanics is to find the trajectory of a particle under the influence of a force. Physically, we control the description of the system by specifying the particle mass, form of the force or potential and boundary conditions (particle starts from rest, particle moves from point $\mathcal{A}$ to point $\mathcal{B}$, *etc.*). Mathematically,

we use the equations of motion derived from the Lagrangian, together with the boundary conditions, to determine the curve $r[t] = x^k[t]\,\gamma_k$ through three-dimensional space with a frame $\{\gamma_k\}$.

Calculus of Variations provides the ordinary differential equation (ODE) structure of interest, a set of three second-order differential equations, the Euler-Lagrange equations of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial v} - \frac{\partial L}{\partial r} = 0 \implies \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{r}} - \frac{\partial L}{\partial r} = 0\,, \tag{12.1}$$

where we identify the particle's momentum $p = L_{,v}$ and force on the particle $f = L_{,r} = -\nabla v$ (note that since $f = -\nabla v$, the force must be conservative; *i.e.*, only a function of position and not velocity). With these identifications, the Euler-Lagrange equations could be written as $\dot{p} = f$, which is precisely Newton's second law of motion.

The advantage of the action approach, and the Lagrangian in particular, is that the equations of motion can be obtained for *any* coordinate representation of the kinetic energy and potential. Although it is easy to define and verify the correctness of the Euler-Lagrange equations in Cartesian coordinates, they are *not necessary* to the formulation of valid equations of motion for systems in which Cartesian coordinates are less physically and mathematically useful.

The Euler-Lagrange equations, in the form eq. (12.1), hold regardless of our association of r, the position vector, with Cartesian coordinates. Suppose we move to cylindrical coordinates $[\rho, \phi, z]$, defined by

$$x = \rho\cos\phi\,, \qquad y = \rho\sin\phi \qquad \text{and} \qquad z = z\,,$$

then the Lagrangian in Cartesian coordinates can be transformed to cylindrical coordinates by making the replacement for $[x, y, z]$ in terms of $[\rho, \phi, z]$ (and associated substitutions for the Cartesian velocities):

$$L[\rho, \phi, z] = L[r[\rho, \phi, z]] = \tfrac{1}{2}m\left(\dot{\rho}^2 + \dot{\phi}^2 + \dot{z}^2\right) - v[\rho, \phi, z]\ .$$

But, the Euler-Lagrange equations require no modification, the variational procedure that gave us eq. (12.1) can be applied in the cylindrical coordinates, giving three equations of motion:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\left(\dfrac{\partial L}{\partial \dot{\rho}}\right) - \dfrac{\partial L}{\partial \rho} = 0\,, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\left(\dfrac{\partial L}{\partial \dot{\phi}}\right) - \dfrac{\partial L}{\partial \phi} = 0\,, \\ \dfrac{\mathrm{d}}{\mathrm{d}t}\left(\dfrac{\partial L}{\partial \dot{z}}\right) - \dfrac{\partial L}{\partial z} = 0\,. \end{cases}$$

The advantage is clear: *coordinate transformation occurs once and only once, in the Lagrangian.* If we were to start with Newton's second law, we would have three equations with acceleration and coordinates coupled together. The decoupling of these would, in the end, return into the last set of equations.

12.11.2. *Metric Tensor.* In the mathematical field of differential geometry, a *metric tensor* is a type of function defined on a manifold (such as a surface in space) which takes as input a pair of tangent vectors v and w and produces a real number (scalar) $g[v, w]$ in a way that generalizes many of the familiar properties of the dot product of vectors in Euclidean space. In the same way as a dot product, metric tensors are used to define the length of, and angle between, *tangent vectors.*

[axioms of metric tensor: `https://en.wikipedia.org/wiki/Metric_tensor#Definition`]

Basically, a generic metric should be bilinear, symmetric and nondegenarated.

Consider a region $\mathcal{V}^n$ in $\mathcal{R}^n$. Then, define the *metric tensor*, denoted g, by

$$g[x, y] \coloneqq \langle xy\rangle_0 = x \cdot y\,.$$

Agree on calling the metric tensor simply *metric.*

Remark. The metric tensor is *not* the same as the Euclidean metric. The Euclidean metric requires that if a vector is inserted in its two slots, the output should be zero! On the contrary, the metric tensor produces the length of a vector in such a case. Thus, the metric tensor behaves more like the Euclidean norm or magnitude, in GA.

Odd the name!

12.11.3. *Geometric Form of the Lagrangian.* Consider three points $\mathcal{P}, \mathcal{Q}, \mathcal{O} \in \mathcal{E}^3$. Relative to $\mathcal{O}$, denote the position of $\mathcal{P}$ by the vector $x[\mathcal{P}]$ and the position of $\mathcal{Q}$ by the vector $x[\mathcal{Q}]$. Then, define the *separation between two points*, denoted s, by

$$\Delta s \coloneqq x[\mathcal{P}] - x[\mathcal{Q}] = \Delta x_{\mathcal{P},\mathcal{Q}} \,.$$

Assume $\mathcal{P}$ and $\mathcal{Q}$ were so close to each other, so to write $\mathrm{d}s = \mathrm{d}x$. Refer to $\mathrm{d}s$ as the *differential separation vector*.

Define, then, the *squared differential length*, denoted $\mathrm{d}s^2$, by

$$\mathrm{d}s^2 \coloneqq \mathrm{d}x^2 = \mathrm{d}x\,\mathrm{d}x = \mathrm{d}x \cdot \mathrm{d}x \,,$$

since $\mathrm{d}x$ is colinear to itself; *i.e.*, $\mathrm{d}x \wedge \mathrm{d}x = 0$. Note that, by the contraction axiom, $\mathrm{d}s^2$ is a scalar.

In terms of the metric, the squared differential length becomes

$$\mathrm{d}s^2 = g[\mathrm{d}x, \mathrm{d}x] \,.$$

Now, consider a particle of mass m moving under the interaction [6] with a *potential* $v[x]$. Then, define the *particle kinetic energy*, denoted k, by

$$k[\dot{x}] \coloneqq \tfrac{1}{2} m\, v^2 = \frac{1}{2} m\, v \cdot v = \frac{1}{2} m\, \dot{x} \cdot \dot{x} = \frac{1}{2} m\, g[\dot{x}, \dot{x}] \,.$$

With this result, define the *particle Lagrangian*, denoted L, by

$$L[x, \dot{x}] = k - v = \tfrac{1}{2} m\, \dot{x} \cdot \dot{x} - v[x] = \tfrac{1}{2} m\, g\dot{x}, \dot{x} - v[x] \,.$$

With the last equation, find the Euler-Lagrange equations, which lead to the equations of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0 \,.$$

Remark. In the above equation, x and $\dot{x}$ represent the particles' position and velocity, both of them are *vectors*.

Refer to the term $L_{,\dot{x}}$ as *momentum* and to the term $L_{,x}$ *force*. See that the force comes from the potential; *i.e.*, $L_{,x} = -\nabla v[x]$.

Finally, note the use of the *geometric principle* applied to physics throughout this section: *no* coordinate systems – thus *no* components, coordinates neither unit vectors – were required to define points, separation vector, distance, metric, mass, potential and kinetic energy and Lagrangian, since they are all geometric objects.

12.11.4. *Equations of Motion in Index Notation.* Consider now a coordinate system with a basis $\mathcal{B}$ for $\mathcal{E}^3$. Then, express vectors by its components onto the basis, say $a = a^i$, where the indices i, j run from 1 to 3. Then, the length of a vector is given by the metric (tensor)

$$g[a, a] = g_{ij} a^i a^j \,,$$

where the numbers $\{g_{ij}\}$ are called the *metric elements onto the basis* $\mathcal{B}$.

By letting x to represent the position vector, by noting that velocity is defined as $v = \dot{x}$ and by using the metric elements, express thus the Lagrangian of a particle of mass m moving under a potential $v[x]$ as

$$L = \tfrac{1}{2} m\, g_{ij} \dot{x}^i\, \dot{x}^j - v[x] \,.$$

Refer to the last equation as the *index form of the Lagrangian*.

Remark. The index form of the Lagrangian does *not* directly express any particular coordinate system – it works for them all.

Finally, the Euler-Lagrange equations, which lead to the equations of motion, are in index notation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i} = 0 \,.$$

Caution. Be careful! It is possible (as in the spherical case) that the metric depends on the coordinates; *i.e.*, $g[x]$. So, use the chain rule and the Leibniz rule where appropriate when deriving the equations of motion from the Euler-Lagrange equations.

[6] Since the interaction comes from a potential, the force on the particle must be conservative; *i.e.*, it should depend exclusively on the particle position $f[x]$. Only then does one have $f = -\nabla v$.

12.12. **Lagrangian Coordinate Transformation Recipe.** It is possible to go from canonical coordinates (cartesians) directly to other coordinate systems and find the form of the Lagrangian without directly transforming it, but, indirectly, using the metric.
Example. Transform the Lagrangian from Cartesian coordinates to spherical coordinates using the metric.

Solution. First, write the differential distance in both coordinate systems:

$$\begin{cases} \mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \,, & \text{[Cartesian coordinates]} \\ \mathrm{d}s^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\,\sin^2\theta\,\mathrm{d}\theta^2 \,. & \text{[spherical coordinates]} \end{cases}$$

Rewrite the differential distance in matrix notation using cartesians

$$\mathrm{d}s^2 = \begin{bmatrix}\mathrm{d}x & \mathrm{d}y & \mathrm{d}z\end{bmatrix} \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \begin{bmatrix}\mathrm{d}x\\ \mathrm{d}y\\ \mathrm{d}z\end{bmatrix} .$$

Rewrite the differential distance in matrix notation using sphericals

$$\mathrm{d}s^2 = \begin{bmatrix}\mathrm{d}r & \mathrm{d}\theta & \mathrm{d}\phi\end{bmatrix} \begin{bmatrix}1 & 0 & 0\\ 0 & r^2 & 0\\ 0 & 0 & r^2\sin^2\theta\end{bmatrix} \begin{bmatrix}\mathrm{d}r\\ \mathrm{d}\theta\\ \mathrm{d}\phi\end{bmatrix} .$$

Next, write the Lagrangian in matrix notation using cartesians

$$L = \tfrac{1}{2}m \begin{bmatrix}\dot{x} & \dot{y} & \dot{z}\end{bmatrix} \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{z}\end{bmatrix} - v[x,y,z] \,.$$

Finally, write the Lagrangian in matrix notation using sphericals

$$L = \tfrac{1}{2}m \begin{bmatrix}\dot{r} & \dot{\theta} & \dot{\phi}\end{bmatrix} \begin{bmatrix}1 & 0 & 0\\ 0 & r^2 & 0\\ 0 & 0 & r^2\sin^2\theta\end{bmatrix} \begin{bmatrix}\dot{r}\\ \dot{\theta}\\ \dot{\phi}\end{bmatrix} - v[x,y,z] \,. \qquad \square$$

Noting the similarities between the equations in the last example, write a prescription to transform the Lagrangian to *any* coordinate system:

(1) Define the coordinate system to work in and express its relationship to Cartesian coordinates. Calculate the differential separation between points and then the squared differential length in such a system. Note the metric elements of the squared differential length. (This is due to the relationship between the metric and the differential length: they are the same! Both measure distances, but have different form. So, if we know the squared differential length, we know the metric elements and, the other way around, if we know the metric elements, we can find the differential distance.)
(2) Write the kinetic energy in Cartesian coordinates and there replace the metric coefficients and replace the coordinate differentials with the coordinate distances; *e.g.*, in Cartesians, the coordinate differentials are $[\mathrm{d}x, \mathrm{d}y, \mathrm{d}z]$ and the coordinate distances are $[\dot{x}, \dot{y}, \dot{z}]$. This means transform the $\mathrm{d}x$'s into $\dot{x}$'s.
(3) Finally, express the Lagrangian in the chosen coordinate system.

The trick works from the realization that the differential distance and the kinetic energy (the part of the Lagrangian we want to transform) have similar *structure*. In index notation:

$$\mathrm{d}s^2 \sim \mathrm{d}x \cdot \mathrm{d}x \sim g[\mathrm{d}x, \mathrm{d}x] \sim g_{ij}\,\mathrm{d}x^i \mathrm{d}x^j \,,$$
$$k \propto v \cdot v \propto \dot{x} \cdot \dot{x} \propto g[\dot{x}, \dot{x}] \propto g_{ij}\,\dot{x}^i \dot{x}^j \,.$$

Besides the $m/2$ factor omitted in the kinetic energy, the only difference between the two equations is that the $\mathrm{d}x$'s in the differential distance changed into $\dot{x}$'s in the kinetic energy. The metric coefficients are, however, the same!

Remember, finally, that the differential distance, the metric, the kinetic energy and the Lagrangian are geometric objects, they need not coordinate systems for their definitions. Nevertheless, for expressing results and calculations, different *representations* of these objects in different coordinate systems are required. The metric is the object that enables such transformations between geometric objects from one representation to another.

12.13. **Lagrangian Coordinate Transformation Recipe – Revisited.** Relationship among the particle position x, the particle kinetic energy k and the metric g in any coordinate system:

Consider an n-dimensional Euclidean space $\mathcal{E}^n$. Consider a frame for $\mathcal{E}^n$ with elements $\{\gamma_k\}$; the frame need not be normal nor orthogonal. Then, project x onto the frame – write x as linear combination of the frame elements:

$$x = \gamma_k x^k \,.$$

Find $\mathrm{d}x$ and then $\mathrm{d}^2x = \mathrm{d}x\mathrm{d}x = \mathrm{d}x \cdot \mathrm{d}x$

$$\mathrm{d}x = \gamma_k \mathrm{d}x^k \implies \mathrm{d}^2x = g_{kl}x^k x^l \,,$$

since $\gamma_k \cdot \gamma_l = g_{kl}$.

Calculate the particle velocity $\dot{x}$ and square it; *i.e.*, $\dot{x}\dot{x}$:

$$\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t} \implies \dot{x}\dot{x} = g_{kl}\dot{x}^k\dot{x}^l$$

and calculate then the particle kinetic energy k

$$k = \frac{1}{2}m\dot{x}\dot{x} = \frac{1}{2}mg_{kl}\dot{x}^k\dot{x}^l \,.$$

Notice, finally, that during the calculation the metric coefficients have not changed when going from d^2x to k! That's why the trick of using the line element to find g_{kl} and using them directly to calculate the kinetic energy works.

12.14. **Lagrangian in Various Coordinate Systems.** While geometric objects have the same form independently on the coordinate system in use, their *representations* onto coordinate systems (*i.e.*, components or coordinates) vary from one system to another. However, sometimes working with one particular coordinate system makes calculations and understanding and calculations easier than with others. Therefore, knowing how to move from one system to another is imperative. This is the topic of this section.

12.14.1. *Lagrangian in Cartesian Coordinate Systems.* In Cartesian coordinates, find the position of a particle by means of the set $\{x, y, z\}$, where all of the set elements measure lengths from the particle to the coordinate axes. Use as a basis for the coordinate system the [7] set $\{\hat{x}, \hat{y}, \hat{z}\}$, where the set elements are normal and orthonormal to each other; *i.e.*, the chosen basis is orthonormal. The position vector is thus $x = x\,\hat{x} + y\,\hat{y} + z\,\hat{z}$; whereas the differential separation is $\mathrm{d}s = \mathrm{d}x\,\hat{x} + \mathrm{d}y\,\hat{y} + \mathrm{d}z\,\hat{z}$. Next, the differential distance is given by

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \,.$$

Next, the metric elements are given in matrix representation by [8] $[g] = g_{ij} = \mathrm{diag}\,[1, 1, 1]$, where i represent the matrix rows and j columns.

Note. By definition of the metric in index notation, read the metric coefficients directly from the differential distance. On the contrary, if the metric coefficients are given, then the differential distance can be directly formed from the metric coefficients.

Finally, the Lagrangian is given by

$$L = \tfrac{1}{2}m\,g_{ij}\dot{x}^i\,\dot{x}^j - v[x, y, z] = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - v[x, y, z] \,.$$

Note. To find the form of the particle kinetic energy, take the differential distance terms and replace the $\{\mathrm{d}x^2\}$'s by $\{\dot{x}^2\}$'s, leaving the coefficients of the metric untouched.

---

[7] The set $\{\hat{x}, \hat{y}, \hat{z}\}$ is traditionally written as $\left\{\hat{\imath}, \hat{\jmath}, \hat{k}\right\}$.

[8] Since Cartesian coordinates directly measure lengths, then expect no change in the metric coefficients.

12.14.2. *Lagrangian in Cylindrical Coordinate Systems.* In cylindrical coordinates, find the position of a particle by means of the set $\{\rho, \phi, z\}$. Transform between Cartesians and cylindricals by

$$x = \rho\cos\phi\,, \qquad y = \rho\sin\phi \qquad \text{and} \qquad z = z\,.$$

Here, ρ and z measure lengths while ϕ angles, so expect the metric coefficients to be different from unity.

Use as a basis for the system the set $\left\{\hat{\rho}, \hat{\phi}, \hat{z}\right\}$. The differential separation is then $\mathrm{d}s = \mathrm{d}\rho\,\hat{\rho} + \rho\,\mathrm{d}\phi\,\hat{\phi} + \mathrm{d}z\,\hat{z}$. The differential distance thus is

$$\mathrm{d}s^2 = \mathrm{d}\rho^2 + \rho^2\,\mathrm{d}\phi^2 + \mathrm{d}z^2\,.$$

Next, the metric elements are given by $[g] = g_{ij} = \mathrm{diag}\left[1, \rho^2, 1\right]$.

Finally, the Lagrangian is

$$L = \tfrac{1}{2}m(\dot{\rho}^2 + \rho^2\,\dot{\phi}^2 + \dot{z}^2) - v[x, y, z]\,.$$

12.14.3. *Lagrangian in Spherical Coordinate Systems.* In spherical coordinates, find the position of a particle by means of the set $\{r, \theta, \phi\}$. Transform between Cartesians and sphericals by

$$x = r\sin\theta\cos\phi\,, \qquad y = r\sin\theta\sin\phi \qquad \text{and} \qquad z = r\cos\theta\,.$$

Here, r measures lengths while θ and ϕ angles, so expect the metric coefficients to be different from unity.

Use as a basis for the system the set $\left\{\hat{r}, \hat{\theta}, \hat{\phi}\right\}$. The differential separation is then $\mathrm{d}s = \mathrm{d}r\,\hat{r} + r\mathrm{d}\theta\,\hat{\theta} + r\sin\theta\,\mathrm{d}\phi\,\hat{\phi}$. The differential distance thus is

$$\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\,.$$

Next, the metric elements are given by $[g] = g_{ij} = \mathrm{diag}\left[1, r^2, r^2\sin^2\theta\right]$.

Finally, the Lagrangian is

$$L = \tfrac{1}{2}m(\dot{r}^2 + r^2\,\dot{\theta}^2 + r^2\sin^2\theta\,\dot{\phi}^2) - v[x, y, z]\,.$$

12.14.4. *Other Coordinate Systems.* The Lagrangian and the Euler-Lagrange equations are independent on the underlying coordinate system used to express them. This is because they were derived by geometric means and by calculus of variations.

In general curvilinear coordinates $\left\{q^i\right\}$, the metric elements are given by

$$g_{ij} = \frac{\partial x}{\partial q^i}\cdot\frac{\partial x}{\partial q^i}\,,$$

so the differential distance becomes

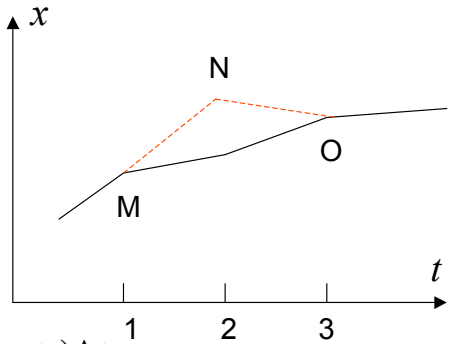
$$\mathrm{d}s^2 = g_{ij}\mathrm{d}q^i\mathrm{d}q^j = \frac{\partial x}{\partial q^i}\cdot\frac{\partial x}{\partial q^i}\,\mathrm{d}q^i\,\mathrm{d}q^j\,,$$

where x is the particle's position (vector).

12.14.5. *Technical Note.* When the coordinate system used is the *Cartesian system*, then the system is called *canonical* and all of the quantities are referred to as such: canonical coordinates, canonical positions, canonical momentum, canonical velocity and so on. On the other hand, when the coordinate system used is not the Cartesian, the the system is called *generalized* and all of the quantities are referred to as such: generalized coordinates, generalized positions, generalized momentum and so forth.

Finally, since the Lagrangian and the Euler-Lagrange equations are independent on the coordinate system, sometimes it is more efficient to choose coordinates that are not Cartesian neither curvilinear, but coordinates based on the geometric properties (quantities) of the system; *i.e.*, the angle that a hanging pendulum forms with the vertical suffices to describe its motion, rather than two Cartesian or polar coordinates. In such cases, the quantities themselves are called *generalized coordinates*.

FIGURE 3. Linearized path of a real path. Notice that the position of the point between $\mathcal{M}$ and $\mathcal{O}$ was varied to occupy the position at $\mathcal{N}$.



12.15. **Geometric Derivation of the Euler-Lagrange Equations.** [least action, lagrangian and hamiltonian mechanics]

For many simple systems, the kinetic energy approximates the potential energy $k \sim v$ when averaged over a path. This leads to the idea that the value of the action A, defined by

$$A \coloneqq \int_{t_a}^{t_b} (k - v)\, \mathrm{d}t\,,$$

evaluated along a path may take a minimum or stationary value. Refer to the integrand of the action as the Lagrangian; *i.e.*, $L = k - v$. The Lagrangian is commonly a function of position x, velocity $\dot{x}$ and time t.

In a more concrete fashion, imagine a particle moving in space. While moving, it traces a smooth curve called the particle trajectory. Such a curve can be approximated by a collection of line segments. This approximated path, as the particle trajectory, will have an action associated with it that can be calculated as the sum of the actions of the line segments:

$$A = \sum_i L\Big[x^i, \dot{x}^i, t\Big]\, \Delta t\,,$$

where the position x^i is taken at the starting point of each segment and the velocity of the particle along the segment is $\dot{x}^i \sim \bar{x}^i = \Delta x/\Delta t$ (average $\dot{x}^i$ for each segment), see section 12.15.

The key observation, due to Euler, is that every section of the action integral, or sum in this case, must be stationary; *i.e.*, it must take either a maximum or a minimum value. To see this, refer to section 12.15. To find this stationary value, refer again to the figure. It shows a linearized path with the points $\mathcal{M}$ and $\mathcal{O}$ on it. The point between them, $\mathcal{N}$ at time t_2, was shifted from the linearized path (approximation of the real path taken by a particle) to a higher position. Note in the figure that the positions of $\mathcal{M}$, $x^{\mathcal{M}}$, and of $\mathcal{O}$, $x^{\mathcal{O}}$, are fixed.

Next, calculate the value of the action between the three points, $A_{\mathcal{MNO}}$, as the sum of the actions of the component line segments, remembering that each x^i is taken at the beginning of each segment

$$A_{\mathcal{MNO}} = L\Big[x^{\mathcal{M}}, \bar{x}^{\mathcal{M}}, t_1\Big]\, \Delta t + L\Big[x^{\mathcal{N}}, \bar{x}^{\mathcal{N}}, t_2\Big]\, \Delta t\,.$$

Without loss of generality, take equal time steps: $\Delta t = t_2 - t_1 = t_3 - t_2$. Then, $\bar{x}^{\mathcal{M}} = (x^{\mathcal{N}} - x^{\mathcal{M}})/\Delta t$ and $\bar{x}^{\mathcal{N}} = (x^{\mathcal{O}} - x^{\mathcal{N}})/\Delta t$.

Require then the action to be stationary noting that A is a function of $x^{\mathcal{N}}$ only, since the positions $x^{\mathcal{M}}$ and $x^{\mathcal{O}}$ are fixed, and using the chain rule for derivatives:

$$\frac{\mathrm{d}A_{\mathcal{MNO}}}{\mathrm{d}x^{\mathcal{N}}} = \left(\frac{\partial L}{\partial \bar{x}^{\mathcal{M}}}\frac{\mathrm{d}\bar{x}^{\mathcal{M}}}{\mathrm{d}x^{\mathcal{N}}} + \frac{\partial L}{\partial x^{\mathcal{N}}} + \frac{\partial L}{\partial \bar{x}^{\mathcal{N}}}\frac{\mathrm{d}\bar{x}^{\mathcal{N}}}{\mathrm{d}x^{\mathcal{N}}}\right)\Delta t = 0\,.$$

Note that $\mathrm{d}\bar{x}^{\mathcal{M}}/\mathrm{d}x^{\mathcal{N}} = 1/\Delta t$ and $\mathrm{d}\bar{x}^{\mathcal{N}}/\mathrm{d}x^{\mathcal{N}} = -1/\Delta t$ to have

$$\frac{\mathrm{d}A_{\mathcal{MNO}}}{\mathrm{d}x^{\mathcal{N}}} = \left(\frac{\partial L}{\partial \bar{x}^{\mathcal{M}}}\frac{1}{\Delta t} + \frac{\partial L}{\partial x^{\mathcal{N}}} - \frac{\partial L}{\partial \bar{x}^{\mathcal{N}}}\frac{1}{\Delta t}\right)\Delta t = \frac{\partial L}{\partial x^{\mathcal{N}}} - \frac{1}{\Delta t}\left(\frac{\partial L}{\partial \bar{x}^{\mathcal{N}}} - \frac{\partial L}{\partial \bar{x}^{\mathcal{M}}}\right) = 0\,.$$

This last condition must hold for any point on the path, so taking $\Delta t \to 0$, one has

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0\,. \tag{12.2}$$

This is to say that the condition that the action be stationary for every point of a particle trajectory implies that the particle's position x and velocity $\dot{x}$ must satisfy eq. (12.2). This equation is called the Euler-Lagrange equation of motion and is the backbone of the Lagrangian description of motion, as opposed to the Newtonian description of motion.

12.16. **Classical test particle with Newtonian gravity.** Suppose we are given a particle with mass m and position $x[t]$ in a Newtonian gravitation field with potential ζ, where $\dim\zeta = [FL/M] = [E/M]$. The particle's world line is parameterized by time t. The particle's kinetic energy is [9]

$$k[t] = \frac{1}{2}m\dot{x}^2[t]$$

and the particle's gravitational potential energy is

$$v[t] = m\zeta[x[t]\,,t]\ .$$

Then the particle's Lagrangian is

$$L[t] = k[t] - v[t] = \frac{1}{2}m\dot{x}^2[t] - m\zeta[x[t]\,,t]\ .$$

Varying x in the integral (equivalent to the Euler-Lagrange differential equation), we get

$$0 = \delta\int L[t]\ \mathrm{d}t = \int \delta L[t]\ \mathrm{d}t = \int \left(m\dot{x}[t]\cdot\delta\dot{x}[t] - m\zeta[x[t]\,,t]\cdot\delta x[t]\right)\mathrm{d}t\,.$$

Integrate the first term by parts and discard the total integral. Then divide out the variation to get

$$0 = -m\ddot{x}[t] - m\nabla\zeta[x[t]\,,t]$$

and thus

$$m\ddot{x}[t] = -m\nabla\zeta[x[t]\,,t]$$

is the equation of motion – two different expressions for the force.

12.17. **Lagrangian in Vector Notation.** Using Lagrange's mechanics and vector calculus, find the equations of motion for a free particle of mass m.

Solution. The particle kinetic energy is $2k[t, x[t]] = m\dot{x}^2[t]$. Then, the particle Lagrangian becomes

$$L[t, x[t]] = \frac{1}{2}m\dot{x}^2[t]\ .$$

Using the particle Langrangian, find the generalized force, generalized momentum and the generalized momentum time derivative

$$\frac{\partial L}{\partial x} = 0\,,\qquad \frac{\partial L}{\partial \dot{x}} = m\dot{x}[t]\qquad\text{and}\qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = m\ddot{x}[t]\ .$$

Euler-Lagrange's equations give finally the particle equations of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = m\ddot{x}[t] = 0 \implies \ddot{x}[t] = 0\,. \qquad \square$$

Consider the last exercise particle to be object of a potential $v[t, x[t]]$. Find the particle equations of motion.

Solution. The particle Lagrangian becomes

$$L[t, x[t]] = \frac{1}{2}m\dot{x}^2[t] - v[t, x[t]]\ .$$

Then, the generalized force, momentum and its time derivative are

$$\frac{\partial L}{\partial x} = -\frac{\partial v[t, x[t]]}{\partial x} = -\nabla v[t, x[t]]\ ,\qquad \frac{\partial L}{\partial \dot{x}} = m\dot{x}[t]\qquad\text{and}\qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} = m\ddot{x}[t]\ .$$

[9] In geometric algebra, if x is a vector, then $x^2 = xx = x\cdot x$, since x is parallel to itself, and thus $x\wedge x = 0$.

Euler-Lagrange's equations give finally the particle equations of motion:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = m\ddot{x}[t] + \nabla v[t, x[t]] = 0 \implies \ddot{x}[t] = -\nabla v[t, x[t]] \ . \qquad \square$$

13. Hamiltonian Mechanics

13.1. **Covariance and contravariance of vectors.** ***[Covariance and contravariance of vectors, wiki]

The terms *covariance* and *contravariance* describe how the quantitative description of certain geometric or physical entities changes with a change of basis. For *holonomic bases*, this is determined by a change from one coordinate system to another. When an *orthogonal basis* is rotated into another orthogonal basis, the distinction between co- and contra-variance is invisible. However, when considering more general coordinate systems such as skew coordinates, curvilinear coordinates and coordinate systems on differentiable manifolds, the distinction is significant.

- For a *vector* (such as a *position* or *velocity*) to be basis-independent, the components of the vector must *contra-vary* with a change of basis to compensate. That is, the components must vary with the inverse transformation to that of the change of basis. The components of vectors (as opposed to those of dual vectors) are said to be *contravariant*. Examples of vectors with contravariant components include the position of an object relative to an observer or any derivative of position with respect to time, including velocity, acceleration, and jerk. In Einstein notation, contravariant components are denoted with upper indices as in
$$v = v^k \gamma_k \, .$$
- For a *dual vector* (also called a *covector*) to be basis-independent, the components of the dual vector must *co-vary* with a change of basis to remain representing the same covector. That is, the components must vary by the same transformation as the change of basis. The components of dual vectors (as opposed to those of vectors) are said to be *covariant*. Examples of covariant vectors generally appear when taking a gradient of a function. In Einstein notation, covariant components are denoted with lower indices as in
$$v = v_k \gamma^k \, .$$

In physics, *vectors often have units of distance or distance times some other unit* (such as the velocity), whereas *covectors have units the inverse of distance or the inverse of distance times some other unit*. The distinction between covariant and contravariant vectors is particularly important for computations with tensors, which can have mixed variance. This means that they have components that are both covariant and contravariant. The valence or type of a tensor gives the number of covariant and contravariant component indices.

Note. Determine the position of the index by noting where the dimensions of length are – up or down. For instance, velocity is a contravariant vector. To determine the position of the index, notice the dimensions of velocity: $\dim v = [L/T]$. Since $v \sim [L]$, then the index goes "up": $v = v^k \gamma_k$. Gradient, on the other hand, is a covariant vector. To determine the position of the index, notice the dimensions of gradient, say on the x-direction: $\dim \nabla = \partial/\partial x = [1/L]$. Since $\nabla \sim [1/L]$, the index goes "down": $\nabla = \nabla_k \gamma^k$.

Example. The components of the *generalized* momentum are covariant.

Solution. By definition of generalized momentum:
$$p_i = \frac{\partial L}{\partial \dot{x}^i} \implies \dim p_i = \dim \frac{\partial L}{\partial \dot{x}^i} = \frac{[E]}{[L/T]} = \frac{[E.T]}{[L]} \, ,$$
where $[[E], [L], [T]]$ denote dimensions of energy, length and time.

Since the components of the generalized momentum are proportional to the dimensions of *inverse* length, $p_i \sim 1/[L]$, then the generalized momentum is a covariant vector, or covector.

Example. The components of the *canonical* momentum are contra-variant or covariant.

Solution. Since *canonical* means using Cartesian coordinates, then there is *no* distinction between contra- and co-variant components.

Contravariant: by the definition of momentum in terms of Cartesian coordinates (momentum is mass times velocity)
$$p^i = m v^i \implies \dim p^i = \dim mv = [M.L/T] \, .$$

Since the components of the canonical momentum are proportional to the dimensions of length, $p^i \sim [L]$, then the canonical momentum is a contravariant vector.

Covariant: by definition of canonical momentum in terms of the Lagrangian:

$$p_i = \frac{\partial L}{\partial \dot{x}^i} \implies \dim p_i = \dim \frac{\partial L}{\partial \dot{x}^i} = \frac{[E]}{[L/T]} = \frac{[E.T]}{[L]} .$$

Since the components of the canonical momentum are proportional to the dimensions of *inverse* length, $p_i \sim 1/[L]$, then the canonical momentum is a covariant vector, or covector.

Note. The ambiguity that coordinate representation brings is one of the reasons why working directly with vectors is preferable than working with their component representation. That is, by definition, momentum (vector) equals mass (scalar) times velocity (vector): $p = mv$. That's all! No indices, no contra-, no co-, no nothing, just vectors.

13.2. **Hamiltonian in Classical Mechanics.** ***[To have Hamiltonian mechanics, the trick is to apply the Legendre Transform to the Lagrangian!]

In classical mechanics, we often start with a Lagrangian, defined as a function of $x[t]$ and $\dot{x}[t]$, say. Then we have, in a sense, two variables in $L[x, \dot{x}]$ and we can replace $\dot{x}$ with an *independent* variable by setting $p = L_{,\dot{x}}$ and by performing a *Legendre transform* to L to eliminate $\dot{x}$ in favor of p. Thus, define the Hamiltonian H by

$$\begin{cases} H[x,p] = p\,\dot{x}[p] - L[x,p] \ , \\ p = \dfrac{\partial L}{\partial \dot{x}} \ , \\ L[x,p] = L[x,\dot{x}[p]] \ . \end{cases} \tag{13.1}$$

Notice that we have performed the transformation on only *one* of the two variables in the Hamiltonian. In words, we construct the Hamiltonian in eq. (13.1) by using the definition of p to find $\dot{x}[p]$ and then writing $p\dot{x} - L[x, \dot{x}]$ entirely in terms of p and x.

Before performing the Legendre transform to any Lagrangian, check section 10.3.

Also note that we worked using the geometric principle: no coordinates were required to define the Hamiltonian! Only geometric objects were used: vectors, Lagrangian and so forth.

Example. Consider a simple harmonic oscillator potential with the Lagrangian given by

$$L[x, \dot{x}] = \tfrac{1}{2} m\dot{x}^2 - \tfrac{1}{2} kx^2 .$$

Find the Hamiltonian for the oscillator.

Solution. Legendre transform the given Lagrangian to replace $\dot{x}$ in favor of p by following the procedure established in section 10.3.

(1) Check the existence/uniqueness conditions:

$$\begin{aligned} &L[x, \dot{x}] \ , && \text{[well behaved function]} \\ &\frac{\partial L}{\partial \dot{x}} = m\dot{x} \, , && \text{[well behaved function]} \\ &\frac{\partial^2 L}{\partial \dot{x}^2} = m \, . && \text{[strictly positive, since } m > 0\text{]} \end{aligned}$$

All of the conditions are met. Proceed.

(2) Define $p[\dot{x}] \coloneqq m\dot{x}$. Invert this to find $\dot{x}[p] = p/m$.

(3) Define

$$H \coloneqq p[\dot{x}]\ \dot{x} - L[x, \dot{x}] = \tfrac{1}{2} m\,\dot{x}\,\dot{x} - \tfrac{1}{2} m\dot{x}^2 + \tfrac{1}{2} kx^2 = \tfrac{1}{2} m\dot{x}^2 + \tfrac{1}{2} kx^2 .$$

(4) Use $\dot{x}[x] = p/m$ to write

$$H = \tfrac{1}{2} m \frac{p^2}{m} + \tfrac{1}{2} kx^2 .$$

(5) Finally, present the Hamiltonian as

$$2H[x, \dot{x}] = p^2/m + kx^2 .$$

We recognize this Hamiltonian as the total energy of the system (numerically).

Consider Cartesian coordinates. Then, the usual Lagrangian in three dimensions takes the form, in geometric (vector) notation

$$L = \tfrac{1}{2} mv^2 - v[x] = \tfrac{1}{2} m\,\dot{x} \cdot \dot{x} - v[x]$$

Now, define the *canonical momentum vector* via [10]

$$\begin{cases} p_x = \dfrac{\partial L}{\partial \dot{x}} \,, \\ p_y = \dfrac{\partial L}{\partial \dot{y}} \,, \\ p_z = \dfrac{\partial L}{\partial \dot{z}} \,. \end{cases}$$

Find, next, the Hamiltonian – once again in geometric notation:

$$H[x,p] = \frac{p^2}{2m} - v[x] \,,$$

where $p^2 = |p|^2$.

This is the starting point for Hamiltonian considerations in classical mechanics. We will begin by looking at some changes that must occur to bring this natural form into usable, tensorial notation.

13.3. **Equipartition theorem.** In classical statistical mechanics,

> the *equipartition theorem* is a general formula that relates the temperature of a system with its average energies.

The equipartition theorem is also known as the law of equipartition, equipartition of energy, or simply equipartition. The original idea of equipartition was that, in thermal equilibrium, energy is shared equally among all of its various forms; for example, the average kinetic energy per degree of freedom in the translational motion of a molecule should equal that of its rotational motions.

The equipartition theorem makes quantitative predictions. Like the virial theorem, it gives the total average kinetic and potential energies for a system at a given temperature, from which the system's heat capacity can be computed. However, equipartition also gives the average values of individual components of the energy, such as the kinetic energy of a particular particle or the potential energy of a single spring. For example, it predicts that every atom in a monoatomic ideal gas has an average kinetic energy of $(3/2)k_{\mathrm{b}}T$ in thermal equilibrium, where k_{b} is the Boltzmann constant and T is the (thermodynamic) temperature. More generally, it can be applied to any classical system in thermal equilibrium, no matter how complicated. The equipartition theorem can be used to derive the ideal gas law and the Dulong-Petit law for the specific heat capacities of solids. It can also be used to predict the properties of stars, even white dwarfs and neutron stars, since it holds even when relativistic effects are considered.

Although the equipartition theorem makes very accurate predictions in certain conditions, it becomes inaccurate when quantum effects are significant, such as at low temperatures. When the thermal energy $k_{\mathrm{b}}T$ is smaller than the quantum energy spacing in a particular degree of freedom, the average energy and heat capacity of this degree of freedom are less than the values predicted by equipartition. Such a degree of freedom is said to be "frozen out" when the thermal energy is much smaller than this spacing. For example, the heat capacity of a solid decreases at low temperatures as various types of motion become frozen out, rather than remaining constant as predicted by equipartition. Such decreases in heat capacity were among the first signs to physicists of the 19th century that classical physics was incorrect and that a new, more subtle, scientific model was required. Along with other evidence, equipartition's failure to model black-body radiation – also known as the ultraviolet catastrophe – led Max Planck to suggest that energy in the oscillators in an object, which emit light, were quantized, a revolutionary hypothesis that spurred the development of quantum mechanics and quantum field theory.

13.3.1. *Basic concept and simple examples.* The name "equipartition" means "equal division". The original concept of equipartition was that the total kinetic energy of a system is shared equally among all of its independent parts, on the average, once the system has reached thermal equilibrium. Equipartition also makes quantitative predictions for these energies. For example, it predicts that every atom of a noble gas, in thermal equilibrium

[10] Since we are using *canonical* (Cartesian) coordinates, then the position of the indices as contra- or co-variant does *not* matter. However, we use the covector version because p is being defined by means of the Lagrangian.

at temperature T, has an average translational kinetic energy of $(3/2)k_{\mathrm{b}}T$, where k_{b} is the Boltzmann constant. As a consequence, since kinetic energy is equal to $1/2(\mathrm{mass})(\mathrm{velocity})^2$, the heavier atoms of xenon have a lower average speed than do the lighter atoms of helium at the same temperature. Figure shows the Maxwell-Boltzmann distribution for the speeds of the atoms in four noble gases.

In this example, the key point is that the kinetic energy is quadratic in the velocity. The equipartition theorem shows that in thermal equilibrium, any degree of freedom (such as a component of the position or velocity of a particle) which appears only quadratically in the energy has an average energy of $(1/2)k_{\mathrm{b}}T$ and therefore contributes $(3/2)k_{\mathrm{b}}$ to the system's heat capacity. This has many applications.

13.3.2. *Translational energy and ideal gases.* The (Newtonian) kinetic energy of a particle of mass m, velocity v is given by

$$H_{\mathrm{kin}} = \frac{1}{2}mv^2 = \frac{1}{2}m\left((v^x)^2 + (v^y)^2 + (v^z)^2\right) ,$$

where v^x, v^y and v^z are the Cartesian components of the velocity v. Here, H is short for Hamiltonian, and used henceforth as a symbol for energy because the Hamiltonian formalism plays a central role in the most general form of the equipartition theorem.

Since the kinetic energy is quadratic in the components of the velocity, by equipartition these three components each contribute $(1/2)k_{\mathrm{b}}T$ to the average kinetic energy in thermal equilibrium. Thus the average kinetic energy of the particle is $(3/2)k_{\mathrm{b}}T$, as in the example of noble gases above.

More generally, in an ideal gas, the total energy consists purely of (translational) kinetic energy: by assumption, the particles have no internal degrees of freedom and move independently of one another. Equipartition therefore predicts that the average total energy of an ideal gas of N particles is $(3/2)Nk_{\mathrm{b}}T$.

It follows that the heat capacity of the gas is $(3/2)Nk_{\mathrm{b}}$ and hence, in particular, the heat capacity of a mole of such gas particles is $(3/2)N_{\mathrm{A}}k_{\mathrm{b}} = (3/2)R$, where N_{A} is the Avogadro constant and R is the gas constant. Since $R \sim 2\,\mathrm{cal}/(\mathrm{mol\,K})$, equipartition predicts that the molar heat capacity of an ideal gas is roughly $3\,\mathrm{cal}/(\mathrm{mol\,K})$. This prediction is confirmed by experiment.

The mean kinetic energy also allows the root mean square speed v_{rms} of the gas particles to be calculated:

$$v_{\mathrm{rms}} = \sqrt{\langle v^2\rangle} = \sqrt{\frac{3k_{\mathrm{b}}T}{m}} = \sqrt{\frac{3RT}{M}} ,$$

where $M = N_{\mathrm{A}}m$ is the mass of a mole of gas particles. This result is useful for many applications such as Graham's law of effusion, which provides a method for enriching uranium.

13.3.3. *Potential energy and harmonic oscillators.* Equipartition applies to potential energies as well as kinetic energies: important examples include harmonic oscillators such as a spring, which has a quadratic potential energy

$$H_{\mathrm{pot}} = \frac{1}{2}aq^2 ,$$

where the constant a describes the stiffness of the spring and q is the deviation from equilibrium. If such a one dimensional system has mass m, then its kinetic energy H_{kin} is

$$H_{\mathrm{kin}} = \frac{1}{2}mv^2 = \frac{p^2}{2m} ,$$

where v and $p = mv$ denote the velocity and momentum of the oscillator. Combining these terms yields the total energy

$$H = H_{\mathrm{kin}} + H_{\mathrm{pot}} = \frac{p^2}{2m} + \frac{1}{2}aq^2 .$$

Equipartition therefore implies that in thermal equilibrium, the oscillator has average energy

$$\langle H\rangle = \langle H_{\mathrm{kin}}\rangle + \langle H_{\mathrm{pot}}\rangle = \frac{1}{2}k_{\mathrm{b}}T + \frac{1}{2}k_{\mathrm{b}}T ,$$

where the angular brackets $\langle\ldots\rangle$ denote the average of the enclosed quantity.

This result is valid for any type of harmonic oscillator, such as a pendulum, a vibrating molecule or a passive electronic oscillator. Systems of such oscillators arise in many situations; by equipartition, each such oscillator receives an average total energy $k_{\mathrm{b}}T$ and hence contributes k_{b} to the system's heat capacity. This can be used to derive the formula for Johnson-Nyquist noise and the Dulong-Petit law of solid heat capacities. The latter application was particularly significant in the history of equipartition.

13.3.4. *Specific heat capacity of solids.* An important application of the equipartition theorem is to the specific heat capacity of a crystalline solid. Each atom in such a solid can oscillate in three independent directions, so the solid can be viewed as a system of $3N$ independent simple harmonic oscillators, where N denotes the number of atoms in the lattice. Since each harmonic oscillator has average energy $k_{\mathrm{b}}T$, the average total energy of the solid is $3Nk_{\mathrm{b}}T$ and its heat capacity is $3Nk_{\mathrm{b}}$.

By taking N to be the Avogadro constant N_{A} and using the relation $R = N_{\mathrm{A}}k_{\mathrm{b}}$ between the gas constant R and the Boltzmann constant k_{b}, this provides an explanation for the Dulong-Petit law of specific heat capacities of solids, which stated that the specific heat capacity (per unit mass) of a solid element is inversely proportional to its atomic weight. A modern version is that the molar heat capacity of a solid is $3R \sim 6\,\mathrm{cal}/(\mathrm{mol\,K})$.

However, this law is inaccurate at lower temperatures, due to quantum effects; it is also inconsistent with the experimentally derived third law of thermodynamics, according to which the molar heat capacity of any substance must go to zero as the temperature goes to absolute zero. A more accurate theory, incorporating quantum effects, was developed by Albert Einstein (1907) and Peter Debye (1911).

Many other physical systems can be modeled as sets of coupled oscillators. The motions of such oscillators can be decomposed into normal modes, like the vibrational modes of a piano string or the resonances of an organ pipe. On the other hand, equipartition often breaks down for such systems, because there is no exchange of energy between the normal modes. In an extreme situation, the modes are independent and so their energies are independently conserved. This shows that some sort of mixing of energies, formally called *ergodicity*, is important for the law of equipartition to hold.

13.3.5. *General formulation of the equipartition theorem.* The most general form of the equipartition theorem states that under suitable assumptions (discussed below), for a physical system with Hamiltonian energy function H and degrees of freedom x_n, the following equipartition formula holds in thermal equilibrium for all indices m and n:

$$\left\langle x_m \frac{\partial H}{\partial x_m} \right\rangle = [m = n]_{\mathrm{iv}}\, k_{\mathrm{b}}T = \delta_{mn} k_{\mathrm{b}}T$$

Here δ_{mn} is the Kronecker delta, which is equal to one if $m = n$ and is zero otherwise. The averaging brackets $\langle \ldots \rangle$ is assumed to be an ensemble average over phase space or, under an assumption of ergodicity, a time average of a single system.

The general equipartition theorem holds in both the microcanonical ensemble, when the total energy of the system is constant, and also in the canonical ensemble, when the system is coupled to a heat bath with which it can exchange energy. Derivations of the general formula are given later in the article. The general formula is equivalent to the following two:

$$\begin{cases} \left\langle x_n \dfrac{\partial H}{\partial x_n} \right\rangle = k_{\mathrm{b}}T & \text{for all } n \\ \left\langle x_m \dfrac{\partial H}{\partial x_n} \right\rangle = 0 & \text{for all } m \neq n \end{cases}$$

If a degree of freedom x_n appears only as a quadratic term $a_n x_n^2$ in the Hamiltonian H, then the first of these formulae implies that

$$k_{\mathrm{b}}T = \left\langle x_n \frac{\partial H}{\partial x_n} \right\rangle = 2\left\langle a_n x_n^2 \right\rangle ,$$

which is twice the contribution that this degree of freedom makes to the average energy $\langle H \rangle$. Thus the equipartition theorem for systems with quadratic energies follows easily from the general formula. A similar argument, with 2 replaced by s, applies to energies of the form $a_n x_n^s$.

The degrees of freedom x_n are coordinates on the phase space of the system and are therefore commonly subdivided into generalized position coordinates q^k and generalized momentum coordinates p_k, where p_k is the conjugate momentum to q^k. In this situation, the first equation means that for all k,

$$\left\langle p_k \frac{\partial H}{\partial p_k} \right\rangle = \left\langle q^k \frac{\partial H}{\partial q^k} \right\rangle = k_{\mathrm{b}} T \,.$$

Using the equations of Hamiltonian mechanics, these formulae may also be written

$$\left\langle p_k \dot{q}^k \right\rangle = - \left\langle q^k \dot{p_k} \right\rangle = k_{\mathrm{b}} T \,.$$

Similarly, one can show using the second equation that

$$\left\langle q^j \frac{\partial H}{\partial p_k} \right\rangle = \left\langle p_j \frac{\partial H}{\partial q^k} \right\rangle = 0 \qquad \text{for all } j, k$$

and

$$\left\langle q^j \frac{\partial H}{\partial q^k} \right\rangle = \left\langle p_j \frac{\partial H}{\partial p_k} \right\rangle = 0 \qquad \text{for all } j \neq k$$

13.4. **Equipartition Theorem, Again.** Degrees of freedom are associated with the kinetic energy of translations, rotation, vibration and the potential energy of vibrations. A result from classical statistical mechanics is the equipartition theorem:

> when a substance is in equilibrium, there is an average energy of $k_{\mathrm{b}}T/2$ per molecule or $RT/2$ per mole associated with each degree of freedom.

Or, equivalently,

> In equilibrium, each degree of freedom contributes $(1/2)k_{\mathrm{b}}T$ to the average energy per molecule.
>
> At temperature T, the average energy of any quadratic degree of freedom is $k_{\mathrm{b}}T$.

13.5. **Microcanonical Ensemble.**

13.6. **Canonical Ensemble.**

13.7. **Ideal Gas Law.**

13.7.1. *Ideal Gas.* An *ideal gas* is a theoretical gas composed of a set of randomly moving, non-interacting point particles. The ideal gas concept is useful because it obeys the ideal gas law, a simplified equation of state, and is amenable to analysis under statistical mechanics.

At normal conditions such as standard temperature and pressure, most real gases behave qualitatively like an ideal gas. Many gases such as nitrogen, oxygen, hydrogen, noble gases, and some heavier gases like carbon dioxide can be treated like ideal gases within reasonable tolerances. Generally, a gas behaves more like an ideal gas at higher temperature and lower density (*i.e.*, lower pressure), as the work which is against intermolecular forces becomes less significant compared with the particles' kinetic energy, and the size of the molecules becomes less significant compared to the empty space between them.

The ideal gas model tends to fail at lower temperatures or higher pressures, when intermolecular forces and molecular size become important. It also fails for most heavy gases, such as many refrigerants and for gases with strong intermolecular forces, notably water vapor. At some point of low temperature and high pressure, real gases undergo a phase transition, such as to a liquid or a solid. The model of an ideal gas, however, does not describe or allow phase transitions. These must be modeled by more complex equations of state.

The ideal gas model has been explored in both the Newtonian dynamics (as in "kinetic theory") and in quantum mechanics (as a "gas in a box"). The ideal gas model has also been used to model the behavior of electrons in a metal (in the Drude model and the free electron model), and it is one of the most important models in statistical mechanics.

13.7.2. *Derivation.* Consider statistical mechanics. Let $q = [q^x, q^y, q^z]$ and $p = [p_x, p_y, p_z]$ denote the position vector and momentum vector of a particle of an ideal gas. Let f denote the net force on that particle. Then the time average momentum of the particle [11] is

$$\langle q \cdot f \rangle = \langle q \cdot \dot{p} \rangle = \langle q^x \dot{p_x} \rangle + \langle q^y \dot{p_y} \rangle + \langle q^z \dot{p_z} \rangle = -\left\langle q^x \frac{\partial H}{\partial q^x} \right\rangle - \left\langle q^y \frac{\partial H}{\partial q^y} \right\rangle - \left\langle q^z \frac{\partial H}{\partial q^z} \right\rangle = 3k_\mathrm{b}T\,,$$

where the first equality is Newton's second law, the third one uses Hamilton's equations, the fourth one uses the equipartition theorem and in the last equality T represents the gas thermodynamic temperature. Summing over a system of N particles yields

$$3Nk_\mathrm{b}T = -\left\langle \sum_{k=1}^{N} q_k \cdot f_k \right\rangle .$$

By Newton's third law and the ideal gas assumption, the net force on the system is the force applied by the walls of their container, and this force is given by the pressure P of the gas. Hence,

$$-\left\langle \sum_{k=1}^{N} q_k \cdot f_k \right\rangle = P \oint_{\text{surface}} q \cdot \mathrm{d}S\,,$$

where $\mathrm{d}S$ is the infinitesimal area element along the walls of the container. Since the divergence of the position vector q is

$$\operatorname{div} q = \nabla \cdot q = \frac{\partial q^x}{\partial q^x} + \frac{\partial q^y}{\partial q^y} + \frac{\partial q^z}{\partial q^z} = 3\,,$$

the divergence theorem implies that

$$P \oint_{\text{surface}} q \cdot \mathrm{d}S = P \int_{\text{volume}} (\nabla \cdot q)\,\mathrm{d}V = 3PV\,,$$

where $\mathrm{d}V$ is an infinitesimal volume within the container and V is the total volume of the container.

Putting these equalities together yields

$$3Nk_\mathrm{b}T = -\left\langle \sum_{k=1}^{N} q_k \cdot f_k \right\rangle = 3PV\,,$$

which immediately implies the ideal gas law for N particles:

$$PV = Nk_\mathrm{b}T = nRT\,,$$

where $n = N/N_\mathrm{A}$ is chemical amount of the gas, N_A Avogadro's constant and $R = N_\mathrm{A}k_\mathrm{b}$ the gas constant.

13.8. **Time Average of a Quantity.** The time average of a quantity $Q[t]$ is defined by

$$\langle Q[t] \rangle = \frac{1}{\tau} \int_{t=0}^{\tau} Q[t]\ \mathrm{d}t\,.$$

13.9. **The Virial Theorem.** [The Virial Theorem, Christopher Palmer]

13.9.1. *Background.* The Virial Theorem has a long history. It can be viewed as an application of a celestial mechanics theorem due to Lagrange, but is usually attributed to Clausius in his "On a Mechanical Theorem Applicable to Heat" (1870):

> The Mean *vis viva* of a system is equal to its *virial*.

The vis viva (living force) is the *double* of what we would now call kinetic energy: $\sum_i m_i v_i^2$.

The virial (or Virial of Clausius) is given by

$$v = \sum_i f_i \cdot x_i\,.$$

Since it involves many interacting particles, the theorem has applications in kinetic theory, celestial mechanics and, as we shall see, atomic physics.

[11] By definition, an ideal gas has only kinetic energy and not potential – since it doesn't interact and since doesn't vibrate or rotate. Therefore, the time average momentum is given by $\langle q \cdot f \rangle$, analogously to the virial: $q \cdot p$.

13.9.2. *Derivation.* We consider a system of particles interacting by forces. Define $G = \sum_i x_i \cdot p_i$. Then,

$$\frac{\mathrm{d}G}{\mathrm{d}t} = \sum_i \frac{\mathrm{d}x_i}{\mathrm{d}t} \cdot p_i + \sum_i x_i \cdot \frac{\mathrm{d}p_i}{\mathrm{d}t} \,,$$

where x_i is the position vector of the ith particle and p_i its linear momentum (vector).

The first term in the last equation is just the vis viva, being $\sum_i m_i v_i^2$, and the second term, from Newton's second law, is v.

We now perform a time-average over a duration τ (integrate over t and divide by τ):

$$\frac{1}{\tau}\left(G[\tau] - G[0]\right) = 2\left\langle k\right\rangle + \left\langle v\right\rangle \,.$$

If the system is periodic we can choose τ equal to the period, so that $G[\tau] = G[0]$. Alternatively, if the system is bounded (neither x_i nor p_i become infinite) then we can make the left side as small as we choose by taking τ large enough.

In either case we can take the left side to be zero, giving the Virial Theorem (with a minus sign for this definition of v):

$$2\left\langle k\right\rangle = -v \,.$$

The periodic case applies to very simple systems like a two-body orbiting system.

The bounded assumption obviously requires that the system does *not* move off as a whole, so that as τ increases the x_i increase without limit. There are three obvious ways of preventing this:

(1) Put the whole system in an effectively infinitely massive box (kinetic theory of gases case).
(2) Introduce a fixed centre of force in addition to the inter-particle forces (solar system, atom with fixed nucleus, galaxy with huge central black hole?). In these cases the 'fixed' centre is assumed to be infinitely massive (or at least hugely more massive that the rest of the system).
(3) Use the theorem in the center-of-mass frame, so that the center of mass of the system is at rest. The bounded assumption then only requires that the system does not fall apart.

The two last cases refer precisely to the atom with fixed nucleus H_0, and the atom with finite mass nucleus $H_0 + H_2$.

13.10. **The Virial Theorem, again.** Consider a collection of particles with masses m_i, $i = 1, 2, \ldots, N$. Let the complete system be in a 'steady state' [12], where the individual particles move around but the overall description of the system does not change qualitatively; *i.e.*, its macroscopic parameters remain within certain bounds. Then we can obtain a relation between the kinetic and potential energies of the system.

The equations of motion for the ith particle are $\dot{p_i} = F_i$. Define, then, the virial $G = \sum_i p_i \cdot x_i$, where x_i is the position vector of the ith particle. Thus, the change of G with respect to time is

$$\dot{G} = \sum_i \dot{p}_i \cdot x_i + \sum_i p_i \cdot \dot{x}_i = \sum_i F_i \cdot x_i + 2k \,,$$

where k is the kinetic energy of the system.

Compute the time average of each quantity in the last equation to find

$$\frac{1}{\tau}\int_0^\tau \dot{G}\,\mathrm{d}t = \left\langle 2k\right\rangle + \left\langle \sum_i F_i \cdot x_i \right\rangle \,.$$

In a steady state, the difference $G[\tau] - G[0]$ will remain finite, so if we take the large τ limit, then we get

$$\frac{1}{\tau}\int_0^\tau \dot{G}\,\mathrm{d}t = \frac{1}{\tau}\left(G[\tau] - G[0]\right) \to 0 \,.$$

[12] This is a call for an application to thermodynamics! Or equilibrium statistical mechanics.

So we find that in steady state

$$2k = -\left\langle \sum_i F_i \cdot x_i \right\rangle ,$$

where the time averages are now assumed to be taken with the limit $\tau \to \infty$.

The RHS of the above equation does not make much physical sense as it stands, but it has to be evaluated for a specific force law.

13.11. **Applications of the Virial Theorem.** [The Virial Theorem and its applications in the teaching of Modern Physics, Celso L. Ladera, Eduardo Alomá y Pilar León]

13.11.1. *Temperature of the interior of a star.* Finding the temperature at the surface of the Sun is a standard example presented in all Modern Physics courses as an application of Planck Quantum Theory of Radiation. Less known is the calculation of the temperature of the interior of a star, a case that is best and most effectively treated using the Virial Theorem. Assuming that a star is a sphere of radius R, and mass M_s, its total gravitational potential energy V is found using a well-known relation of general physics courses:

$$v = -\frac{3}{5}\frac{GM_\mathrm{s}^2}{R} .$$

With the safe assumption that a single atom moving in the interior of the star has a mean kinetic energy $\langle k \rangle$ given by energy equipartition by

$$\langle k \rangle = \frac{3}{2} k_\mathrm{b} \langle T_\mathrm{s} \rangle ,$$

where $\langle T_\mathrm{s} \rangle$ is the mean temperature over the interior of the star, and k_b is Boltzmann Constant.

If N is the total number of atoms in the star then the application of the Virial Theorem gives

$$-\frac{1}{2}\langle v \rangle \sim -\frac{3}{10}\frac{GM_\mathrm{s}^2}{R} = \frac{3}{2} N k_\mathrm{b} \langle T_\mathrm{s} \rangle ,$$

therefore, we have

$$\langle T_\mathrm{s} \rangle \sim \frac{1}{5}\frac{GM_\mathrm{s}^2}{k_\mathrm{b} N R} = \frac{1}{5}\frac{GM_\mathrm{s} m}{k_\mathrm{b} R} ,$$

where $m = M_\mathrm{s}/N$ is the average mass of an atom of the star. Typical stars such as our Sun contain mostly hydrogen atoms ($\sim 61\%$) and helium atoms (38%), and we may therefore approximate the atom mass $m = 2.2 \times 10^{-27}\,\mathrm{kg}$. The mass of the Sun is about $M_\mathrm{s} = 2 \times 10^{30}\,\mathrm{kg}$ and its radius may be taken as $R = 70 \times 10^6\,\mathrm{km}$. Introducing these constants in the last equation, we get an estimate of our Sun interior temperature: $10^7\,\mathrm{K}$, which coincides with estimates using other physics phenomena that take place in the star (*e.g.*, nucleo synthesis). As Kittel et al. comment, this is a remarkable result given the simple calculation required, and the small amount of experimental data demanded, all of which is readily available from measurements in our own planet: not necessary to get close to the Sun!

13.11.2. *Kinetic theory of gases.* If we consider a gas confined into a recipient of volume V, the interactions between molecules of the gas will be bound by the walls of the recipient. Let us then evaluate the terms in the r.h.s. of the virial theorem.

Taking a force differential on the gas molecules, defined by the pressure P exerted by the wall of the recipient in a differential area $\mathrm{d}A$ we may write $\mathrm{d}F = P\,\mathrm{d}A\,\hat{n}$, so the total force will be $F = \int P\,\mathrm{d}A\,\hat{n}$. Then, the term $F \cdot x$ in the virial theorem is, together with the total force,

$$F \cdot x = P \int x \cdot \mathrm{d}A\,\hat{n} .$$

By applying to this the well-known Gauss theorem of vector calculus, thus we get

$$\int x \cdot \mathrm{d}A\,\hat{n} = \int (\nabla \cdot x)\,\mathrm{d}V = 3V .$$

The remaining term of of the virial theorem, that is mv^2, is twice the value of kinetic energy. Again from the theorem of energy equipartition the average kinetic energy of an

ideal gas is given by $\langle k \rangle = (3/2) k_\mathrm{b} \langle T \rangle$. If we now take $\left\langle \dot{G} \right\rangle \to 0$, we get $\langle F \cdot x \rangle + \left\langle m v^2 \right\rangle = 0$. Replacing the values of the precedent equations into the virial and eliminating the common factor $3/2$, we arrive to the well-known equation of the ideal gases:

$$\langle P \rangle V = N k_\mathrm{b} \langle T \rangle \ ,$$

where N is the number of molecules and $\langle P \rangle$ and $\langle T \rangle$ the average macroscopic pressure and average macroscopic thermodynamic temperature.

14. Classical Dynamics

[Classical Dynamics, Dr David Tong, 2004-2005]

14.1. **Newtonian Mechanics: A Single Particle.** Basic concepts: a particle is defined to be an object of insignificant size; *e.g.*, an electron, a tennis ball or a planet. Obviously the validity of this statement depends on the context: to first approximation, the earth can be treated as a particle when computing its orbit around the sun. But if you want to understand its spin, it must be treated as an extended object.

The motion of a particle of mass m at the position x is governed by Newton's Second Law $f = ma$ or, more precisely,

$$f[x, \dot{x}] = \dot{p}\,,$$

where f is the force which, in general, can depend on both the position x as well as the velocity $\dot{x}$, (for example, friction forces depend on $\dot{x}$) and $p = m\dot{x}$ is the momentum. Both f and p are 3-vectors. The last equation reduces to $f = ma$ if $\dot{m} = 0$. But if $m = m[t]$ (*e.g.*, in rocket science), then the form with $\dot{p}$ is correct.

General theorems governing differential equations guarantee that if we are given x and $\dot{x}$ at an initial time $t = t_0$, we can integrate the last equation to determine $x[t]$ for all t (as long as f remains finite). This is the *goal of classical dynamics*.

The last is not quite correct as stated: we must add the caveat that it holds *only* in an *inertial frame*. This is defined to be a frame in which a free particle with $\dot{m} = 0$ travels in a straight line,

$$x = x_0 + \dot{x}t\,,$$

Newton's first law is the statement that such frames exist.

Angular momentum: we define the angular momentum l of a particle and the torque τ acting upon it as

$$l = x \times p \qquad \text{and} \qquad \tau = x \times f\,.$$

Note that, unlike linear momentum p, both l and τ depend on where we take the origin: *we measure angular momentum with respect to a particular point*. Let us cross both sides of the last equation with x. Using the fact that $\dot{x}$ is parallel to p, we can write

$$\frac{\mathrm{d}(x \times p)}{\mathrm{d}t} = x \times \dot{p}\,.$$

Then we get a version of Newton's second law that holds for angular momentum:

$$\tau = \dot{l}\,.$$

Conservation Laws: From the last equations, two important conservation laws follow immediately.

- If $f = 0$, then p is constant throughout the motion;
- If $\tau = 0$, then l is constant throughout the motion.

Notice that $\tau = 0$ does not require $f = 0$, but only $x \times f = 0$. This means that f *must* be parallel to x. This is the definition of a *central force*. An example is given by the gravitational force between the earth and the sun: the earth's angular momentum about the sun is constant. As written above in terms of forces and torques, these conservation laws appear trivial.

Energy: Let's now recall the definitions of energy. We firstly define the kinetic energy k as

$$k = \frac{1}{2}m\dot{x} \cdot \dot{x}\,.$$

Suppose from now on that the mass is constant. We can compute the change of kinetic energy with time: $\dot{k} = \dot{p} \cdot \dot{x} = f \cdot \dot{x}$. If the particle travels from position x_1 at time t_1 to position x_2 at time t_2 then this change in kinetic energy is given by

$$k[t_2] - k[t_1] = \int_{t_1}^{t_2} \dot{k}\,\mathrm{d}t = \int_{t_1}^{t_2} f \cdot \dot{x}\,\mathrm{d}t = \int_{x_1}^{x_2} f \cdot \mathrm{d}x\,,$$

where the final expression involving the integral of the force over the path is called the *work done by the force*. So we see that

> the work done is equal to the change in kinetic energy.

From now on we will mostly focus on a very special type of force known as a *conservative force*. Such a force depends only on position x rather than velocity $\dot{x}$ and is such that the work done is *independent of the path taken*. In particular, for a closed path, the work done vanishes:

$$\oint f \cdot \mathrm{d}x \iff \nabla \times f = 0\,.$$

It is a deep property of flat space $\mathcal{E}^3$ that this property implies we may write the force as

$$f = -\nabla v[x]$$

for some potential $v[x]$. Systems which admit a potential of this form include gravitational, electrostatic and interatomic forces. When we have a conservative force, we necessarily have a conservation law for energy. To see this, return to the change of kinetic energy equation which now reads

$$k[t_2] - k[t_1] = \int_{x_1}^{x_2} f \cdot \mathrm{d}x = -\int_{x_1}^{x_2} \nabla v \cdot \mathrm{d}x = -v[t_2] + v[t_1]\ ,$$

or, rearranging things,

$$k[t_1] + v[t_1] = k[t_2] + v[t_2] = e\,.$$

So $e = k + v$ is also a *constant of motion*. It is the energy. When the energy is considered to be a function of position x and momentum p it is referred to as the *Hamiltonian* H.

Example 1: The Simple Harmonic Oscillator. This is a one-dimensional system with a force proportional to the distance x to the origin: $f[x] = -kx$. This force arises from a potential $2v = kx^2$. Since $f \neq 0$, momentum is *not* conserved (the object oscillates backwards and forwards) and, since the system lives in only one dimension, angular momentum is *not defined*. But energy

$$e = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2$$

is conserved.

Example 2: The Damped Simple Harmonic Oscillator. We now include a friction term so that $f[x, \dot{x}] = -kx - \gamma\dot{x}$. Since f is not conservative, energy is *not* conserved. This system loses energy until it comes to rest. (The last statement is not entirely true. Mechanical energy is not conserved, but total energy is, because mechanical energy transforms into heating and then the system comes to rest.)

Example 3: Particle Moving Under Gravity. Consider a particle of mass m moving in three dimensions under the gravitational pull of a much larger particle of mass M. The force is $f = -(GmM/x^2)\hat{x}$ which arises from the potential $v = -GmM/x$. Again, the linear momentum p of the smaller particle is *not* conserved, but the force is both central and conservative, ensuring the particle's total energy e and the angular momentum l are conserved.

14.2. **The Principle of Least Action - The Lagrangian Formalism.** Firstly, let's get our notation right. Part of the power of the Lagrangian formulation over the Newtonian approach is that

> the Lagrangian formulation does away with vectors in favor of more general coordinates.

We start by doing this trivially. Let's rewrite the positions of N particles with coordinates x_i as q^i where $i = 1, \ldots, 3N$. Then Newton's equations read

$$\dot{p_i} = -\frac{\partial v}{\partial q^i}\,, \tag{14.1}$$

where $p_i = m_i\dot{q}^i$. The number of degrees of freedom of the system is said to be 3N. These parameterize a $3N$-dimensional space known as the *configuration space* C. Each point in C specifies a configuration of the system (*i.e.*, the positions of all N particles). Time evolution gives rise to a curve in C.

Define the *Lagrangian* to be a function of the positions q^i and the velocities $\dot{q}^i$ of all the particles, given by [13]

$$L\Big[q^i, \dot{q}^i\Big] = k\Big[\dot{q}^i\Big] - v\Big[q^i\Big] \ ,$$

where $2k = \sum_i m_i(\dot{q}^i)^2$ is the kinetic energy and $v\big[q^i\big]$ is the potential energy. Note the minus sign between k and v! To describe the principle of least action, we consider all smooth paths $q^i[t]$ in C with fixed end points so that

$$q^i[t_\mathrm{i}] = q^i_\mathrm{initial} \qquad \text{and} \qquad q^i[t_\mathrm{f}] = q^i_\mathrm{final} \, .$$

Of all these possible paths, only one is the true path taken by the system. Which one? To each path, let us assign a number called the action A defined as

$$A\Big[q^i[t]\Big] = \int_{t_\mathrm{initial}}^{t_\mathrm{final}} L\Big[q^i, \dot{q}^i\Big] \, \mathrm{d}t \, .$$

The action is a *functional* (*i.e.*, a function of the path which is itself a function). The principle of least action is the following result:

Theorem (Principle of Least Action): The actual path taken by the system is an *extremum* of A.

[proof: omitted]

This requirement holds if and only if

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0 \qquad \text{for each } i = 1, \ldots, 3N \, . \tag{14.2}$$

These are known as *Lagrange's equations* (or sometimes as the *Euler-Lagrange equations*).

Lagrange's equations are equivalent to Newton's. From the definition of the Lagrangian, we have $\partial L/\partial q^i = -\partial v/q^i$, while $\partial L/\partial \dot{q}^i = p_i$. It's then easy to see that eq. (14.2) are indeed equivalent to eq. (14.1).

Note. The q^i are called *generalized coordinates*, while the $\dot{q}^i$ are called *generalized velocities*. Both have *contravariant* components:

$$\dim q^i = [L] \implies \text{contravariant components} \, ,$$
$$\dim \dot{q}^i = [L/T] \implies \text{contravariant components} \, .$$

The first terms in eq. (14.2) are called *generalized forces*, f_i,

$$f_i = \frac{\partial L}{\partial q^i} \, ,$$

while the second terms are called *generalized momenta*, p_i,

$$p_i = \frac{\partial L}{\partial \dot{q}^i} \, .$$

Thus, Euler-Lagrange equations, eq. (14.2), can be written in a form analogous to Newton's second law of motion:

$$f_i = \dot{p}_i \, ,$$

this is, generalized forces equal the time change of generalized momenta.

Additionally, notice that both generalized forces and momenta have *covariant components*:

$$\dim f_i = \frac{\partial L}{\partial q^i} = [E/L] \implies \text{covariant components} \, ,$$
$$\dim p_i = \frac{\partial L}{\partial \dot{q}^i} = [E.T/L] \implies \text{covariant components} \, ,$$

where $[E]$ stands for the dimension of energy.

Finally, note that generalized momenta are the *conjugate* of generalized coordinates:

$$p_i = \frac{\partial L}{\partial \dot{q}^i} \, .$$

Some remarks on this important result:

- This is an example of a variational principle.

[13] The Lagrangian is defined as the difference of two energies, therefore it has the dimensions of energy $[E]$; *i.e.*, $\dim L = [E]$.

- The principle of least action is a slight misnomer. The proof only requires that $\delta A = 0$ and does not specify whether it is a maxima or minima of A. Since $L = k - v$, we can always increase A by taking a very fast, wiggly path with $k \gg 0$, so the true path is never a maximum. However, it may be either a minimum or a saddle point. So *Principle of stationary action* would be a more accurate, but less catchy, name. It is sometimes called *Hamilton's principle*.
- All the fundamental laws of physics can be written in terms of an action principle. This includes electromagnetism, general relativity, the standard model of particle physics and attempts to go beyond the known laws of physics such as string theory.
- There is a beautiful generalization of the action principle to quantum mechanics due to Feynman in which the particle takes all paths with some probability determined by A.
- Back to classical mechanics, there are two very important reasons for working with Lagrange's equations rather than with Newton's. The first is that Lagrange's equations hold in any coordinate system, while Newton's are restricted to an inertial frame. The second is the ease with which we can deal with constraints in the Lagrangian system.

14.3. **Changing Coordinate Systems.** Lagrange's equations hold in *any* coordinate system. This follows immediately from the action principle, which is a statement about paths and not about coordinates. So the *form* of Lagrange's equations holds in any coordinate system. This is in contrast to Newton's equations which are only valid in an inertial frame. Let's illustrate the power of this fact with an example.

Example: Rotating Coordinate Systems: Consider a free particle with Lagrangian given by

$$L = \frac{1}{2} m \dot{x}^2 \,,$$

with $x = [x, y, z]$. Now measure the motion of the particle with respect to a coordinate system which is rotating with angular velocity $\omega = [0, 0, \omega]$ about the z axis. If $x' = [x', y', z']$ are the coordinates in the rotating system, we have the relationship

$$\begin{aligned} x' &= x\cos[\omega t] + y\sin[\omega t] \,, \\ y' &= y\cos[\omega t] - x\sin[\omega t] \,, \\ z' &= z \,. \end{aligned}$$

Then we can substitute these expressions into the Lagrangian to find L in terms of the rotating coordinates,

$$L = \frac{1}{2} m \left((\dot{x}' - \omega y')^2 + (\dot{y}' + \omega x')^2 + \dot{z}^2 \right) = \frac{1}{2} m (\dot{x}' + \omega \times x')^2 \,.$$

In this rotating frame, we can use Lagrange's equations to derive the equations of motion. Taking derivatives, we have

$$\begin{aligned} \frac{\partial L}{\partial x'} &= m \left(\dot{x}' \times \omega - \omega \times (\omega \times x') \right) \,, \\ \frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{x}'} &= m \left(\ddot{x}' + \omega \times \dot{x}' \right) \,, \end{aligned}$$

so Lagrange's equation reads

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L}{\partial \dot{x}'} - \frac{\partial L}{\partial x'} = m \left(\ddot{x}' + \omega \times (\omega \times x') + 2\omega \times \dot{x}' \right) = 0 \,.$$

The second and third terms in this expression are the *centrifugal* and *coriolis forces*. These are examples of the *fictitious forces*. They're called fictitious because they're a consequence of the reference frame, rather than any interaction. But don't underestimate their importance just because they're "fictitious"! According to Einstein's theory of general relativity, the force of gravity is on the same footing as these fictitious forces.

14.4. **Constraints and Generalized Coordinates.** Define the operator $\partial_i = \partial/\partial q^i$.

Now we turn to the second advantage of the Lagrangian formulation. In writing $f_i = \dot{p}_i = -\partial_i v$, we implicitly assume that each particle can happily roam anywhere in space $\mathcal{E}^3$. What if there are constraints? In Newtonian mechanics, we introduce *constraint forces*. These are things like the tension of ropes and normal forces applied by surfaces. In the Lagrangian formulation, we don't have to worry about such things.

An Example: The Pendulum. The simple pendulum has a single dynamical degree of freedom θ, the angle the pendulum makes with the vertical. The position of the mass m in the plane is described by two Cartesian coordinates x and y subject to a constraint $x^2 + y^2 = l^2$. We can parameterize this as $x = l\sin[\theta]$ and $y = l\cos[\theta]$. Employing the Newtonian method to solve this system, we introduce the tension T and resolve the force vectors to find

$$m\ddot{x} = -Tx/l \qquad \text{and} \qquad m\ddot{y} = mg - Ty/l\,.$$

To determine the motion of the system, we impose the constraints at the level of the equation of motion, and then easily find

$$\ddot{\theta} = -(g/l)\sin[\theta] \qquad \text{and} \qquad T = ml\dot{\theta}^2 + mg\cos[\theta]\ .$$

While this example was pretty straightforward to solve using Newtonian methods, things get rapidly harder when we consider more complicated constraints (and we'll see plenty presently). Moreover, you may have noticed that half of the work of the calculation went into computing the tension T. On occasion we'll be interested in this. (For example, we might want to know how fast we can spin the pendulum before it breaks). But often we won't care about these constraint forces, but will only want to know the motion of the pendulum itself. In this case it seems like a waste of effort to go through the motions of computing T. We'll now see how we can avoid this extra work in the Lagrangian formulation. Firstly, let's define what we mean by constraints more rigorously.

14.4.1. *Holonomic Constraints. Holonomic Constraints* [14] are relationships between the coordinates of the form

$$f_\alpha\Big[x^i, t\Big] = 0 \qquad \alpha = 1, \ldots, 3N - n\,.$$

In general the constraints can be time dependent and our notation above allows for this. Holonomic constraints can be solved in terms of n *generalised coordinates* $\big\{q^i : i = 1, \ldots, n\big\}$. So

$$x^i = x^i\big[q^1, \ldots, q^n\big]\ .$$

The system is said to have n degrees of freedom. For the pendulum example above, the system has a single degree of freedom, $q^1 = q = \theta$.

One method to use the Lagrangian formulation is to introduce constraints of this form: introduce $3N - n$ new variables λ_α, called *Lagrange multipliers* and define a new Lagrangian. So we can incorporate constraint forces into the Lagrangian setup using Lagrange multipliers. But the big news is that we don't have to! Often we don't care about constraint forces, but only want to know what the generalized coordinates q^i are doing. In this case we have the following useful theorem:

Theorem: For constrained systems, we may derive the equations of motion directly in generalized coordinates q^i

$$L\Big[q^i, \dot{q}^i, t\Big] = L\Big[x^i\Big[q^i, t\Big], \dot{x}^i\Big[q^i, t\Big]\Big]\ .$$

If we are only interested in the dynamics of the generalized coordinates q^i, we may ignore the Lagrange multipliers and work entirely with the unconstrained Lagrangian $L\big[q^i, \dot{q}^i, t\big]$ defined in the last equation where we just substitute in $x^i = x^i\big[q^i, t\big]$.

Let's see how this works in the simple example of the pendulum. We can parameterize the constraints in terms of the generalized coordinate θ so that $x = l\sin[\theta]$ and $y = l\cos[\theta]$. We now substitute this directly into the Lagrangian for a particle moving in the plane under the effect of gravity, to get

$$L = \frac{1}{2}m\left(\dot{x}^2 + \dot{y}^2\right) + mgy = \frac{1}{2}ml^2\dot{\theta}^2 + mgl\cos[\theta]\ .$$

[14] Introduced by H. Hertz in 1894, the term *holonomic* comes from the Greek and means whole law.

From which we may derive Lagrange's equations using the coordinate θ directly

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} = ml^2\ddot{\theta} + mgl\sin[\theta] = 0\,,$$

which indeed reproduces the equation of motion for the pendulum; *viz.*, $\ddot{\theta} = -(g/l)\sin[\theta]$. Note that, as promised, we haven't calculated the tension T using this method. This has the advantage that we've needed to do less work. If we need to figure out the tension, we have to go back to the more laborious Lagrange multiplier or Newton methods.

14.4.2. *Non-Holonomic Constraints.* For completeness, let's quickly review a couple of non-holonomic constraints. There's no general theory to solve systems of this type, although it turns out that both of the examples we describe here can be solved with relative ease using different methods.

Inequalities: Consider a particle moving under gravity on the outside of a sphere of radius r. It is constrained to satisfy $z^2 + y^2 + z^2 \geq r^2$. This type of constraint, involving an inequality, is non-holonomic. When the particle lies close to the top of the sphere, we know that it will remain in contact with the surface and we can treat the constraint effectively as holonomic. But at some point the particle will fall off. To determine when this happens requires different methods from those above (although it is not particularly difficult).

Velocity Dependent Constraints: Constraints of the form $g\left[x^i, \dot{x^i}, t\right] = 0$ which cannot be integrated to give $f\left[x^2, t\right] = 0$ are non-holonomic. For example, consider a coin of radius r rolling down a slope. The coordinates $[x, y]$ fix the coin's position on the slope. But the coin has other degrees of freedom as well: the angle θ it makes with the path of steepest descent, and the angle ϕ that a marked point on the rim of the coin makes with the vertical. If the coin rolls without slipping, then there are constraints on the evolution of these coordinates. We must have that the velocity of the rim is $v_{\mathrm{rim}} = r\dot{\phi}$. So, in terms of our four coordinates, we have the constraint

$$x = r\dot{\phi}\sin[\theta] \qquad \text{and} \qquad y = r\dot{\phi}\cos[\theta]\ .$$

But these cannot be integrated to give constraints of the form $f[x, y, \theta, \phi] = 0$. They are non-holonomic.

14.5. **Summary.** Let's review what we've learnt so far. A system is described by n generalized coordinates q^i which define a point in an n-dimensional configuration space C. Time evolution is a curve in C governed by the Lagrangian

$$L\left[q^i, \dot{q}^i, t\right]$$

such that the q^i obey

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0\,.$$

These are n coupled 2nd order (usually) non-linear differential equations.

Before we move on, let's take this opportunity to give an important definition. The quantity

$$p_i = \frac{\partial L}{\partial \dot{q}^i}$$

is called the *generalised momentum conjugate to* q^i. (It only coincides with the real momentum in Cartesian coordinates). We can now rewrite Lagrange's equations as

$$\dot{p}_i = \frac{\partial L}{\partial q^i}\,.$$

14.6. **Noether's Theorem and Symmetries.** In this subsection we shall discuss the appearance of conservation laws in the Lagrangian formulation and, in particular, a beautiful and important theorem due to Noether relating conserved quantities to symmetries.

Let's start with a definition. A function $f\left[q^i, \dot{q}^i, t\right]$ of the coordinates, their time derivatives and (possibly) time t is called a *constant of motion* (or a *conserved quantity*) if

the total time derivative vanishes (an application of the chain rule!)

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial q^j}\dot{q}^j + \frac{\partial f}{\partial \dot{q}^j}\ddot{q}^j \right) + \frac{\partial f}{\partial t}\,,$$

whenever $q^i[t]$ satisfy Lagrange's equations. This means that f remains constant along the path followed by the system. Here's a couple of examples:

Claim: If L does not depend explicitly on time t (*i.e.*, $\partial L/\partial t = 0$), then

$$H = \sum_j \dot{q}^j \frac{\partial L}{\partial \dot{q}^j} - L$$

is constant. When H is written as a function of q^i and p_i, it is known as the *Hamiltonian*. It is usually identified with the total energy of the system.

Claim: Suppose $\partial L/\partial q^j = 0$ for some q^j. Then, q^j is said to be *ignorable* (or *cyclic*). We have the conserved quantity

$$p_j = \frac{\partial L}{\partial \dot{q}^j}\,.$$

14.6.1. *Noether's Theorem.* Consider a one-parameter family of maps

$$q^i[t] \to Q^i[s,t] \qquad \text{with} \qquad s \in \mathcal{R}$$

such that $Q^i[0,t] = q^i[t]$. Then, this transformation is said to be a *continuous symmetry of the Lagrangian* L if

$$\frac{\partial}{\partial s} L\Big[Q^i[s,t]\,, \dot{Q}^i[s,t]\,, t\Big] = 0\,.$$

Noether's theorem states that

> for each such symmetry there exists a conserved quantity.

Homogeneity of Space:

> Homogeneity of Space implies Translation Invariance of L implies Conservation of Total Linear Momentum.

This statement should be intuitively clear. One point in space is much the same as any other. So why would a system of particles speed up to get over there, when here is just as good? This manifests itself as conservation of linear momentum.

Isotropy of Space:

> Isotropy of Space implies Rotational Invariance of L implies Conservation of Total Angular Momentum.

Homogeneity of Time: What about homogeneity of time? In mathematical language, this means L is invariant under $t \to t + s$ or, in other words, $\partial L/\partial t = 0$. But we already saw earlier in this section that this implies

$$H = \sum_i \dot{q}^i \left(\frac{\partial L}{\partial \dot{q}^i} \right) - L$$

is conserved. In the systems we're considering, this is simply the total energy. We see that the existence of a conserved quantity which we call energy can be traced to the homogeneous passage of time. Or

> Time is to Energy as Space is to Momentum.

Recall from your course on special relativity that energy and 3-momentum fit together to form a 4-vector which rotates under spacetime transformations. Here we see that the link between energy-momentum and time-space exists even in the non-relativistic framework of Newtonian physics. You don't have to be Einstein to see it. You just have to be Emmy Noether.

Note. It turns out that *all* conservation laws in nature are related to symmetries through Noether's theorem. This includes the conservation of electric charge and the conservation of particles such as protons and neutrons (known as baryons).

14.7. **Applications.**

14.7.1. *Bead on a Rotating Hoop.* This is an example of a system with a time dependent holonomic constraint. The hoop is of radius a and rotates with frequency ω. The bead, of mass m, is threaded on the hoop and moves without friction. We want to determine its motion. There is a single degree of freedom ϕ, the angle the bead makes with the vertical. In terms of Cartesian coordinates $[x, y, z]$ the position of the bead is

$$x = a\sin[\phi]\cos[\omega t]\ , \quad y = a\sin[\phi]\sin[\omega t] \quad \text{and} \quad z = a - a\cos[\phi]\ .$$

To determine the Lagrangian in terms of the generalized coordinate ϕ we must substitute these expressions into the Lagrangian for the free particle. For the kinetic energy k we have

$$k = \frac{1}{2}m\left(x^2 + y^2 + z^2\right) = \frac{1}{2}ma\left(\dot{\phi}^2 + \omega^2\sin^2[\phi]\right)\ ,$$

while the potential energy v is given by (ignoring an overall constant)

$$v = mgz = -mga\cos[\phi]\ .$$

So, replacing x, y and z by ϕ, we have the Lagrangian

$$L = ma^2\left(\frac{1}{2}\dot{\phi}^2 - v_{\text{eff}}\right)$$

where the effective potential is

$$v_{\text{eff}} = \frac{1}{ma^2}\left(-mga\cos[\phi] - \frac{1}{2}ma^2\omega^2\sin^2[\phi]\right)\ .$$

We can now derive the equations of motion for the bead simply from Lagrange's equations which read

$$\ddot{\phi} = -\frac{\partial v_{\text{eff}}}{\partial\phi}\ .$$

Let's look for stationary solutions of these equations in which the bead doesn't move (*i.e.*, solutions of the form $\ddot{\phi} = \dot{\phi} = 0$). From the equation of motion, we must solve $\partial v_{\text{eff}}/\partial\phi = 0$ to find that the bead can remain stationary at points satisfying

$$g\sin[\phi] = a\omega^2\sin[\phi]\cos[\phi]\ .$$

There are at most three such points: $\phi = 0$, $\phi = \pi$ or $\cos[\phi] = g/a\omega^2$. Note that the first two solutions always exist, while the third stationary point is only there if the hoop is spinning fast enough so that $\omega^2 \geq g/a$. Which of these stationary points is stable depends on whether $v_{\text{eff}}[\phi]$ has a local minimum (stable) or maximum (unstable). This in turn depends on the value of ω.

14.7.2. *Spherical Pendulum.* The spherical pendulum is allowed to rotate in three dimensions. The system has two degrees of freedom which cover the range $0 \leq \theta < \pi$ and $0 \leq \phi < 2\pi$. In terms of cartesian coordinates, we have

$$x = l\cos[\phi]\sin[\theta]\ , \quad y = l\sin[\phi]\sin[\theta] \quad \text{and} \quad z = -l\cos[\theta]\ .$$

We substitute these constraints into the Lagrangian for a free particle to get

$$L = \frac{1}{2}m\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right) = \frac{1}{2}ml^2\left(\dot{\theta}^2 + \dot{\phi}^2\sin^2[\theta]\right) + mgl\cos[\theta]\ .$$

Notice that the coordinate ϕ is ignorable (it does not appear explicitly in the Lagrangian). From Noether's theorem, we know that the quantity

$$J = \frac{\partial L}{\partial\dot{\phi}} = ml^2\dot{\phi}\sin^2[\theta]\ .$$

is constant. This is the component of angular momentum in the ϕ direction. The equation of motion for θ follows from Lagrange's equations and is

$$ml^2\ddot{\theta} = ml^2\dot{\phi}^2\sin[\theta]\cos[\theta] - mgl\sin[\theta]\ .$$

We can substitute $\dot{\phi}$ for the constant J in this expression to get an equation entirely in terms of θ which we chose to write as

$$\ddot{\theta} = -\frac{\partial v_{\text{eff}}}{\partial\theta}\ ,$$

where the effective potential is defined to be

$$v_{\text{eff}} = -\frac{g}{l}\cos[\theta] + \frac{J^2}{2m^2l^4}\frac{1}{\sin^2[\theta]}\,.$$

An important point here: we must substitute for J into the equations of motion. If you substitute J for $\dot{\phi}$ directly into the Lagrangian, you will derive an equation that looks like the one above, but you'll get a minus sign wrong! This is because Lagrange's equations are derived under the assumption that θ and ϕ are independent.

As well as the conservation of angular momentum J, we also have $\partial L/\partial t = 0$ so energy is conserved. This is given by

$$e = \frac{1}{2}\dot{\theta}^2 + v_{\text{eff}}[\theta]\ ,$$

where e is a constant. In fact we can invert this equation for e to solve for θ in terms of an integral

$$t - t_0 = \frac{1}{\sqrt{2}}\int \frac{\mathrm{d}\theta}{\sqrt{e - v[\theta]}}\,.$$

If we succeed in writing the solution to a problem in terms of an integral like this then we say we've "reduced the problem to quadrature". It's kind of a cute way of saying we can't do the integral. But at least we have an expression for the solution that we can play with or, if all else fails, we can simply plot on a computer.

Once we have an expression for $\theta[t]$ we can solve for $\phi[t]$ using the expression for J,

$$\phi = \int \frac{J}{ml^2}\frac{\mathrm{d}t}{\sin^2[\theta]} = \frac{J}{\sqrt{2}ml^2}\int \frac{1}{\sqrt{e - v_{\text{eff}}}}\,\mathrm{d}\theta\,,$$

which gives us $\phi = \phi[\theta] = \phi[t]$.

14.7.3. *Purely Kinetics Lagrangians.* Often in physics, one is interested in systems with only kinetic energy and no potential energy. For a system with n dynamical degrees of freedom q^a, $a = 1, \ldots, n$, the most general form of the Lagrangian with just a kinetic term is

$$L = \frac{1}{2}g_{ab}[q_c]\,\dot{q}^a\dot{q}^b\,. \tag{14.3}$$

The functions $g_{ab} = g_{ba}$ depend on all the generalized coordinates. Assume that $\det g_{ab} \neq 0$ so that the inverse matrix g^{ab} exists ($g^{ab}g_{bc} = \delta^a_c$). It is a short exercise to show that Lagrange's equation for this system are given by

$$\ddot{q}^a + \Gamma^a_{bc}\dot{q}^b\dot{q}^c = 0\,, \tag{14.4}$$

where

$$\Gamma^a_{bc} = \frac{1}{2}g_{ad}\left(\frac{\partial g_{bd}}{\partial q^c} + \frac{\partial g_{cd}}{\partial q^b} - \frac{\partial g_{bc}}{\partial q^d}\right)\,.$$

The functions g_{ab} define a *metric* on the configuration space and the equations eq. (14.4) are known as the *geodesic equations*. They appear naturally in general relativity where they describe a particle moving in curved spacetime. Lagrangians of the form eq. (14.3) also appear in many other areas of physics, including the condensed matter physics, the theory of nuclear forces and string theory. In these contexts, the systems are referred to as *sigma models*.

14.7.4. *Particles in Electromagnetic Fields.* We saw from the beginning that the Lagrangian formulation works with conservative forces which can be written in terms of a potential. It is no good at dealing with friction forces which are often of the type $f = -k\dot{x}$. But there are other velocity dependent forces which arise in the fundamental laws of Nature. It's a crucial fact about Nature that all of these can be written in Lagrangian form. Let's illustrate this in an important example.

Recall that the electric field E and the magnetic field (magnetic induction!) B can be written in terms of a vector potential $A[x,t]$ and a scalar potential $\phi[x,t]$

$$B = \nabla \times A \qquad \text{and} \qquad E = -\nabla\phi - \frac{1}{c}\frac{\partial A}{\partial t}\,,$$

where c is the speed of light. Let's study the Lagrangian for a particle of electric charge e of the form,

$$L = \frac{1}{2}m\dot{x}^2 - e\left(\phi - \frac{1}{c}\dot{x}\cdot A\right) .$$

The momentum conjugate to x is

$$p = \frac{\partial L}{\partial \dot{x}} = m\dot{x} + \frac{e}{c}A .$$

Notice that the momentum is not simply $m\dot{x}$; it's modified in the presence of electric and magnetic fields. Now we can calculate Lagrange's equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = \frac{\mathrm{d}}{\mathrm{d}t}\left(m\dot{x} + \frac{e}{c}A\right) + e\nabla\phi - \frac{e}{c}\nabla\left(\dot{x}\cdot A\right) = 0 .$$

To disentangle this, let's work with indices $a, b = 1, 2, 3$ on the Cartesian coordinates and rewrite the equation of motion as

$$m\ddot{x}^a = -e\left(\frac{\partial\phi}{\partial x^a} + \frac{1}{c}\frac{\partial A_a}{\partial t}\right) + \frac{e}{c}\left(\frac{\partial A_b}{\partial x^a} - \frac{\partial A_a}{\partial x^b}\right)\dot{x}^b .$$

Now we use our definitions of the E and B fields which, in terms of indices, read

$$E_a = -\frac{\partial\phi}{\partial x^a} - \frac{1}{c}\frac{\partial A_a}{\partial t} \qquad B_c = \epsilon_{cab}\frac{\partial A_a}{\partial x^b} ,$$

so the equation of motion can be written as

$$m\ddot{x}^a = eE_a + \frac{e}{c}\,\epsilon_{cab}B_c\dot{x}^b ,$$

or, reverting to vector notation,

$$m\ddot{x} = e\left(E + \frac{1}{c}\,\dot{x}\times B\right) ,$$

which is the *Lorentz force law*.

Gauge Invariance: The scalar and vector potentials are not unique. We may make a change of the form

$$\phi \to \phi - \frac{\partial\Lambda}{\partial t} \qquad \text{and} \qquad A \to A + c\nabla\Lambda .$$

These give the same E and B fields for any function Λ. This is known as a *gauge transformation*. Under this change, we have

$$L \to L + e\frac{\partial\Lambda}{\partial t} + e\dot{x}\cdot\nabla\Lambda = L + e\frac{\mathrm{d}\Lambda}{\mathrm{d}t} ,$$

but we know that the equations of motion remain invariant under the addition of a total derivative to the Lagrangian. This concept of gauge invariance underpins much of modern physics.

14.8. **The Hamiltonian Formalism.** We'll now move onto the next level in the formalism of classical mechanics, due initially to Hamilton around 1830. While we won't use Hamilton's approach to solve any further complicated problems, we will use it to reveal much more of the structure underlying classical dynamics. If you like, it will help us understands what questions we should ask.

14.8.1. *Hamilton's Equations.* Recall that in the Lagrangian formulation, we have the function $L\left[q^i, \dot{q}^i, t\right]$ where q^i $(i = 1, \ldots, n)$ are n generalized coordinates. The equations of motion are

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0 .$$

These are n 2nd order differential equations which require $2n$ initial conditions, say $q^i[t=0]$ and $\dot{q}^i[t=0]$. The basic idea of

> Hamilton's approach is to try and place q^i and $\dot{q}^i$ on a more symmetric footing.

More precisely, we'll work with the n generalized momenta that we introduced earlier,

$$p_i = \frac{\partial L}{\partial \dot{q}^i} \qquad \text{where} \qquad i = 1, \ldots, n\,,$$

so $p_i = p_i\big[q^i, \dot{q}^i, t\big]$. *This coincides with what we usually call momentum* only if *we work in Cartesian coordinates* [so the kinetic term is $1/2\, m_i(\dot{q}^i)^2$]. If we rewrite Lagrange's equations using the definition of the momentum (the last equation [15]), they become

$$\dot{p_i} = \frac{\partial L}{\partial q^i}\,.$$

The plan will be to eliminate $\dot{q}^i$ in favor of the momenta p_i and then to place q^i and p_i on equal footing.

Let's start by thinking pictorially. Recall that $\{q^i\}$ defines a point in n-dimensional configuration space C. Time evolution is a path in C. However, the state of the system is defined by $\{q^i\}$ and $\{p_i\}$ in the sense that this information will allow us to determine the state at all times in the future. The pair $\big[q^i, p_i\big]$ defines a point in $2n$-dimensional *phase space*. Note that since a point in phase space is sufficient to determine the future evolution of the system, paths in phase space can never cross. We say that evolution is governed by a *flow* in phase space.

14.8.2. *The Legendre Transform.* We want to find a function on phase space that will determine the unique evolution of q^i and p_i. This means it should be a function of q^i and p_i (and not of $\dot{q}^i$), but must contain the same information as the Lagrangian $L\big[q^i, \dot{q}^i, t\big]$. There is a mathematical trick to do this, known as the *Legendre transform*.

To describe this, consider an arbitrary function $f[x, y]$ so that the total derivative is

$$\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y\,.$$

Now define a function $g[x, y, u] = ux - f[x, y]$, which depends on three variables, x, y and also u. If we look at the total derivative of g, we have

$$\mathrm{d}g = \mathrm{d}(ux) - \mathrm{d}f = u\mathrm{d}x + x\mathrm{d}u - \frac{\partial f}{\partial x}\mathrm{d}x - \frac{\partial f}{\partial y}\mathrm{d}y\,.$$

At this point u is an independent variable. But suppose we choose it to be a specific function of x and y, defined by

$$u[x, y] = \frac{\partial f}{\partial x}\,.$$

Then the term proportional to $\mathrm{d}x$ in $\mathrm{d}g$ vanishes and we have

$$\mathrm{d}g = x\mathrm{d}u - \frac{\partial f}{\partial y}\mathrm{d}y\,.$$

Or, in other words, g is to be thought of as a function of u and y: $g = g[u, y]$. If we want an explicit expression for $g[u, y]$, we must first invert $u[x, y]$ to get $x = x[u, y]$ and then insert this into the definition of g so that

$$g[u, y] = ux[u, y] - f\big[x[u, y]\,, y\big]\ .$$

This is the Legendre transform. It takes us from one function $f[x, y]$ to a different function $g[u, y]$ where $u = f_{,x}$. The key point is that we haven't lost any information. Indeed, we can always recover $f[x, y]$ from $g[u, y]$ by noting that

$$\frac{\partial g}{\partial u}|_y = -x[u, y] \qquad \text{and} \qquad \frac{\partial g}{\partial y}|_u = -v[u, y]\ ,$$

which assures us that the inverse Legendre transform $f = -g_{,u}u - g$ takes us back to the original function.

The geometrical meaning of the Legendre transform is captured in the diagram [:)]. For fixed y, we draw the two curves $f[x, y]$ and ux. For each slope u, the value of $g[u]$ is the maximal distance between the two curves. To see this, note that extremising this distance means

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(ux - f[x]\right) = 0 \implies u = \frac{\partial f}{\partial x}\,.$$

[15] Note that $\dim p_i = [E.T]\,/\,[L] \sim 1/\,[L]$. For this reason, generalized momentum is a covector.

This picture also tells us that we can *only apply the Legendre transform to convex functions for which this maximum exists.* Now, armed with this tool, let's return to dynamics

14.8.3. *Hamilton's Equations.* The Lagrangian $L\big[q^i, \dot{q}^i, t\big]$ is a function of the coordinates q^i, their time derivatives $\dot{q}^i$ and (possibly) time. We define the Hamiltonian to be the Legendre transform of the Lagrangian with respect to the $\dot{q}^i$ variables

$$H\Big[q^i, p_i, t\Big] = \sum_{i=1}^{n} p_i \dot{q}^i - L\Big[q^i, \dot{q}^i, t\Big] \,,$$

where $\dot{q}^i$ is eliminated from the right hand side in favor of p_i by using

$$p_i = \frac{\partial L}{\partial \dot{q}^i} = p_i\Big[q^j, \dot{q}^j, t\Big]$$

and inverting to get $\dot{q}^i = \dot{q}^i\big[q^j, \dot{q}^j, t\big]$. Now look at the variation of H:

$$\begin{aligned} \mathrm{d}H &= \Big(\mathrm{d}p_i\, \dot{q}^i + p_i\, \mathrm{d}\dot{q}^i\Big) - \left(\frac{\partial L}{\partial q^i}\, \mathrm{d}q^i + \frac{\partial L}{\partial \dot{q}^i}\, \mathrm{d}\dot{q}^i + \frac{\partial L}{\partial t}\, \mathrm{d}t\right) \,, \\ &= \mathrm{d}p_i\, \dot{q}^i - \frac{\partial L}{\partial q^i}\, \mathrm{d}q^i - \frac{\partial L}{\partial t}\, \mathrm{d}t \,. \end{aligned}$$

but we know that this can be rewritten as

$$\mathrm{d}H = \frac{\partial H}{\partial q^i}\, \mathrm{d}q^i + \frac{\partial H}{\partial p_i}\, \mathrm{d}p_i + \frac{\partial H}{\partial t}\, \mathrm{d}t \,.$$

So we can equate terms. So far this is repeating the steps of the Legendre transform. The new ingredient that we now add is Lagrange's equation which reads $\dot{p}_i = \partial L/\partial q^i$. We find

$$\dot{p}_i = -\frac{\partial H}{\partial q^i} \,, \qquad \dot{q}^i = \frac{\partial H}{\partial p_i} \qquad \text{and} \qquad -\frac{\partial L}{\partial t} = \frac{\partial H}{\partial t} \,.$$

These are *Hamilton's equations*. We have replaced n 2nd order differential equations by $2n$ 1st order differential equations for q^i and p_i. In practice, for solving problems, this isn't particularly helpful. But, as we shall see, conceptually it's very useful!

14.8.4. *Examples.* A Particle in a Potential: Let's start with a simple example: a particle moving in a potential in 3-dimensional space. The Lagrangian is simply

$$L = \frac{1}{2} m \dot{q}^2 - v[q] \,,$$

where q is the generalized position vector.

We calculate the momentum by taking the derivative with respect to $\dot{q}$:

$$p = \frac{\partial L}{\partial \dot{q}} = m\dot{q} \,,$$

which, in this case, coincides with what we usually call momentum. The Hamiltonian is then given by

$$H = p \cdot \dot{q} - L = \frac{1}{2m} p^2 + v[q] \,,$$

where, in the end, we've eliminated $\dot{q}$ in favor of p and written the Hamiltonian as a function of p and q. Hamilton's equations are simply

$$\dot{q} = \frac{\partial H}{\partial p} = \frac{1}{m} p \quad \text{and} \quad \dot{p} = -\frac{\partial H}{\partial q} = -\nabla v \,,$$

which are familiar: the first is the definition of momentum in terms of velocity; the second is Newton's equation for this system.

Note. The process for calculating the equations of motion using Hamilton's formalism can be summarized as follows.

Given a Lagrangian in terms of the generalized coordinates q^i and generalized velocities $\dot{q}^i$ and time:

- The momenta are calculated by differentiating the Lagrangian with respect to the (generalized) velocities:
$$p_i\big[q^i, \dot{q}^i, t\big] = \frac{\partial L}{\partial \dot{q}^i} \,.$$
- The velocities $\dot{q}^i$ are expressed in terms of the momenta p_i by inverting the expressions in the previous step.

- The Hamiltonian is calculated using the usual definition of H as the Legendre transformation of L:
$$H = \sum_i \dot{q}^i \frac{\partial L}{\partial \dot{q}^i} - L = \sum_i \dot{q}^i p_i - L\,.$$
Then the velocities are substituted for using the previous results.
- Hamilton's equations are applied, to obtain the equations of motion of the system.
$$\dot{p_i} = -\frac{\partial H}{\partial q^i} \qquad \text{and} \qquad \dot{q}^i = \frac{\partial H}{\partial p_i}\,.$$
(Note the symmetry between p and q; they are the conjugate to the other.)

A Particle in an Electromagnetic Field: We saw earlier that the Lagrangian for a charged particle moving in an electromagnetic field is

$$L = \frac{1}{2}m\dot{q}^2 - e\left(\phi - \frac{1}{c}\dot{q}\cdot A\right)\,.$$

From this we compute the momentum conjugate to the position

$$p = \frac{\partial L}{\partial \dot{q}} = m\dot{q} + \frac{e}{c}A\,,$$

which now differs from what we usually call momentum by the addition of the vector potential A. Inverting, we have

$$\dot{q} = \frac{1}{m}\left(p - \frac{e}{c}A\right)\,.$$

So we calculate the Hamiltonian to be

$$H[p,q] = p\cdot\dot{q} - L = \frac{1}{2m}\left(p - \frac{e}{c}A\right)^2 + e\phi\,.$$

Now Hamilton's equations read

$$\dot{q} = \frac{\partial H}{\partial p} = \frac{1}{m}\left(p - \frac{e}{c}A\right)\,,$$

while the $\dot{p} = -\partial H/\partial x$ equation is best expressed in terms of components

$$\dot{p_a} = -\frac{\partial H}{\partial q^a} = -e\frac{\partial \phi}{\partial q^a} + \frac{e}{2m}\left(p_b - \frac{e}{c}A_b\right)\frac{\partial A_b}{\partial q^a}\,.$$

To show that this is equivalent to the Lorentz force law requires some rearranging of the indices, but it's not too hard.

An Example of the Example: Let's illustrate the dynamics of a particle moving in a magnetic field by looking at a particular case. Imagine a uniform magnetic field pointing in the z-direction: $B = [0,0,B]$. We can get this from a vector potential $B = \nabla\times A$ with

$$A = [-By, 0, 0]\,.$$

This vector potential isn't unique: we could choose others related by a gauge transform as described earlier. But this one will do for our purposes. Consider a particle moving in the $[x,y]$-plane. Then the Hamiltonian for this system is

$$H = \frac{1}{2m}\left(p_x + \frac{eB}{c}y\right)^2 + \frac{1}{2m}p_y^2\,.$$

From which we have four, first order differential equations which are Hamilton's equations

$$\begin{aligned}
\dot{p_x} &= 0\,,\\
\dot{x} &= \frac{1}{m}\left(\dot{p_x} + \frac{eB}{c}y\right)\,,\\
\dot{p_y} &= -\frac{eB}{cm}\left(\dot{p_x} + \frac{eB}{c}y\right) \quad \text{and}\\
\dot{y} &= \frac{p_y}{m}\,.
\end{aligned}$$

If we add these together in the right way, we find that

$$p_y + \frac{eB}{c}x = a = \text{const.}\,,$$
$$p_x = m\dot{x} - \frac{eB}{c}y = b = \text{const.}\,,$$

which is easy to solve: we have

$$x = \frac{ac}{eB} + R\sin[\omega(t-t_0)]\ ,$$
$$y = -\frac{bc}{eB} + R\cos[\omega(t-t_0)]\ ,$$

with a, b, R and t_0 integration constants. So we see that the particle makes circles in the $[x,y]$-plane with frequency

$$\omega = \frac{eB}{cm}\,.$$

This is known as the *Larmor frequency*.

14.8.5. *Some Conservation Laws.* Previously, we saw the importance of conservation laws in solving a given problem. The conservation laws are often simple to see in the Hamiltonian formalism. For example,

Claim: If $H_{,t} = 0$ (*i.e.*, H does not depend on time explicitly), then H itself is a constant of motion.

Claim: If an ignorable coordinate q doesn't appear in the Lagrangian, then, by construction, it also doesn't appear in the Hamiltonian. The conjugate momentum p_q is then conserved.

14.8.6. *The Principle of Least Action.* Recall that earlier we saw the principle of least action from the Lagrangian perspective. This followed from defining the action

$$A = \int_{t_1}^{t_2} L\Big[q^i, \dot{q}^i, t\Big]\ \mathrm{d}t\,.$$

Then we could derive Lagrange's equations by insisting that $\delta A = 0$ for all paths with fixed end points so that $\delta q^i[t_1] = \delta q^i[t_2] = 0$. How does this work in the Hamiltonian formalism? It's quite simple! We define the action

$$A = \int_{t_1}^{t_2} \Big(p_i\dot{q}^i - H\Big)\ \mathrm{d}t\,.$$

where, of course, $\dot{q}^i = \dot{q}^i\big[q^i, p_i\big]$. Now we consider varying q^i and p_i *independently.* Notice that this is different from the Lagrangian set-up, where a variation of q^i automatically leads to a variation of $\dot{q}^i$. But remember that the whole point of the Hamiltonian formalism is that we treat q^i and p_i on equal footing. So we vary both. We have

$$\begin{aligned}\delta A &= \int_{t_1}^{t_2} \left\{\delta p_i\dot{q}^i + p_i\delta\dot{q}^i - \frac{\partial H}{\partial p_i}\delta p_i - \frac{\partial H}{\partial q^i}\delta q^i\right\}\ \mathrm{d}t\,,\\ &= \int_{t_1}^{t_2} \left\{\left[\dot{q}^i - \frac{\partial H}{\partial p_i}\right]\delta p_i + \left[-\dot{p}_i - \frac{\partial H}{\partial q^i}\right]\delta q^i\right\}\ \mathrm{d}t + \Big[p_i\delta q^i\Big]_{t_1}^{t_2}\ .\end{aligned}$$

and there are Hamilton's equations waiting for us in the square brackets. If we look for extrema $\delta A = 0$ for all δp_i and δq^i we get Hamilton's equations

$$\dot{q}^i = \frac{\partial H}{\partial p_i} \qquad \text{and} \qquad q^i = -\frac{\partial H}{\partial q^i}\,.$$

Except there's a very slight subtlety with the boundary conditions. We need the last term in the action, $\big[p_i\delta q^i\big]_{t_1}^{t_2}$, to vanish, and so require only that

$$\delta q^i[t_1] = \delta q^i[t_2] = 0\,,$$

while δp_i can be free at the end points $t = t_1$ and $t = t_2$. So, despite our best efforts, q^i and p_i are not quite symmetric in this formalism.

Note that we could simply impose $\delta p_i[t_1] = \delta p_i[t_2] = 0$ if we really wanted to and the above derivation still holds. It would mean we were being more restrictive on the types of paths we considered. But it does have the advantage that it keeps q^i and p_i on a symmetric

footing. It also means that we have the freedom to add a function to consider actions of the form

$$A = \int_{t_1}^{t_2} \left(p_i \dot{q}^i - H[q,p] + \frac{\mathrm{d}F[q,p]}{\mathrm{d}t} \right) ,$$

so that what sits in the integrand differs from the Lagrangian. For some situations this may be useful.

14.8.7. *Poisson Brackets.* In this section, we'll present a rather formal, algebraic description of classical dynamics which makes it look almost identical to quantum mechanics!

We start with a definition. Let $f[q,p]$ and $f[q,p]$ be two functions on phase space. Then, the *Poisson bracket* is defined to be

$$[f,g]_{\mathrm{pb}} = \frac{\partial f}{\partial q^i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q^i} .$$

Since this is a kind of weird definition, let's look at some of the properties of the Poisson bracket to get a feel for it. We have

- anti-commutativity: $[f,g]_{\mathrm{pb}} = -\,[g,f]_{\mathrm{pb}}$.
- linearity: $[\alpha f + \beta g, h]_{\mathrm{pb}} = \alpha\,[f,h]_{\mathrm{pb}} + \beta\,[g,h]_{\mathrm{pb}}$ for all $\alpha, \beta \in \mathcal{R}$.
- Leibniz rule: $[fg,h]_{\mathrm{pb}} = f\,[g,h]_{\mathrm{pb}} + [f,h]_{\mathrm{pb}}\,g$ which follows from the chain rule in differentiation.
- Jacobi identity:

  $$\Big[f,[g,h]_{\mathrm{pb}}\Big]_{\mathrm{pb}} + \Big[g,[h,f]_{\mathrm{pb}}\Big]_{\mathrm{pb}} + \Big[h,[f,g]_{\mathrm{pb}}\Big]_{\mathrm{pb}} = 0 .$$

  To prove this you need a large piece of paper and a hot cup of coffee. Expand out all 24 terms and watch them cancel one by one.

What we've seen above is that the Poisson bracket $[,]_{\mathrm{pb}}$ satisfies the same algebraic structure as matrix commutators $[,]$ and the differentiation operator d. This is related to Heisenberg's and Schrödinger's viewpoints of quantum mechanics respectively. (You may be confused about what the Jacobi identity means for the derivative operator d. Strictly speaking, the Poisson bracket is like a "Lie derivative" found in differential geometry, for which there is a corresponding Jacobi identity).

The relationship to quantum mechanics is emphasized even more if we calculate

$$\Big[q^i,q^j\Big]_{\mathrm{pb}} = 0 , \qquad [p_i,p_j]_{\mathrm{pb}} = 0 , \qquad \Big[q^i,p_j\Big]_{\mathrm{pb}} = \delta^i_j .$$

Claim: For any function $f[q,p,t]$,

$$\frac{\mathrm{d}f}{\mathrm{d}t} = [f,H]_{\mathrm{pb}} + \frac{\partial f}{\partial t} .$$

Proof.

$$\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}t} &= \frac{\partial f}{\partial p_i}\dot{p}_i + \frac{\partial f}{\partial q^i}\dot{q}^i + \frac{\partial f}{\partial t} , \\
&= -\frac{\partial f}{\partial p_i}\frac{\partial H}{\partial q^i} + \frac{\partial f}{\partial q^i}\frac{\partial H}{\partial p_i} + \frac{\partial f}{\partial t} , \\
&= [f,H]_{\mathrm{pb}} + \frac{\partial f}{\partial t} .
\end{aligned}$$

Isn't this a lovely equation! One consequence is that if we can find a function $i[p,q]$ which satisfy

$$[i,H]_{\mathrm{pb}} = 0 ,$$

then i is a constant of motion. We say that i and H *Poisson commute*. As an example of this, suppose that q^i is ignorable (*i.e.*, it does not appear in H), then

$$[p_i,H]_{\mathrm{pb}} = 0 ,$$

which is the way to see the relationship between ignorable coordinates and conserved quantities in the Poisson bracket language.

Note that if I and J are constants of motion, then

$$\Big[[I,J]_{\mathrm{pb}},H\Big]_{\mathrm{pb}} = \Big[I,[J,H]_{\mathrm{pb}}\Big]_{\mathrm{pb}} + \Big[[I,H]_{\mathrm{pb}},J\Big]_{\mathrm{pb}} = 0 ,$$

which means that $[I, J]_{\text{pb}}$ is also a constant of motion. We say that the constants of motion form a closed algebra under the Poisson bracket.

An Example: Angular Momentum and Runge-Lenz: Consider the angular momentum $l = q \times p$ which, in component form, reads

$$l_1 = q^2 p_3 - q^3 p_2 \,, \qquad l_2 = q^3 p_1 - q^1 p_3 \qquad \text{and} \qquad l_3 = q^1 p_2 - q^2 p_1$$

and let's look at the Poisson bracket structure. We have

$$[l_1, l_2]_{\text{pb}} = \left[q^2 p_3 - q^3 p_2, q^3 p_1 - q^1 p_3\right]_{\text{pb}} = \left[q^2 p_3, q^3 p_1\right]_{\text{pb}} + \left[q^3 p_2, q^1 p_3\right]_{\text{pb}} = q^1 p_2 - q^2 p_1 = l_3 \,.$$

So if l_1 and l_2 are conserved, we see that l_3 must also be conserved. Or, in other words, the whole vector l is conserved if any two components are. Similarly, one can show that

$$\left[l^2, l_i\right]_{\text{pb}} = 0 \,,$$

where $l^2 = \sum_i l_i^2$. This should all be looking familiar from quantum mechanics.

15. Poisson Bracket and Hamiltonian Mechanics

Sometimes laws of physics are just guessed using a bit of intuition and a gut feeling that nature must be beautiful or elegantly simple (though occasionally awesomely complex in beauty).

— JOHN C. BAEZ, Lectures on Classical Mechanics, Lectures, 2005

15.1. **Momentum.** In classical mechanics, *linear momentum* or translational momentum is the product of the mass and velocity of an object. For example, a heavy truck moving fast has a large momentum – it takes a large and prolonged force to get the truck up to this speed, and it takes a large and prolonged force to bring it to a stop afterwards. If the truck were lighter, or moving more slowly, then it would have less momentum.

Like velocity, linear momentum is a vector quantity, possessing a direction as well as a magnitude:

$$p = mv\,.$$

Linear momentum is also a *conserved quantity*, meaning that if a closed system is not affected by external forces, its total linear momentum cannot change. In classical mechanics, conservation of linear momentum is implied by Newton's laws; but it also holds in special relativity (with a modified formula) and, with appropriate definitions, a (generalized) linear momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory and general relativity.

15.1.1. *Newtonian mechanics.* Momentum has a direction as well as magnitude. Quantities that have both a magnitude and a direction are known as vector quantities. Because momentum has a direction, it can be used to predict the resulting direction of objects after they collide, as well as their speeds. Below, the basic properties of momentum are described in one dimension. The vector equations are almost identical to the scalar equations.

Single particle: The momentum of a particle is traditionally represented by the letter p. It is the product of two quantities, the mass (represented by the letter m) and velocity (v):

$$p = mv\,.$$

The units of momentum are the product of the units of mass and velocity. In SI units, if the mass is in kilograms and the velocity in meters per second, then the momentum is in kilograms meters/second (kg m/s). Being a vector, momentum has magnitude and direction. For example, a model airplane of 1 kg, traveling due north at 1 m/s in straight and level flight, has a momentum of 1 kg m/s due north measured from the ground.

Many particles: The momentum of a system of particles is the sum of their momenta. If two particles have masses m_1 and m_2 and velocities v_1 and v_2, the total momentum is

$$p = p_1 + p_2 = m_1 v_1 + m_2 v_2\,.$$

The momenta of more than two particles can be added in the same way.

A system of particles has a *center of mass*, a point determined by the weighted sum of their positions:

$$x_{\mathrm{cm}} = \frac{m_1 x_1 + m_2 x_2 + \cdots}{m_1 + m_2 + \cdots}\,.$$

If all the particles are moving, the center of mass will generally be moving as well. If the center of mass is moving at velocity v_{cm}, the momentum is:

$$p = m v_{\mathrm{cm}}\,.$$

This is known as Euler's first law.

Relation to force: If a force f is applied to a particle for a time interval Δt, the momentum of the particle changes by an amount

$$\Delta p = f \Delta t\,.$$

In differential form, this gives Newton's second law: the rate of change of the momentum of a particle is equal to the force f acting on it:

$$f = \frac{\mathrm{d}p}{\mathrm{d}t}\,.$$

If the force depends on time, the change in momentum (or impulse) between times t_1 and t_2 is

$$\Delta p = \int_{t_1}^{t_2} f[t]\ \mathrm{d}t\,.$$

The second law only applies to a particle that does not exchange matter with its surroundings, and so it is equivalent to write

$$f = m\frac{\mathrm{d}v}{\mathrm{d}t} = ma\,,$$

so the force is equal to mass times acceleration.

Example: a model airplane of 1 kg accelerates from rest to a velocity of 6 m/s due north in 2 s. The thrust required to produce this acceleration is 3 N. The change in momentum is 6 kg m/s. The rate of change of momentum is $3\,(\mathrm{kg\,m/s})/\mathrm{s} = 3\,\mathrm{N}$.

Conservation: In a closed system (one that does not exchange any matter with the outside and is not acted on by outside forces) the total momentum is constant. This fact, known as the law of conservation of momentum, is implied by Newton's laws of motion. Suppose, for example, that two particles interact. Because of the third law, the forces between them are equal and opposite. If the particles are numbered 1 and 2, the second law states that $f_1 = \mathrm{d}p_1/\mathrm{d}t$ and $f_2 = \mathrm{d}p_2/\mathrm{d}t$. Therefore

$$\frac{\mathrm{d}p_1}{\mathrm{d}t} = \frac{\mathrm{d}p_2}{\mathrm{d}t}$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(p_1 + p_2\right)\,.$$

If the velocities of the particles are u_1 and u_2 before the interaction, and afterwards they are v_1 and v_2, then

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2\,.$$

This law holds no matter how complicated the force is between particles. Similarly, if there are several particles, the momentum exchanged between each pair of particles adds up to zero, so the total change in momentum is zero. This conservation law applies to all interactions, including collisions and separations caused by explosive forces.[5] It can also be generalized to situations where Newton's laws do not hold, for example in the theory of relativity and in electrodynamics.

Objects of variable mass: The concept of momentum plays a fundamental role in explaining the behavior of variable-mass objects such as a rocket ejecting fuel or a star accreting gas. In analyzing such an object, one treats the object's mass as a function that varies with time: $m[t]$. The momentum of the object at time t is therefore $p[t] = m[t]\,v[t]$. One might then try to invoke Newton's second law of motion by saying that the external force f on the object is related to its momentum $p[t]$ by $f = \mathrm{d}p/\mathrm{d}t$, but this is *incorrect*, as is the related expression found by applying the product rule to $\mathrm{d}(mv)/\mathrm{d}t$:

$$f = m[t]\,\frac{\mathrm{d}v}{\mathrm{d}t} + v[t]\,\frac{\mathrm{d}m}{\mathrm{d}t}\,. \qquad \text{[wrong!]}$$

This equation does *not* correctly describe the motion of variable-mass objects. The correct equation is

$$f = m[t]\,\frac{\mathrm{d}v}{\mathrm{d}t} - u\frac{\mathrm{d}m}{\mathrm{d}t}\,.$$

where u is the velocity of the ejected/accreted mass *as seen in the object's rest frame*. This is distinct from v, which is the velocity of the object itself as seen in an inertial frame.

This equation is derived by keeping track of both the momentum of the object as well as the momentum of the ejected/accreted mass. When considered together, the object and the mass constitute a closed system in which total momentum is conserved.

15.1.2. *Generalized coordinates.* Newton's laws can be difficult to apply to many kinds of motion because the motion is limited by constraints. For example, a bead on an abacus is constrained to move along its wire and a pendulum bob is constrained to swing at a fixed distance from the pivot. Many such constraints can be incorporated by changing the normal Cartesian coordinates to a set of generalized coordinates that may be fewer in number. Refined mathematical methods have been developed for solving mechanics problems in

generalized coordinates. They introduce a generalized momentum, also known as the canonical or conjugate momentum, that extends the concepts of both linear momentum and angular momentum. To distinguish it from generalized momentum, the product of mass and velocity is also referred to as mechanical, kinetic or kinematic momentum. The two main methods are described below.

Lagrangian mechanics: In Lagrangian mechanics, a Lagrangian is defined as the difference between the kinetic energy k and the potential energy v:

$$L = k - v\,.$$

If the generalized coordinates are represented as a vector $q = \left[q^1, q^2, \ldots, q^n\right]$ and time differentiation is represented by a dot over the variable, then the equations of motion (known as the Lagrange or Euler-Lagrange equations) are a set of N equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^j}\right) - \frac{\partial L}{\partial q^j} = 0\,.$$

If a coordinate q^i is *not* a Cartesian coordinate, the associated generalized momentum component p_i does *not* necessarily have the dimensions of linear momentum. Even if q^i is a Cartesian coordinate, p_i will *not* be the same as the mechanical momentum if the potential depends on velocity. Some sources represent the kinematic momentum by the symbol Π.

In this mathematical framework, a generalized momentum is associated with the generalized coordinates. Its components are defined as

$$p_j = \frac{\partial L}{\partial q^j}\,.$$

Each component p_j is said to be the conjugate momentum for the coordinate q^j.

Now if a given coordinate q^i does *not* appear in the Lagrangian (although its time derivative might appear), then

$$p_j = \text{constant}\,.$$

This is the *generalization of the conservation of momentum*.

Even if the generalized coordinates are just the ordinary spatial coordinates, the conjugate momenta are *not* necessarily the ordinary momentum coordinates.

15.1.3. *Hamiltonian mechanics.* In Hamiltonian mechanics, the Lagrangian (a function of generalized coordinates and their derivatives) is replaced by a Hamiltonian that is a function of generalized coordinates and momentum. The Hamiltonian is defined as

$$H[q, p, t] = p \cdot \dot{q} - L[q, \dot{q}, t]\ ,$$

where the momentum is obtained by differentiating the Lagrangian as above. The Hamiltonian equations of motion are

$$\dot{q}^i = \frac{\partial H}{\partial p_i}\,, \qquad -\dot{p}_i = \frac{\partial H}{\partial q^i} \qquad \text{and} \qquad -\frac{\partial L}{\partial t} = \frac{\mathrm{d}H}{\mathrm{d}t}\,.$$

(Note the signs and the partial and ordinary derivatives!)

As in Lagrangian mechanics, if a generalized coordinate does *not* appear in the Hamiltonian, its conjugate momentum component is *conserved*.

Symmetry and conservation: Conservation of momentum is a mathematical consequence of the homogeneity (shift symmetry) of space (position in space is the *canonical conjugate quantity to momentum*). That is, conservation of momentum is a consequence of the fact that the laws of physics do not depend on position; this is a special case of Noether's theorem.

15.2. **Poisson Bracket.** In mathematics and classical mechanics, the *Poisson bracket* is an important binary operation in Hamiltonian mechanics, playing a central role in Hamilton's equations of motion, which govern the time-evolution of a Hamiltonian dynamical system. The Poisson bracket also distinguishes a certain class of coordinate-transformations, called *canonical transformations*, which maps canonical coordinate systems into canonical coordinate systems. (A "canonical coordinate system" consists of canonical position and momentum variables that satisfy canonical Poisson-bracket relations.) Note that the set of possible canonical transformations is always very rich. For instance, often it is possible

to choose the Hamiltonian itself $H = H[q, p; t]$ as one of the new canonical momentum coordinates.

In a more general sense: the Poisson bracket is used to define a Poisson algebra, of which the algebra of functions on a Poisson manifold is a special case.

15.2.1. *Canonical coordinates.* In canonical coordinates (also known as Darboux coordinates) $\left[q^i, p_j\right]$ on the phase space, given two functions $f\left[p_i, q^i, t\right]$ and $g\left[p_i, q^i, t\right]$, then the Poisson bracket takes the form

$$[f, g]_{\mathrm{pb}} = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial q^i} \frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i} \frac{\partial g}{\partial q^i} \right) .$$

15.2.2. *Hamilton's Equations of Motion.* The Hamilton's equations of motion have an equivalent expression in terms of the Poisson bracket. This may be most directly demonstrated in an explicit coordinate frame. Suppose that $f[p, q, t]$ is a function on the manifold. Then from the multivariable chain rule, one has

$$\frac{\mathrm{d}f}{\mathrm{d}t}[p, q, t] = \frac{\partial f}{\partial p} \frac{\mathrm{d}p}{\mathrm{d}t} + \frac{\partial f}{\partial q} \frac{\mathrm{d}q}{\mathrm{d}t} + \frac{\partial f}{\partial t} .$$

Further, one may take $p = p[t]$ and $q = q[t]$ to be solutions to Hamilton's equations; that is,

$$\dot{q} = \frac{\partial H}{\partial p} = [q, H]_{\mathrm{pb}} \qquad \text{and} \qquad \dot{p} = \frac{\partial H}{\partial q} = [p, H]_{\mathrm{pb}} .$$

Then one has

$$\frac{\mathrm{d}f}{\mathrm{d}t}[p, q, t] = [f, H]_{\mathrm{pb}} + \frac{\partial f}{\partial t} .$$

Thus, the time evolution of a function f on a symplectic manifold can be given as a one-parameter family of symplectomorphisms (*i.e.*, canonical transformations, area-preserving diffeomorphisms), with the time t being the parameter: Hamiltonian motion is a canonical transformation generated by the Hamiltonian. That is, Poisson brackets are preserved in it, so that *any* time t in the solution to Hamilton's equations,

$$q[t] = \exp -t \, [H, \cdot]_{\mathrm{pb}} q[0] \qquad \text{and} \qquad p[t] = \exp -t \, [H, \cdot]_{\mathrm{pb}} p[0] ,$$

can serve as the bracket coordinates. Poisson brackets are *canonical invariants*.

Dropping the coordinates, one has

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \left(\frac{\partial}{\partial t} - [H, \cdot]_{\mathrm{pb}} \right) f .$$

The operator in the convective part of the derivative, $i\hat{L} = -\left[H, \cdot\right]_{\mathrm{pb}}$ is sometimes referred to as the Liouvillian.

15.2.3. *Constants of Motion.* An integrable dynamical system will have constants of motion in addition to the energy. Such constants of motion will commute with the Hamiltonian under the Poisson bracket. Suppose some function $f[p, q]$ is a constant of motion. This implies that if $p[t]$, $q[t]$ is a trajectory or solution to the Hamilton's equations of motion, then one has that

$$0 = \frac{\mathrm{d}f}{\mathrm{d}t}$$

along that trajectory. Then one has

$$0 = \frac{\mathrm{d}f}{\mathrm{d}t}[p, q] = [f, H]_{\mathrm{pb}} + \frac{\partial f}{\partial t} ,$$

where, as above, the intermediate step follows by applying the equations of motion. This equation is known as the *Liouville equation*. The content of Liouville's theorem is that the time evolution of a measure (or "distribution function" on the phase space) is given by the above.

If the Poisson bracket of f and g vanishes ($[f, g]_{\mathrm{pb}} = 0$), then f and g are said *to be in involution*. In order

> for a Hamiltonian system to be completely integrable, all of the constants of motion must be in mutual involution.

15.3. **Lagrangian versus Hamiltonian Approaches.** [Lectures on Classical Mechanics – John C. Baez]

I am not sure where to mention this, but before launching into the history of the Lagrangian approach may be as good a time as any. In later chapters we will describe another approach to classical mechanics: the Hamiltonian approach. Why do we need two approaches, Lagrangian and Hamiltonian?

They both have their own advantages. In the simplest terms, the Hamiltonian approach focuses on *position and momentum*, while the Lagrangian approach focuses on *position and velocity*. The Hamiltonian approach focuses on energy, which is a function of position and momentum – indeed, 'Hamiltonian' is just a fancy word for energy. The Lagrangian approach focuses on the *Lagrangian*, which is a function of position and velocity. Our first task in understanding Lagrangian mechanics is to get a gut feeling for what the Lagrangian means. The key is to understand the integral of the Lagrangian over time – the 'action', A. We shall see that this describes the 'total amount that happened' from one moment to another as a particle traces out a path. And, peeking ahead to quantum mechanics, the quantity $\exp[iA/\hbar]$, where $\hbar$ is Planck's constant, will describe the 'change in phase' of a *quantum* system as it traces out this path.

In short, while the Lagrangian approach takes a while to get used to, it provides invaluable insights into classical mechanics and its relation to quantum mechanics.

15.4. **Prehistory of the Lagrangian Approach.** We've seen that a particle going from point a at time t_0 to a point b at time t_1 follows a path that is a critical point of the action,

$$A = \int_{t_0}^{t_1} (k - v)\, \mathrm{d}t\,,$$

so that slight changes in its path do not change the action (to first order). Often, though not always, the action is minimized, so this is called the *Principle of Least Action*.

Suppose we did not have the hindsight afforded by the Newtonian picture. Then we might ask, "Why does nature like to minimize the action? And why *this* action $(k - v)\, \mathrm{d}t$? Why not some other action?"

'Why' questions are always tough. Indeed, some people say that scientists should never ask 'why'. This seems too extreme: a more reasonable attitude is that we should only ask a 'why' question if we expect to learn something scientifically interesting in our attempt to answer it.

There are certainly some interesting things to learn from the question "why is action minimized?" First, note that total energy is conserved, so energy can slosh back and forth between kinetic and potential forms. The Lagrangian $L = k - v$ is big when most of the energy is in kinetic form, and small when most of the energy is in potential form. Kinetic energy measures how much is 'happening' – how much our system is moving around. Potential energy measures how much *could* happen, but isn't yet – that's what the word 'potential' means. (Imagine a big rock sitting on top of a cliff, with the potential to fall down.) So, the Lagrangian measures something we could vaguely refer to as the 'activity' or 'liveliness' of a system: the higher the kinetic energy the more lively the system, the higher the potential energy the less lively. So, we're being told that nature likes to minimize the total of 'liveliness' over time: that is, the total action. In other words,

nature is as lazy as possible!

For example, consider the path of a thrown rock in the Earth's gravitational field. The rock traces out a parabola, and we can think of it as doing this in order to minimize its action. On the one hand, it wants to spend a lot much time near the top of its trajectory, since this is where the kinetic energy is least and the potential energy is greatest. On the other hand, if it spends *too* much time near the top of its trajectory, it will need to really rush to get up there and get back down, and this will take a lot of action. The perfect compromise is a parabolic path!

Here we are anthropomorphizing the rock by saying that it 'wants' to minimize its action. This is okay if we don't take it too seriously. Indeed, one of the virtues of the Principle of Least Action is that it lets us put ourselves in the position of some physical system and imagine what we would do to minimize the action.

There is another way to make progress on understanding 'why' action is minimized: history. Historically there were two principles that were fairly easy to deduce from observations of nature: (i) the principle of minimum energy used in statics, and (ii) the principle of least time, used in optics. By putting these together, we can guess the principle of least action.

15.5. **Hamilton's Equations.** [Hamilton's equations. Dr. M Ramegowda]

15.5.1. *The Hamilton equations of motion.* Lagrange formulation is in terms of generalized coordinates q^i and generalized velocities $\dot{q}^i$ gives equations of motion, which are second order in time. Instead if we regard N generalized coordinates q^i and N generalized momenta p_i as independent variables, and again $q[t]$ and $p[t]$ at every instant of time t, we will get $2N$ first order equations. Hence the $2N$ equations of motion describe the behavior of the system in a phase space whose coodinates are the $2N$ independent variables. These are called *canonical coordinates* and *canonical momenta*. This new formulation is by the Hamiltonian and is known as Hamiltonian formulation.

The Lagrange equations for a free particle can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0\,,$$

where

$$L[q,\dot{q},t] = k - v = \frac{1}{2}\sum_i m_i \dot{q}^i \dot{q}^i - v$$

and

$$\frac{\partial L}{\partial \dot{q}^i} = m_i \dot{q}^i = p_i\,.$$

The p_i are called *generalized or conjugate momenta*. Replacing last equation in Lagrange equations gives,

$$\dot{p_i} = \frac{\partial L}{\partial q^i}\,.$$

The differential of the Lagrangian can be written as

$$\mathrm{d}L = \sum_i \frac{\partial L}{\partial q^i}\mathrm{d}q^i + \sum_i \frac{\partial L}{\partial \dot{q}^i}\mathrm{d}\dot{q}^i + \frac{\partial L}{\partial t}\mathrm{d}t\,.$$

Applying the previous definitions into the last equations, one has

$$\mathrm{d}L = \sum_i \dot{p_i}\mathrm{d}q^i + \sum_i p_i \mathrm{d}\dot{q}^i + \frac{\partial L}{\partial t}\mathrm{d}t\,.$$

If we define the Hamiltonian $H[q,p,t]$ as a function of generalized coordinates q^i and generalized momenta p_i, the Legendre transformation generate the Hamiltonian

$$H[q,p,t] = \sum_i \dot{q}^i p_i - L[q,\dot{q},t]\ .$$

Finding the differential of the Hamiltonian and plugging it into the last equation, one has

$$\dot{q}^i = \frac{\partial H}{\partial p_i}\,, \qquad -\dot{p_i} = \frac{\partial H}{\partial q^i} \qquad \text{and} \qquad -\frac{\partial L}{\partial t} = \frac{\partial H}{\partial t}\,.$$

(Note the signs!)

The first two equations are known as the *canonical equations of Hamilton*. They constitute the desired set of $2N$ first order equations of motion replacing the N second order Lagrange equations.

If $[x,y,z]$ are the Cartesian coordinates at time t of a free material point of mass m moving in a potential field $v[x,y,z] = v\left[q^i\right]$, we may take $q^1 = x$, $q^2 = y$ and $q^3 = z$.

The kinetic energy k is given by

$$k = \frac{1}{m}\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right) = \frac{1}{2}m\sum_i (\dot{q}^i)^2\,.$$

The Lagrangian for the particle is

$$k - v = L = \frac{1}{2}m\sum_i (\dot{q}^i)^2 - v\left[q^i\right]\ ,$$

which implies that

$$\frac{\partial L}{\partial \dot{q}^i} = m\dot{q}^i \qquad \text{and} \qquad p_i = m\dot{q}^i \,.$$

On substituting for L and p_i in the definition of the Hamiltonian, one has

$$H = \sum_i \dot{q}^i p_i - L = m\sum_i (\dot{q}^i)^2 - (k - v) = k + v \,.$$

Thus the Hamiltonian becomes the total energy of the system.

15.5.2. *Hamiltonian for a free particle in different coordinates.* Using Cartesian coordinates: $[x, y, z]$ are the Cartesian coordinates at time t of a free material point of mass m moving in a potential field $v[x, y, z]$. The kinetic energy k is given by $2k = m\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right)$. Thus the Hamiltonian for the particle is

$$H = k + v = \frac{1}{2}m\left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right) + v[x, y, z] = \frac{1}{2m}\left(p_x^2 + p_y^2 + p_z^2\right) + v[x, y, z] \,.$$

Using cylindrical polar coordinates: $[r, \theta, z]$ are the cylindrical coordinates at time t of a free material point of mass m in the potential field $v[r]$. The kinetic energy k is

$$k = \frac{1}{2}m\left(\dot{r}^2 + (r\dot{\theta})^2 + \dot{z}^2\right) = \frac{1}{2m}\left((m\dot{r})^2 + 1/r^2(mr^2\dot{\theta})^2 + (m\dot{z})^2\right) = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2} + p_z^2\right) \,.$$

Thus,

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2} + p_z^2\right) + v[r] \,.$$

Using spherical polar coordinates: $[r, \theta, \phi]$ are the spherical polar coordinates at time t of a free material point of mass m in the potential field $v[r]$.

Following a procedure analogous to the one used to find the Hamiltonian in cylindrical polar coordinates, then the Hamiltonian in spherical polar coordinates becomes

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2} + \frac{p_\phi^2}{r^2\,\sin[\theta]}\right) + v[r] \,.$$

15.5.3. *Hamiltonian for an electron in a Coulomb field.* When an electron revolving about the charge e, its potential energy is given by $v = -e^2/r$. Then, the Hamiltonian is

$$H = \frac{1}{2m}\left(p_r^2 + \frac{p_\theta^2}{r^2} + \frac{p_\phi^2}{r^2\,\sin[\theta]}\right) - \frac{e^2}{r} \,.$$

15.5.4. *Hamiltonian for the simple harmonic oscillator.* The Lagrangian for a simple harmonic oscillator can be written as

$$L = \frac{1}{2}m\sum_i (\dot{q}^i)^2 - \frac{1}{2}m\omega^2\sum_i (q^i)^2 \,.$$

The generalized momentum is

$$p_i = \frac{\partial L}{\partial \dot{q}^i} = m\dot{q}^i \implies \dot{q}^i = \frac{p_i}{m} \,.$$

Then, the Hamiltonian becomes

$$H = \sum_i p_i\dot{q}^i - L = \frac{1}{2m}\sum_i p_i^2 + \frac{1}{2}m\omega^2\sum_i (q^i)^2 \,.$$

15.5.5. *Hamiltonian for an electron in electromagnetic field.* Consider a particle of mass m and charge e moving in an electromagnetic field. Lagrangian for the particle is

$$L = k - v = \frac{1}{2}m\sum_i (\dot{q}^i)^2 - e\left(\phi - A\cdot\sum_i q^i\right) \,,$$

where $e\left(\phi - A\cdot\sum_i q^i\right)$ is the velocity dependent potential.

The generalized momentum is

$$p_i = \frac{\partial L}{\partial \dot{q}^i} = m\dot{q}^i + eA \,.$$

And thus the Hamiltonian becomes

$$H = \sum_i p_i \dot{q}^i - L = \frac{1}{2m}\left(p - eA\right)^2 + e\phi\,.$$

15.5.6. *Cyclic Coordinates.* Consider a system of N degrees of freedom described by q^i generalized coordinates. Then, the Lagrangian of the system is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}^i}\right) - \frac{\partial L}{\partial q^i} = 0\,.$$

If the Lagrangian of the system does not contain a given coordinate q^i even though it may contain corresponding velocity $\dot{q}^i$, then the coordinate q^i is said to be *cyclic* or *ignorable*. Then,

$$\frac{\partial L}{\partial q^i} = 0 \implies \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}^i} = 0 \implies \frac{\mathrm{d}p}{\mathrm{d}t} = 0 \implies p_i = \text{constant}\,.$$

Therefore, *the generalized momentum conjugate to a cyclic coordinate is conserved.*

Example: In a planetary motion, θ is cyclic. Therefore, the angular momentum $p_\theta = mr^2\dot{\theta}^2$ is constant or, equivalently, the angular momentum is conserved.

15.5.7. *Poisson brackets.* Poisson brackets are a powerful and sophisticated tool in the Hamiltonian formalism of Classical Mechanics. They also happen to provide a direct link between classical and quantum mechanics. A classical system with N degrees of freedom, say a set of $N/3$ particles in three dimensions, is described by $2N$ *phase space* coordinates. These are the N generalized coordinates $\left\{q^1, q^2, \ldots, q^N\right\}$ and N conjugate momenta $\{p_1, p_2, \ldots, p_n\}$. The Hamiltonian of the system depends on these $2N$ variables and possibly on time t as well, and it can be expressed as

$$H\left[q^1, \ldots, q^n; p_1, \ldots, p_N; t\right] H\left[q^i, p_i, t\right]\,.$$

The Poisson bracket is an operation which takes *two* functions of phase space and time, call them $f\left[q^i, p_i, t\right]$ and $g\left[q^i, p_i, t\right]$ and produces a *new* function. With respect to canonical coordinates $\left[q^i, p_i\right]$, it is defined as

$$[f, g]_{\mathrm{pb}} = \sum_i^N \left(\frac{\partial f}{\partial q^i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q^i}\right)\,.$$

In vector notation, Hamilton's equations can be expressed in a more symmetric fashion using Poisson brackets:

$$\dot{q} = [q, H]_{\mathrm{pb}} \qquad \text{and} \qquad \dot{p} = [p, H]_{\mathrm{pb}}\,.$$

Proof. Expand the last equations using the definition of Poisson brackets and note that, in Hamilton formalism, q and p are *independent of each other* (*i.e.*, $q_{,p} = p_{,q} = 0$) to have

$$\dot{q} = \frac{\partial q}{\partial q}\frac{\partial H}{\partial p} - \frac{\partial q}{\partial p}\frac{\partial H}{\partial q} = \frac{\partial H}{\partial p} \qquad \text{and} \qquad \dot{p} = \frac{\partial p}{\partial q}\frac{\partial H}{\partial p} - \frac{\partial p}{\partial p}\frac{\partial H}{\partial q} = -\frac{\partial H}{\partial q}\,. \qquad \square$$

In the case of a single degree of freedom, $N = 1$, phase space is 2-dimensional, $[q, p]$, and the Poisson bracket has only two terms

$$[f, g]_{\mathrm{pb}} = \frac{\partial f}{\partial q}\frac{\partial g}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial g}{\partial q}\,.$$

The time derivative of the function $f\left[q^i, p_i, t\right]$ is

$$\frac{\mathrm{d}f}{\mathrm{d}t} = [f, H]_{\mathrm{pb}} + \frac{\partial f}{\partial t}\,.$$

This is the equation of motion of the function f expressed in terms of Poisson bracket.

15.5.8. *Constants of the Motion.* A constant of the motion is some function of phase space, independent of time, $f\big[q^i, p_i\big]$, whose value is constant for any particle. In other words, f is a constant of the motion if

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \dot{f} = 0\,.$$

Since we specified that f does not depend explicitly in time it follows that $f_{,t} = 0$. Then,

$$[f, H]_{\mathrm{pb}} = 0\,.$$

Thus f is a constant of the motion if and only if $[f, H]_{\mathrm{pb}} = 0$ for all points in phase space.

Energy: Due to the anti-symmetry of the Poisson bracket $[H, H]_{\mathrm{pb}} = 0$. Using this we find

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial H}{\partial t}\,.$$

If the Hamiltonian does not depend on time explicitly, then $H_{,t} = 0$. Therefore $\dot{H} = 0$ and $H\big[q^i, p_i\big]$ is constant. In other words,

> energy is conserved in cases where the Hamiltonian is time independent.

Linear Momentum: In a case where the Hamiltonian does not contain a particular coordinate, q^i, explicitly it is said to be cyclic in that coordinate. Then

$$[p_i, H]_{\mathrm{pb}} = -\frac{\partial H}{\partial q^i}\,,$$

Since q^i is cyclic, then $H_{,q^i} = 0$ and $[p_i, H]_{\mathrm{pb}} = 0$. Therfore, p_i is a constant of the motion. In other words,

> momentum is conserved if it is conjugate to a cyclic coordinate.

Angular Momentum: Consider a particle in three dimension, $[x, y, z]$, object to a central force potential $v[r] = v[x, y, z]$. [... maths here to show the following]

> for a particle moving in a central force potential all three components of angular momentum are conserved.

16. Probability

The probable is what usually happens.

— ARISTOTLE,

16.1. **Wiki.** Probability is a measure or estimation of how likely it is that something will happen or that a statement is true. Probabilities are given a value between 0 (0% chance or *will not happen*) and 1 (100% chance or *will happen*). The higher the degree of probability, the more likely the event is to happen, or, in a longer series of samples, the greater the number of times such event is *expected* to happen.

These concepts have been given an axiomatic mathematical derivation in probability theory, which is used widely in such areas of study as mathematics, statistics, finance, gambling, science, artificial intelligence/machine learning and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

16.1.1. *Interpretations.* When dealing with experiments that are random and well-defined in a purely theoretical setting (like tossing a fair coin), probabilities describe the statistical number of outcomes considered divided by the number of all outcomes (tossing a fair coin twice will yield HH with probability 1/4, because the four outcomes HH, HT, TH and TT are possible). When it comes to practical application, however, the word probability does not have a singular direct definition. In fact, there are two major categories of probability interpretations, whose adherents possess conflicting views about the fundamental nature of probability:

- Objectivists assign numbers to describe some objective or physical state of affairs. The most popular version of objective probability is frequentist probability, which claims that the probability of a random event denotes the relative frequency of occurrence of an experiment's outcome, when repeating the experiment. This interpretation considers probability to be the relative frequency "in the long run" of outcomes. A modification of this is propensity probability, which interprets probability as the tendency of some experiment to yield a certain outcome, even if it is performed only once.
- Subjectivists assign numbers per subjective probability, *i.e.*, as a degree of belief. The most popular version of subjective probability is Bayesian probability, which includes expert knowledge as well as experimental data to produce probabilities. The expert knowledge is represented by some (subjective) prior probability distribution. The data is incorporated in a likelihood function. The product of the prior and the likelihood, normalized, results in a posterior probability distribution that incorporates all the information known to date.[6] Starting from arbitrary, subjective probabilities for a group of agents, some Bayesians claim that all agents will eventually have sufficiently similar assessments of probabilities, given enough evidence.

16.1.2. *Etymology.* The modern meaning of probability: probability is a measure of the weight of empirical evidence that is arrived at from inductive reasoning and statistical inference.

16.2. **Theory.** Like other theories, the theory of probability is a representation of probabilistic concepts in formal terms – that is, in terms that can be considered separately from their meaning. These formal terms are manipulated by the rules of mathematics and logic and any results are interpreted or translated back into the problem domain.

There have been at least two successful attempts to formalize probability, namely the Kolmogorov formulation and the Cox formulation. In Kolmogorov's formulation (see probability space), sets are interpreted as events and probability itself as a measure on a class of sets. In Cox's theorem, probability is taken as a primitive (that is, not further analyzed) and the emphasis is on constructing a consistent assignment of probability values

to propositions. In both cases, the laws of probability are the same, except for technical details.

There are other methods for quantifying uncertainty, such as the Dempster-Shafer theory or possibility theory, but those are essentially different and not compatible with the laws of probability as usually understood.

16.3. **Mathematical treatment.** Consider an *experiment* that can produce a number of *results*. The collection of all results is called the *sample space* of the experiment. The power set of the sample space is formed by considering all different collections of possible results. For example, rolling a die can produce six possible results. One collection of possible results gives an odd number on the die. Thus, the subset $\{1, 3, 5\}$ is an element of the power set of the sample space of die rolls. These collections are called *events*. In this case, $\{1, 3, 5\}$ is the event that the die falls on some odd number. If the results that actually occur fall in a given event, the event is said to have occurred.

A probability is a way of assigning every event a value between zero and one, with the requirement that the event made up of all possible results (in our example, the event $\{1, 2, 3, 4, 5, 6\}$) is assigned a value of one. To qualify as a probability, the assignment of values must satisfy the requirement that if you look at a collection of *mutually exclusive events* (events with no common results, *e.g.*, the events $\{1, 6\}$, $\{3\}$ and $\{2, 4\}$ are all mutually exclusive), the probability that at least one of the events will occur is given by the sum of the probabilities of all the individual events.

The probability of an event A is written as $P[A]$, $p[A]$ or $Pr[A]$. This mathematical definition of probability can extend to infinite sample spaces, and even uncountable sample spaces, using the concept of a measure.

The *opposite or complement of an event* A is the event [not A] (that is, the event of A not occurring); its probability is given by $p[\neg A] = 1 - p[A]$. As an example, the chance of not rolling a six on a six-sided die is 1: p[change of rolling a six] $= 1 - 1/6 = 1/5$.

If both events A and B occur on a single performance of an experiment, this is called the *intersection* or *joint probability of* A *and* B, denoted as $p[A \cap B]$.

16.4. **Independent probability.** If two events, A and B are *independent*, then the joint probability is

$$p[A \wedge B] = p[A \cap B] = p[A]\, p[B] \ ,$$

for example, if two coins are flipped the chance of both being heads is $1/2 \times 1/2 = 1/4$.

16.5. **Mutually exclusive.** If either event A or event B or both events occur on a single performance of an experiment this is called the *union of the events* A *and* B denoted as $p[A \cup B]$. If two events are mutually exclusive then the probability of either occurring is

$$p[A \vee B] = p[A \cup B] = p[A] + p[B] \ .$$

For example, the chance of rolling a 1 or 2 on a six-sided die is $1/6 + 1/6 = 1/3$.

16.6. **Not mutually exclusive.** If the events are *not* mutually exclusive then

$$p[A \vee B] = p[A] + p[B] - p[A \wedge B] \ .$$

For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J,Q,K) (or one that is both) is $13/52 + 12/52 - 3/52 = 11/26$, because of the 52 cards of a deck 13 are hearts, 12 are face cards, and 3 are both: here the possibilities included in the "3 that are both" are included in each of the "13 hearts" and the "12 face cards" but should only be counted once.

16.6.1. *Conditional probability.* Conditional probability is the probability of some event A, given the occurrence of some other event B. Conditional probability is written $A \,|\, B$, and is read "the probability of A, given B". It is defined by

$$p[A \,|\, B] = \frac{p[A \cap B]}{p[B]} \ .$$

If $p[B] = 0$, then $p[A \,|\, B]$ is formally undefined by this expression.

For example, in a bag of 2 red balls and 2 blue balls (4 balls in total), the probability of taking a red ball is $1/2$; however, when taking a second ball, the probability of it being

either a red ball or a blue ball depends on the ball previously taken, such as, if a red ball was taken, the probability of picking a red ball again would be since only 1 red and 2 blue balls would have been remaining.

16.6.2. *Summary of Probabilities.*

- event: probability
- A: $p[A] \in [0,1]$.
- not A: $p[\neg A] = 1 - p[A]$.
- A or B: $p[A \vee B] = p[A \cup B] = p[A] + p[B] - p[A \cap B]$.
- A or B (mutually exclusive): $p[A \vee B] = p[A \cup B] = p[A] + p[B]$.
- A and B: $p[A \wedge B] = p[A \cap B] = p[A \,|\, B]\, p[B] = p[B \,|\, A]\, p[A]$.
- A and B (independent): $p[A \wedge B] = p[A \cap B] = p[A]\, p[B]$.
- A given B (conditional prob.): $p[A \,|\, B] = p[A \cap B]\, / p[B]$.

16.7. **Probability and Probability Experiments.** [Probability demystified – Allan G. Bluman]

Probability can be defined as the mathematics of chance.

A *probability experiment* is a chance process that leads to well-defined outcomes or results. For example, tossing a coin can be considered a probability experiment since there are two well-defined outcomes: heads and tails.

An *outcome* of a probability experiment is the result of a *single* trial of a probability experiment. A *trial* means flipping a coin once or drawing a single card from a deck. A trial could also mean rolling two dice at once, tossing three coins at once or drawing five cards from a deck at once. *A single trial of a probability experiment means to perform the experiment one time.*

The set of all outcomes of a probability experiment is called a *sample space*. Some sample spaces for various probability experiments are

- experiment: sample space;
- toss a coin: $\{H, T\}$: head, tail;
- roll a die: $\{1, 2, 3, 4, 5, 6\}$;
- Toss two coins: $\{HH, HT, TH, TT\}$.

It should be mentioned that each outcome of a probability experiment occurs *at random*. This means you cannot predict with certainty which outcome will occur when the experiment is conducted. Also,

each outcome of the experiment is equally likely unless otherwise stated.

That means that each outcome has the same probability of occurring.

When finding probabilities, it is often necessary to consider several outcomes of the experiment. For example, when a single die is rolled, you may want to consider obtaining an even number; that is, a two, four, or six. This is called an event. An event then usually consists of one or more outcomes of the sample space. (Note: It is sometimes necessary to consider an event which has no outcomes. More later.)

An event with one outcome is called a *simple event*. For example, a die is rolled and the event of getting a four is a simple event since there is only one way to get a four. When an event consists of two or more outcomes, it is called a *compound event*. For example, if a die is rolled and the event is getting an odd number, the event is a compound event since there are three ways to get an odd number, namely, 1, 3, or 5.

Simple and compound events should *not* be confused with the number of times the experiment is repeated. For example, if two coins are tossed, the event of getting two heads is a simple event since there is only one way to get two heads, whereas the event of getting a head and a tail in either order is a compound event since it consists of two outcomes, namely head, tail and tail, head.

16.7.1. *Classical Probability.* Sample spaces are used in classical probability to determine the numerical probability that an event will occur. The formula for determining the probability of an event E is

$$p[E] = \frac{\text{number of outcomes contained in the event } E}{\text{total number of outcomes in the sample space}}\,.$$

Example: Two coins are tossed; find the probability that both coins land heads up.

Solution: The sample space for tossing two coins is HH, HT, TH, and TT. Since there are 4 events in the sample space, and only one way to get two heads (HH), the answer is $p[HH] = 1/4$.

Probabilities can be expressed as reduced fractions, decimals, or percents. For example, if a coin is tossed, the probability of getting heads up is 1/2 or 0.5 or 50%. (Note: Some mathematicians feel that probabilities should be expressed only as fractions or decimals. However, probabilities are often given as percents in everyday life. For example, one often hears, "There is a 50% chance that it will rain tomorrow".)

Probability problems use a certain language. For example, suppose a die is tossed. An event that is specified as "getting at least a 3" means getting a 3, 4, 5, or 6. An event that is specified as "getting at most a 3" means getting a 1, 2, or 3.

16.7.2. *Probability Rules.* There are certain rules that apply to classical probability theory. They are presented next.

(1) Rule 1: The probability of any event will always be a number from zero to one.
    This can be denoted mathematically as $0 \leq p[E] \leq 1$. What this means is that all answers to probability problems will be numbers ranging from zero to one. Probabilities cannot be negative nor can they be greater than one.
    Also, when the probability of an event is close to zero, the occurrence of the event is relatively *unlikely*. For example, if the chances that you will win a certain lottery are 0.001 or one in one thousand, you probably won't win, unless of course, you are very "lucky". When the probability of an event is 0.5 or 12, there is a 50-50 chance that the event will happen – the same probability of the two outcomes when flipping a coin. When the probability of an event is close to one, the event is *almost sure to occur*. For example, if the chance of it snowing tomorrow is 90%, more than likely, you'll see some snow.
(2) Rule 2: When an event cannot occur, the probability will be zero.
(3) Rule 3: When an event is certain to occur, the probability is 1.
(4) Rule 4: The sum of the probabilities of all of the outcomes in the sample space is 1.
(5) Rule 5: The probability that an event will not occur is equal to 1 minus the probability that the event will occur.
    If an event E consists of certain outcomes, then event $\bar{E}$ is called the *complement of event* E and consists of the outcomes in the sample space which are not outcomes of event E.
    Now rule five can be stated mathematically as
$$p\big[\bar{E}\big] = 1 - E\,.$$

16.7.3. *Empirical Probability.* Probabilities can be computed for situations that do not use sample spaces. In such cases, *frequency distributions* are used and the probability is called *empirical probability*. For example, suppose a class of students consists of 4 freshmen, 8 sophomores, 6 juniors, and 7 seniors. The information can be summarized in a frequency distribution in a table.

From a frequency distribution, probabilities can be computed using the following formula:
$$p[E] = \frac{\text{frequency of E}}{\text{sum of the frequencies}}\,.$$
Empirical probability is sometimes called relative frequency probability.

Example: Using the frequency distribution shown previously, find the probability of selecting a junior student at random. Solution: Since there are 6 juniors and a total of 25 students, p[junior] $= 6/25$.

Another aspect of empirical probability is that if a large number of subjects (called a *sample*) is selected from a particular group (called a *population*), and the probability of a specific attribute is computed, then when another subject is selected, we can say that the probability that this subject has the same attribute is the same as the original probability computed for the group. For example, a Gallup Poll of 1004 adults surveyed found that 17% of the subjects stated that they considered Abraham Lincoln to be the

greatest President of the United States. Now if a subject is selected, the probability that he or she will say that Abraham Lincoln was the greatest president is also 17%.

Several things should be explained here. First of all, the 1004 people constituted a sample selected from a larger group called the population. Second, the exact probability for the population can never be known unless every single member of the group is surveyed. This does not happen in these kinds of surveys since the population is usually very large. Hence, the 17% is only an estimate of the probability. However, if the sample is *representative* of the population, the estimate will usually be fairly close to the exact probability. Statisticians have a way of computing the accuracy (called the *margin of error*) for these situations. For the present, we shall just concentrate on the probability.

Also, by a representative sample, we mean the subjects of the sample have similar characteristics as those in the population. There are statistical methods to help the statisticians obtain a representative sample. These methods are called sampling methods and can be found in many statistics books.

16.7.4. *Law of Large Numbers.* We know from classical probability that if a coin is tossed one time, we cannot predict the outcome, but the probability of getting a head is 12 and the probability of getting a tail is 12 if everything is fair. But what happens if we toss the coin 100 times? Will we get 50 heads? Common sense tells us that most of the time, we will not get exactly 50 heads, but we should get close to 50 heads. What will happen if we toss a coin 1000 times? Will we get exactly 500 heads? Probably not. However, as the number of tosses increases, the ratio of the number of heads to the total number of tosses will get closer to 12. This phenomenon is known as the law of large numbers. This law holds for any type of gambling game such as rolling dice, playing roulette, etc.

16.7.5. *Subjective Probability.* A third type of probability is called subjective probability. Subjective probability is based upon an educated guess, estimate, opinion, or inexact information. For example, a sports writer may say that there is a 30% probability that the Pittsburgh Steelers will be in the Super Bowl next year. Here the sports writer is basing his opinion on subjective information such as the relative strength of the Steelers, their opponents, their coach, etc. Subjective probabilities are used in everyday life; however, they are beyond the scope of this book.

16.7.6. *Summary.* Probability is the mathematics of chance. There are three types of probability: classical probability, empirical probability, and subjective probability. Classical probability uses sample spaces. A sample space is the set of outcomes of a probability experiment. The range of probability is from 0 to 1. If an event cannot occur, its probability is 0. If an event is certain to occur, its probability is 1. Classical probability is defined as the number of ways (outcomes) the event can occur divided by the total number of outcomes in the sample space.

Empirical probability uses frequency distributions, and it is defined as the frequency of an event divided by the total number of frequencies.

Subjective probability is made by a person's knowledge of the situation and is basically an educated guess as to the chances of an event occurring.

16.8. **Sample Spaces.**

16.8.1. *Introduction.* In order to compute classical probabilities, you need to find the sample space for a probability experiment. In the previous chapter, sample spaces were found by using common sense. In this chapter two specific devices will be used to find sample spaces for probability experiments. They are tree diagrams and tables.

16.8.2. *Tree Diagrams.* A tree diagram consists of branches corresponding to the outcomes of two or more probability experiments that are done in sequence.

In order to construct a tree diagram, use branches corresponding to the outcomes of the first experiment. These branches will emanate from a single point. Then from each branch of the first experiment draw branches that represent the outcomes of the second experiment. You can continue the process for further experiments of the sequence if necessary.

Example: Three coins are tossed. Draw a tree diagram and find the sample space.

Solution: Each coin can land either heads up (H) or tails up (T); therefore, the tree diagram will consist of three parts and each part will have two branches. See Figure. Hence the sample space is HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

Once the sample space is found, probabilities can be computed.

Example: Three coins are tossed. Find the probability of getting

(1) Two heads and a tail in any order.
(2) Three heads.
(3) No heads.
(4) At least two tails.
(5) At most two tails.

Solution:

(1) There are eight outcomes in the sample space, and there are three ways to get two heads and a tail in any order. They are HHT, HTH, and THH; hence,

$$p[\text{2 heads and a tail}] = 3/8\,.$$

(2) Three heads can occur in only one way; hence

$$p[\text{HHH}] = 1/8\,.$$

(3) The event of getting no heads can occur in only one way – namely, TTT; hence,

$$p[\text{TTT}] = 1/8\,.$$

(4) The event of at least two tails means two tails and one head or three tails. There are four outcomes in this event – namely, TTH, THT, HTT, and TTT; hence,

$$p[\text{at least two tails}] = 4/8 = 1/2\,.$$

(5) The event of getting at most two tails means zero tails, one tail, or two tails. There are seven outcomes in this event – HHH, THH, HTH, HHT, TTH, THT, and HTT; hence,

$$p[\text{at most two tails}] = 7/8\,.$$

When selecting more than one object from a group of objects, it is important to know whether or not the object selected is replaced before drawing the second object. Consider the next two examples.

Example: A box contains a red ball (R), a blue ball (B), and a yellow ball (Y). Two balls are selected at random in succession. Draw a tree diagram and find the sample space if the first ball is replaced before the second ball is selected.

Solution: There are three ways to select the first ball. They are a red ball, a blue ball, or a yellow ball. Since the first ball is replaced before the second one is selected, there are three ways to select the second ball. They are a red ball, a blue ball, or a yellow ball. The tree diagram is shown in figure.

The sample space consists of nine outcomes. They are RR, RB, RY, BR, BB, BY, YR, YB, YY. Each outcome has a probability of 1/9.

Now what happens if the first ball is not replaced before the second ball is selected?

Example: A box contains a red ball (R), a blue ball (B), and a yellow ball (Y). Two balls are selected at random in succession. Draw a tree diagram and find the sample space if the first ball is not replaced before the second ball is selected.

Solution: There are three outcomes for the first ball. They are a red ball, a blue ball, or a yellow ball. Since the first ball is not replaced before the second ball is drawn, there are only two outcomes for the second ball, and these outcomes depend on the color of the first ball selected. If the first ball selected is blue, then the second ball can be either red or yellow, *etc.* The tree diagram is shown in Figure.

The sample space consists of six outcomes, which are RB, RY, BR, BY, YR, YB. Each outcome has a probability of 1/6.

16.8.3. *Tables.* Another way to find a sample space is to use a table.

16.8.4. *Summary.* Two devices can be used to represent sample spaces. They are tree diagrams and tables.

A tree diagram can be used to determine the outcome of a probability experiment. A tree diagram consists of branches corresponding to the outcomes of two or more probability experiments that are done in sequence.

Sample spaces can also be represented by using tables. For example, the outcomes when selecting a card from an ordinary deck can be represented by a table. When two dice are rolled, the 36 outcomes can be represented by using a table. Once a sample space is found, probabilities can be computed for specific events.

16.9. **The Addition Rules.**

16.9.1. *Introduction.* In this chapter, the theory of probability is extended by using what are called the addition rules. Here one is interested in finding the probability of one event or another event occurring. In these situations, one must consider whether or not both events have common outcomes. For example, if you are asked to find the probability that you will get three oranges or three cherries on a slot machine, you know that these two events cannot occur at the same time if the machine has only three windows. In another situation you may be asked to find the probability of getting an odd number or a number less than 500 on a daily three-digit lottery drawing. Here the events have common outcomes. For example, the number 451 is an odd number and a number less than 500. The two addition rules will enable you to solve these kinds of problems as well as many other probability problems.

16.9.2. *Mutually Exclusive Events.* Many problems in probability involve finding the probability of two or more events. For example, when a card is selected at random from a deck, what is the probability that the card is a king or a queen? In this case, there are two situations to consider. They are: 1. The card selected is a king and 2. The card selected is a queen.

Now consider another example. When a card is selected from a deck, find the probability that the card is a king or a diamond. In this case, there are three situations to consider: 1. The card is a king, 2. The card is a diamond, 3. The card is a king and a diamond. That is, the card is the king of diamonds.

The difference is that in the first example, a card cannot be both a king and a queen at the same time, whereas in the second example, it is possible for the card selected to be a king and a diamond at the same time. In the first example, we say the two events are *mutually exclusive*. In the second example, we say the two events are *not mutually exclusive*. Two events then are mutually exclusive if they cannot occur at the same time. In other words,

> mutually exclusive events have *no* common outcomes.

16.9.3. *Addition Rule I.* The probability of two or more events occurring can be determined by using the addition rules. The first rule is used when the events are mutually exclusive.

> Addition Rule I: When two events are mutually exclusive,
>
> $$p[A \vee B] = p[A] + p[B] \ .$$

Example: In a committee meeting, there were 5 freshmen, 6 sophomores, 3 juniors, and 2 seniors. If a student is selected at random to be the chairperson, find the probability that the chairperson is a sophomore or a junior.

Solution: There are 6 sophomores and 3 juniors and a total of 16 students.

$$p[\text{sophomore} \vee \text{junior}] = p[\text{sophomore}] + p[\text{junior}] = 6/16 + 3/16 = 9/16 \, .$$

16.9.4. *Addition Rule II.* When two events are not mutually exclusive, you need to add the probabilities of each of the two events and subtract the probability of the outcomes that are common to both events. In this case, addition rule II can be used.

> Addition Rule II: If A and B are two events that are not mutually exclusive,
> then
>
> $$p[A \vee B] = p[A] + p[B] - p[A \wedge B] \ ,$$

where $p[A \wedge B]$ means the number of outcomes that event A and event B have in common.

Example: A card is selected at random from a deck of 52 cards. Find the probability that it is a 6 or a diamond.

Solution: Let A the event of getting a 6; then $p[a] = 4/52$, since there are four 6s. Let B the event of getting a diamond; then $p[B] = 13/52$, since there are 13 diamonds. Since there is one card that is both a 6 and a diamond (*i.e.*, the 6 of diamonds), $p[A \wedge B] = 1/52$. Hence,

$$p[A \vee B] = p[A] + p[B] - p[A \wedge B] = 4/52 + 13/52 - 1/52 = 16/52 = 4/13\,.$$

The key word for addition is *or*, and it means that one event or the other occurs. If the events are not mutually exclusive, the probability of the outcomes that the two events have in common must be subtracted from the sum of the probabilities of the two events. For the mathematical purist, only one addition rule is necessary, and that is

$$p[A \vee B] = p[A] + p[B] - p[A \wedge B]\ .$$

The reason is that when the events are mutually exclusive, $p[A \wedge B]$ is equal to zero because mutually exclusive events have no outcomes in common.

16.9.5. *Summary.* Many times in probability, it is necessary to find the probability of two or more events occurring. In these cases, the addition rules are used. When the events are mutually exclusive, addition rule I is used, and when the events are not mutually exclusive, addition rule II is used. If the events are mutually exclusive, they have no outcomes in common. When the two events are not mutually exclusive, they have some common outcomes. The key word in these problems is *or*, and it means to *add*.

16.10. **The Multiplication Rules.**

16.10.1. *Introduction.* The previous chapter showed how the addition rules could be used to solve problems in probability. This chapter will show you how to use the multiplication rules to solve many problems in probability. In addition, the concept of independent and dependent events will be introduced.

16.10.2. *Independent and Dependent Events.* The multiplication rules can be used to find the probability of two or more events that occur in sequence. For example, we can find the probability of selecting three jacks from a deck of cards on three sequential draws. Before explaining the rules, it is necessary to differentiate between *independent* and *dependent events*.

Two events, A and B, are said to be *independent* if the fact that event A occurs does *not* affect the probability that event B occurs. For example, if a coin is tossed and then a die is rolled, the outcome of the coin in no way affects or changes the probability of the outcome of the die. Another example would be selecting a card from a deck, replacing it, and then selecting a second card from a deck. The outcome of the first card, as long as it is replaced, has no effect on the probability of the outcome of the second card.

On the other hand, when the occurrence of the first event in some way changes the probability of the occurrence of the second event, the two events are said to be *dependent*. For example, suppose a card is selected from a deck and not replaced, and a second card is selected. In this case, the probability of selecting any specific card on the first draw is 1/52, but since this card is not replaced, the probability of selecting any other specific card on the second draw is 1/52, since there are only 51 cards left.

Another example would be parking in a no parking zone and getting a parking ticket. Again, if you are legally parked, the chances of getting a parking ticket are pretty close to zero (as long as the meter does not run out). However, if you are illegally parked, your chances of getting a parking ticket dramatically increase.

16.10.3. *Multiplication Rule I.* Before explaining the first multiplication rule, consider the example of tossing two coins. The sample space is HH, HT, TH, TT. From classical probability theory, it can be determined that the probability of getting two heads is 1/4, since there is only one way to get two heads and there are four outcomes in the sample space. However, there is another way to determine the probability of getting two heads. In this case, the probability of getting a head on the first toss is 1/2, and the probability of getting a head on the second toss is also 1/2. So the probability of getting two heads can be determined by multiplying $1/2 \times 1/2 = 1/4$. This example illustrates the first multiplication rule.

> Multiplication Rule I: For two independent events A and B,
>
> $$p[A \wedge B] = p[A]\, p[B] \ .$$
>
> In other words, when two independent events occur in sequence, the probability that both events will occur can be found by multiplying the probabilities of each individual event.

The word *and* is the key word and means that both events occur in sequence and to multiply.

Example: A coin is tossed and a die is rolled. Find the probability of getting a tail on the coin and a 5 on the die.

Solution: Since $p[\text{tail}] = 1/2$ and $p[5] = 1/6$, then $p[\text{tail} \wedge 5] = 1/21/6 = 1/12$. Note that the events are independent.

16.10.4. *Multiplication Rule II.* When two sequential events are dependent, a slight variation of the multiplication rule is used to find the probability of both events occurring. For example, when a card is selected from an ordinary deck of 52 cards the probability of getting a specific card is 1/52, but the probability of getting a specific card on the second draw is 1/51, since 51 cards remain.

When the two events A and B are dependent, the probability that the second event B occurs after the first event A has already occurred is written as $p[B \,|\, A]$. This does not mean that B is divided by A; rather, it means *and* is read as "the probability that event B occurs given that event A has already occurred". $p[B \,|\, A]$ also means the conditional probability that event B occurs given event A has occurred. The second multiplication rule follows.

> Multiplication Rule II: When two events are dependent, the probability of both events occurring is
>
> $$p[A \wedge B] = p[A]\, p[B \,|\, A] \ .$$

Example: A box contains 24 toasters, 3 of which are defective. If two toasters are selected and tested, find the probability that both are defective.

Solution: Since there are 3 defective toasters out of 24, the probability that the first toaster is defective is $3/24 = 1/8$. Since the second toaster is selected from the remaining 23 and there are two defective toasters left, the probability that it is defective is 2/23. Hence, the probability that both toasters are defective is

$$p[D_1 \wedge D_2] = p[D_1]\, p[D_2 \,|\, D_1] = 3/24 \times 2/23 = 1/92 \,.$$

Remember that the key word for the multiplication rule is *and*. It means to *multiply*.

When two events are dependent, the probability that the second event occurs must be adjusted for the occurrence of the first event. For the mathematical purist, only one multiplication rule is necessary for two events, and that is

$$p[A \wedge B] = p[A]\, p[B \,|\, A] \ .$$

The reason is that when the events are independent $p[B \,|\, A] = p[B]$, since the occurrence of the first event A has no effect on the occurrence of the second event B.

16.10.5. *Conditional Probability.* Previously, conditional probability was used to find the probability of sequential events occurring when they were dependent. Recall that $p[B \mid A]$ means the probability of event B occurring given that event A has already occurred. Another situation where conditional probability can be used is when additional information about an event is known. Sometimes it might be known that some outcomes in the sample space have occurred or that some outcomes cannot occur. When conditions are imposed or known on events, there is a possibility that the probability of the certain event occurring may change. For example, suppose you want to determine the probability that a house will be destroyed by a hurricane. If you used all houses in the United States as the sample space, the probability would be very small. However, if you used only the houses in the states that border the Atlantic Ocean as the sample space, the probability would be much higher. Consider the following examples.

Example: A die is rolled; find the probability of getting a 4 if it is known that an even number occurred when the die was rolled.

Solution: If it is known that an even number has occurred, the sample space is reduced to 2, 4, or 6. Hence the probability of getting a 4 is 13 since there is one chance in three of getting a 4 if it is known that the result was an even number.

The previou examples of conditional probability was solved using classical probability and reduced sample spaces; however, they can be solved by using the following formula for conditional probability.

The conditional probability of two events A and B is

$$p[A \mid B] = \frac{p[A \wedge B]}{p[B]} .$$

$p[A \wedge B]$ means the probability of the outcomes that events A and B have in common.

16.10.6. *Summary.* When two events occur in sequence, the probability that both events occur can be found by using one of the multiplication rules. When two events are independent, the probability that the first event occurs does not affect or change the probability of the second event occurring. If the events are independent, multiplication rule I is used. When the two events are dependent, the probability of the second event occurring is changed after the first event occurs. If the events are dependent, multiplication rule II is used. The key word for using the multiplication rule is *and*. Conditional probability is used when additional information is known about the probability of an event.

16.11. **Expectation.** When a person plays a slot machine, sometimes the person wins and other times – most often – the person loses. The question is, "How much will the person win or lose in the long run?" In other words, what is the person's expected gain or loss? Although an individual's exact gain or exact loss cannot be computed, the overall gain or loss of all people playing the slot machine can be computed using the concept of mathematical expectation.

Expectation or *expected value* is a long run average. The expected value is also called the *mean*, and it is used in games of chance, insurance, and in other areas such as decision theory. The outcomes must be numerical in nature. The expected value of the outcome of a probability experiment can be found by multiplying each outcome by its corresponding probability and adding the results.

Formally defined, the expected value for the outcomes of a probability experiment is

$$e[x] = x_1 p[x_1] + x_2 p[x_2] + \cdots + x_n p[x_n] ,$$

where the x corresponds to an outcome and the $p[x]$ to the corresponding probability of the outcome.

16.12. **The Counting Rules.**

16.12.1. *Introduction.* Since probability problems require knowing the total number of ways one or more events can occur, it is necessary to have a way to compute the number of outcomes in the sample spaces for a probability experiment. This is especially true when the number of outcomes is large. For example, when finding the probability of a specific poker hand, it is necessary to know the number of different possible ways five cards can be dealt from a 52-card deck. (This computation will be shown later in this chapter.)

In order to do the computation, we use the fundamental counting rule, the permutation rules, and the combination rule. The rules then can be used to compute the probability for events such as winning lotteries, getting a specific hand in poker, *etc.*

16.12.2. *The Fundamental Counting Rule.* The first rule is called the Fundamental Counting Rule.

For a sequence of n events in which the first event can occur in k_1 ways and the second event can occur in k_2 ways and the third event can occur in k_3 ways, and so on, the total number of ways the sequence can occur is $k_1 k_2 k_3 \cdots k_n$.

Example: In order to paint a room, a person has a choice of four colors: white, light blue, yellow, and light green; two types of paint: oil or latex; and three types of texture: flat, semi-glass, or satin. How many different selections can be made?

Solution: There are four colors, two types of paint, and three textures, so the total number of ways a paint can be selected is $4 \times 2 \times 3 = 24$ ways.

When determining the number of different ways a sequence of events can occur, it is necessary to know whether or not repetitions are permitted. The next two examples show the difference between the two situations.

Example: The employees of a company are given a 4-digit identification number. How many different numbers are available if repetitions are permitted?

Solution: There are 10 digits (zero through nine), so each of the four digits can be selected in ten different ways since repetitions are permitted. Hence the total number of identification numbers is $10 \times 10 \times 10 \times 10 = 10\,000$.

Example: The employees of a company are given 4-digit identification numbers; however, repetitions are not allowed. How many different numbers are available?

Solution: In this case, there are 10 ways to select the first digit, 9 ways to select the second digit, 8 ways to select the third digit, and 7 ways to select the fourth digit, so the total number of ways is $10 \times 9 \times 8 \times 7 = 5040$.

16.12.3. *Factorials.* In mathematics there is a notation called *factorial notation*, which uses the exclamation point. Some examples of factorial notation are $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$.

Notice that factorial notation means to start with the number and find its product with all of the whole numbers less than the number and stopping at one. Formally defined,

$$n! = n \times (n-1) \times (n-2) \cdots 3 \times 2 \times 1\,.$$

Factorial notation can be stopped at any time. For example, $6! = 6 \times 5! = 6 \times 5 \times 4 \times 3!$

In order to use the formulas in the rest of the chapter, it is necessary to know how to multiply and divide factorials. In order to multiply factorials, it is necessary to multiply them out and then multiply the products. For example,

$$3! \times 4! = 3 \times 2 \times 1 \times 4 \times 3 \times 2 \times 1 = 144\,.$$

Notice that $3! \times 4! \neq 12!$, since $12! = 479\,001\,600$.

Division of factorials is somewhat tricky. You can always multiply them out and then divide the top number by the bottom number. For example,

$$\frac{8!}{6!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1}\,.$$

or you can cancel out, as shown:

$$\frac{8!}{6!} = \frac{8 \times 7 \times 6!}{6!} = 8 \times 7 = 56\,.$$

You cannot divide factorials directly.

Also $0! = 1$ by definition.

16.12.4. *The Permutation Rules.* The second way to determine the number of outcomes of an event is to use the *permutation rules*. An arrangement of n distinct objects in a specific order is called a *permutation*. For example, if an art dealer had 3 paintings, say A, B, and C, to arrange in a row on a wall, there would be 6 distinct ways to display the paintings. They are ABC, BAC, CAB, ACB, BCA and CBA.

The total number of different ways can be found using the fundamental counting rule. There are 3 ways to select the first object, 2 ways to select the second object, and 1 way to

select the third object. Hence, there are $3 \times 2 \times 1 = 6$ different ways to arrange three objects in a row on a shelf. Another way to solve this kind of problem is to use permutations.

The number of permutations of n objects using all the objects is $n!$.

Example: In how many different ways can 6 people be arranged in a row for a photograph?

Solution: This is a permutation of 6 objects. Hence $6! = 720$ ways.

In the previous example, all the objects were used; however, in many situations only some of the objects are used. In this case, the *permutation rule* can be used.

The arrangement of n objects in a specific order using r objects at a time is called a permutation of n objects taking r objects at a time. It is written as $pr[n, k]$nPr and the formula is

$$pr[n, k] = \frac{n!}{(n-k)!}\,.$$

Example: In how many different ways can 3 people be arranged in a row for a photograph if they are selected from a group of 5 people?

Solution: Since 3 people are being selected from 5 people and arranged in a specific order, $n = 5$, $r = 3$. Hence, there are

$$pr[5, 3] = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2!}{2!} = 5 \times 4 \times 3 = 60\,\text{ways}\,.$$

Example: How many different signals can be made from seven different flags if four flags are displayed in a row?

Solution: Hence, $n = 7$ and $r = 4$, so $pr[7, 4] = 840$.

In the preceding examples, all the objects were different, but when some of the objects are identical, the second permutation rule can be used.

The number of permutations of n objects when r_1 objects are identical, r_2 objects are identical, *etc.* is

$$\frac{n!}{r_1!r_2!\cdots r_p!}\,,$$

where $r_1 + r_2 + \cdots + r_p = n$.

Example: How many different permutations can be made from the letters of the word Mississippi?

Solution: There are 4s, 4i, 2p, and 1m; hence, $n = 11$, $r_1 = 4$, $r_2 = 4$, $r_3 = 2$ and $r_4 = 1$. Then,

$$\frac{11!}{4! \times 4! \times 2! \times 1} = 34\,650\,.$$

16.12.5. *Combinations.* Sometimes when selecting objects, the order in which the objects are selected is *not* important. For example, when five cards are dealt in a poker game, the order in which you receive the cards is not important. When 5 balls are selected in a lottery, the order in which they are selected is not important. These situations differ from permutations in which order is important and are called combinations. A *combination* is a selection of objects without regard to the order in which they are selected.

The combination rule is used to find the number of ways to select objects without regard to order.

The number of ways of selecting r objects from n objects without regard to order is

$$c[n, r] = \binom{n}{r} = \frac{n!}{(n-r)!r!}\,,$$

where $\binom{n}{r}$ is the binomial coefficient.

Example: In a classroom, there are 8 women and 5 men. A committee of 3 women and 2 men is to be formed for a project. How many different possibilities are there?

Solution: In this case, you must select 3 women from 8 women and 2 men from 5 men. Since the word *and* is used, multiply the answers:

$$\binom{8}{3}\binom{5}{2} = 56 \times 10 = 560\,.$$

16.12.6. *Probability and the Counting Rules.* A wide variety of probability problems can be solved using the counting rules and the probability rule.

Example: Find the probability of getting a flush (including a straight flush) when 5 cards are dealt from a deck of 52 cards.

Solution: A flush consists of 5 cards of the same suit. That is, either 5 clubs or 5 spades or 5 hearts or 5 diamonds, and includes straight flushes. Since there are 13 cards in a suit, there are $\binom{13}{5}$ ways to get a flush in one suit, and there are 4 suits, so the number of ways to get a flush is

$$4\binom{13}{5} = 5148\,.$$

There are $\binom{52}{5}$ ways to select 5 cards.

$$\binom{52}{5} = 2598960\,.$$

Therefore, the probability of getting a flush is

$$p[\text{flush}] = \frac{5148}{2598960} \sim 0.00198\,,$$

which is about one chance in 500.

16.12.7. *Summary.* In order to determine the number of outcomes of events, the fundamental counting rule, the permutation rules, and the combination rule can be used. The difference between a permutation and a combination is that for a permutation, the order or arrangement of the objects is important. For example, order is important in phone numbers, identification tags, social security numbers, license plates, *etc*. Order is not important when selecting objects from a group. Many probability problems can be solved by using the counting rules to determine the number of outcomes of the events that are used in the problems.

16.13. **The Binomial Distribution.**

16.13.1. *Introduction.* Many probability problems involve assigning probabilities to the outcomes of a probability experiment. These probabilities and the corresponding outcomes make up a *probability distribution*. There are many different probability distributions. One special probability distribution is called the *binomial distribution*. The binomial distribution has many uses such as in gambling, in inspecting parts, and in other areas.

16.13.2. *Discrete Probability Distributions.* In mathematics, a *variable* can assume different values. For example, if one records the temperature outside every hour for a 24-hour period, temperature is considered a variable since it assumes different values. Variables whose values are due to chance are called random variables. When a die is rolled, the value of the spots on the face up occurs by chance; hence, the number of spots on the face up on the die is considered to be a *random variable*. The outcomes of a die are 1, 2, 3, 4, 5, and 6, and the probability of each outcome occurring is 1/6. The outcomes and their corresponding probabilities can be written in a table, as shown, and make up what is called a *probability distribution*.

- value, x: 1, 2, 3, 4, 5, 6.
- probability, $p[x]$: 1/6, 1/6, 1/6, 1/6, 1/6, 1/6.

A *probability distribution* consists of the values of a random variable and their corresponding probabilities.

There are two kinds of probability distributions. They are *discrete* and *continuous*. A *discrete* variable has a countable number of values (countable means values of zero, one, two, three, *etc*.). For example, when four coins are tossed, the outcomes for the number of heads obtained are zero, one, two, three, and four. When a single die is rolled, the outcomes are one, two, three, four, five, and six. These are examples of discrete variables.

A *continuous* variable has an infinite number of values between any two values. Continuous variables are measured. For example, temperature is a continuous variable since the variable can assume any value between 108 and 208 or any other two temperatures or values for that matter. Height and weight are continuous variables. Of course, we are

limited by our measuring devices and values of continuous variables are usually "rounded off".

Example: Construct a discrete probability distribution for the number of heads when three coins are tossed.

Solution: Recall that the sample space for tossing three coins is TTT, TTH, THT, HTT, HHT, HTH, THH, and HHH.

The outcomes can be arranged according to the number of heads, as shown.

- 0 heads TTT
- 1 head TTH, THT, HTT
- 2 heads THH, HTH, HHT
- 3 heads HHH

Finally, the outcomes and corresponding probabilities can be written in a table, as shown.

- Outcome, x: 0, 1, 2, 3;
- Probability, $p[x]$: 1/8, 3/8, 3/8, 1/8

The sum of the probabilities of a probability distribution must be 1.

A discrete probability distribution can also be shown graphically by labeling the x axis with the values of the outcomes and letting the values on the y axis represent the probabilities for the outcomes. The graph for the discrete probability distribution of the number of heads occurring when three coins are tossed is shown in Figure.

There are many kinds of discrete probability distributions; however, the distribution of the number of heads when three coins are tossed is a special kind of distribution called a *binomial distribution*.

A binomial distribution is obtained from a probability experiment called a *binomial experiment*. The experiment must satisfy these conditions:

(1) Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. The outcomes are usually considered as a success or a failure.
(2) There is a fixed number of trials.
(3) The outcomes of each trial are independent of each other.
(4) The probability of a success must remain the same for each trial.

Now consider rolling a die. Since there are six outcomes, it cannot be considered a binomial experiment. However, it can be made into a binomial experiment by considering the outcome of getting five spots (for example) a success and every other outcome a failure.

In order to determine the probability of a success for a single trial of a probability experiment, the following formula can be used.

$$c[n, x]\, p^x (1-p)^{n-x}\,,$$

where n are the total number of trials, x the number of successes $(1, 2, 3, \ldots, n)$, p the probability of a success.

The formula has three parts: $c[n, x]$ determines the number of ways a success can occur, $(p)^x$ is the probability of getting x successes and $(1-p)^{n-x}$ is the probability of getting $n - x$ failures.

Example: A coin is tossed 3 times. Find the probability of getting two heads and a tail in any given order.

Solution: Since the coin is tossed 3 times, $n = 3$. The probability of getting a head (success) is 1/2, so $p = 1/2$ and the probability of getting a tail (failure) is $1 - 1/2 = 1/2$; $x = 2$ since the problem asks for 2 heads. $(n - x) = 3 - 2 = 1$.

Hence,

$$p[2 \text{ heads}] = c[3, 2]\,(1/2)^2(1/2) = 3(1/4)(1/2) = 3/8\,.$$

Notice that there were $c[3, 2]$ or 3 ways to get two heads and a tail. The answer 3/8 is also the same as the answer obtained using classical probability that was shown in the first example in this chapter.

In order to construct a probability distribution, the following formula is used:

$$c[n, x]\, p^x (1-p)^{n-x}\,,$$

where $x = 1, 2, 3, \ldots, n$.

The next example shows how to use the formula.

Example: A die is rolled 3 times. Construct a probability distribution for the number of fives that will occur.

Solution: In this case, the die is tossed 3 times, so $n = 3$. The probability of getting a 5 on a die is $1/6$, and one can get $x = 0, 1, 2$ or 3 fives.

- For $x = 0$, $c[3, 0]\,(1/6)^0(5/6)^3 = 0.5787$.
- For $x = 1$, $c[3, 1]\,(1/6)^1(5/6)^2 = 0.3472$.
- For $x = 2$, $c[3, 2]\,(1/6)^2(5/6)^1 = 0.0694$.
- For $x = 3$, $c[3, 3]\,(1/6)^3(5/6)^0 = 0.0046$.

Hence, the probability distribution is

- Number of fives, x: 0, 1, 2, 3.
- Probability, $p[x]$: 0.5787, 0.3472, 0.0694, 0.0046.

Note: Most statistics books have tables that can be used to compute probabilities for binomial variables.

16.13.3. *The Mean and Standard Deviation for a Binomial Distribution.* Suppose you roll a die many times and record the number of threes you obtain. Is it possible to predict ahead of time the average number of threes you will obtain? The answer is "Yes". It is called *expected value* or the *mean* of a binomial distribution. This mean can be found by using the formula mean $\langle\mu\rangle = np$, where n is the number of times the experiment is repeated and p is the probability of a success. The symbol for the mean is the Greek letter μ, (mu).

Example: A die is tossed 180 times and the number of threes obtained is recorded. Find the mean or expected number of threes.

Solution: $n = 180$ and $p = 1/6$, since there is one chance in 6 to get a three on each roll. Then, $\mu = np = 180(1/6) = 30$. Hence, one would expect on average 30 threes.

Statisticians are not only interested in the average of the outcomes of a probability experiment but also in how the results of a probability experiment vary from trial to trial. Suppose we roll a die 180 times and record the number of threes obtained. We know that we would expect to get about 30 threes. Now what if the experiment was repeated again and again? In this case, the number of threes obtained each time would not always be 30 but would vary about the mean of 30. For example, we might get 28 threes one time and 34 threes the next time, etc. How can this variability be explained? Statisticians use a measure called the *standard deviation*. When the standard deviation of a variable is large, the individual values of the variable are spread out from the mean of the distribution. When the standard deviation of a variable is small, the individual values of the variable are close to the mean.

The formula for the standard deviation for a binomial distribution is $\sigma = \sqrt{np(1-p)}$. The symbol for the standard deviation is the Greek letter σ (sigma).

Example: A die is rolled 180 times. Find the standard deviation of the number of threes.

Solution: $n = 180$, $p = 1/6$, $1 - p = 1 - 1/6 = 5/6$. Thus,

$$\sigma = \sqrt{np(1-p)} = \sqrt{(180)(1/6)(5/6)} = 5\,.$$

The standard deviation is 5. Now what does this tell us?

Roughly speaking, most of the values fall within two standard deviations of the mean:

$$\mu - 2\sigma < \text{most values} < \mu + 2\sigma\,.$$

In the die example, we can expect most values will fall between $20 <$ most values < 40. In this case, if we did the experiment many times we would expect between 20 and 40 threes most of the time. This is an approximate *range of values*. Suppose we rolled a die 180 times and we got only 5 threes, what can be said? It can be said that this is an unusually small number of threes. It can happen by chance, but not very often. We might want to consider some other possibilities. Perhaps the die is loaded or perhaps the die has been manipulated by the person rolling it!

16.14. **Other Probability Distributions.**

16.14.1. *Introduction.* The last chapter explained the concepts of the binomial distribution. There are many other types of commonly used discrete distributions. A few are the multinomial distribution, the hypergeometric distribution, the Poisson distribution, and the geometric distribution. This chapter briefly explains the basic concepts of these distributions.

16.14.2. *The Multinomial Distribution.* Recall that for a probability experiment to be binomial, two outcomes are necessary. But if each trial of a probability experiment has more than two outcomes, a distribution that can be used to describe the experiment is called a *multinomial distribution*. In addition, there must be a fixed number of independent trials, and the probability for each success must remain the same for each trial.

A short version of the multinomial formula for three outcomes is given next. If X consists of events E_1, E_2 and E_3, which have corresponding probabilities of p_1, p_2 and p_3 of occurring, where x_1 is the number of times E_1 will occur, x_2 is the number of times E_2 will occur and x_3 is the number of times E_3 will occur, then the probability of X is

$$\frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}\ ,$$

where $x_1 + x_2 + x_3 = n$ and $p_1 + p_2 + p_3 = 1$.

17. Kinetic Theory

The *kinetic theory of gases* describes a gas as a large number of small particles (atoms or molecules), all of which are in constant, random motion. The rapidly moving particles constantly collide with each other and with the walls of the container. Kinetic theory explains macroscopic properties of gases, such as pressure, temperature, and volume, by considering their molecular composition and motion. Essentially, the theory posits that pressure is due not to static repulsion between molecules, as was Isaac Newton's conjecture, but due to collisions between molecules moving at different velocities through Brownian motion.

While the particles making up a gas are too small to be visible, the jittering motion of pollen grains or dust particles which can be seen under a microscope, known as Brownian motion, results directly from collisions between the particle and gas molecules. As pointed out by Albert Einstein in 1905, this experimental evidence for kinetic theory is generally seen as having confirmed the existence of atoms and molecules.

17.1. **Postulates.** The theory for ideal gases makes the following assumptions:

- The gas consists of very small particles known as molecules. This smallness of their size is such that the total volume of the individual gas molecules added up is negligible compared to the volume of the smallest open ball containing all the molecules. This is equivalent to stating that the average distance separating the gas particles is large compared to their size.
- These particles have the same mass.
- The number of molecules is so large that statistical treatment can be applied.
- These molecules are in constant, random, and rapid motion.
- The rapidly moving particles constantly collide among themselves and with the walls of the container. All these collisions are perfectly elastic. This means, the molecules are considered to be perfectly spherical in shape, and elastic in nature.
- Except during collisions, the interactions among molecules are negligible. (That is, they exert no forces on one another.)
- This implies:
  (1) Relativistic effects are negligible.
  (2) Quantum-mechanical effects are negligible. This means that the inter-particle distance is much larger than the thermal de Broglie wavelength and the molecules are treated as classical objects.
  (3) Because of the above two, their dynamics can be treated classically. This means, the equations of motion of the molecules are time-reversible.
- The average kinetic energy of the gas particles depends only on the temperature of the system.
- The time during collision of molecule with the container's wall is negligible as compared to the time between successive collisions.
- Because they have mass, the gas molecules will be affected by gravity.

More modern developments relax these assumptions and are based on the Boltzmann equation. These can accurately describe the properties of dense gases, because they include the volume of the molecules. The necessary assumptions are the absence of quantum effects, molecular chaos and small gradients in bulk properties. Expansions to higher orders in the density are known as virial expansions. The definitive work is the book by Chapman and Enskog but there have been many modern developments and there is an alternative approach developed by Grad based on moment expansions. In the other limit, for extremely rarefied gases, the gradients in bulk properties are not small compared to the mean free paths. This is known as the Knudsen regime and expansions can be performed in the Knudsen number.

17.2. **Properties.**

17.2.1. *Pressure and Kinetic Energy.* Pressure is explained by kinetic theory as arising from the force exerted by molecules or atoms impacting on the walls of a container. Consider a gas of N molecules, each of mass m, enclosed in a cuboidal container of volume $V = L^3$. When a gas molecule collides with the wall of the container perpendicular to the

x coordinate axis and bounces off in the opposite direction with the same speed (an elastic collision), then the momentum lost by the particle and gained by the wall is:

$$\Delta p = p_{\mathrm{i},x} - p_{\mathrm{f},x} = p_{\mathrm{i},x} - (-p_{\mathrm{i},x}) = 2p_{\mathrm{i},x} = 2mv_x\,.$$

where v_x is the x-component of the initial velocity of the particle.

The particle impacts *one specific side wall* once every $\Delta t = 2L/v_x$, where L is the distance between opposite walls. Then, the force due to this particle is $f = \Delta p/\Delta t = mv_x^2/L$. Therefore, the total force on the wall is $F = Nm\left\langle v_x^2\right\rangle/L$, where $\langle\ldots\rangle$ denotes an average over the N particles. Since the assumption is that the particles move in random directions, we will have to conclude that if we divide the velocity vectors of all particles in three mutually perpendicular directions, the average value along each direction must be same [16]: $\left\langle v_x^2\right\rangle = \left\langle v^2\right\rangle/3$.

We can thus rewrite the force as $F = Nm\left\langle v^2\right\rangle/3L$. This force is exerted on an area L^2. Therefore, the pressure of the gas is $P = f/L^2$ or

$$P = \frac{1}{3}N\frac{m\left\langle v^2\right\rangle}{V}\,, \tag{17.1}$$

where $V = L^3$ is the volume of the box. The fraction $n = N/V$ is the number density of the gas (the mass density $\rho = nm$ is less convenient for theoretical derivations on atomic level). (Note that $\dim n = \left[molecule/L^3\right]$.) Using n, we can rewrite the pressure as

$$P = \frac{1}{3}nm\left\langle v^2\right\rangle\,.$$

This is a first non-trivial result of the kinetic theory because it relates pressure, a macroscopic property, to the average (translational) kinetic energy [17] per molecule $m\left\langle v^2\right\rangle/2$, which is a microscopic property.

17.2.2. *Temperature and Kinetic Energy.* From the ideal gas law

$$PV = Nk_{\mathrm{b}}T\,, \tag{17.2}$$

where k_{b} is the Boltzmann constant and T the absolute (thermodynamic) temperature, and from eq. (17.1), we have $PV = Nm\left\langle v^2\right\rangle/3$, and thus $Nk_{\mathrm{b}}T = Nm\left\langle v^2\right\rangle/3$. Therefore, the temperature takes the form

$$T = \frac{1}{3}\frac{m\left\langle v^2\right\rangle}{k_{\mathrm{b}}}\,, \tag{17.3}$$

which leads to the expression of the kinetic energy of a molecule: $(1/2)m\left\langle v^2\right\rangle = (3/2)k_{\mathrm{b}}T$. Then, the kinetic energy of the system is N times that of a molecule: $2k = Nm\left\langle v^2\right\rangle$. Temperature thus becomes

$$T = \frac{2}{3}\frac{k}{Nk_{\mathrm{b}}}\,. \tag{17.4}$$

Equation (17.4) is one important result of the kinetic theory:

> The average molecular kinetic energy is proportional to the absolute temperature.

From eq. (17.2) and eq. (17.4), we have

$$PV = \frac{2}{3}k\,. \tag{17.5}$$

Thus, the product of pressure and volume per mole is proportional to the average (translational) molecular kinetic energy.

Equation (17.2) and eq. (17.5) are called the *classical results*, which could also be derived from statistical mechanics.

Since there are $3N$ degrees of freedom in a monoatomic-gas system with N particles, the kinetic energy per degree of freedom per molecule is $k/3N = k_{\mathrm{b}}T/2$.

In the kinetic energy per degree of freedom, the constant of proportionality of temperature is $1/2$ times Boltzmann constant. In addition to this, the temperature will decrease

[16] This does not mean that each particle always travel in 45 degrees to the coordinate axes.

[17] Since $\dim n = \left[molecule/L^3\right]$, then $m\left\langle v^2\right\rangle$ must have dimensions of $[E/molecule]$, if the product $nm\left\langle v^2\right\rangle$ is to have dimensions of pressure.

when the pressure drops to a certain point. This result is related to the equipartition theorem.

As noted in the article on heat capacity, diatomic gases should have 7 degrees of freedom, but the lighter gases act as if they have only 5. Thus the kinetic energy per kelvin (monatomic ideal gas) is per mole: 12.47 J and per molecule: 20.7 yJ = 129 µeV.

At standard temperature (273.15 K), we get that per mole: 3406 J and per molecule: 5.65 zJ = 35.2 meV.

17.2.3. *Collisions with container.* One can calculate the number of atomic or molecular collisions with a wall of a container per unit area per unit time.

Assuming an ideal gas, a derivation results in an equation for total number of collisions per unit time per area:

$$A = \frac{1}{4}\frac{N}{V}v_{\text{avg}} = \frac{n}{4}\sqrt{\frac{8k_{\text{b}}T}{\pi m}}\,.$$

This quantity is also known as the "impingement rate" in vacuum physics.

17.2.4. *Speed of molecules.* From the kinetic energy formula it can be shown that

$$v_{\text{rms}}^2 = \frac{3RT}{\text{molar mass}}\,,$$

with v in m/s, T in kelvins and R is the gas constant. The molar mass is given as kg/mol. The most probable speed is 81.6% of the rms speed and the mean speeds 92.1% (isotropic distribution of speeds).

17.2.5. *Numeric Values of the Constants.*

- Avogadro constant: $6.022\,141\,29 \times 10^{23}\,\text{mol}^{-1}$;
- Boltzmann constant: $1.380\,648\,8 \times 10^{-23}\,\text{J/K}$;
- molar gas constant: $8.314\,462\,1\,\text{J/mol K}$.

18. Modeling – Applied Mathematics

18.1. **Mathematical Modeling: Introductory Remarks.** - Applied mathematics deals with problems arising in the sciences, engineering and social sciences. Starting with a *world* problem, the goal is to give it a mathematical structure, mostly in terms of equations, analyze these equations, set them in a computational framework, and come up with quantitative results on the original problem. A *validation process* should be put in place to evaluate whether the results obtained accurately reflect the original problem.

The task of the applied mathematician may be summarized as follows:

- Consider problems emerging from science, engineering, medicine, social sciences, and, in general, from *real life*.
- Give them a mathematical structure as appropriate, for instance, using the laws of physics (such as balance laws, mass, linear momentum, energy, *etc.*), or make reasonable assumptions motivated by the experiments in question, or by whatever information is available on the problem. Once the model is built, it is very important to examine how it can be *transported* to problems that have emerged from very different situations.
- Apply methods of analysis to study the mathematical model at hand. These methods may relate to differential equations (ordinary, partial, stochastic, and so on), linear algebra, statistics, and so forth. In many occasions, new mathematics have emerged from the process of solving a real life problems. For instance, calculus emerged from the the study of gravity and planetary motion; the Maxwell equations and their analysis resulted from the study of electromagnetic phenomena and its applications.
- Cast the mathematical models in a computer amenable form. In problems formulated as systems of differential equations, this process typically involves discretization of space and time. Such discrete models are then analyzed by numerical methods, that are subsequently processed in a computer.
- Validation and revision of the computer generated data in terms of the original problem, for instance, comparing the results to experimental measurements.

18.1.1. *Examples: Harmonic Oscillator.* The equation of the harmonic oscillator shown in figure is

$$m\ddot{x} + kx = 0\,.$$

The general solution is

$$x[t] = A\cos\left[\sqrt{\frac{k}{m}}\right]t + B\sin\left[\sqrt{\frac{k}{m}}\right]t\,,$$

where A and B are constants that depend on the prescribed initial data.

The equation for the harmonic oscillator can be generalized to include friction ($c > 0$ denotes the friction coefficient), and also the presence of an external force $F = F[t]$:

$$m\ddot{x} + c\dot{x} + kx = F\,.$$

18.1.2. *Heat Equation.* Let D be a bounded domain in $\mathcal{R}^3$, with smooth boundary ∂D. The equation giving the distribution of temperature $u = u[x,t]$ in D is [18]

$$\rho c\frac{\partial T}{\partial t} = k\nabla^2 T\,,$$

ρ denotes the density of the material, c the specific heat capacity, k the conductivity. The independent variables are space $x \in D$, and time $t \geq 0$. The unknown function $T = T[x,t]$ denotes temperature.

To solve this equation, initial and boundary conditions need to be specified. The latter could be *isothermal* conditions; *i.e.*, the temperature is prescribed on the boundary, or *flux* conditions when the amount of heat going through the boundary is given.

Both, the linear oscillator equation and the heat equation are *linear*.

Topics on ordinary differential equations that we will study:

[18] In compact notation: $\rho c T_{,t} = k\nabla^2 T$.

- Initial value problems for *nonlinear*, second order, and also special higher order equations. Analyze the evolution of the solution with time and its meaning. In the former case, we will study energy methods and the *phase plane*. One of the models that we will analyze is the *nonlinear pendulum* equation. We will also study some equations of third order, such as the *Lorenz* system, and *population models* such as the evolution of HIV.
- In many applications, the equations governing phenomena of interest contain one or more parameters. Consequently, solutions will also depend on such parameters. When there is a *scale* separation among parameters, *perturbation* methods are called for. We will study *regular* and *singular* perturbation methods. Here is a simple example of an ordinary differential equation that can be explicitly solved (rare!!):
$$\frac{\mathrm{d}y}{\mathrm{d}x} = 1 + y^2 \,, \qquad y[0] = 0 \,,$$
We can see that the solution of this initial value problem is
$$y = \tan[x] \,.$$
Can we use this information to solve the modified equations
$$\frac{\mathrm{d}y}{\mathrm{d}x} = 1 + (1+\epsilon)y^2 \,, \qquad |\epsilon| \ll 1 \,?$$
Again, this latter problem has also an exact solution. How does the solution depend on the parameter ϵ? How do we solve problems for which there is no exact solution? The answer is provided by *perturbation methods*.
- Perturbation methods and stability, such as the normal mode analysis, and eigenvalue problems.
- Boundary value problems and bifurcation.

The heat equation is a statement of *balance of energy*. Balance equations are very important in physics. We will present a derivation of the heat equation in terms of balance of energy.

The heat equation is also associated with *diffusive processes* (*e.g.*, as when salt is dissolved in water). From this point of view, the equation is associated with *stochastic* phenomenon. We will also study the heat equation in such a context.

Prior to developing mathematical methods to solve certain problems, we will explore information that can be obtained on a problem from purely common sense.

18.2. **Dimensional Analysis and Scaling Laws.**

18.2.1. *Drag Force.* Let us discuss the following example. When we ride a bike, we notice that the *force of air resistance*, *aka* drag force, is positively related to the speed and to the cross-sectional area (skinny versus broad rider). We want to find an equation that relates the force F_d with the velocity v and the area A.

We could write a prototype equation such as $F_\mathrm{d} = f[A, v]$. However, since the force involves mass, the equation cannot depend on v and A only. So, let us write a new prototype equation:
$$F_\mathrm{d} = f[\rho_\mathrm{a}, A, v] \,,$$
where ρ_a denotes air density, and with f the relation to be determined. To find f, perform dimensional analysis.

To begin with, write the dimensions of the quantities in the MLT system:
$$\dim F_\mathrm{d} = \left[ML/T^2\right] \,, \quad \dim \rho_\mathrm{a} = \left[M/L^3\right] \,, \quad \dim A = \left[L^2\right] \quad \text{and} \quad \dim v = [L/T] \,.$$

We have a system with four physical quantities and three dimensions. According to the Buckingham theorem, one dimensionless quantity Π is enough to find f: $\Pi = f[F_\mathrm{d}, \rho_\mathrm{a}, A, v]$.

The simplest dimensionful combination of the four quantities is $F_\mathrm{d}/\rho_\mathrm{a} A v^2$. Then, the prototype equation becomes $\Pi = F_\mathrm{d}/\rho_\mathrm{a} A v^2$. This, finally, yields the equation of the drag force:
$$F_\mathrm{d} = \Pi \rho_\mathrm{a} A v^2 \,.$$

Remark: Understanding scaling can help us to build small scale models of large phenomenon, such as wind tunnels to model airplanes.

18.2.2. *The yield of a nuclear explosion by G.I. Taylor.* G.I.Taylor (1940's, Cambridge University) computed the energy yield of the first atomic explosion (New Mexico, 1945) after viewing the photographs of the spread of the fireball. He assumed that there exists a physical law of the form

$$g[t, r, \rho, E] = 0\,.$$

Here

- r denotes the radius of the front at time t,
- ρ is the initial air density,
- E is the energy released by the explosion.

We first ask how many dimensionless groups we can form with the quantities $\{t, r, \rho, E\}$? We find that

$$\frac{r^5\rho}{t^2E}$$

is dimensionless and that there are no other independent dimensionless quantities that we can form with $\{t, r, \rho, E\}$.

By the Pi-Theorem (*any physical law has a dimensionless form*), we rewrite the original equation as

$$f\left[\frac{r^5\rho}{t^2E}\right] = 0\,,$$

that is, f is a function of a single variable. Note that the solution corresponds to a root Π (constant) of the previous equation. So,

$$\Pi = \frac{r^5\rho}{t^2E}\,,$$

which implies that

$$r = \left(\Pi\frac{Et^2}{\rho}\right)^{1/5}\,.$$

This last relation is known as a *scaling law* and it states how the radius of the fireball grows with time: $r \propto t^{2/5}$. This is confirmed by experiments and photographs.

18.3. **Mass Balance.** A *mass balance*, also called a material balance, is an application of conservation of mass to the analysis of physical systems. By accounting for material entering and leaving a system, mass flows can be identified which might have been unknown, or difficult to measure without this technique. The exact conservation law used in the analysis of the system depends on the context of the problem but all revolve around mass conservation, i.e. that matter cannot disappear or be created spontaneously.

Therefore, mass balances are used widely in engineering and environmental analyses. For example, mass balance theory is used to design chemical reactors, analyze alternative processes to produce chemicals as well as in pollution dispersion models and other models of physical systems. Closely related and complementary analysis techniques include the population balance, energy balance and the somewhat more complex entropy balance. These techniques are required for thorough design and analysis of systems such as the refrigeration cycle.

In environmental monitoring the term budget calculations is used to describe mass balance equations where they are used to evaluate the monitoring data (comparing input and output, *etc.*). In biology the dynamic energy budget theory for metabolic organization makes explicit use of mass and energy balances.

18.3.1. *Introduction.* The general form quoted for a mass balance is

> the mass that enters a system must, by conservation of mass, either leave the system or accumulate within the system.

Mathematically the mass balance for a system *without* a chemical reaction is as follows:

> input = output + accumulation.

Strictly speaking the above equation holds also for systems with chemical reactions if the terms in the balance equation are taken to refer to total mass; *i.e.*, the sum of all the chemical species of the system. In the absence of a chemical reaction the amount of any chemical species flowing in and out will be the same. This gives rise to an equation for each

species in the system. However, if this is not the case then the mass balance equation must be amended to allow for the generation (formation) or depletion (consumption) of each chemical species. Some use one term in this equation to account for chemical reactions, which will be negative for depletion and positive for generation. However, the conventional form of this equation is written to account for both a positive generation term (*i.e.*, product of reaction) and a negative consumption term (the reactants used to produce the products). Although overall one term will account for the total balance on the system, if this balance equation is to be applied to an individual species and then the entire process, both terms are necessary. This modified equation can be used not only for reactive systems, but for population balances such as occur in particle mechanics problems. The equation is given below – note that it simplifies to the earlier equation in the case that the generation term is zero:

$$\text{input} + \text{formation} = \text{output} + \text{accumulation} + \text{consumption}.$$

- In the absence of a nuclear reaction the number of atoms flowing in and out are the same, even in the presence of a chemical reaction.
- To perform a balance the boundaries of the system must be well defined.
- Mass balances can be taken over physical systems at multiple scales.
- Mass balances can be simplified with the assumption of *steady state*, where the accumulation term is zero.

18.3.2. *Illustrative example.* A simple example can illustrate the concept. Consider the situation in which a slurry is flowing into a settling tank to remove the solids in the tank, solids are collected at the bottom by means of a conveyor belt partially submerged in the tank, and water exits via an overflow outlet.

In this example, there are two substances, solids and water. The water-overflow outlet carries an increased concentration of water relative to solids, as compared to the slurry inlet, and the exit of the conveyor belt carries an increased concentration of solids relative to water.

Assumptions

- Steady state.
- Non-reactive system.

Analysis: The slurry inlet composition (by mass) is 50% solid and 50% water, with a mass flow of 100 kg/min. The tank is assumed to be operating at steady state, and as such accumulation is zero, so input and output must be equal for both the solids and water. If we know that the removal efficiency for the slurry tank is 60%, then the water outlet will contain 20 kg/min of solids (40% times 100 kg/min times 50% solids). If we measure the flow-rate of the combined solids and water, and the water outlet is shown to be 60 kg/min, then the amount of water exiting via the conveyor belt is 10 kg/min. This allows us to completely determine how the mass has been distributed in the system with only limited information and using the mass balance relations across the system boundaries.

18.3.3. *Mass feedback (recycle).* Mass balances can be performed across systems which have cyclic flows. In these systems output streams are fed back into the input of a unit, often for further reprocessing.

Such systems are common in grinding circuits, where materials are crushed then sieved to only allow a particular size of particle out of the circuit and the larger particles are returned to the grinder. However recycle flows are by no means restricted to solid mechanics operations, they are used in liquid and gas flows as well. One such example is in cooling towers, where water is pumped through the cooling tower many times, with only a small quantity of water drawn off at each pass (to prevent solids build up) until it has either evaporated or exited with the drawn off water.

The use of the recycle aids in increasing overall conversion of input products, which is useful for low per-pass conversion processes, for example the Haber process.

18.3.4. *Differential mass balances.* A mass balance can also be taken differentially. The concept is the same as for a large mass balance, however it is performed in the context of a limiting system (*e.g.*, one can consider the limiting case in time or, more commonly,

volume). The use of a differential mass balance is to generate differential equations that can be used to provide an understanding and effective modeling tool for the target system.

The differential mass balance is usually solved in two steps, firstly a set of governing differential equations must be obtained, and then these equations must be solved, either analytically or, for less tractable problems, numerically.

A good example of the applications of differential mass balance are shown in the following systems:

- Ideal (stirred) Batch reactor.
- Ideal tank reactor, also named Continuous Stirred Tank Reactor (CSTR).
- Ideal Plug Flow Reactor (PFR).

Ideal batch reactor: the ideal completely mixed batch reactor is a closed system. Isothermal conditions are assumed, and mixing prevents concentration gradients as reactant concentrations decrease and product concentrations increase over time. Many chemistry textbooks implicitly assume that the studied system can be described as a batch reactor when they write about reaction kinetics and chemical equilibrium. The mass balance for a substance A becomes

$$\text{in} + \text{form.} = \text{out} + \text{acc.} \implies 0 + r_\text{A} V = 0 + \frac{\mathrm{d}n_\text{A}}{\mathrm{d}t}\,,$$

where r_A denotes the rate at which substance A is produced, V is the volume (which may be constant or not), n_A the chemical amount (n) of substance A.

In a fed-batch reactor some reactants/ingredients are added continuously or in pulses (compare making porridge by either first blending all ingredients and the let it boil, which can be described as a batch reactor, or by first mixing only water and salt and making that boil before the other ingredients are added, which can be described as a fed-batch reactor). Mass balances for fed-batch reactors become a bit more complicated.

Reactive system: In this example we will use the law of mass action to derive the expression for a chemical equilibrium constant.

Assume we have a closed reactor in which the following liquid phase reversible reaction occurs:

$$\mathrm{aA} + \mathrm{bB} \longleftrightarrow \mathrm{cC} + \mathrm{dD}\,.$$

The mass balance for substance A becomes

$$\text{in} + \text{form.} = \text{out} + \text{acc.} \implies 0 + r_\text{A} V = 0 + \dot{n}_\text{A}\,.$$

As we have a liquid phase reaction, then we can (usually) assume a constant volume and, since $n_\text{A} = V c_\text{A}$, where c_A is the concentration of A, therefore we get

$$r_\text{A} V = V \dot{c}_\text{A} \implies r_\text{A} = \dot{c}_\text{A}\,.$$

> In many text books this is given as the "definition of reaction rate" without specifying the implicit assumption that we are talking about reaction rate in a closed system with only one reaction. This is an unfortunate mistake that has confused many students over the years.

According to the law of mass action the forward reaction rate can be written as

$$r_{+1} = k_{+1} \left[\mathrm{A}\right]^a \left[\mathrm{B}\right]^b$$

and the backward reaction rate as

$$r_{-1} = k_{-1} \left[\mathrm{C}\right]^c \left[\mathrm{D}\right]^d\,.$$

The rate at which substance A is produced is thus

$$r_\text{A} = r_{-1} - r_{+1}\,.$$

and since, at equilibrium, the concentration of A is constant we get

$$r_\text{A} = r_{-1} - r_{+1} = \dot{c}_\text{A} = 0$$

or, rearranged

$$\frac{k_{+1}}{k_{-1}} = \frac{\left[\mathrm{C}\right]^c \left[\mathrm{D}\right]^d}{\left[\mathrm{A}\right]^a \left[\mathrm{B}\right]^b} = K_\text{eq}\,.$$

Ideal tank reactor/continuously stirred tank reactor: the continuously mixed tank reactor is an open system with an influent stream of reactants and an effluent stream of

products. A lake can be regarded as a tank reactor and lakes with long turnover times (*e.g.*, with a low flux to volume ratio) can for many purposes be regarded as continuously stirred (*e.g.*, homogeneous in all respects). The mass balance becomes

$$\text{in} + \text{form.} = \text{out} + \text{acc.} \implies q[0]\, c_{\mathrm{A}}[0] + r_{\mathrm{A}} V = q c_{\mathrm{A}} + \dot{n}_{\mathrm{A}}\,,$$

where $q[0]$ and q denote the *volumetric flow* in and out of the system and $c_{\mathrm{A}}[0]$ and c_{A} the concentration of A in the inflow and outflow. In an open system we can never reach a chemical equilibrium. We can, however, reach a steady state where all state variables (temperature, concentrations, *etc.*) remain constant (acc. = 0).

Example: Consider a bathtub in which there is some bathing salt dissolved. We now fill in more water, keeping the bottom plug in. What happens?

Since there is no reaction, form. = 0, and, since there is no outflow, $q = 0$. The mass balance becomes

$$\text{in} + \text{form.} = \text{out} + \text{acc.} \implies q[0]\, c_{\mathrm{A}}[0] + 0 = 0 c_{\mathrm{A}} + \dot{n}_{\mathrm{A}}$$

or

$$q[0]\, c_{\mathrm{A}}[0] = \frac{\mathrm{d} c_{\mathrm{A}} V}{\mathrm{d}t} = V \dot{c}_{\mathrm{A}} + c_{\mathrm{A}} \dot{V}\,.$$

Using a mass balance for total volume, however, it is evident that $\dot{V} = q[0]$ and that $V = V_{t=0} + q[0]\, t$. Thus we get

$$\dot{c}_{\mathrm{A}} = \frac{q[0]}{V_{t=0} + q[0]\, t} \left(c_{\mathrm{A}}[0] - c_{\mathrm{A}} \right)\,.$$

Note that there is no reaction and hence no reaction rate or rate law involved, and yet $\dot{c}_{\mathrm{A}} \neq 0$. We can thus draw the conclusion that reaction rate can not be defined in a general manner using $\dot{c}_{\mathrm{A}}$.

> One *must* first write down a mass balance before a link between $\dot{c}_{\mathrm{A}}$ and the reaction rate can be found.

Many textbooks, however, define reaction rate as $v = \dot{c}_{\mathrm{A}}$, *without* mentioning that this definition implicitly assumes that the system is closed, has a constant volume and that there is only one reaction.

Ideal plug flow reactor (PFR): The idealized plug flow reactor is an open system resembling a tube with no mixing in the direction of flow but perfect mixing perpendicular to the direction of flow. Often used for systems like rivers and water pipes if the flow is turbulent. When a mass balance is made for a tube, one first considers an infinitesimal part of the tube and make a mass balance over that using the ideal tank reactor model. That mass balance is then integrated over the entire reactor volume to obtain:

$$\frac{\mathrm{d}(q c_{\mathrm{A}})}{\mathrm{d}V} = r_{\mathrm{A}}\,.$$

In numeric solutions, *e.g.*, when using computers, the ideal tube is often translated to a series of tank reactors, as it can be shown that a PFR is equivalent to an infinite number of stirred tanks in series, but the latter is often easier to analyze, especially at steady state.

18.4. **Models Derived from Balance Laws.**

18.4.1. *Mass and Energy Conservation.* The conservation equations are derived using two basic principles:

- the conservation laws and
- the constitutive relations.

The conservation laws are based on the law of conservation of mass, which states that mass is conserved, and the Newton's law for the conservation of momentum, which states that the rate of change of momentum is equal to the sum of the applied forces. However, there is a complication when these are applied to flow systems, because fluids are transported with the mean flow, and so it is necessary to apply the conservation principles in a reference frame moving with the fluid. Therefore, the time derivatives used in the conservation equations have to be defined a little more carefully. So we will first consider the concept of 'substantial derivatives' before we proceed to deriving the conservation equations.

Substantial derivatives will be illustrated using a position dependent concentration field as an example.

Partial derivative: The partial time derivative of the concentration is the rate of change of concentration c at a *fixed* location in space. Fix the location of observation, and determine the change in the concentration with time at this position. If the concentration at the position x at time t is $c[t, x[t]]$ and the concentration at position x at time $t + \Delta t$ is $c[t + \Delta t, x[t]]$, the 'partial derivative' is written as

$$\frac{\partial c}{\partial t}[t, x[t]] = \lim_{\Delta t \to 0} \frac{1}{\Delta t} c[x[t]\,, t + \Delta t] - c[t, x[t]] \ .$$

18.4.2. *Substantial Derivative.* Though the partial derivative is defined as the change in the value of the concentration at a point in the fluid, this does not reflect the change in the concentration in material volumes, because these material volumes are convected with the flow. Therefore, the volume of fluid which was located at $\left[x^1, x^2, x^3\right]$ at time t would have moved to a new position $\left[x^1 + \dot{x}^1 \Delta t, x^2 + \dot{x}^2 \Delta t, x^3 + \dot{x}^3 \Delta t\right]$ at time $t + \Delta t$.

> The substantial derivative determines the change in concentration on material volumes that are moving with the fluid.

In a three dimensional flow, there are three components of the velocity field $v = \dot{x}^k = v^k$, and the substantial derivative contains terms due to each of these three components $\{v_1, v_2, v_3\}$:

$$\mathrm{D}_t c = v \cdot \nabla c = v \cdot \operatorname{grad} c \,.$$

We mentioned that some mathematical models, especially those coming from mechanics, can be formulated in terms of balance laws. The next example presents a statement of balance of energy leading to the heat equation.

18.4.3. *Equation of Balance of Energy.* We derive an equation governing the flow of heat in a homogeneous, isotropic and continuous solid. This picture represents a bounded domain $\mathcal{D} \subset \mathcal{R}^3$, with smooth boundary, $\partial\mathcal{D}$. The vector n denotes the unit outward normal to the boundary, and q represents the heat flux vector. In addition to q, we introduce the energy density $E[x, t]$ (energy per unit volume at a point x and at time t). This energy is associated with random molecular motion. Recall that $q \cdot n$ represents the amount of energy (heat) going out of the domain across the boundary per unit area and per unit time. (So, $-q \cdot n$ is the influx).

The following equation is the statement of balance of energy [19] in the body $\mathcal{D}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{D}} E[x, t] \ \mathrm{d}\mathcal{V} = - \int_{\partial\mathcal{D}} q \cdot n \, \mathrm{d}\mathcal{S} \,,$$

where $\mathcal{V} \subset \mathcal{D}$ is the control volume and $\mathcal{S}$ the outward surface of the control volume.

Differentiating under the integral sign and applying the divergence theorem to the surface integral gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{D}} E[x, t] \ \mathrm{d}\mathcal{V} + \int_{\mathcal{D}} \nabla \cdot q \, \mathrm{d}\mathcal{V} = \int_{\mathcal{D}} \left(\frac{\partial E}{\partial t}[x, t] + \nabla \cdot q \right) \mathrm{d}\mathcal{V} = 0 \,.$$

Note that this statement of balance of energy can be applied to any part of the body $\mathcal{D}$. It, then, follows that the integrand is identically zero. (Here we assume that the integrand is continuous, in which case, the localization theorem applies). Hence,

$$\frac{\partial E}{\partial t}[x, t] + \nabla \cdot q = 0 \,.$$

(Note: a bit of dim analysis is in rigor now: $\dim E_{,t} = [E/TV]$, since $\dim E = [E/V]$, and $\dim \nabla \cdot q = [1/L]\,[E/AT] = [E/TV]$. Hence, the equation is dimensionally homogeneous.)

We observe that this equation has more unknowns than variables. So, we need to specify constitutive equations, that is, relations between E and q so as to get a single unknown field.

[19] Rate of input energy plus rate of energy release equals rate of output energy plus rate of energy accumulation within the body and plus rate of energy consumption. In the present case, only rate of input energy and of energy accumulation are considered.

Constitutive equations also specify the type of material under consideration. In this case, we assume that

$$E[x,t] = \rho c T[x,t] \qquad \text{and} \qquad q[x,t] = -k\nabla T[x,t]\ .$$

The first equation gives the energy of the body as function of the absolute temperature. This is consistent with temperature as measure of random molecular motion. The second equation is Fourier Law of heat conduction expressing the fact that heat flows from hot to cold. Here,

- $\rho > 0$ denotes the material mass density, and $c > 0$ the specific heat capacity, the amount of heat required to raise the temperature of unit of mass of the material, at temperature T, by one degree,
- $k > 0$ represents the heat conductivity.

So, substituting the previous constitutive relations into the equation of balance of energy (local form), we get the heat equation:

$$\frac{\partial T}{\partial t} = \nabla\cdot(\kappa\nabla T)\,, \quad \text{where} \quad \kappa = \frac{k}{\rho c}\,.$$

(Sanity check: $\dim T_{,t} = [\Theta/T]$ and $\dim \nabla\cdot(\kappa\nabla T) = [1/L]\left[L^2/T\right][1/L][\Theta] = [\Theta/T]$.)

(Nomenclature: since T is a function of the position vector and returns a scalar, then it is known as a scalar field or, more specifically, a temperature field; *i.e.*, it's a function that assigns temperature to every point of $\mathcal{D}$.)

The quantity κ is called the *thermal diffusivity of the material*. Examples of thermal conductivity values in m^2/s:

- water: 1.4×10^{-7};
- air: 2.2×10^{-5};
- gold: 1.27×10^{-4} (best heat conductor).

Finally, if $\mathcal{D}$ represents an homogeneous material, then κ is constant throughout the body. Thus, the heat equation can be written as

$$\frac{\partial T}{\partial t} = \kappa\nabla\cdot\nabla T = \kappa\nabla^2 T = \kappa\nabla^2 T\,,$$

where ∇^2 is called the Laplace operator.

Note that the appearance of material properties such as c, k and κ is a sure sign that we have introduced a constitutive relation, and it should be stressed that these relations between E, q and T are material-dependent and experimentally determined. There is no *a priori* reason for them to have the nice linear form given above, and indeed for some materials one or other may be strongly nonlinear.

18.5. **Yet another derivation of the continuity equation for energy.** Consider a body whose center temperature is greater than its outer surface temperature. Experience states that the energy in the center must flow to the outer surface. The task is then to find the energy distribution inside the body. We do so by applying the conservation of energy principle.

Let e be the internal energy density of the body and let $\mathrm{d}v$ be the volume of a non-moving control volume inside the body of volume v. Since the energy contained in the control volume is $e\,\mathrm{d}v$, then the total internal energy of the body is $\int_v e\,\mathrm{d}v$. Now the rate at which the internal energy decreases is thus

$$-\frac{\mathrm{d}}{\mathrm{d}t}\int_v e\,\mathrm{d}v = -\int_v \frac{\partial e}{\partial t}\,\mathrm{d}v\,,$$

where the dimensions of the last equation are those of energy flow [20], E/T, *aka* thermal power.

On the other hand, the energy flowing out of the control volume through its oriented surface boundary ∂v is $j\cdot n\,\mathrm{d}s$, where j is the energy flux, n a normal vector pointing out the control volume of surface $\mathrm{d}s$. Thus, the total energy flux out of the body is

$$+\int_{\partial v} j\cdot n\,\mathrm{d}s = +\int_v \operatorname{div} j\,\mathrm{d}v\,,$$

[20] We have chosen the energy, length, time and temperature, $ELT\Theta$, dimensional system.

where the dimensions of the last equation are also those of energy flow, E/T.

Since according to the conservation of energy principle, the two energy flows must be equal to one another, we find therefore that

$$-\int_v \frac{\partial e}{\partial t}\,\mathrm{d}v = +\int_v \operatorname{div} j\,\mathrm{d}v \implies \int_v \left(\frac{\partial e}{\partial t} + \operatorname{div} j\right)\,\mathrm{d}v = 0\,.$$

The last equation must hold for the whole of the body, which implies a vanishing integrand, or

$$\frac{\partial e}{\partial t} + \operatorname{div} j = 0\,,$$

whose dimensions are those of energy density flow, E/L^3T.

We need next a way to relate the body internal energy to the energy flowing from the body center to its outer surface. We use two experimentally based relationships, *aka* constitutive equations.

On the one hand, experimental evidence suggests that the internal energy of a body is proportional to its temperature: $e \propto T$ or $e = \rho cT$, where ρ is the body mass density, c a thermal property of the body material called *specific heat capacity*, $\dim c = E/M\Theta$, and T body temperature. The term ρc is called *volumetric heat capacity*, $\dim \rho c = E/L^3\Theta$. Introducing $e = \rho cT$, we find

$$\frac{\partial e}{\partial t} = \frac{\partial \rho cT}{\partial t} = \rho c\frac{\partial T}{\partial t}\,,$$

where, in the last equality, we assumed that the body is homogeneous, so ρ and c are constant.

Fourier, on the other hand, proposed, after experimental analysis, that the energy flux is proportional to the temperature gradient: $j \propto \operatorname{grad} T$ or $j = -k \operatorname{grad} T$, where the negative sign reflects the fact that the flow occurs in the direction of decreasing temperature and where k is a thermal property of the body material called *thermal conductivity*, $\dim k = E/LT\Theta$. Introducing this relation, called Fourier's law, in the place of $\nabla \cdot j$, we have

$$\operatorname{div} j = \operatorname{div}(-k \operatorname{grad} T) = -k \operatorname{lap} T\,,$$

where, in the last equality, it was assumed that the body is homogeneous and isotropic, so that k is constant throughout the body and independent on the flow direction. Besides, Laplace operator, $\operatorname{lap} T = \operatorname{div}\operatorname{grad} T$, was used.

Equating again both fluxes yields

$$\rho c\frac{\partial T}{\partial t} = -k \operatorname{lap} T \implies \frac{\partial T}{\partial t} = -\lambda\nabla^2 T\,,$$

where the body thermal property $\lambda = k/\rho c$ is called *thermal diffusivity*, $\dim \lambda = L^2/T$, and ∇^2 is Laplace operator in terms of the geometric derivative, ∇. Last equation is called *heat equation*.

Finally, the heat equation can be alternatively written using index notation and Einstein summation convention once a coordinate system has been chosen. In the case of Cartesian coordinates: $\partial_t T = -\lambda\partial_{xx} T$ or using the comma derivative notation as $T_{,t} = -\lambda T_{,xx}$. In the case of general curvilinear coordinates, say $\left[\xi^1, \xi^2, \xi^3\right]$, one must replace Laplace operator by

$$\nabla^2 = \nabla\xi^m \cdot \nabla\xi^n \frac{\partial^2}{\partial\xi^m \xi^n} + \nabla^2\xi^m \frac{\partial}{\partial\xi^m}\,,$$

where the summation over repeated indices is implied.

18.6. **Mass continuity equation.** Consider a mass of non-reactive solute placed into the center of a solvent of volume v_s, given a solution volume v. As the solute dissolves into the solvent, the solute concentration varies spatially and temporally in such a way that its mass flows from more concentrated zones to less concentrated ones. This phenomenon can be mathematically described in the same fashion as the case of a hot-center body, but using the mass conservation principle instead of the energy conservation principle.

Using similar arguments that those used in the energy continuity equation derivation, one finds

$$\frac{\partial c_\mathrm{s}}{\partial t} + \operatorname{div} j = 0\,,$$

where c_s is the solute s concentration, solute mass per unit solution volume, j the concentration flux. Note that the equation has dimensions of concentration flow, M/L^3T. (The dimensions of the divergence are the same as the geometric derivative; *i.e.*, $\dim \operatorname{div} = 1/L$.)

Now, we need a way to relate c_s and j; *i.e.*, a constitutive equation. Such a relation is given by Fick's first law of diffusion: experimentation suggests that the concentration flux is proportional to the concentration gradient; that is, $j = -d \operatorname{grad} c_s$, where the minus sign reflects the fact that the flow occurs in the direction of decreasing concentration and where d is a molecular property of the solute called *diffusivity*. Diffusivity is proportional to the squared velocity of the diffusing particles, which, in turns, depends on solvent temperature, solvent viscosity and particle size. Note that $\dim d = L^2/T$. Then, using Fick's law, we find that

$$\frac{\partial c_s}{\partial t} = -\operatorname{div}(d \operatorname{grad} c_s) = -d \operatorname{lap} c_s = -d\nabla^2 c_s \,,$$

where in the last two equations it was assumed that the body is homogeneous and isotropic; *i.e.*, d is constant and independent on the flow direction.

18.7. **Chemical Kinetics.** [A. Cornish-Bowden, Fundamentals of Enzyme Kinetics, Fourth Edition]

18.7.1. *First-order kinetics.* The rate v of a first-order reaction A $\longrightarrow$ P can be expressed as

$$v = \dot{p} = -\dot{a} = ka = k(a_0 - p) \,,$$

in which a and p are the concentrations of A and P respectively at any time t, k is a first-order rate constant and a_0 is a constant. As we shall see throughout this book, the idea of a *rate constant* [21] is fundamental in all varieties of chemical kinetics. The first two equality signs in the equation represent alternative definitions of the rate v: because every molecule of A that is consumed becomes a molecule of P, it makes no difference to the mathematics whether the rate is defined in terms of the appearance of product or disappearance of reactant. It may make a difference experimentally, however, because experiments are not done with perfect accuracy, and in the early stages of a reaction the relative changes in p are much larger than those in a (Figure 1.2). For this reason it will usually be more accurate to measure increases in p than decreases in a.

The third equality sign in the equation is the one that specifies that this is a first-order reaction, because it states that the rate is proportional to the concentration of reactant A.

Finally, if the time zero is defined in such a way that $a = a_0$ and $p = 0$ when $t = 0$, the stoichiometry allows the values of a and p at any time to be related according to the equation $a + p = a_0$, thereby allowing the last equality in the equation.

The last equation can readily be integrated by separating the two variables p and t, bringing all terms in p to the left-hand side and all terms in t to the right-hand side:

$$\int \frac{\mathrm{d}p}{a_0 - p} = \int k\mathrm{d}t \,, \implies -\ln[a_0 - p] = kt + \alpha \,,$$

in which α, the constant of integration, can be evaluated by noting that there is no product at the start of the reaction, so $p = 0$ when $t = 0$. Then $\alpha = -\ln[a_0]$ and so

$$\ln[1 - p/a_0] = -kt \,.$$

Taking exponentials of both sides and rearranging terms, we have

$$p = a_0 \left(1 - \exp[-kt]\right) \,.$$

Notice that the constant of integration α was included in this derivation, evaluated and found to be nonzero. Constants of integration must always be included and evaluated when integrating kinetic equations; they are rarely found to be zero.

Inserting $p = 0.5a$ into the last equation at a time $t = t_{0.5}$ known as the *half-time* allows us to calculate $kt_{0.5} = ln2 = 0.693$, so $t_{0.5} = 0.693/k$. This value is independent of the

[21] Some authors, especially those with a strong background in physics, object to the term "rate constant" (preferring "rate coefficient") for quantities like k in the last equation and for many similar quantities that will occur in this book, on the perfectly valid grounds that they are not constant, because they vary with temperature and with many other conditions. However, the use of the word "constant" to refer to quantities that are constant only under highly restricted conditions is virtually universal in biochemical kinetics (and far from unknown in chemical kinetics), and it is hardly practical to abandon this usage in this book.

value of a_0, so the time required for the concentration of reactant to decrease by half is a constant, for a first-order process, as illustrated in Figure 1.3. The half-time is not a constant for other orders of reaction.

18.7.2. *Second-order kinetics.* The commonest type of bimolecular reaction is one of the form $\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{P} + \mathrm{Q}$, in which two different kinds of molecule A and B react to give products. In this example the rate is likely to be given by a second-order expression of the form

$$v = \dot{p} = kab = k\,(a_0 - p)\,(b_0 - p)\ ,$$

in which k is now a *second-order rate constant* [22]. Again, integration is readily achieved by separating the two variables p and t, with a solution:

$$\frac{a_0\,(b_0 - p)}{b_0\,(a_0 - p)} = \exp[(b_0 - a_0)\,kt]\ .$$

A special case of this result is important: if a_0 is negligible compared with b_0, then $(b_0 - a_0) \sim b_0$; p can never exceed a_0, on account of the stoichiometry of the reaction, and so $(b0 - p) \sim b_0$. Introducing both approximations, the last equation can be simplified as follows:

$$p = a_0\,(1 - \exp[-kb_0t])\ ,$$

which has exactly the same form as the equation for a first-order reaction. This type of reaction is known as a *pseudo-first-order reaction*, and kb_0 is a *pseudo-first-order rate constant*. Pseudo-first-order conditions occur naturally when one of the reactants is the solvent, as in most hydrolysis reactions, but it is also advantageous to create them deliberately, to simplify evaluation of the rate constant.

18.7.3. *Dimensions of rate constants.* Dimensional analysis provides a quick and versatile technique for detecting algebraic mistakes and checking results. It depends on the existence of a few simple rules governing the permissible ways of combining quantities of different dimensions, and on the frequency with which algebraic errors result in dimensionally inconsistent expressions. Concentrations can be expressed in M (or mol/L), and reaction rates in M/s. In an equation that expresses a rate v in terms of a concentration a as $v = ka$, therefore, the rate constant k must be expressed in s^{-1} if the left- and right- hand sides of the equation are to have the same dimensions. All first-order rate constants have the dimensions of time^{-1}, and by a similar argument second-order rate constants have the dimensions of $\mathrm{concentration}^{-1}\,\mathrm{time}^{-1}$ (Figure 1.7), third-order rate constants have the dimensions of $\mathrm{concentration}^{-2}\,\mathrm{time}^{-1}$, and zero-order rate constants have the dimensions of $\mathrm{concentration}\,\mathrm{time}^{-1}$.

18.7.4. *Reversible reactions.* All chemical reactions are reversible in principle, and for many the reverse reaction is readily observable in practice as well, and must be allowed for in the rate equation:

$$\mathrm{A} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P}\,,$$

where the concentration of $\mathrm{A} = a_0 - p$ and that of $\mathrm{P} = p$. In this case,

$$v = \dot{p} = k_1\,(a_0 - p) - k_{-1}p = k_1a_0 - (k_1 + k_{-1})\,p\,.$$

This differential equation is of exactly the same form as equation 1.1, and can be solved in the same way:

$$\int \frac{\mathrm{d}p}{k_1a_0 - (k_1 + k_{-1})\,p} = \int \mathrm{d}t\,,\ \Longrightarrow\ \frac{\ln[k_1a_0 - (k_1 + k_{-1}p)]}{-\,(k_1 + k_{-1})} = t + \alpha\,.$$

Setting $p = 0$ when $t = 0$, gives $\alpha = -\ln[k_1a_0]\,/\,(k_1 + k_{-1})$. Replacing α in the above equation and, after rearranging, we have that

$$p = p_\infty\,(1 - \exp[-\,(k_1 + k_{-1})\,t])\ ,$$

[22] Conventional symbolism does not indicate the order of a rate constant. For example, it is common practice to illustrate simple enzyme kinetics with a mechanism in which k_1 is a second-order rate constant and k_2 is a first-order rate constant: there is no way to know this from the symbols alone, it is important to define each rate constant when it is first used.

where $p_\infty = k_1 a_0 / (k_1 + k_{-1})$. This is the value of p after infinite time, because the exponential term approaches zero as t becomes large. The expected behavior is illustrated in Figure 1.9.

18.7.5. *Reaction rates.* For a general chemical reaction

$$\mathrm{aA} + \mathrm{bB} + \cdots \longrightarrow \mathrm{yY} + \mathrm{zZ} + \cdots ,$$

we define *specific reaction rates* with respect to each reactant or product:

$$r_\Gamma = \pm \frac{1}{\nu} \frac{\mathrm{d}\Gamma}{\mathrm{d}t} ,$$

where ν is the stochiometric coefficient for species Γ in the balanced equation. The $+$ sign is used if Γ is a product, the $-$ sign if it is a reactant. Thus,

$$r_\mathrm{A} = -\frac{1}{a} \frac{\mathrm{d}c_\mathrm{A}}{\mathrm{d}t} .$$

The rate always has units of concentration/time. For solution reactions the usual units are mol/ls, while for gas phase reactions the most common unit is $1/\mathrm{cm}^3\mathrm{s}$.

These specific rates are not necessarily the same for different species. If there are no reaction intermediates of significant concentrations, then

$$r_\mathrm{A} = r_\mathrm{B} = r_\mathrm{Y} = r_\mathrm{Z} = v ,$$

the rate of the reaction. For very many systems, intermediates are important, all the specific rates are different, and it is then necessary to specify which specific rate is being discussed.

18.7.6. *Rate laws.* For most reactions, the rate(s) depend on the concentrations of one or more reactants or products. Then we write

$$r_\Gamma = f\left(c_\mathrm{A}, c_\mathrm{B}, c_\mathrm{Y}, c_\mathrm{I}, c_\mathrm{C}, T, p, \dots\right) ,$$

where the list shows explicitly that r might depend on the concentrations of species other than those in the balanced equation, as well as on temperature T, pressure p, and so on. Often the dependence on variables other than concentrations is suppressed (a set of conditions is implied or specified), so that we write

$$r_\Gamma = f\left(c_\mathrm{A}, c_\mathrm{B}, c_\mathrm{Y}, c_\mathrm{I}, c_\mathrm{C}, \dots\right) .$$

This kind of expression, giving the rate of the reaction as a function of the concentrations of various chemical species, is called a *rate law*. Notice that the rate law is a differential equation: it gives the derivative (with respect to time) of one of the concentrations in terms of all the concentrations. The solution to such a differential equation is a function that gives the concentration of species Γ as a function of time.

19. Environmental Modeling

[environmental modeling - ekkehard holzbecher]

Control volume: a fixed volume that does not change in size in time.

19.1. **Continuity Equation for Mass.** Consider the change of mass during the time Δt within a control volume with spacing Δx, Δy and Δz, each for one direction in $\mathcal{E}^3$. On the one hand, consider the mass within the control volume at the beginning and at the end of the time period and then calculate the difference between the two. On the other hand, balance all fluxes across the boundaries of the volume; *i.e.*, fluxes into the volume have to be taken as positive, while those leaving the volume are negative. In $\mathcal{E}^3$, six faces of the control volume have to be taken into account.

Mass at the beginning and at the end of the period $[t, t+\Delta t]$ is given by [23]

$$\theta c[x,t]\,\Delta x\Delta y\Delta z \qquad \text{and} \qquad \theta c[x,t+\Delta t]\,\Delta x\Delta y\Delta z\,,$$

where θ denotes the share on the total volume. In case of a saturated porous medium, θ denotes porosity. In the unsaturated zone, within a soil, for example, θ is the volumetric water saturation, when the aqueous phase is concerned. In the situation in which two fluids occupy the space (say water and oil), the share of each phase has to be taken into account too. $\Delta x\Delta y\Delta z$ stands for volume and c denotes mass concentration, $\dim c = [M]\,/\,[V]$ (mass per unit volume). The change of mass per unit time is then

$$\theta\frac{c[x,t+\Delta t]-c[x,t]}{\Delta t}\Delta x\Delta y\Delta z\,.$$

Fluxes in x-direction are given across faces of the control volume:

$$\theta j_{x-}[x,t]\,\Delta y\Delta z \qquad \text{and} \qquad \theta j_{x+}[x,t]\,\Delta y\Delta z\,,$$

where j_{x-} denotes mass flux across the left face of the volume, in negative x-direction. Analogously, j_{x+} denotes the mass flux in x-direction across the right face, in positive x-direction. Fluxes may change spatially and temporally which do the brackets indicate. Both fluxes are positive, if they add mass to the control volume and negative otherwise. The physical dimension of mass flux is $\dim j = [M]\,/\,[A.T]$. The term $\theta\Delta y\Delta z$ denotes the area through which flow takes place.

The balance between both flux terms is thus given by

$$\theta\left(j_{x-}[x,t]-j_{x+}[x,t]\right)\Delta y\Delta z\,.$$

For the sake of simplicity, the fluxes across the four other faces are neglected during the derivation at this point*i.e.*, assume here that the flux components in the y- and z-directions are both zero. As previously stated, both formulations measure mass change and thus need be equal:

$$\theta\frac{c[x,t+\Delta t]-c[x,t]}{\Delta t}\Delta x\Delta y\Delta z=\theta\left(j_{x-}[x,t]-j_{x+}[x,t]\right)\Delta y\Delta z\,.$$

Divide the last equation through the volume and porosity to have

$$\frac{c[x,t+\Delta t]-c[x,t]}{\Delta t}=\frac{j_{x-}[x,t]-j_{x+}[x,t]}{\Delta x}\,.$$

From this equation a differential equation can be derived by the transition of the finite grid spacing Δx and time step Δt to infinitesimal expressions; *viz.*., by the limits $\Delta x\to 0$ and $\Delta t\to 0$. It follows

$$\frac{\partial c}{\partial t}[x,t]=-\frac{\partial j_x}{\partial x}[x,t]\;,$$

which is a differential formulation for the principle of mass conservation. The presumption for the differentiation procedure is that the functions c and j_x are smooth, *i.e.*, differentiable. The last equation is valid for one-dimensional transport and s the basis for the mathematical analysis of transport processes. The dimensions of the equation are $[M]\,/\,\left[L^3.T\right]$.

This formulation is also valid if there are no internal mass "sources" or "sinks". Sources and sinks are here understood in the most general sense: each process, which creates or destroys some species, as measured by c, can contribute to such a source or sink.

[23] In this derivation, Euler description of motion is implicitly used: the control volume is not only fixed in size, but also fixed in space, while the fluid passes across it.

To extend the formulation so to consider sources and sinks: if these are described by a source- or sink-rate $q[x,t]$ with dimensions $[M]\,/\,\left[L^3.T\right]$, which may vary spatially and temporally, add the integral term

$$\int_{\Delta x}\int_{\Delta t} q[x,t]\ \mathrm{d}t\mathrm{d}x\,.$$

The term is positive if mass as added (source) and negative if mass is removed (sink). In the derivation of the mass conservation equation, the integral term has to be differentiated, this leads to the general transport equation in one space dimension:

$$\theta\frac{\partial c}{\partial t} = -\frac{\partial \theta j_x}{\partial x} + q\,.$$

Flux components in y- and z-directions can also be taken into account, based on formulae analogous to the formula for the x-direction. The fluxes j_{y-}, j_{y+}, j_{z-} and j_{y+} have to be introduced, balanced and the balances added to the derivation. Take the limits $\Delta y \to 0$ and $\Delta z \to 0$ to obtain

$$\theta\frac{\partial c}{\partial t} = -\nabla\cdot\theta j + q$$

or, equivalently,

$$\theta\partial_t c = \theta c_{,t} = -\operatorname{div} j + q$$

The last equations are geometric objects, so the vector equations are valid in any coordinate system. Besides, they are written in a compact (elegant) form.

The derived equation for mass conservation alone is not yet sufficient for a complete mathematical formulation. There are too many unknown variables, namely concentration and the components of the flux vector j. In order to reduce the number of unknowns, use a formulation that connects concentration and flux, resulting in an equation where the concentration is the only unknown variable.

The advective flux is given by the product of the concentration and the fluid velocity:

$$j = cv = vc\,,$$

since c is a scalar-valued function.

20. Some Lingo about Approximate Solutions

The art of being wise is the art of knowing what to overlook.

— WILLIAM JAMES,

20.1. **Spherical Cow.** Spherical cow is a metaphor for highly simplified scientific models of complex real life phenomena.

The phrase comes from a joke about theoretical physicists:

> Milk production at a dairy farm was low, so the farmer wrote to the local university, asking for help from academia. A multidisciplinary team of professors was assembled, headed by a theoretical physicist, and two weeks of intensive on-site investigation took place. The scholars then returned to the university, notebooks crammed with data, where the task of writing the report was left to the team leader. Shortly thereafter the physicist returned to the farm, saying to the farmer "I have the solution, but it only works in the case of spherical cows in a vacuum".

The point of the joke is that physicists will often reduce a problem to the simplest form they can imagine in order to make calculations more feasible, even though such simplification may hinder the model's application to reality.

20.2. **Fermi Problem.** In science, particularly in physics or engineering education, a Fermi problem, Fermi question, or Fermi estimate is an estimation problem designed to teach dimensional analysis, approximation, and the importance of clearly identifying one's assumptions. Named after physicist Enrico Fermi, such problems typically involve making justified guesses about quantities that seem impossible to compute given limited available information.

Fermi was known for his ability to make good approximate calculations with little or no actual data, hence the name. One example is his estimate of the strength of the atomic bomb detonated at the Trinity test, based on the distance traveled by pieces of paper dropped from his hand during the blast. Fermi's estimate of 10 kilotons of TNT was remarkably close to the now-accepted value of around 20 kilotons, a difference of less than one order of magnitude.

20.2.1. *Examples of Fermi problems.* The classic Fermi problem, generally attributed to Fermi, is "How many piano tuners are there in Chicago?" A typical solution to this problem involves multiplying a series of estimates that yield the correct answer if the estimates are correct. For example, we might make the following assumptions:

- There are approximately 5,000,000 people living in Chicago.
- On average, there are two persons in each household in Chicago.
- Roughly one household in twenty has a piano that is tuned regularly.
- Pianos that are tuned regularly are tuned on average about once per year.
- It takes a piano tuner about two hours to tune a piano, including travel time.
- Each piano tuner works eight hours in a day, five days in a week, and 50 weeks in a year.

From these assumptions, we can compute that the number of piano tunings in a single year in Chicago is:

(5,000,000 persons in Chicago) / (2 persons/household) × (1 piano/20 households) × (1 piano tuning per piano per year) = 125,000 piano tunings per year in Chicago.

We can similarly calculate that the average piano tuner performs:

(50 weeks/year)×(5 days/week)×(8 hours/day)/(2 hours to tune a piano) = 1000 piano tunings per year per piano tuner.

Dividing gives:

(125,000 piano tunings per year in Chicago) / (1000 piano tunings per year per piano tuner) = 125 piano tuners in Chicago.

A famous example of a Fermi-problem-like estimate is the Drake equation, which seeks to estimate the number of intelligent civilizations in the galaxy. The basic question of why,

if there is a significant number of such civilizations, ours has never encountered any others is called the Fermi paradox.

20.2.2. *Advantages and scope.* Scientists often look for Fermi estimates of the answer to a problem before turning to more sophisticated methods to calculate a precise answer. This provides a useful check on the results: where the complexity of a precise calculation might obscure a large error, the simplicity of Fermi calculations makes them far less susceptible to such mistakes. (Performing the Fermi calculation first is preferable because the intermediate estimates might otherwise be biased by knowledge of the calculated answer.)

Fermi estimates are also useful in approaching problems where the optimal choice of calculation method depends on the expected size of the answer. For instance, a Fermi estimate might indicate whether the internal stresses of a structure are low enough that it can be accurately described by linear elasticity; or if the estimate already bears significant relationship in scale relative to some other value, for example, if a structure will be over-engineered to withstand loads several times greater than the estimate.

Although Fermi calculations are often not accurate, as there may be many problems with their assumptions, this sort of analysis does tell us what to look for to get a better answer. For the above example, we might try to find a better estimate of the number of pianos tuned by a piano tuner in a typical day, or look up an accurate number for the population of Chicago. It also gives us a rough estimate that may be good enough for some purposes: if we want to start a store in Chicago that sells piano tuning equipment, and we calculate that we need 10,000 potential customers to stay in business, we can reasonably assume that the above estimate is far enough below 10,000 that we should consider a different business plan (and, with a little more work, we could compute a rough upper bound on the number of piano tuners by considering the most extreme reasonable values that could appear in each of our assumptions).

20.2.3. *Explanation.* Fermi estimates generally work because the estimations of the individual terms are often close to correct, and overestimates and underestimates help cancel each other out. That is, if there is no consistent bias, a Fermi calculation that involves the multiplication of several estimated factors (such as the number of piano tuners in Chicago) will probably be more accurate than might be first supposed.

In detail, multiplying estimates corresponds to adding their logarithms; thus one obtains a sort of Wiener process or random walk on the logarithmic scale, which diffuses as $\sqrt{n}$ (in number of terms n). In discrete terms, the number of overestimates minus underestimates will have a binomial distribution. In continuous terms, if one makes a Fermi estimate of n steps, with standard deviation σ units on the log scale from the actual value, then the overall estimate will have standard deviation $\sigma\sqrt{n}$, since the standard deviation of a sum scales as $\sqrt{n}$ in the number of summands.

For instance, if one makes a 9-step Fermi estimate, at each step overestimating or underestimating the correct number by a factor of 2 (or with a standard deviation 2), then after 9 steps the standard error will have grown by a logarithmic factor of $\sqrt{9} = 3$, so $2^3 = 8$. Thus one will expect to be within 1/8 to 8 times the correct value – within an order of magnitude, and much less than the worst case of erring by a factor of (about 2.7 orders of magnitude). If one has a shorter chain or estimates more accurately, the overall estimate will be correspondingly better.

20.3. **Back-of-the-envelope calculation.** A back-of-the-envelope calculation is a rough calculation, typically jotted down on any available scrap of paper such as the actual back of an envelope. It is more than a guess but less than an accurate calculation or mathematical proof.

The defining characteristic of back-of-the-envelope calculations is the use of simplified assumptions.

A similar phrase is "back of a napkin", which is also used in the business world to describe sketching out a quick, rough idea of a business or product.

20.4. **Sanity testing.** A sanity test or sanity check is a basic test to quickly evaluate whether a claim or the result of a calculation can possibly be true. It is a simple check to see if the produced material is rational (that the material's creator was thinking rationally, applying sanity). The point of a sanity test is to rule out certain classes of obviously false results, not to catch every possible error. A rule-of-thumb may be checked to perform the test. The advantage of a sanity test, over performing a complete or rigorous test, is speed.

In arithmetic, for example, when multiplying by 9, using the divisibility rule for 9 to verify that the sum of digits of the result is divisible by 9 is a sanity test – it will not catch every multiplication error, however it's a quick and simple method to discover many possible errors.

In computer science, a sanity test is a very brief run-through of the functionality of a computer program, system, calculation, or other analysis, to assure that part of the system or methodology works roughly as expected. This is often prior to a more exhaustive round of testing.

When talking about quantities in physics, the claim of a power output of a car cannot be 700 kJ since that is a unit of energy, not power (energy per unit time).

20.5. **Heuristic.** Heuristic refers to experience-based techniques for problem solving, learning, and discovery. Where the exhaustive search is impractical, heuristic methods are used to speed up the process of finding a satisfactory solution; mental shortcuts to ease the cognitive load of making a decision. Examples of this method include using a rule of thumb, an educated guess, an intuitive judgment, or common sense.

The most fundamental heuristic is trial and error, which can be used in everything from matching nuts and bolts to finding the values of variables in algebra problems.

Here are a few other commonly used heuristics, from George Pólya's 1945 book, How to Solve It:

- If you are having difficulty understanding a problem, try drawing a picture.
- If you can't find a solution, try assuming that you have a solution and seeing what you can derive from that ("working backward").
- If the problem is abstract, try examining a concrete example.
- Try solving a more general problem first (the "inventor's paradox": the more ambitious plan may have more chances of success).

In engineering, a heuristic is an experience-based method that can be used as an aid to solve process design problems, varying from size of equipment to operating conditions. By using heuristics, time can be reduced when solving problems. Several methods are available to engineers. These include Failure mode and effects analysis and Fault tree analysis. The former relies on a group of qualified engineers to evaluate problems, rank them in order of importance and then recommend solutions. The methods of forensic engineering are an important source of information for investigating problems, especially by elimination of unlikely causes and using the weakest link principle. Because heuristics are fallible, it is important to understand their limitations. They are aids that facilitate quick estimates and preliminary process designs.

Heuristic (engineering): In engineering, heuristics are experience-based methods used to reduce the need for calculations pertaining to equipment size, performance, or operating conditions. Heuristics are fallible and do not guarantee a correct solution. It is important to understand their limitations when applying them to different equipment and processes. Though heuristics are limited, they may be of value. This is because they offer time saving approximations in preliminary process design.

Problem solving methods are intrinsic to forensic engineering methods, where failures are analyzed for the root cause or causes. Only when failures have been investigated with conclusive results can remedial action be taken with confidence.

Example: Storage Vessels: These heuristics were taken from Turton's "Analysis, Synthesis, and Design of Chemical Processes".

- Use vertical tanks on legs when the tank is less than $3.8\,\mathrm{m}^3$.
- Use horizontal tanks on concrete supports when the tank is between 3.8 and $38\,\mathrm{m}^3$,
- Use vertical tanks on concrete pads when the tank is beyond $38\,\mathrm{m}^3$,

- Liquids subject to breathing losses may be stored in tanks with floating or expansion roofs for conservation.
- Freeboard is 15% below $1.9\,\mathrm{m}^3$ and 10% above $1.9\,\mathrm{m}^3$.
- Thirty day capacity often is specified for raw materials and products, but depends on connecting transportation equipment schedules.

20.6. **Orders of approximation.** In science, engineering, and other quantitative disciplines, orders of approximation refer to formal or informal terms for how precise an approximation is, and to indicate progressively more refined approximations: in increasing order of precision, a zeroth order approximation, a first order approximation, a second order approximation, and so forth.

Formally, an nth order approximation is one where the order of magnitude of the error is at most x^n, or in terms of big O notation, the error is $O[x^n]$. In suitable circumstances, approximating a function by a Taylor polynomial of degree n yields an nth order approximation, by Taylor's theorem: a first order approximation is a linear approximation, and so forth.

20.7. **Handwaving.** Handwaving arguments often include order-of-magnitude estimates and dimensional consistency. Competent, well-intentioned researchers and professors rely on handwaving when, given a limited time, a large result must be shown and minor technical details cannot be given much attention; *e.g.*, "It can be shown that z is even".

Back-of-the-envelope calculations are approximate ways to get an answer by oversimplification and are compatible with handwaving.

21. Some Numeric Methods

21.1. **Back-of-the-envelop Calculations.** *Back-of-the-envelope calculations* use rough estimates for important factors and correction. Calculations are done in two parts:

- The "big part": the most important factor in a back-of-the-envelope product usually comes from the powers of 10, so evaluate this big part first; and
- The correction, the "small part": after taking out the big part, the remaining part is a correction factor. Normally, this product too is simplified by taking out its big part. To perform the calculation, round each factor to the closest number among *three* choices: 1, "few" or 10. The invented number few lies midway between 1 and 10: It is the geometric mean of 1 and 10, so $(\text{few})^2 = 10$ and $\text{few} \sim 3$.

21.2. **Chinese Multiplication.** Cool trick: it replaces multiplication by counting dots! Easy to apply!

`https://www.youtube.com/watch?v=OAREm_4Z8fs`.

21.3. **Lumping.** Approximate methods are robust: They almost always provide a reasonable answer. And the least accurate but most robust method is lumping. Instead of dividing a changing process into many tiny pieces (as done in calculus), group or lump it into one or two pieces. This simple approximation and its advantages are illustrated using examples ranging from demographics to nonlinear differential equations.

21.3.1. *Estimating populations: How many babies?* The first example is to estimate the number of babies in the United States. For definiteness, call a child a baby until he or she turns 2 years old. An exact calculation requires the birth dates of every person in the United States. This, or closely similar, information is collected once every decade by the US Census Bureau.

As an approximation to this voluminous data, the Census Bureau publishes the number of people at each age. The data for 1991 is a set of points lying on a yr wiggly line $N[t]$, where t is age. Then

$$N_{\text{babies}} = \int_0^{2\,\text{yr}} N[t]\ \mathrm{d}t\,.$$

This method has several problems. First, it depends on the huge resources of the US Census Bureau, so it is not usable on a desert island for back-of-the-envelope calculations. Second, it requires integrating a curve with no analytic form, so the integration must be done numerically. Third, the integral is of data specific to this problem, whereas mathematics should be about generality. An exact integration, in short, provides little insight and has minimal transfer value. Instead of integrating the population curve exactly, approximate it – lump the curve into one rectangle.

What are the height and width of this rectangle? The rectangle's width is a time, and a plausible time related to populations is the life expectancy. It is roughly 80 years, so make 80 years the width by pretending that everyone dies abruptly on his or her 80th birthday. The rectangle's height can be computed from the rectangle's area, which is the US population – conveniently 300 million in 2008. Therefore,

$$\text{height} = \frac{\text{area}}{\text{width}} \sim \frac{3\times 10^8}{75\,\text{yr}}\,.$$

Why did the life expectancy drop from 80 to 75 years? Fudging the life expectancy simplifies the mental division: 75 divides easily into 3 and 300. The inaccuracy is no larger than the error made by lumping, and it might even cancel the lumping error. Using 75 years as the width makes the height approximately $4\times 10^6\,\text{yr}^{-1}$.

Integrating the population curve over the range $t = 0, \ldots, 2\,\text{yr}$ becomes just multiplication:

$$N_{\text{babies}} \sim 4\times 10^6\,\text{yr}^{-1} \times 2\,\text{yr} = 8\times 10^6\,.$$

The Census Bureau's figure is very close: 7.980×10^6. The error from lumping canceled the error from fudging the life expectancy to 75 years!

21.4. **Estimating Integrals.** The US population curve was difficult to integrate partly because it was unknown. But even well-known functions can be difficult to integrate. In such cases, two lumping methods are particularly useful: the $1/e$ heuristic and the full width at half maximum (FWHM) heuristic.

21.4.1. $1/e$ *heuristic.* Electronic circuits, atmospheric pressure and radioactive decay contain the ubiquitous exponential and its integral (given here in dimensionless form)

$$\int_0^\infty e^{-t}\,\mathrm{d}t\,.$$

To approximate its value, let's lump the e^{-t} curve into one rectangle.

What values should be chosen for the width and height of the rectangle? A reasonable height for the rectangle is the maximum of e^{-t}, namely 1. To choose its width, use significant change as the criterion (a method used again later): Choose a significant change in e^{-t}; then find the width Δt that produces this change. In an exponential decay, a simple and natural significant change is when e^{-t} becomes a factor of e closer to its final value (which is 0 here because t goes to ∞). With this criterion, $\Delta t = 1$. The lumping rectangle then has unit area – which is the exact value of the integral!

Encouraged by this result, let's try the heuristic on the difficult integral.

$$\int_\infty^\infty e^{-x^2}\,\mathrm{d}x\,.$$

Again lump the area into a single rectangle. Its height is the maximum of e^{-x^2}, which is 1. Its width is enough that e^{-x^2} falls by a factor of e. This drop happens at $x = \pm 1$, so the width is $\Delta x = 2$ and its area is 1×2. The exact area is $\sqrt{\pi} \sim 1.77$, so lumping makes an error of only 13%: For such a short derivation, the accuracy is extremely high.

21.4.2. *Full width at half maximum.* Another reasonable lumping heuristic arose in the early days of spectroscopy. As a spectroscope swept through a range of wavelengths, a chart recorder would plot how strongly a molecule absorbed radiation of that wavelength. This curve contains many peaks whose location and area reveal the structure of the molecule (and were essential in developing quantum theory). But decades before digital chart recorders existed, how could the areas of the peaks be computed?

They were computed by lumping the peak into a rectangle whose height is the height of the peak and whose width is the full width at half maximum (FWHM). Where the $1/e$ heuristic uses a factor of e as the significant change, the FWHM heuristic uses a factor of 2.

Try this recipe on the Gaussian integral $\int_\infty^\infty e^{-x^2}\,\mathrm{d}x$. The maximum height of e^{-x^2} is 1, so the half maxima [24] are at $x = \pm\sqrt{\ln[2]}$ and the full width is $2\sqrt{\ln[2]}$. The lumped rectangle therefore has area $2\sqrt{\ln[2]} \sim 1.665$. The exact area is $\sqrt{\pi} \sim 1.77$: The FWHM heuristic makes an error of only 6%, which is roughly one-half the error of the $1/e$ heuristic.

21.5. **Estimating Derivatives.** In the preceding examples, lumping helped estimate integrals. Because integration and differentiation are closely related, lumping also provides a method for estimating derivatives. The method begins with a dimensional observation about derivatives. A derivative is a ratio of differentials; for example, $\mathrm{d}f/\mathrm{d}x$ is the ratio of $\mathrm{d}f$ to $\mathrm{d}x$. Because d is dimensionless, the dimensions of $\mathrm{d}f/\mathrm{d}x$ are the dimensions of f/x. This useful, surprising conclusion is worth testing with a familiar example: Differentiating height y with respect to time t produces velocity $\mathrm{d}y/\mathrm{d}t$, whose dimensions of $[L/T]$ are indeed the dimensions of y/t.

21.5.1. *Secant Approximation.* As $\mathrm{d}f/\mathrm{d}x$ and f/x have identical dimensions, perhaps their magnitudes are similar:

$$\frac{\mathrm{d}f}{\mathrm{d}x} \sim \frac{f}{x}\,.$$

[24] $f[x] = e^{-x^2}$, then $f'[x] = -2xe^{-x^2}$. $\max f[x] = 1$, so the half-max occurs at $1/2$. Therefore, $e^{-x^2} = 1/2 \implies -x^2 = \ln[1/2] \implies x^2 = \ln[2]$. Finally, $x = \pm\sqrt{\ln[2]}$.

Geometrically, the derivative $\mathrm{d}f/\mathrm{d}x$ is the slope of the *tangent* line, whereas the approximation f/x is the slope of the *secant* line. By replacing the curve with the secant line, we make a lumping approximation.

Let's test the approximation on an easy function such as $f[x] = x^2$. Good news – the secant and tangent slopes differ only by a factor of 2:

$$\frac{\mathrm{d}f}{\mathrm{d}x}[x] = 2x \qquad \text{and} \qquad \frac{f}{x}[x] = x\,.$$

How accurate is the secant approximation for $f[x] = x^2+100$? The secant approximation is quick and useful but can make large errors. When $f[x] = x^2 + 100$, for example, the secant and tangent at $x = 1$ have dramatically different slopes. The tangent slope $\mathrm{d}f/\mathrm{d}x$ is 2, whereas the secant slope $f[1]/1$ is 101. The ratio of these two slopes, although dimensionless, is distressingly large.

The large discrepancy in replacing the derivative $\mathrm{d}f/\mathrm{d}x$, which is

$$\lim_{\Delta x\to 0} \frac{f[x] - f[x - \Delta x]}{\Delta x}\,,$$

with the secant slope $f[x]/x$ is due to two approximations. The first approximation is to take $\Delta x = x$ rather than $\Delta x = 0$. Then $\mathrm{d}f/\mathrm{d}x \sim (f[x] - f[0])/x$. This first approximation produces the slope of the line from $[0, f[0]]$ to $[x, f[x]]$. The second approximation replaces $f[0]$ with 0, which produces $\mathrm{d}f/\mathrm{d}x \sim f/x$; that ratio is the slope of the secant from $[0, 0]$ to $[x, f[x]]$.

21.5.2. *Improved Secant Approximation.* The second approximation is fixed by starting the secant at $[x, f[x]]$ instead of $[0, 0]$.

With that change, what are the secant and tangent slopes when $f[x] = x^2 + C$? Call the secant starting at $[0, 0]$ the origin secant; call the new secant the $x = 0$ secant. Then the $x = 0$ secant always has one-half the slope of the tangent, no matter the constant C. The $x = 0$ secant approximation is robust against – is *unaffected* by – vertical translation.

How robust is the x =0 secant approximation against horizontal translation? To investigate how the $x = 0$ secant handles horizontal translation, translate $f[x] = x^2$ rightward by 100 to make $f[x] = (x - 100)^2$. At the parabola's vertex $x = 100$, the $x = 0$ secant, from $[0, 104]$ to $[100, 0]$, has slope -100; however, the tangent has zero slope. Thus the $x = 0$ secant, although an improvement on the origin secant, is *affected* by horizontal translation.

21.5.3. *Significant Change Approximation.* The derivative itself is unaffected by horizontal and vertical translation, so a derivative suitably approximated might be translation invariant. An approximate derivative is

$$\frac{\mathrm{d}f}{\mathrm{d}x} \sim \frac{f[x + \Delta x] - f[x]}{\Delta x}\,,$$

where Δx is not zero but is still small.

How small should Δx be? Is $\Delta x = 0.01$ small enough? The choice $\Delta x = 0.01$ has two defects. First, it cannot work when x has dimensions. If x is a length, what length is small enough? Choosing $\Delta x = 1\,\mathrm{mm}$ is probably small enough for computing derivatives related to the solar system, but is probably too large for computing derivatives related to falling fog droplets. Second, no fixed choice can be scale invariant. Although $\Delta x = 0.01$ produces accurate derivatives when $f[x] = \sin[x]$, it fails when $f[x] = \sin[1000x]$, the result of simply rescaling x to $1000x$.

These problems suggest trying the following significant-change approximation:

$$\frac{\mathrm{d}f}{\mathrm{d}x} \sim \frac{\text{significant } \Delta f \text{ (change in } f\text{) at } x}{\Delta x \text{ that produces a significant } \Delta f}\,.$$

Because the Δx here is defined by the properties of the curve at the point of interest, without favoring particular coordinate values or values of Δx, the approximation is *scale and translation invariant.*

To illustrate this approximation, let's try $f[x] = \cos[x]$ and estimate f' at $x = 3\pi/2$ with the three approximations: the origin secant, the $x = 0$ secant, and the significant-change approximation. The origin secant goes from $[0, 0]$ to $[3\pi/2, 0]$, so it has zero slope. It is a

poor approximation to the exact slope of 1. The $x = 0$ secant goes from $[0, 1]$ to $[3\pi/2, 0]$, so it has a slope of $-2/3\pi$, which is worse than predicting zero slope because even the sign is wrong!

The significant-change approximation might provide more accuracy. What is a significant change in $f[x] = \cos[x]$? Because the cosine changes by 2 (from -1 to 1), call $1/2$ a significant change in $f[x]$. That change happens when x changes from $3\pi/2$, where $f[x] = 0$, to $3\pi/2 + \pi/6$, where $f[x] = 1/2$. In other words, Δx is $\pi/6$. The approximate derivative is therefore

$$\frac{\mathrm{d}f}{\mathrm{d}x} \sim \frac{\text{significant } \Delta f \text{ near } x}{\Delta x} \sim \frac{1/2}{\pi/6} = 3/\pi\,.$$

This estimate is approximately 0.955 – amazingly close to the true derivative of 1.

21.6. **Analyzing differential equations: The spring-mass system.** Estimating derivatives reduces differentiation to division; it thereby reduces differential equations to algebraic equations.

To produce an example equation to analyze, connect a block of mass m to an ideal spring with spring constant (stiffness) k, pull the block a distance x_0 to the right relative to the equilibrium position $x = 0$, and release it at time $t = 0$. The block oscillates back and forth, its position x described by the ideal-spring differential equation

$$m\ddot{x} + kx = 0\,.$$

Let's approximate the equation and thereby estimate the oscillation frequency.

21.6.1. *Checking Dimensions.* Upon seeing any equation, first check its dimensions. If all terms do not have identical dimensions, the equation is not worth solving – a great saving of effort. If the dimensions match, the check has prompted reflection on the meaning of the terms; this reflection helps prepare for solving the equation and for understanding any solution.

What are the dimensions of the two terms in the spring equation? Look first at the simple second term kx. It arises from Hooke's law, which says that an ideal spring exerts a force kx where x is the extension of the spring relative to its equilibrium length. Thus the second term kx is a force. Is the first term also a force?

The first term $m\ddot{x}$ contains the second derivative $\ddot{x} = \mathrm{d}^2x/\mathrm{d}t^2$, which is familiar as an acceleration. Many differential equations, however, contain unfamiliar derivatives. The Navier-Stokes equations of fluid mechanics are an example.

To practice for later handling such complicated terms, let's now find the dimensions of $\mathrm{d}^2x/\mathrm{d}t^2$ by hand. Because $\mathrm{d}^2x/\mathrm{d}t^2$ contains two exponents of 2, and x is length and t is time, $\mathrm{d}^2x/\mathrm{d}t^2$ might plausibly have dimensions of $\left[L^2/T^2\right]$.

Are $\left[L^2/T^2\right]$ the correct dimensions? To decide, use the idea that the differential symbol d means "a little bit of". The numerator d^2x, meaning d of dx, is "a little bit of a little bit of x". Thus, it is a length. The denominator $\mathrm{d}t^2$ could plausibly mean $(\mathrm{d}t)^2$ or $\mathrm{d}(t^2)$. [It turns out to mean $(\mathrm{d}t)^2$.] In either case, its dimensions are $\left[T^2\right]$. Therefore, the dimensions of the second derivative are $\left[L/T^2\right]$:

$$\dim \frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \frac{[L]}{[T^2]}\,.$$

This combination is an acceleration, so the spring equation's first term $m\ddot{x}$ is mass times acceleration – giving it the same dimensions as the kx term.

21.6.2. *Estimating the magnitudes of the terms.* The spring equation passes the dimensions test, so it is worth analyzing to find the oscillation frequency. The method is to replace each term with its approximate magnitude. These replacements will turn a complicated differential equation into a simple algebraic equation for the frequency.

To approximate the first term $m\ddot{x}$, use the significant-change approximation to estimate the magnitude of the acceleration $\ddot{x}$:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} \sim \frac{\text{significant } \Delta x}{(\Delta t \text{ that produces a significant } \Delta x)}\,.$$

To evaluate this approximate acceleration, first decide on a significant Δx – on what constitutes a significant change in the mass's position. The mass moves between the points $x = -x_0$ and $x = +x_0$, so a significant change in position should be a significant fraction of the peak-to-peak amplitude $2x_0$. The simplest choice is $\Delta x = x_0$.

Now estimate Δt: the time for the block to move a distance comparable to Δx. This time – called the characteristic time of the system – is related to the oscillation period T. During one period, the mass moves back and forth and travels a distance $4x_0$ – much farther than x_0. If Δt were, say, $T/4$ or $T/2\pi$, then in the time Δt the mass would travel a distance comparable to x_0. Those choices for Δt have a natural interpretation as being approximately $1/\omega$, where the angular frequency ω is connected to the period by the definition $\omega \coloneqq 2\pi/T$. With the preceding choices for Δx and Δt, the $m\ddot{x}$ term is roughly $mx_0\omega^2$.

What does "is roughly" mean? The phrase cannot mean that $mx_0\omega^2$ and $m\ddot{x}$ are within, say, a factor of 2, because $m\ddot{x}$ varies and mx_0/T^2 is constant. Rather, "is roughly" means that a typical or characteristic magnitude of $m\ddot{x}$ – for example, its root-mean-square value – is comparable to $mx_0\omega^2$. Let's include this meaning within the twiddle notation $\sim$. Then the typical-magnitude estimate can be written

$$m\ddot{x} \sim mx_0\omega^2 \,.$$

With the same meaning of "is roughly", namely that the typical magnitudes are comparable, the spring equation's second term kx is roughly kx_0. The two terms must add to zero – a consequence of the spring equation $m\ddot{x} + kx = 0$.

Therefore, the magnitudes of the two terms are comparable:

$$mx_0\omega^2 \sim kx_0 \,.$$

The amplitude x_0 divides out! With x_0 gone, the frequency ω and oscillation period $T = 2\pi/\omega$ independent of amplitude. [This reasoning uses several approximations, but this conclusion is exact.] The approximated angular frequency ω is then $\sqrt{k/m}$.

For comparison, the exact solution of the spring differential equation is

$$x = x_0 \cos[\omega t] \,,$$

where ω is $\sqrt{k/m}$. The approximated angular frequency is also exact!

21.7. **Predicting the period of a pendulum.** Lumping not only turns integration into multiplication, it turns nonlinear into linear differential equations. Our example is the analysis of the period of a pendulum, for centuries the basis of Western timekeeping.

How does the period of a pendulum depend on its amplitude? The amplitude θ_0 is the maximum angle of the swing; for a loss-less pendulum released from rest, it is also the angle of release. The effect of amplitude is contained in the solution to the pendulum differential equation:

$$m\ddot{\theta} + g/l \sin[\theta] = 0 \,.$$

The analysis will use all our tools: dimensions, easy cases and lumping.

21.7.1. *Dimensions.* Since the term g/l has no dimensions nor do angles, the differential equation is manifestly dimensionless. So we can proceed to analyze the equation.

21.7.2. *Small amplitudes: Applying extreme cases.* The pendulum equation is difficult because of its nonlinear factor $\sin[\theta]$. Fortunately, the factor is easy in the small-amplitude extreme case $\theta \to 0$. In that limit, the height of the triangle, which is $\sin[\theta]$, is almost exactly the arclength θ. Therefore, for small angles, $\sin[\theta] \sim \theta$.

In the small-amplitude extreme, the pendulum equation becomes linear:

$$\ddot{\theta} + g/l\theta = 0 \,.$$

Compare this equation to the spring-mass equation:

$$m\ddot{x} + kx = 0 \,.$$

The equations correspond with x analogous to θ and k/m analogous to g/l. The frequency of the spring-mass system is $\omega = \sqrt{k/m}$ k/m, and its period is $T = 2\pi/\omega = 2\pi m/k$. For the pendulum equation, the corresponding period is

$$T = 2\pi\sqrt{\frac{l}{g}}\,. \qquad \text{[for small amplitudes]}$$

(This analysis is a preview of the method of analogy.)

21.7.3. *Arbitrary amplitudes: Applying dimensional analysis.* The preceding results might change if the amplitude θ_0 is no longer small.

As θ_0 increases, does the period increase, remain constant, or decrease? Any analysis becomes cleaner if expressed using dimensionless groups. This problem involves the period T, length l, gravitational strength g, and amplitude θ. Therefore, T can belong to the dimensionless group $T/\sqrt{l/g}$. Because angles are dimless, θ_0 is itself a dimless group. The two groups are independent.

An instructive contrast is the ideal spring-mass system. The period T, spring constant k and mass m can form the dimensionless group $T/\sqrt{m/k}$; but the amplitude x_0, as the only quantity containing a length, cannot be part of any dimensionless group and cannot therefore affect the period of the spring-mass system. In contrast, the pendulum's amplitude θ_0 is already a dimensionless group, so it can affect the period of the system.

Two dimensionless groups produce the general dimensionless form

$$\text{one group} = \text{function of the other group},$$

so

$$\frac{T}{\sqrt{l/g}} = \text{function of } \theta_0\,.$$

Because $T/\sqrt{l/g} = 2\pi$, when $\theta_0 = 0$ (the small-amplitude limit), factor out the 2π to simplify the subsequent equations, and define a dimensionless period h as follows:

$$\frac{T}{\sqrt{l/g}} = 2\pi h[\theta_0]\ .$$

The function h contains all information about how amplitude affects the period of a pendulum. Using h, the original question about the period becomes the following: Is h an increasing, constant, or decreasing function of amplitude? This question is answered in the following section.

21.7.4. *Large amplitudes: Extreme cases again.* For guessing the general behavior of h as a function of amplitude, useful clues come from evaluating h at two amplitudes. One easy amplitude is the extreme of zero amplitude, where $h[0] = 1$h. A second easy amplitude is the opposite extreme of large amplitudes.

How does the period behave at large amplitudes? As part of that question, what is a large amplitude? An interesting large amplitude is $\pi/2$, which means releasing the pendulum from horizontal. However, at $\pi/2$ the exact h is the following awful expression [25]

$$h[\pi/2] = \frac{\sqrt{2}}{\pi}\int_0^{\pi/2}\frac{\mathrm{d}\theta}{\sqrt{\cos[\theta]}}\,.$$

Is this integral less than, equal to or more than 1? Who knows? The integral is likely to have no closed form and to require numerical evaluation.

Because $\theta_0 = \pi/2$ is not a helpful extreme, be even more extreme. Try $\theta_0 = \pi$, which means releasing the pendulum bob from vertical. If the bob is connected to the pivot point by a string, however, a vertical release would mean that the bob falls straight down instead of oscillating. This novel behavior is neither included in nor described by the pendulum differential equation.

Fortunately, a thought experiment is cheap to improve: Replace the string with a massless steel rod. Balanced perfectly at $\theta_0 = \pi$, the pendulum bob hangs upside down forever, so $T[\pi] = \infty$ and $h[\pi] = \infty$. Thus, $h[\pi] > 1$ and $h[0] = 1$. From these data, the

[25] According to Wolfram Alpha, the value of the integral is $\sqrt{\pi}/\Gamma[3/4]^2 \sim 1.180340599\ldots$, where Γ is the gamma function.

most likely conjecture is that h increases monotonically with amplitude. Although h could first decrease and then increase, such twists and turns would be surprising behavior from such a clean differential equation.

21.7.5. *Moderate amplitudes: Applying lumping.* The conjecture that h increases monotonically was derived using the extremes of zero and vertical amplitude, so it should apply at intermediate amplitudes. Before taking that statement on faith, recall a proverb from arms-control negotiations: "Trust, but verify".

At moderate (small but nonzero) amplitudes, does the period, or its dimensionless cousin h, increase with amplitude? In the zero-amplitude extreme, $\sin[\theta]$ is close to θ. That approximation turned the nonlinear pendulum equation $\ddot{\theta} + g/l \sin[\theta] = 0$ into the linear, ideal-spring equation – in which the period is independent of amplitude.

At nonzero amplitude, however, θ and $\sin[\theta]$ differ and their difference affects the period. To account for the difference and predict the period, split $\sin[\theta]$ into the tractable factor θ and an adjustment factor $f[\theta]$. The resulting equation is

$$\ddot{\theta} + \frac{g}{l}\theta\frac{\sin[\theta]}{\theta} = 0\,,$$

where the adjustment factor: $f[\theta] = \sin[\theta]/\theta$.

The nonconstant $f[\theta]$ encapsulates the nonlinearity of the pendulum equation. When θ is tiny, $f[\theta] \sim 1$: The pendulum behaves like a linear, ideal-spring system. But when θ is large, $f[\theta]$ falls significantly below 1, making the ideal-spring approximation significantly inaccurate. As is often the case, a changing process is difficult to analyze. As a countermeasure, make a lumping approximation by replacing the changing $f[\theta]$ with a constant.

The simplest constant is $f[0]$. Then the pendulum differential equation becomes $\ddot{\theta} + g/l = 0$. This equation is, again, the ideal-spring equation. In this approximation, period does not depend on amplitude, so $h = 1$ for all amplitudes. For determining how the period of an unapproximated pendulum depends on amplitude, the $f[\theta] \to f[0]$ lumping approximation discards too much information.

Therefore, replace $f[\theta]$ with the other extreme $f[\theta_0]$. Then the pendulum equation becomes

$$\ddot{\theta} + g/l\,\theta\, f[\theta_0] = 0\,.$$

Is this equation linear? What physical system does it describe? Because $f[\theta_0]$ is a constant, this equation is linear! It describes a zero-amplitude pendulum on a planet with gravity g_{eff} that is slightly weaker than earth gravity – as shown by the following slight regrouping:

$$\ddot{\theta} + \frac{g f[\theta_0]}{l}\theta = 0\,,$$

where $g_{\text{eff}} = g f[\theta_0]$.

Because the zero-amplitude pendulum has period $T = 2\pi\sqrt{l/g}$, the zero-amplitude, low-gravity pendulum has period

$$T[\theta_0] \sim 2\pi\sqrt{\frac{l}{g_{\text{eff}}}} = 2\pi\sqrt{\frac{l}{g f[\theta_0]}}\,.$$

Using the dimensionless period h avoids writing the factors of 2π, l and g, and it yields the simple prediction

$$h[\theta_0] \sim f[\theta_0]^{-1/2} = \left(\frac{\sin[\theta_0]}{\theta_0}\right)^{-1/2}\,.$$

At moderate amplitudes the approximation closely follows the exact dimensionless period (dark curve, in figure). As a bonus, it also predicts $h[\pi] = \infty$, so it agrees with the thought experiment of releasing the pendulum from upright.

How much larger than the period at zero amplitude is the period at 10° amplitude? A 10° amplitude is roughly 0.17 rad, a moderate angle, so the approximate prediction for h can itself accurately be approximated using a Taylor series. The Taylor series for $\sin[\theta]$ begins $\theta - \theta^3/6$, so

$$f[\theta_0] \sim 1 - \frac{\theta_0^2}{6}\,.$$

Then $h[\theta_0]$, which is roughly $f[\theta_0]^{-1/2}$, becomes

$$h[\theta_0] = \left(1 - \frac{\theta_0^2}{6}\right)^{-1/2} .$$

Another Taylor series yields $(1+x)^{-1/2} \sim 1 - x/2$ (for small x). Therefore,

$$h[\theta_0] \sim 1 + \frac{\theta_0^2}{12} .$$

Restoring the dimensioned quantities gives the period itself:

$$T \sim 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{\theta_0^2}{12}\right) .$$

Compared to the period at zero amplitude, a 10° amplitude produces a fractional increase of roughly $\theta_0^2/12 \sim 0.0025$ or 0.25%. Even at moderate amplitudes, the period is nearly independent of amplitude!

Does our lumping approximation underestimate or overestimate the period? The lumping approximation simplified the pendulum differential equation by replacing $f[\theta]$ with $f[\theta_0]$. Equivalently, it assumed that the mass always remained at the endpoints of the motion where $|\theta| = \theta_0$. Instead, the pendulum spends much of its time at intermediate positions where $|\theta| < \theta_0$ and $f[\theta] > f[\theta_0]$. Therefore, the average f is greater than $f[\theta_0]$. Because h is inversely related to f ($h = f^{-1/2}$), the $f[\theta] \to f[\theta_0]$ lumping approximation overestimates h and the period.

The $f[\theta] \to f[0]$ lumping approximation, which predicts $T = 2\pi\sqrt{l/g}$, underestimates the period. Therefore, the true coefficient of the θ_0^2 term in the period approximation

$$T \sim 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{\theta_0^2}{12}\right)$$

lies between 0 and 1/12. A natural guess is that the coefficient lies halfway between these extremes – namely, 1/24. However, the pendulum spends more time toward the extremes (where $f[\theta] = f[\theta_0]$) than it spends near the equilibrium position (where $f[\theta] = f[0]$. Therefore, the true coefficient is probably closer to 1/12 – the prediction of the $f[\theta] \to f[\theta_0]$ approximation – than it is to 0. An improved guess might be two-thirds of the way from 0 to 1/12, namely 1/18.

In comparison, a full successive-approximation solution of the pendulum differential equation gives the following period

$$T = 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{1}{16}\theta_0^2 + \frac{11}{3072}\theta_0^4 + \cdots\right) .$$

Our educated guess of 1/18 is very close to the true coefficient of 1/16!

21.7.6. *Summary.* Lumping turns calculus on its head. Whereas calculus analyzes a changing process by dividing it into ever finer intervals, lumping simplifies a changing process by combining it into one unchanging process. It turns curves into straight lines, difficult integrals into multiplication, and mildly nonlinear differential equations into linear differential equations.

21.8. **Newton's Method.** In numerical analysis, *Newton's method*, *aka* Newton-Raphson method, is a method for finding successively better approximations to the roots (or zeroes) of a real-valued function: $x : f[x] = 0$.

21.8.1. *Algorithm.* The algorithm in one variable is the following:

- Given a function f defined over the reals x and given its derivative f', then guess x_0 for a root of the function f.
- Provided the function satisfies all the assumptions made in the derivation formula, a better approximation x_1 is

$$x_1 = x_0 - \frac{f[x_0]}{f'[x_0]} .$$

- Geometrically, $[x_1, 0]$ is the intersection with the x-axis of a line tangent to f at $[x_0, f[x_0]]$.
- The process is repeated as

$$x_{n+1} = x_n - \frac{f[x_n]}{f'[x_n]}\,,$$

until a sufficiently accurate value is reached.

21.8.2. *Description.* The idea of the method is as follows: one starts with an initial guess which is reasonably close to the true root, then the function is approximated by its tangent line (which can be computed using the tools of calculus) and one computes the x-intercept of this tangent line. This x-intercept will typically be a better approximation to the function's root than the original guess and the method can be iterated.

Suppose $f : [a, b] \to \mathcal{R}$ is a differentiable function defined on the interval $[a, b]$ with values in the real numbers $\mathcal{R}$. The formula for converging on the root can be easily derived. Suppose we have some current approximation x_n. Then, we can derive the formula for a better approximation, x_{n+1}. We know from the definition of the derivative at a given point that it is the slope of a tangent at that point. That is,

$$f'[x_n] = \frac{\Delta y}{\Delta x} = \frac{f[x_n] - 0}{x_n - x_{n+1}}\,.$$

Here, f' denotes the derivative of the function f. Then, by algebra, we can derive

$$x_{n+1} = x_n - \frac{f[x_n]}{f'[x_n]}\,.$$

We start the process off with some arbitrary initial value x_0. (The closer to the zero, the better. But, in the absence of any intuition about where the zero might lie, a "guess and check" method might narrow the possibilities to a reasonably small interval by appealing to the intermediate value theorem.) The method will usually converge, provided this initial guess is close enough to the unknown zero and that $f'[x_0] \neq 0$. Furthermore, for a zero of multiplicity 1, the convergence is at least quadratic in a neighborhood of the zero, which intuitively means that the number of correct digits roughly at least doubles in every step.

21.8.3. *Failure Analysis.* [read wiki article ;)]

21.8.4. *Applications.* Minimization and maximization problems: Newton's method can be used to find a minimum or maximum of a function. The derivative is zero at a minimum or maximum, so minima and maxima can be found by applying Newton's method to the derivative. The iteration becomes:

$$x_{n+1} = x_n - \frac{f'[x_n]}{f''[x_n]}\,.$$

Solving transcendental equations: many transcendental equations can be solved using Newton's method. Given the equation

$$g[x] = h[x]\,,$$

with $g[x]$ or $h[x]$ a transcendental equation, one writes

$$f[x] = g[x] - h[x]\,.$$

The values of x that solves the original equation are then the roots of $f[x]$, which may be found via Newton's method.

21.8.5. *Examples.* Square root of a number: find the square root of 612. This is equivalent to finding the solution to

$$x^2 = 612\,,$$

with derivative $f'[x] = 2x$.

With an initial guess of 10, the sequence given by Newton's method is ... With only 5 iterations, one can obtain a solution accurate to many decimal places: 24.738 633 753 767.

Solution of $\cos[x] = x^3$. Rephrase this problem to find the roots of the equation $f[x] = \cos[x] - x^3$. We have $f'[x] = -\sin[x] - 3x^2$. Since $\cos x \leq 1$ for all x and $x^3 > 1$ for $x > 1$, then we know that our zero lies between 0 and 1. We try a starting value of

$x_0 = 0.5$. (Note that a starting value of 0 will lead to a undefined result, showing the importance of using a starting point that is close to the zero.)...

The correct digits are underlined in the above example. In particular, x_6 is correct to the number of decimal places given (0.865 474 033 102). We see that the number of correct digits after the decimal point increases from 2 (for x_3) to 5 and 10, illustrating the quadratic convergence.

21.8.6. *Pseudocode.* [wiki article ;)]

21.9. **Rectangle Method.** The *rectangle method* computes an approximation to a definite integral, made by finding the area of a collection of rectangles whose heights are determined by the values of the function.

21.9.1. *Formula.* Specifically, the interval $]a, b[$ over which the function is to be integrated is divided into N equal subintervals of length $h = (b - a)/N$. The rectangles are then drawn so that either their left or right corners, or the middle of their top line lies on the graph of the function, with bases running along the x-axis. The approximation to the integral is then calculated by adding up the areas (base multiplied by height) of the N rectangles, giving the formula

$$\int_a^b f[x] \, \mathrm{d}x \sim h \sum_{n=0}^{N-1} f[x_n] \, ,$$

where $h = (b - a)/N$ and $x_n = a + nh$.

The formula for x_n above gives x_n for the top-left corner approximation.

As N gets larger, this approximation gets more accurate. In fact, this computation is the spirit of the definition of the Riemann integral and the limit of this approximation as $n \to \infty$ is defined and equal to the integral of f on $]a, b[$ if this Riemann integral is defined. Note that this is true regardless of which i is used, however, the midpoint approximation tends to be more accurate for finite n.

21.9.2. *Error.* For a function f which is twice differentiable, the approximation error in each section $]a, a + \Delta[$ of the midpoint rule decays as the cube of the width of the rectangle.

$$E_i \leq \frac{\Delta^3}{24} f''[\xi] \, ,$$

for some ξ in $]a, a + \Delta[$. Summing this, the approximation error for n intervals with width Δ is less than or equal to $n = 1, 2, 3, \ldots$, where $n + 1$ is the number of nodes

$$E \leq \frac{n\Delta^3}{24} f''[\xi]$$

in terms of the total interval, we know that $n\Delta = b - a$, so we can rewrite the expression:

$$E \leq \frac{(b - a)\Delta^2}{24} f''[\xi] \, ,$$

for some ξ in $]a, b[$.

21.10. **Trapezoidal Rule.** The *trapezoidal rule*, *aka* the trapezoid rule or trapezium rule, is a technique for approximating the definite integral

$$\int_a^b f[x] \, \mathrm{d}x \, .$$

The trapezoid rule works by approximating the region under the graph of the function $f[x]$ as a trapezoid and calculating its area. It follows that

$$\int_a^b f[x] \, \mathrm{d}x \sim (b - a) \frac{f[a] + f[b]}{2} \, .$$

21.10.1. *Applicability and alternatives.* The trapezoidal rule is one of a family of formulas for numerical integration called *Newton-Cotes formulas*, of which the midpoint rule is similar to the trapezoid rule. Simpson's rule is another member of the same family and, in general, has faster convergence than the trapezoidal rule for functions which are twice continuously differentiable, though not in all specific cases. However, for various classes of rougher function (ones with weaker smoothness conditions), the trapezoidal rule has faster convergence in general than Simpson's rule.

Moreover, the trapezoidal rule tends to become extremely accurate when periodic functions are integrated over their periods, which can be analyzed in various ways.

For non-periodic functions, however, methods with unequally spaced points such as Gaussian quadrature and Clenshaw-Curtis quadrature are generally far more accurate.

21.10.2. *Numerical Implementation.* For a domain discretized into N equally spaced panels, or $N+1$ grid points $(1, 2, 3, \ldots, N+1)$, where the grid spacing is $h = (b-a)/N$, the approximation to the integral becomes

$$\begin{aligned}\int_a^b f[x]\ \mathrm{d}x &\sim \frac{h}{2}\sum_{k=1}^{N}\left(f[x_{k+1}] + f[x_k]\right)\\ &= \frac{b-a}{2N}\left(f[x_1] + 2f[x_2] + 2f[x_3] + \cdots + 2f[x_N] + 2f[x_{N+1}]\right)\ .\end{aligned}$$

21.10.3. *Error Analysis.* The error of the composite trapezoidal rule is the difference between the value of the integral and the numerical result:

$$E = \int_a^b f[x]\ \mathrm{d}x - \frac{b-a}{2N}\left(\frac{f[a]+f[b]}{2} + \sum_{k=1}^{N-1} f\left[a + k\frac{b-a}{2N}\right]\right)\ .$$

There exists a number ξ between a and b, such that

$$E = -\frac{(b-a)^3}{12N^2} f''[\xi]\ .$$

21.11. **Simpson's Rule.** *Simpson's rule* is a method for numerical integration, the numerical approximation of definite integrals.

Simpson's rule also corresponds to the 3-point Newton-Cotes quadrature rule.

Simpson's rule is a staple of scientific data analysis and engineering. Ti si widely used, for instance, by naval architects to numerically integrate hull offsets and cross-sectional areas to determine volumes and centroids of ships or lifeboats.

21.11.1. *Formula.* Specifically, Simpson's rule is the following approximation

$$\int_a^b f[x]\ \mathrm{d}x \sim \frac{b-a}{6}\left(f[a] + 4f\left[\frac{a+b}{2}\right] + f[b]\right)\ .$$

21.11.2. *Error.* The error in approximating an integral by Simpson's rule is

$$\frac{1}{90}\left(\frac{b-a}{2}\right)^5 \left|f^{(4)}[\xi]\right|,$$

where ξ is some number between a and b.

The error is asymptotically proportional to $(b-a)^5$. Since the error term is proportional to the fourth derivative of f at ξ, then this shows that Simpson's rule provides exact results for any polynomial f degree three or less, since the fourth derivative of such a polynomial is zero at all points.

21.12. **Perturbation Theory.** *Perturbation theory* comprises mathematical methods that are used to find an approximate solution to a problem which cannot be solved exactly, by starting from the exact solution of a related problem. Perturbation theory is applicable if the problem at hand can be formulated by adding a "small" term to the mathematical description of the exactly solvable problem.

Perturbation theory leads to an expression for the desired solution in terms of a formal power series in some "small" parameter – known as a *perturbation series* – that quantifies the deviation from the exactly solvable problem. The leading term in this power series is the solution of the exactly solvable problem, while further terms describe the deviation

in the solution, due to the deviation from the initial problem. Formally, we have for the approximation to the full solution A, a series in the small parameter (here called ϵ), like the following:

$$A = A_0 + \epsilon^1 A_1 + \epsilon^2 A_2 + \cdots .$$

In this example, A_0 would be the known solution to the exactly solvable initial problem and , $A_1, A_2, \ldots$ represent the *higher-order terms* which may be found iteratively by some systematic procedure. For small ϵ these higher-order terms in the series become successively smaller. An approximate "perturbation solution" is obtained by truncating the series, usually by keeping only the first two terms, the initial solution and the "first-order" perturbation correction:

$$A \sim A_0 + \epsilon A_1 .$$

21.12.1. *Perturbation Theory for Algebraic Equations.* Consider the quadratic equation

$$x^2 - 1 = \epsilon x .$$

The two roots of this equation are

$$x_1 = \epsilon/2 + \sqrt{1 + \epsilon^2/4} \qquad \text{and} \qquad x_2 = \epsilon/2 - \sqrt{1 + \epsilon^2/4} .$$

For small ϵ, these roots are well approximated by the first few terms of their Taylor series expansion

$$x_1 = 1 + \epsilon/2 + \epsilon^2/8 + O\big[\epsilon^3\big] \qquad \text{and} \qquad x_2 = -1 + \epsilon/2 - \epsilon^2/8 + O\big[\epsilon^3\big] .$$

Can we obtain the last equations without prior knowledge of the exact solutions of the quadratic? Yes, using regular perturbation theory. The technique involves four steps.

(1) Assume that the solution(s) of the quadratic equation can be Taylor expanded in ϵ. Then we have

$$x = X_0 + \epsilon X_1 + \epsilon^2 X_2 + O\big[\epsilon^3\big] ,$$

for X_0, X_1, X_2 to be determined.

(2) Substitute the last equation into that of the quadratic written as $x^2 - 1 - \epsilon x = 0$ and expand the left hand side of the resulting equation in power series of ϵ. Using

$$x^2 = X_0^2 + 2\epsilon X_0 X_1 + \epsilon^2 (X_1^2 + 2X_0 X_2) + O\big[\epsilon^3\big] \implies \epsilon x = \epsilon X_0 + \epsilon^2 X_1 + O\big[\epsilon^3\big] .$$

this gives

$$X_0^2 - 1 + \epsilon(2X_0 X_1 - X_0) + \epsilon^2 (X_1^2 + 2X_0 X_1 - X_1) + O\big[\epsilon^3\big] = 0 .$$

(3) Equate to zero the successive terms of the series in the left hand side of the last equation:

$$\begin{aligned} O\big[\epsilon^0\big] &: & X_0^2 - 1 &= 0 , \\ O\big[\epsilon^1\big] &: & 2X_0 X_1 - X_0 &= 0 , \\ O\big[\epsilon^2\big] &: & X_1^2 + 2X_0 X_2 - X_1 &= 0 , \\ O\big[\epsilon^3\big] &: & \ldots & . \end{aligned}$$

(4) Successively solve the sequence of equations obtained in in the last step. Since $X_0^2 - 1 = 0$ has two roots, $X_0 = \pm 1$, one obtains

$$\begin{aligned} X_0 = 1, X_1 &= 1/2, X_2 = 1/8 , \\ X_0 = -1, X_1 &= 1/2, X_2 = -1/8 \end{aligned}$$

It can be checked that substituting the last equations into the roots of the quadratic one recovers the quadratic.

From the previous example it might not be clear what the advantage of regular perturbation theory is, since one can obtain the approximations to the quadratic roots more directly by Taylor expansion of the roots themselves. To see the strength of regular perturbation theory, consider the following equation

$$x^2 - 1 = \epsilon e^x .$$

The solutions of this equation are not available; therefore the direct method is inapplicable here. However, the Taylor series expansion of these solutions can be obtained

by perturbation theory. We introduce the expansion $x = X_0 + \epsilon X_1 + \epsilon^2 X_2 + O\left[\epsilon^3\right]$ as in the first step of the solution to the last example. In the second step, we use (recall that $e^z = 1 + z + z^2/2 + O\left[z^3\right]$)

$$\epsilon e^x = \epsilon e^{X_0+\epsilon X_1+\epsilon^2 X_2+O\left[\epsilon^3\right]} = \epsilon e^{X_0} e^{\epsilon X_1+\epsilon^2 X_2+O\left[\epsilon^3\right]} = \epsilon e^{X_0} + \epsilon^2 X_1 e^{X_0} + O\left[\epsilon^3\right] .$$

Substituting this expression in the original equation written as $x^2 - 1 - \epsilon e^x = 0$, we obtain

$$X_0^2 - 1 + \epsilon \left(2X_0X_1 - e^{X_0}\right) + \epsilon^2 \left(X_1^2 + 2X_0X_1 - X_1e^{X_0}\right) + O\left[\epsilon^3\right] = 0 .$$

Then, after having obtained the roots to the last system of equations, the expansion can be written as

$$x_1 = 1 + \epsilon e/2 + \epsilon^2 e^2/8 + O\left[\epsilon^3\right] ,$$
$$x_2 = -1 - \epsilon/(2e) - \epsilon^2/(8e^2) + O\left[\epsilon^3\right] .$$

21.13. **Normalization of Algebraic Equations.** [Use of nondim. and approx. techniques]

A simple example concerns finding roots of a simultaneous system of algebraic equations:

$$\begin{cases} x + 10y &= 21 , \\ 5x + 5y &= 7 . \end{cases}$$

In the first equation, the coefficient of x is small compared to the coefficient of y, so it is tempting to ignore x and approximate the value of y:

$$y \sim 2.1$$

Substituting the y approximation into the second equation approximates the value of x:

$$x \sim \frac{7 - 2.1}{5} \sim 0.98 .$$

To check the approximation's validity, $[x, y] \sim [0.98, 2.1]$ is substituted into the first equation of the system, producing the ratio of the first term to the second term:

$$\frac{x}{10y} \sim \frac{0.98}{2.1} \sim 0.05 \ll 1 .$$

This ratio validates $y = 2.1$. In fact, approximate roots $x = 0.98$ and $y = 2.1$ are close to exact roots:

$$x = 1 \qquad \text{and} \qquad y = 2 .$$

Notice the cycle: assume, derive, calculate, check!

22. Curve Fitting

22.1. **Outliers.**

22.1.1. *Chauvenet's criterion.* In statistical theory, Chauvenet's criterion (named for William Chauvenet) is a means of assessing whether one piece of experimental data – an outlier – from a set of observations, is likely to be spurious.

To apply Chauvenet's criterion, first calculate the mean and standard deviation of the observed data. Based on how much the suspect datum differs from the mean, use the normal distribution function (or a table thereof) to determine the probability that a given data point will be at the value of the suspect data point. Multiply this probability by the number of data points taken. If the result is less than 0.5, the suspicious data point may be discarded, *i.e.*, a reading may be rejected if the probability of obtaining the particular deviation from the mean is less than $1/(2n)$.

Example: For instance, suppose a value is measured experimentally in several trials as 9, 10, 10, 10, 11, and 50. The mean is 16.7 and the standard deviation 14.92. 50 differs from 16.7 by 33.3, slightly more than two standard deviations. The probability of taking data more than two standard deviations from the mean is roughly 0.05. Six measurements were taken, so the statistic value (data size multiplied by the probability) is $(0.05)(6) = 0.3$. Because $0.3 < 0.5$, according to Chauvenet's criterion, the measured value of 50 should be discarded (leaving a new mean of 10, with standard deviation 0.7).

Criticism: Deletion of outlier data is a controversial practice frowned on by many scientists and science instructors; while Chauvenet's criterion provides an objective and quantitative method for data rejection, it does not make the practice more scientifically or methodologically sound, especially in small sets or where a normal distribution cannot be assumed. Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known.

22.2. **Data Normalization.** Given a set of n data points $\{x_i\}$, called the *raw scores*, define the *sample mean*, denoted $\langle x \rangle$, by

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \,,$$

which is the arithmetic mean of the $\{x_i\}$.

Then, define the *sample standard deviation*, denoted s_x, by

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2} \,.$$

Finally, transform the raw scores into *standard scores*, denoted z_{x_i}, for every x_i by applying

$$z_{x_i} = \frac{x_i - \langle x \rangle}{s_x} \,.$$

22.3. **Least Square Fitting.**

22.3.1. *Theory.* A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets ("the residuals") of the points from the curve. The sum of the squares of the offsets is used instead of the offset absolute values because this allows the residuals to be treated as a continuous differentiable quantity. However, because squares of the offsets are used, outlying points can have a disproportionate effect on the fit, a property which may or may not be desirable depending on the problem at hand.

In practice, the vertical offsets from a line (polynomial, surface, hyperplane, *etc.*) are almost always minimized instead of the perpendicular offsets. This provides a fitting function for the independent variable that estimates for a given (most often what an experimenter wants), allows uncertainties of the data points along the x- and y-axes to be incorporated simply, and also provides a much simpler analytic form for the fitting parameters than would be obtained using a fit based on perpendicular offsets. In addition,

the fitting technique can be easily generalized from a best-fit line to a best-fit polynomial when sums of vertical distances are used. In any case, for a reasonable number of noisy data points, the difference between vertical and perpendicular fits is quite small.

The linear least squares fitting technique is the simplest and most commonly applied form of linear regression and provides a solution to the problem of finding the best fitting straight line through a set of points. In fact, if the functional relationship between the two quantities being graphed is known to within additive or multiplicative constants, it is common practice to transform the data in such a way that the resulting line is a straight line, say by plotting T *vs.* $\sqrt{l}$ instead of T *vs.* l in the case of analyzing the period T of a pendulum as a function of its length l. For this reason, standard forms for exponential, logarithmic, and power laws are often explicitly computed. The formulas for linear least squares fitting were independently derived by Gauss and Legendre.

Vertical least squares fitting proceeds by finding the sum of the squares of the vertical deviations r^2 of a set of n data points $[x_i, y_i]$

$$r^2 = \sum \left(y_i - f[x_i, a_i]\right)^2 \tag{22.1}$$

from a function f. Note that this procedure does not minimize the actual deviations from the line (which would be measured perpendicular to the given function). In addition, although the unsquared sum of distances might seem a more appropriate quantity to minimize, use of the absolute value results in discontinuous derivatives which cannot be treated analytically. The square deviations from each point are therefore summed, and the resulting residual is then minimized to find the best fit line. This procedure results in outlying points being given disproportionately large weighting.

The condition for r^2 to be a minimum is that

$$\frac{\partial(r^2)}{\partial a_i} = (r^2)_{,a_i} = \partial_i(r^2) = 0\,, \tag{22.2}$$

for $i = 1, \ldots, n$.

22.3.2. *Procedure.* To fit a curve to a set of n data points $\{[x_i, y_i]\}$:

- Plot data in a Cartesian coordinate system to help choosing the curve to be used to fit data. Then, choose the formula of the curve: $y_i = f[x_i, a_i]$, where the $\{x_i, y_i\}$ are the data points and the $\{a_i\}$ the parameters to be found.
- Find the sum of the squares of the vertical deviations r^2: $r^2 = \sum \left(y_i - f[x_i, a_i]\right)^2$.
- Apply the condition for r^2 to be a minimum, $(r^2)_{,a_i} = 0$. This yields n equations with n unknowns ($\{a_i\}$).
- Replace the actual data $[x_i, y_i]$ to find the numerical values of the $\{a_i\}$.
- Finally, find the parameters $\{a_i\}$ by solving the system of equations and replace the parameters in the formula of the curve.

Note. Additionally, the data points can be normalized before fitting the curve. In this case, the $\{x_i\}$, $\{y_i\}$ or both maybe normalize to z_x or z_y. To have the fitting function, then, transform the z to x and y.

22.4. **Examples.**

22.4.1. *Linear Fit Without Normalization.* For a linear fit, $f[a_1, a_2] = a_1 + a_2 x$. Thus, eq. (22.1) becomes

$$r^2[a_1, a_2] = \sum_{i=1}^{n} \left(y_i - (a_1 + a_2 x_i)\right)^2 \,.$$

Applying eq. (22.2), find

$$\begin{cases} {r^2}_{,a_1} = -2\sum_{i=1}^{n} \left(y_i - (a_1 + a_2 x_i)\right) = 0\,, \\ {r^2}_{,a_2} = -2\sum_{i=1}^{n} \left(y_i - (a_1 + a_2 x_i)\right) x_i = 0\,, \end{cases}$$

which leads to the system of equations

$$\begin{cases} n a_1 + a_2 \sum_i x_i = \sum_i y_i\,, \\ a_1 \sum_i x_i + a_2 \sum_i x_i^2 = \sum_i x_i y_i\,. \end{cases}$$

Finally, use actual data to replace in the system and then find $\{a_1, a_2\}$.

i	x_i	y_i	x_i^2	$x_i y_i$
1	0.1	9.9	0.01	0.99
2	0.2	9.2	0.04	1.84
3	0.3	8.4	0.09	2.52
4	0.4	6.6	0.16	2.64
5	0.5	5.9	0.25	2.95
6	0.6	5.0	0.36	3.00
7	0.7	4.1	0.49	2.87
8	0.8	3.1	0.64	2.48
9	0.9	1.9	0.81	1.71
10	1.0	1.1	1.00	1.10
Sum	5.5	55.2	3.85	22.10

TABLE 1. Numerical linear fit without normalization

i	z_{x_i}	z_{y_i}	$z_{x_i}^2$	$z_{x_i} z_{y_i}$
1	-1.49	1.44	2.21	-2.14
2	-1.16	1.21	1.34	-1.40
3	-0.83	0.95	0.68	-0.78
4	-0.50	0.36	0.25	-0.18
5	-0.17	0.13	0.03	-0.02
6	0.17	-0.17	0.03	-0.03
7	0.50	-0.47	0.25	-0.23
8	0.83	-0.80	0.68	-0.66
9	1.16	-1.19	1.34	-1.38
10	1.49	-1.45	2.21	-2.16
Sum	0.0	0.0	9.00	-8.98

TABLE 2. Numerical linear fit with normalization

22.4.2. *Numerical Linear Fit Without Normalization.* Consider the set of 10 data points shown in the three first columns of table 1.

The equations for the linear fit require to find $\{n, \sum x, \sum x^2, \sum y, \sum xy\}$. This is done with the aid of table 1: $n = 10$, $\sum x = 5.50$, $\sum x^2 = 3.85$, $\sum y = 55.2$ and $\sum xy = 22.1$.

Replace the numerical values in the equations for linear fit to have

$$\begin{cases} 10a_1 + 5.50a_2 = 55.2\,, \\ 5.50a_1 + 3.85a_2 = 22.10\,. \end{cases}$$

Solving the last system of equations yields $a_0 \sim 11.0$ and $a_1 \sim -10.0$, which, in turn, gives the fitting curve:

$$f[x] = 11.0x - 10.0\,.$$

22.4.3. *Numerical Linear Fit With Normalization.* Consider the data presented in table 1. Normalize both the $\{x_i\}$ and the $\{y_i\}$.

To begin the normalization, find the average values of the variables:

$$\langle x \rangle = \frac{1}{10} 5.5 = 0.55 \qquad \text{and} \qquad \langle y \rangle = \frac{1}{10} 55.2 = 5.52\,.$$

Then, calculate their standard deviations:

$$s_x = \sqrt{\frac{1}{9} \sum_i (x_i - 0.55)^2} = 0.303 \qquad \text{and} \qquad s_y = \sqrt{\frac{1}{9} \sum_i (x_i - 5.52)^2} = 3.04\,.$$

With the average values and the standard deviations, find the standard scores:

$$z_{x_i} = \frac{x_i - 0.55}{0.303} \qquad \text{and} \qquad z_{y_i} = \frac{y_i - 5.52}{3.04}\,.$$

The results of normalization are presented in table 2.

Then, apply the linear fit method to the data presented in table 2. The result of such a procedure is the equation for the linear fit:

$$z_y = -0.9975 z_x \,.$$

Transform next the standard scores into the raw scores:

$$z_y = -0.9975 z_x \implies \frac{y - \langle y \rangle}{s_y} = -0.9975 \frac{x - \langle x \rangle}{s_x} \implies \frac{y - 5.52}{3.04} = -0.9975 \frac{x - 0.55}{0.303} \,,$$

which finally leads to the linear fit function

$$f[x] = 11.0x - 10.0 \,.$$

This result agrees with the one found with non-normalized data.

22.5. **Problem Statement in Vector Notation.** Given a set of m empirical datum pairs of independent and dependent variables, $[x_i, y_i]$, optimize the parameters $\{\beta\}$ of the model curve $f[x_i, \beta_i]$ so that the sum of the squares of the deviations

$$S[\beta_i] = \sum_{i=1}^{m} (y_i - f[x_i, \beta_i])^2$$

becomes minimal.

The condition to be minimal is

$$\frac{\partial S[\beta_i]}{\partial \beta_i} = 0 \,.$$

In vector notation, the problem statement can be changed into vector notation as

$$S[\beta] = \sum_{i=1}^{m} (y - f[x, \beta])^2 \,,$$

where β represents the vector of the parameters, x the vector of the independent variables and y the vector of the dependent variables.

The solution of the problem is then

$$\frac{\partial S}{\partial \beta}[\beta] = 0 \,.$$

In index notation, the solution of the problem can be written as

$$\frac{\partial S}{\partial \beta_i}[\beta_i] = 0 \,,$$

which becomes into

$$\frac{\partial}{\partial \beta_i} \sum_i (y_i - f[x_i, \beta_i])^2 = 2 \sum_i (y_i - f[x_i, \beta_i]) \frac{\partial f}{\partial \beta_i}[x_i, \beta_i] = 0 \,.$$

The last equation leads to the system of equations:

$$\sum_i (y_i - f[x_i, \beta_i]) \frac{\partial f}{\partial \beta_i}[x_i, \beta_i] = 0 \,.$$

23. Uncertainties

[A Beginner's Guide to Uncertainty of Measurement, Stephanie Bell]

23.1. **Basic Definitions.** What is a measurement? A measurement tells us about a property of something. It might tell us how heavy an object is, or how hot, or how long it is. A measurement gives a number to that property. Measurements are always made using an instrument of some kind. Rulers, stopwatches, weighing scales, and thermometers are all measuring instruments. The result of a measurement is normally in two parts: a number and a unit of measurement, *e.g.*. 'How long is it? ... 2 m'.

What is not a measurement? There are some processes that might seem to be measurements, but are not. For example, comparing two pieces of string to see which is longer is not really a measurement. Counting is not normally viewed as a measurement. Often, a test is not a measurement: tests normally lead to a 'yes/no' answer or a 'pass/fail' result. (However, measurements may be part of the process leading up to a test result.)

What is uncertainty of measurement? The uncertainty of a measurement tells us something about its quality.

> Uncertainty of measurement is the doubt that exists about the result of any measurement.

You might think that well-made rulers, clocks and thermometers should be trustworthy, and give the right answers. But for every measurement – even the most careful – there is always a margin of doubt. In everyday speech, this might be expressed as 'give or take': *e.g.*, a stick might be two meters long 'give or take a centimeter'.

23.1.1. *Expressing uncertainty of measurement.* Since there is always a margin of doubt about any measurement, we need to ask 'How big is the margin?' and 'How bad is the doubt?' Thus, two numbers are really needed in order to quantify an uncertainty. One is the width of the margin, or interval. The other is a confidence level, and states how sure we are that the 'true value' is within that margin. For example: We might say that the length of a certain stick measures 20 cm plus or minus 1 cm, at the 95 percent confidence level. This result could be written: $20\,\text{cm} \pm 1\,\text{cm}$, at a level of confidence of 95%.

The statement says that we are 95 percent sure that the stick is between 19 cm and 21 cm long. There are other ways to state confidence levels.

23.1.2. *Error versus uncertainty.* It is important not to confuse the terms 'error' and 'uncertainty'.

Error is the difference between the measured value and the 'true value' of the thing being measured.

Uncertainty is a quantification of the doubt about the measurement result.

Whenever possible we try to correct for any known errors: for example, by applying corrections from calibration certificates. But any error whose value we do not know is a source of uncertainty.

23.2. **How many readings should you average?** Broadly speaking, the more measurements you use, the better the estimate you will have of the 'true' value. The ideal would be to find the mean from an infinite set of values. The more results you use, the closer you get to that ideal estimate of the mean. But performing more readings takes extra effort, and yields 'diminishing returns'. What is a good number? Ten is a popular choice because it makes the arithmetic easy. Using 20 would only give a slightly better estimate than 10. Using 50 would be only slightly better than 20. As a rule of thumb usually between 4 and 10 readings is sufficient.

23.3. **How many readings do you need to find an estimated standard deviation?** Again, the more readings you use, the better the estimate will be. In this case it is the estimate of uncertainty that improves with the number of readings (not the estimate of the mean or 'end result'). In ordinary situations 10 readings is enough. For a more thorough estimate, the results should be adjusted to take into account the number of readings.

23.4. **Where do errors and uncertainties come from?** Many things can undermine a measurement. Flaws in the measurement may be visible or invisible. Because real measurements are never made under perfect conditions, errors and uncertainties can come from:

- The measuring instrument: instruments can suffer from errors including bias, changes due to aging, wear, or other kinds of drift, poor readability, noise (for electrical instruments) and many other problems.
- The item being measured: which may not be stable. (Imagine trying to measure the size of an ice cube in a warm room.)
- The measurement process: the measurement itself may be difficult to make. For example measuring the weight of small but lively animals presents particular difficulties in getting the subjects to co-operate.
- 'Imported' uncertainties: calibration of your instrument has an uncertainty which is then built into the uncertainty of the measurements you make. (But remember that the uncertainty due to not calibrating would be much worse.)
- Operator skill: some measurements depend on the skill and judgment of the operator. One person may be better than another at the delicate work of setting up a measurement, or at reading fine detail by eye. The use of an instrument such as a stopwatch depends on the reaction time of the operator. (But gross mistakes are a different matter and are not to be accounted for as uncertainties.)
- Sampling issues: the measurements you make must be properly representative of the process you are trying to assess. If you want to know the temperature at the work-bench, don't measure it with a thermometer placed on the wall near an air conditioning outlet. If you are choosing samples from a production line for measurement, don't always take the first ten made on a Monday morning.
- The environment: temperature, air pressure, humidity and many other conditions can affect the measuring instrument or the item being measured.

Where the size and effect of an error are known (*e.g.*. from a calibration certificate) a correction can be applied to the measurement result. But, in general, uncertainties from each of these sources, and from other sources, would be individual 'inputs' contributing to the overall uncertainty in the measurement.

23.5. **The general kinds of uncertainty in any measurement.**

23.5.1. *Random or systematic.* The effects that give rise to uncertainty in measurement can be either:

- random: where repeating the measurement gives a randomly different result. If so, the more measurements you make, and then average, the better estimate you generally can expect to get.
- systematic: where the same influence affects the result for each of the repeated measurements (but you may not be able to tell). In this case, you learn nothing extra just by repeating measurements. Other methods are needed to estimate uncertainties due to systematic effects, *e.g.*. different measurements, or calculations.

23.5.2. *Distribution - the 'shape' of the errors.* The spread of a set of values can take different forms, or probability distributions.

Normal distribution: In a set of readings, sometimes the values are more likely to fall near the average than further away. This is typical of a normal or Gaussian distribution. You might see this type of distribution if you examined the heights of individuals in a large group of men. Most men are close to average height; few are extremely tall or short.

Figure 2 shows a set of 10 'random' values in an approximately normal distribution. A sketch of a normal distribution is shown in Figure 3.

Uniform or rectangular distribution: When the measurements are quite evenly spread between the highest and the lowest values, a rectangular or uniform distribution is produced. This would be seen if you examined how rain drops fall on a thin, straight telephone wire, for example. They would be as likely to fall on any one part as on another.

Figure 4 shows a set of 10 'random' values in an approximately rectangular distribution. A sketch of a rectangular distribution is shown in Figure 5.

23.6. **How to calculate uncertainty of measurement.** To calculate the uncertainty of a measurement, firstly you must identify the sources of uncertainty in the measurement. Then you must estimate the size of the uncertainty from each source. Finally the individual uncertainties are combined to give an overall figure.

There are clear rules for assessing the contribution from each uncertainty, and for combining these together.

23.6.1. *The two ways to estimate uncertainties.* No matter what are the sources of your uncertainties, there are two approaches to estimating them: 'Type A' and 'Type B' evaluations. In most measurement situations, uncertainty evaluations of both types are needed.

(1) Type A evaluations: uncertainty estimates using statistics (usually from repeated readings).
(2) Type B evaluations: uncertainty estimates from any other information. This could be information from past experience of the measurements, from calibration certificates, manufacturer's specifications, from calculations, from published information, and from common sense.

There is a temptation to think of 'Type A' as 'random' and 'Type B' as 'systematic', but this is *not* necessarily true.

How to use the information from Type A and Type B evaluations is described below.

23.7. **Eight main steps to evaluating uncertainty.** The main steps to evaluating the overall uncertainty of a measurement are as follows.

(1) Decide what you need to find out from your measurements. Decide what actual measurements and calculations are needed to produce the final result.
(2) Carry out the measurements needed.
(3) Estimate the uncertainty of each input quantity that feeds into the final result. Express all uncertainties in similar terms. (See below).
(4) Decide whether the errors of the input quantities are independent of each other. If you think not, then some extra calculations or information are needed. (See correlation below.)
(5) Calculate the result of your measurement (including any known corrections for things such as calibration).
(6) Find the combined standard uncertainty from all the individual aspects. (See below.)
(7) Express the uncertainty in terms of a coverage factor (see below), together with a size of the uncertainty interval, and state a level of confidence.
(8) Write down the measurement result and the uncertainty, and state how you got both of these. (See below.)

23.8. **Other things you should know before making an uncertainty calculation.** Uncertainty contributions must be expressed in similar terms before they are combined. Thus, all the uncertainties must be given in the same units, and at the same level of confidence.

23.8.1. *Standard uncertainty.* All contributing uncertainties should be expressed at the same confidence level, by converting them into standard uncertainties. A standard uncertainty is a margin whose size can be thought of as 'plus or minus one standard deviation'. The standard uncertainty tells us about the uncertainty of an average (not just about the spread of values). A standard uncertainty is usually shown by the symbol u (small u), or $u[y]$ (the standard uncertainty in y).

- Calculating standard uncertainty for a Type A evaluation: When a set of several repeated readings has been taken (for a Type A estimate of uncertainty), the mean, $\langle x \rangle$, and estimated standard deviation, s, can be calculated for the set. From these, the estimated standard uncertainty, u, of the mean is calculated from:

$$u = \frac{s}{\sqrt{n}} ,$$

where n was the number of measurements in the set. (The standard uncertainty of the mean has historically also been called the standard deviation of the mean, or the standard error of the mean.)

- Calculating standard uncertainty for a Type B evaluation: Where the information is more scarce (in some Type B estimates), you might only be able to estimate the upper and lower limits of uncertainty. You may then have to assume the value is equally likely to fall anywhere in between, *i.e.*, a rectangular or uniform distribution. The standard uncertainty for a rectangular distribution is found from:

$$\frac{a}{\sqrt{3}}\,,$$

where a is the semi-range (or half-width) between the upper and lower limits.

Rectangular or uniform distributions occur quite commonly, but if you have good reason to expect some other distribution, then you should base your calculation on that. For example, you can usually assume that uncertainties 'imported' from the calibration certificate for a measuring instrument are normally distributed.

- Converting uncertainties from one unit of measurement to another: Uncertainty contributions must be in the same units before they are combined. As the saying goes, you cannot 'compare apples with pears'.

23.9. **Combining standard uncertainties.** Individual standard uncertainties calculated by Type A or Type B evaluations can be combined validly by 'summation in quadrature' (also known as 'root sum of the squares'). The result of this is called the combined standard uncertainty, shown by u_c or $u_c[y]$.

Summation in quadrature is simplest where the result of a measurement is reached by addition or subtraction. The more complicated cases are also covered below for the multiplication and division of measurements, as well as for other functions.

23.10. **Correlation.** The equations given above to calculate the combined standard uncertainty are only correct if the input standard uncertainties are not inter-related or correlated. This means we usually need to question whether all the uncertainty contributions are independent. Could a large error in one input cause a large error in another? Could some outside influence, such as temperature, have a similar effect on several aspects of uncertainty at once - visibly or invisibly? Often individual errors are independent. But if they are not, extra calculations are needed. These are not detailed here, but can be found in some of the further reading.

23.11. **Coverage factor k.** Having scaled the components of uncertainty consistently, to find the combined standard uncertainty, we may then want to re-scale the result. The combined standard uncertainty may be thought of as equivalent to 'one standard deviation', but we may wish to have an overall uncertainty stated at another level of confidence, *e.g.*, 95 percent. This re-scaling can be done using a coverage factor, k. Multiplying the combined standard uncertainty, u_c, by a coverage factor gives a result which is called the expanded uncertainty, usually shown by the symbol U; *i.e.*,

$$U = ku_c\,.$$

A particular value of coverage factor gives a particular confidence level for the expanded uncertainty.

Most commonly, we scale the overall uncertainty by using the coverage factor $k = 2$, to give a level of confidence of approximately 95 percent. ($k = 2$ is correct if the combined standard uncertainty is normally distributed. This is usually a fair assumption, but the reasoning behind this is explained elsewhere, in the references.)

Some other coverage factors (for a normal distribution) are:

- $k = 1$ for a confidence level of approximately 68 percent.
- $k = 2.58$ for a confidence level of 99 percent.
- $k = 3$ for a confidence level of 99.7 percent.

Other, less common, shapes of distribution have different coverage factors.

Conversely, wherever an expanded uncertainty is quoted with a given coverage factor, you can find the standard uncertainty by the reverse process, *i.e.*, by dividing by the

appropriate coverage factor. This means that expanded uncertainties given on calibration certificates, if properly expressed, can be 'decoded' into standard uncertainties.

23.12. **How to express the answer.** It is important to express the answer so that a reader can use the information. The main things to mention are:

- The measurement result, together with the uncertainty figure, *e.g.*, 'The length of the stick was 20 cm ± 1 cm' or 'The length of the stick was 20(1) cm'.
- The statement of the coverage factor and the level of confidence. A recommended wording is: 'The reported uncertainty is based on a standard uncertainty multiplied by a coverage factor $k = 2$, providing a level of confidence of approximately 95%'.
- How the uncertainty was estimated (you could refer to a publication where the method is described, *e.g.*, UKAS Publication M 3003).

23.13. **Example - a basic calculation of uncertainty.** Below is a worked example of a simple uncertainty analysis. It is not realistic in every detail, but it is meant to be simple and clear enough to illustrate the method. First the measurement and the analysis of uncertainty are described. Secondly, the uncertainty analysis is shown in a table (a 'spreadsheet model' or 'uncertainty budget').

23.13.1. *The measurement: how long is a piece of string?* Suppose you need to make a careful estimate of the length of a piece of string. Following the steps listed in above, the process is as follows.

Step 1. Decide what you need to find out from your measurements. Decide what actual measurements and calculations are needed to produce the final result. You need to make a measurement of the length, using a tape measure. Apart from the actual length reading on the tape measure, you may need to consider:

- Possible errors of the tape measure: Does it need any correction, or has calibration shown it to read correctly and what is the uncertainty in the calibration? Is the tape prone to stretching? Could bending have shortened it? How much could it have changed since it was calibrated? What is the resolution, *i.e.* how small are the divisions on the tape (*e.g.*, milimetres)?
- Possible errors due to the item being measured: Does the string lie straight? Is it under- or over-stretched? Does the prevailing temperature or humidity (or anything else) affect its actual length? Are the ends of the string well-defined, or are they frayed?
- Possible errors due to the measuring process, and the person making the measurement: How well can you line up the beginning of the string with the beginning of the tape measure? Can the tape be laid properly parallel with the string? How repeatable is the measurement? Can you think of any others?

Step 2. Carry out the measurements needed. You make and record your measurements of length. To be extra thorough, you repeat the measurement a total of 10 times, aligning the tape measure freshly each time (probably not very likely in reality!). Let us suppose you calculate the mean to be 5.017 m, and the estimated standard deviation to be 0.0021 m (*i.e.*, 2.1 mm).

For a careful measurement you might also record:

- when you did it;
- how you did it, *e.g.*, along the ground or vertically, reversing the tape measure or not, and other details of how you aligned the tape with the string - which tape measure you used;
- environmental conditions (if you think these could affect your results);
- anything else that could be relevant.

Step 3. Estimate the uncertainty of each input quantity that feeds into the final result. Express all uncertainties in similar terms (standard uncertainty, u). You would look at all the possible sources of uncertainty and estimate the magnitude of each. Let us say that in this case:

- The tape measure has been calibrated. It needs no correction, but the calibration uncertainty is 0.1 percent of reading, at a coverage factor $k = 2$ (for a normal

distribution). In this case, 0.1 percent of 5.017 m is close to 5 mm. Dividing by 2 gives the standard uncertainty (for $k = 1$) to be $u = 2.5$ mm.

- The divisions on the tape are millimetres. Reading to the nearest division gives an error of no more than ± 0.5 mm. We can take this to be a uniformly distributed uncertainty (the true readings could lie variously anywhere in the 1 mm interval; *i.e.*, ± 0.5 mm). To find the standard uncertainty, u, we divide the half-width (0.5 mm) by $\sqrt{3}$, giving $u = 0.3$ mm, approximately.
- The tape lies straight, but let us suppose the string unavoidably has a few slight bends in it. Therefore the measurement is likely to underestimate the actual length of the string. Let us guess that the underestimate is about 0.2 percent, and that the uncertainty in this is also 0.2 percent at most. That means we should correct the result by adding 0.2 percent (*i.e.*, 10 mm). The uncertainty is assumed to be uniformly distributed, in the absence of better information. Dividing the half-width of the uncertainty (10 mm) by $\sqrt{3}$ gives the standard uncertainty $u = 5.8$ mm (to the nearest 0.1 mm).

The above are all Type B estimates. Below is a Type A estimate.

The standard deviation tells us about how repeatable the placement of the tape measure is, and how much this contributes to the uncertainty of the mean value. The estimated standard deviation of the mean of the 10 readings is found using the formula above:

$$\frac{s}{\sqrt{n}} = \frac{2.1}{10} = 0.7\,\text{mm} \qquad [\text{to one decimal place}]\,.$$

Let us suppose that no other uncertainties need to be counted in this example. (In reality, other things would probably need to be included.)

Step 4. Decide whether the errors of the input quantities are independent of each other. (If you think not, then some extra calculations or information are needed.) In this case, let us say that they are all independent.

Step 5. Calculate the result of your measurement (including any known corrections for things such as calibration). The result comes from the mean reading, together with the correction needed for the string lying slightly crookedly, *i.e.*,

$$5.017\,\text{m} + 0.010\,\text{m} = 5.027\,\text{m}\,.$$

Step 6. Find the combined standard uncertainty from all the individual aspects. The only calculation used in finding the result was the addition of a correction, so summation in quadrature can be used in its simplest form (using the equation above). The standard uncertainties are combined as

$$u_c = \sqrt{2.5^2 + 0.3^2 + 5.8^2 + 0.7^2} = 6.4\,\text{mm} \qquad [\text{to one decimal place}]\,.$$

Step 7. Express the uncertainty in terms of a coverage factor (see above), together with a size of the uncertainty interval, and state a level of confidence. For a coverage factor $k = 2$, multiply the combined standard uncertainty by 2, to give an expanded uncertainty of 12.8 mm (*i.e.*, 0.0128 m). This gives a level of confidence of about 95 percent.

Step 8. Write down the measurement result and the uncertainty, and state how you got both of these. You might record:

> The length of the string was 5.027(13) m. The reported expanded uncertainty is based on a standard uncertainty multiplied by a coverage factor $k = 2$, providing a level of confidence of approximately 95%.
>
> The reported length is the mean of 10 repeated measurements of the string laid horizontally. The result is corrected for the estimated effect of the string not lying completely straight when measured. The uncertainty was estimated according to the method in 'A Beginner's Guide to Uncertainty of Measurement'.

23.13.2. *Analysis of uncertainty: spreadsheet model.* To help in the process of calculation, it can be useful to summarize the uncertainty analysis or 'uncertainty budget' in a spreadsheet as in table 3 below.

Source of uncertainty	Value (±)	Prob. distribution	Divisor	Std. uncertainty
Calibration uncertainty	5.0 mm	Normal	2	2.5 mm
Resolution (size of divisions)	0.5 mm*	Rectangular	$\sqrt{3}$	0.3 mm
String not lying perfectly straight	10.0 mm*	Rectangular	$\sqrt{3}$	5.8 mm
Std. uncer. of mean (10 repeated reads)	0.7 mm	Normal	1	0.7 mm
Combined standard uncertainty		Assumed normal		6.4 mm
Expanded uncertainty		Assumed normal ($k = 2$)		12.8 mm

TABLE 3. Spreadsheet model showing the 'uncertainty budget'. The * means that the (±) half-width divided by $\sqrt{3}$ was used.

23.14. **How to reduce uncertainty in measurement.** Always remember that it is usually as important to minimize uncertainties as it is to quantify them. There are some good practices which can help to reduce uncertainties in making measurements generally. A few recommendations are:

- Calibrate measuring instruments (or have them calibrated for you) and use the calibration corrections which are given on the certificate.
- Make corrections to compensate for any (other) errors you know about.
- Make your measurements traceable to national standards – by using calibrations which can be traced to national standards via an unbroken chain of measurements. You can place particular confidence in measurement traceability if the measurements are quality-assured through a measurement accreditation (UKAS in the UK).
- Choose the best measuring instruments, and use calibration facilities with the smallest uncertainties.
- Check measurements by repeating them, or by getting someone else to repeat them from time to time, or use other kinds of checks. Checking by a different method may be best of all.
- Check calculations, and where numbers are copied from one place to another, check this too.
- Use an uncertainty budget to identify the worst uncertainties, and address these.
- Be aware that in a successive chain of calibrations, the uncertainty increases at every step of the chain.

23.15. **Some other good measurement practices.** Overall, use recognized good practices in measurements, for example:

- Follow the maker's instructions for using and maintaining instruments.
- Use experienced staff, and provide training for measurement.
- Check or validate software, to make sure it works correctly.
- Use rounding correctly in your calculations.
- Keep good records of your measurements and calculations. Write down readings at the time they are made. Keep a note of any extra information that may be relevant. If past measurements are ever called into doubt, such records can be very useful.

Many more good measurement practices are detailed elsewhere, for example in the international standard ISO/IEC 17025 'General requirements for the competence of testing and calibration laboratories'.

23.16. **Rounding.** Calculators and spreadsheets can give an answer to many decimal places. There are some recommended practices for rounding the results:

- Use a meaningful degree of rounding in calculations. The uncertainty in a measurement result may define how many decimal places you should report. For example, if the uncertainty in your result is in the first decimal place, then the measurement result should probably also be stated to one decimal place, *e.g.*,

$$20.1(2)\,\mathrm{cm}\,.$$

- Make your calculations to at least one more significant figure than you eventually require. Be aware of how many significant figures you need to use when multiplying or dividing or carrying out more complex calculations.
- Rounding of values should be carried out only at the end of the calculation, to avoid rounding errors. For example, if 2.346 is rounded up to 2.35 at an early stage in a calculation, it could later be rounded up to 2.4. But if 2.346 is used throughout a calculation it would be correctly rounded to 2.3 at the *final* stage.
- Although results are finally rounded either up or down, depending on which is the nearest figure, the rule for rounding uncertainties is different.

  The final uncertainty is rounded up to the next largest figure, not down.

23.17. **Words of warning.** Uncertainty analysis is an evolving subject area. There have been subtle changes in approach over the years. What is more, the rules given in this Beginner's Guide are not 'absolute'. There are plenty of special cases where slightly different rules apply. There is even room for debate on the finer points of how to account for particular uncertainties. But still the advice given in this publication represents normal good practice.

What is given here is not the full story. Special cases have not been dealt with in this Guide. Extra rules apply:

- if you use statistics on very small sets of data (less than about 10),
- if one component of uncertainty is much bigger than all the others involved,
- if some inputs to the calculation are correlated,
- if the spread or distribution is unusual in shape,
- if the uncertainty is not for a single result, but for fitting a curve or line to a number of points.

These cases are covered by some other texts.

24. Propagation of Uncertainties

[Analysis of experiments for the physical sciences, J. Mitroy]

24.1. **The need for uncertainties.** The whole structure and application of science depends on measurements which, however carefully made, are subject to uncertainties, or "errors". While it can take a great deal of knowledge and expertise to obtain accurate data it is just as important to analyze the data correctly. A significant part of the analysis of any experiment is to identify sources of uncertainties, reduce them whenever possible and determine the overall accuracy of the results. Unfortunately this aspect of scientific work is not treated with the importance it deserves, it is often done carelessly and in the worst cases ignored.

In order to know whether two measurements of a physical quantity, say the local value of the gravitational acceleration, are the same, some idea of the reliability of each measurement is needed. Suppose the values of g obtained using a sophisticated experiment that uses a laser to time the fall of a body down an evacuated tube are $9.7831(2)\,\mathrm{m/s^2}$ and $9.7848(2)\,\mathrm{m/s^2}$. Even though the difference between the two values of g is very small, being only $0.0017\,\mathrm{m/s^2}$, this difference is much larger than the stated uncertainty in the experiment and you would have to conclude that the results were mutually incompatible. On the other hand, we could hang a metal bob on the end of a string, and use this as a pendulum to measure g. If we obtained $9.97(18)\ \mathrm{m/s^2}$ and $9.80(20)\,\mathrm{m/s^2}$ for two measurements of g, we would conclude that the measurements were consistent with each other. Despite the fact that the difference between the two measurements is quite large, $0.17\,\mathrm{m/s^2}$, the two results lie within their mutual uncertainties so they are compatible. The moral of the story is simple,

> unless the accuracy of the individual measurements is known, it is impossible to make a sensible comparison between the two measurements.

The basic aim of most experiments is to obtain a value for a quantity, which I will call x, and determine its uncertainty $\mathrm{d}x$, so that the result can be stated as

$$x_{\mathrm{best}} \pm \mathrm{d}x\,.$$

This statement means that the experimenter's best estimate for the quantity is the value x_{best}, and in addition denotes the fact that the value of x could in addition lie somewhere between $(x_{\mathrm{best}} - \mathrm{d}x)$ and $(x_{\mathrm{best}} + \mathrm{d}x)$. The standard convention is to define the uncertainty $\mathrm{d}x$ to be positive so $(x_{\mathrm{best}} + \mathrm{d}x)$ is always the highest probable value of the quantity and $(x_{\mathrm{best}} - \mathrm{d}x)$ the lowest. In this chapter various sources of uncertainties will be discussed, then we will consider how the various uncertainties combine to give the final uncertainty. In the sections that follow, we use x rather than the cumbersome expression x_{best} to denote the best estimate of the experimental results, and we use the best estimate of x when evaluating expressions.

For low precision work, it is acceptable to quote the uncertainty with one significant figure. In high precision work uncertainties are stated with two significant figures. As will be seen below, there is no point in quoting uncertainties with three significant figures.

24.2. **Relative and absolute uncertainties.** There are two ways the uncertainty in an experimental quantity can be expressed. First as an absolute uncertainty, and second as a relative uncertainty. When an experimental measurement is written in the form

$$x \pm \mathrm{d}x$$

the quantity, $\mathrm{d}x$ is the absolute uncertainty and is always positive by convention.

It is also possible to express the uncertainty as a ratio. The relative (or fractional) uncertainty is defined as

$$\mathrm{d}x_r = \frac{\mathrm{d}x}{|x|}\,.$$

The relative uncertainty is always a positive number and, since $\mathrm{d}x$ and x have the same dimensions, the relative uncertainty, (or fractional uncertainty) is *dimensionless*.

The relative uncertainty is sometimes multiplied by 100 and quoted as a percentage error. While it is sometimes convenient to use the term "percentage error" when discussing the sizes of relative error, the use of percentage errors in written reports is discouraged.

In a number of situations, such as when combining uncertainties from two different measurements, it is often more convenient to work with the relative uncertainty rather than the absolute uncertainty.

24.3. **Uncertainties in Experimental Measurements.** The theory which estimates the uncertainties in experimental results is sometimes called the "Theory of Errors". In the context of data analysis, the concept of error is used for several different things, such as spread of data resulting from random fluctuations in the length of a scale due to temperature changes, systematic errors resulting from the calibration errors in an electronic balance, or to a spread in data due to sloppy measurement technique. The term error is unfortunate, since it would seem to imply that experiments are plagued by mistakes whereas this is not always the case. A better term to use is uncertainty, and this will be used throughout this chapter and succeeding chapters. Uncertainties likely to be encountered arise from a number of causes including,

- mistakes, gross uncertainties,
- reading uncertainties,
- calibration errors,
- random uncertainties and
- systematic or regular uncertainties.

The first three items discuss some of the physical reasons causes result in uncertainties. The last two items divide the uncertainties into the two classes of random errors and systematic errors. As will be discussed in below, different procedures should be used when either random errors or statistical errors have to be combined to determine the total uncertainty.

24.4. **Estimating the uncertainty in a single measurement.** Having talked about the types of uncertainties that can arise, some discussion of how these uncertainties can be estimated immediately arises. There are no hard rules that can be used but the following methods are useful. Read The Manual or RTFM (as often abbreviated on the internet). If you are using an instrument such as a voltmeter or a Cathode Ray Oscilloscope or something more complicated it will likely come with an instruction manual. Somewhere in the manual will be a description of the reading error or the calibration error. Some simple instruments like rulers do not come with instruction manuals, but it is expected that students should be able to estimate the reading error of a ruler.

One way to estimate the error is to repeat the measurement. If you are measuring the elapsed time for some process with a stopwatch, take a series of additional measurements for identical experimental conditions. The spread in the different measurements of the elapsed time will give you some idea of the uncertainty. If you are measuring the width of the bar, you might consider taking measurements at a number of different points along the length of the bar. When you have made a series of measurements, you might take the arithmetic mean as your best estimate for the experimental parameter. You also might take

$$\mathrm{d}x = \frac{|x_{\text{biggest}} - x_{\text{smallest}}|}{2}$$

as your *estimate* for the uncertainty in x.

24.5. **Precision versus accuracy.** When the words precision and accuracy are used in everyday language, they are usually interpreted as having the same meaning. (The word-processor used to create this manuscript has a built-in thesaurus. In this thesaurus, precision and accuracy are listed as synonyms).

In the sciences these two words have different meanings. The best way to explain this is with an example. Suppose a series of measurements have been made with a Voltmeter which has a display showing 4 digits. However, according to the manual, the experimental uncertainty of the voltmeter might be $\pm 1.5\,\%$. The precision of the reading refers to the accuracy of the voltmeter display (4 digits), while the accuracy of the reading is only $\pm 1.5\,\%$.

24.6. **Propagation of Uncertainties.** After a number of direct measurements the final experimental result is usually calculated, via a graph and/or mathematical expression which uses all the measured quantities. The uncertainty in the final result must then be determined from measured uncertainties, that is, we must find out how measured uncertainties "propagate" through the calculations to produce the uncertainty in the final answer.

Note: in Stephanie Bell's book, "A Beginner's Guide to Uncertainty of Measurement", this is called "Combining standard uncertainties".

24.6.1. *Sums and Differences.* Suppose we have measured quantities x, y, with corresponding uncertainties $\mathrm{d}x$, $\mathrm{d}y$, and we wish to know the uncertainty in p where $p = x + y$. To estimate the uncertainty in p, it is only necessary to decide the highest and lowest values.

Given $p = x + y$ the highest probable value of p is $p = (x + y) + (\mathrm{d}x + \mathrm{d}y)$ and the lowest probable value is of p is $p = (x + y) - (\mathrm{d}x + \mathrm{d}y)$. The best estimate of p is $p = x + y$ and its uncertainty is $\mathrm{d}p = \mathrm{d}x + \mathrm{d}y$.

If p is a function of n variables, $\{x_i\}$, then it is easy to generalize the last argument to

$$p = \left(\sum_{i=1}^{n} x_i\right) \pm \left(\sum_{i=1}^{n} \mathrm{d}x_i\right) .$$

24.6.2. *Arbitrary functions (use of differential calculus).* When more complicated relationships such as log, sin, *etc*, occur uncertainties can be calculated by the application of differential calculus. If x is measured with an uncertainty $\mathrm{d}x$ and is used to calculate a function $p = p[x]$, then the uncertainty $\mathrm{d}p$ can be derived from differential calculus. Using Taylor's theorem to expand the function $p[x]$ in the neighborhood of a point, x

$$p[x + \mathrm{d}x] = p[x] + \frac{\mathrm{d}p}{\mathrm{d}x}\,\mathrm{d}x + \cdots .$$

Rearranging this, it is easy to see that

$$p[x + \mathrm{d}x] - p[x] = \frac{\mathrm{d}p}{\mathrm{d}x}\,\mathrm{d}x + \cdots .$$

Interpreting, $(p[x + \mathrm{d}x] - p[x])$ as the change in p resulting from a small change in x, gives

$$\mathrm{d}p = \left|\frac{\mathrm{d}p}{\mathrm{d}x}\right|\mathrm{d}x\,.$$

(Sometimes $\mathrm{d}p/\mathrm{d}x$ is written as $p'[x]$). A graphical depiction showing how a small change in x leads to a change in p is shown in Figure. The change in p can be determined from the slope of the graph at the point of interest.

Example. A classic method of determining the height of a tall building involves a barometer. The method involves the experimenter taking the barometer to the top of the building, dropping it off the side and recording the time it takes to hit the ground. According to standard theory, the time taken for the barometer to reach the base is related to the height by the equation, $2h = gt^2$. Given that the local value of $g = 9.783(1)\,\mathrm{m/s^2}$ and the elapsed time is $3.4(1)\,\mathrm{s}$, what is the height of the building and its uncertainty?

Solution. Replace the values of g and t in the expression for h to have

$$h = \frac{1}{2}gt^2 = \frac{1}{2}9.783 \times 3.4^2 = 56.5457\,\mathrm{m}\,.$$

To find the absolute uncertainty in h, derivate the expression for h using the product rule for derivatives:

$$2\mathrm{d}h = \mathrm{d}g t^2 + g2t\mathrm{d}t \implies \mathrm{d}h = \frac{1}{2}t^2\mathrm{d}g + gt\mathrm{d}t\,.$$

Replace numeric values for the quantities in the last equation to find

$$\mathrm{d}h = (0.5)(3.4)^2(0.001) + (9.783)(3.4)(0.1) = 3.332\,\mathrm{m}\,.$$

Therefore, the height of the building can be expressed as

$$h = 56.5(33)\,\mathrm{m}\,.$$ □

An alternative solution uses the relative uncertainty instead of the absolute uncertainty:

Solution. Find $\mathrm{d}h$ as in the previous solution,

$$\mathrm{d}h = \frac{1}{2}t^2\mathrm{d}g + gt\mathrm{d}t\,.$$

Then, divide the last equation by $2h = gt^2$ to have the relative uncertainty:

$$\frac{\mathrm{d}h}{h} = \frac{\mathrm{d}g}{g} + 2\frac{\mathrm{d}t}{t}\,.$$

Note that the last equation is easier to apply and uses less operations to yield $\mathrm{d}h$; *i.e.*, it's less error prone!

Next, replace numeric values in the last equation to yield:

$$\frac{\mathrm{d}h}{h} = \frac{0.001}{9.783} + 2\frac{0.1}{3.4} = 0.0589257 \implies \mathrm{d}h = (56.5457)(0.0589257) = 3.331997 \sim 3.3\,\mathrm{m}\,,$$

which agrees with the previous result.

Example. Suppose the diagonal length of a rectangle, c, has to be computed from the horizontal and vertical dimensions. Given $a = 0.760(1)\,\mathrm{m}$ and $b = 0.246(1)\,\mathrm{m}$, compute c using the identity $c^2 = a^2 + b^2$ and determine uncertainty in c.

Solution. Calculate c as

$$c = \sqrt{a^2 + b^2} = \sqrt{0.760^2 + 0.246^2} = 0.79882\,.$$

Then, find the absolute uncertainty by

$$c\mathrm{d}c = a\mathrm{d}a + b\mathrm{d}b \implies \mathrm{d}c = \frac{a}{c}\mathrm{d}a + \frac{b}{c}\mathrm{d}b \implies \mathrm{d}c = \frac{0.760}{0.79882}0.001 + \frac{0.246}{0.79882}0.001 \sim 0.001259\,.$$

Therefore, the diagonal length of the rectangle is

$$c = 0.7988(12)\,\mathrm{m}\,.$$ □

24.6.3. *General Expression for Error Propagation.* In this section a general technique will be developed from which all the previous rules can be derived. The general technique discussed in this section is often easier to apply in situations where there are quite complicated functional relations between the experimental variables. The general expression avoids the calculation of the uncertainty in a number of steps.

Suppose two quantities x and y have been measured and then used to calculate some function $p = p[x, y]$. This function could be as simple as $p[x, y] = x + y$ or something more complicated like $p[x, y] = \exp x + \log[y]$.

Using Taylor's theorem generalized to two dimensions

$$p[x + \mathrm{d}x, y + \mathrm{d}y] = p[x, y] + \frac{\partial p}{\partial x}\mathrm{d}x + \frac{\partial p}{\partial y}\mathrm{d}y + \cdots\,,$$

where $\mathrm{d}x$ and $\mathrm{d}y$ are any small changes in x and y and $p_{,x}$ and $p_{,y}$ are the partial derivatives of p with respect to x and y. Identifying the uncertainty in p with $(p[x + \mathrm{d}x, y + \mathrm{d}y] - p[x, y])$ we see that

$$\mathrm{d}p = \left|\frac{\partial p}{\partial x}\right|\mathrm{d}x + \left|\frac{\partial p}{\partial y}\right|\mathrm{d}y\,.$$

This result can be generalized to a system with n measured variables $\{x_i\}$ with uncertainties $\{\mathrm{d}x_i\}$. When the measured values $\{x_i\}$ are used to compute the function $p[x_i]$, then the general expression for the uncertainty is

$$\mathrm{d}p = \sum_{i=1}^{n}\left|\frac{\partial p}{\partial x_i}\right|\mathrm{d}x_i\,. \tag{24.1}$$

You should note that the last equation does not represent that last word in error formulae. When individual errors are statistical in nature, the manner in which errors are added together should be modified. A detailed discussion of this topic is postponed until a further section.

24.7. **Examples.** Find the volume of 0.25 mol of a gas at 200 kPa and 300 K.

24.7.1. *Volume calculation.* Assume the gas to be an ideal gas. Then, model its properties by the ideal gas law: $pv = nrt$. Using such a model, the volume the gas occupies becomes

$$v = \frac{nrt}{p} = \frac{(0.25)(8.3144621)(300)}{200} \sim 3.12\,\mathrm{L},$$

where the value for the molar gas constant, $r = 8.314\,462\,1(75)\,\mathrm{J/mol\,K}$, was taken from the NIST website.

24.7.2. *Estimation of uncertainties.* Since no uncertainties were given for the data but for r, assume the following:

- Type A uncertainties: molar gas constant: std. unc.: $\mathrm{d}r = 7.5 \times 10^{-6}\,\mathrm{J/mol\,K}$. (According to the NIST website, all the values for fundamental constants are quoted with standard uncertainties.)
- Type B uncertainties: all the rest of physical quantities are assumed to be rectangular distributed; *i.e.*, their standard uncertainties are given by $a/\sqrt{3}$, where a is an estimated uncertainty. Specifically, assume the following values for the standard uncertainties:
  - amount of gas, $\mathrm{d}n = 0.05/\sqrt{3} \sim 0.029\,\mathrm{mol}$,
  - pressure, $\mathrm{d}p = 1/\sqrt{3} \sim 0.58\,\mathrm{kPa}$,
  - thermodynamic temperature, $\mathrm{d}t = 1/\sqrt{3} \sim 0.58\,\mathrm{K}$.

Then, the combined uncertainty for v can be calculated by

$$\frac{\mathrm{d}v}{v} = \frac{\mathrm{d}n}{n} + \frac{\mathrm{d}r}{r} + \frac{\mathrm{d}t}{t} - \frac{\mathrm{d}p}{p}\,,$$

which comes from the application of eq. (24.1) divided by $v = nrt/p$.

Introduce numeric values in the last equation to have

$$\frac{\mathrm{d}v}{v} = \frac{0.029}{0.25} + \frac{7.5 \times 10^{-6}}{8.3144626} + \frac{0.58}{300} - \frac{0.58}{200}\,,$$

which yields the combined standard uncertainty for the volume: $\mathrm{d}v \sim 0.36\,\mathrm{L}$.

Next, use a coverage factor of 2 to find the expanded uncertainty for the volume with a confidence level of 95%; *i.e.*, $u_v = (2)(0.36) = 0.72\,\mathrm{L}$.

Finally, report the value for the volume and its uncertainty:

> The volume of 0.25 mol of a gas at 200 kPa and 300 K is 3.12(72) L. The reported value assumes the gas to be ideal and, thus, comes directly from the application of the ideal gas law.
>
> The reported expanded uncertainty, on the other hand, results from a combined standard uncertainty multiplied by a coverage factor of 2, providing then a level of confidence of *ca.* 95%.

25. Differential Geometry

25.1. **Comma and Semi-colon Derivatives.** Comma-derivative: The components of the gradient of the one-form $\mathrm{d}A$ are denoted $A_{,k}$, or sometimes $\partial_k A$, and are given by

$$A_{,k} = \partial_k A = \frac{\partial A}{\partial x^k} \,.$$

Semi-colon or covariant derivative: The covariant derivative of a *contravariant* tensor A^a (also called the "semicolon derivative" since its symbol is a semicolon) is given by

$$A^a{}_{;b} = \frac{\partial A^a}{\partial x^k} + \Gamma^a_{bk} A^k = A^a{}_{,b} + \Gamma^a_{bk} A^k \,,$$

where Γ^k_{ij} is a Christoffel symbol, Einstein summation has been used in the last term, and $A^a{}_{,b}$ is a comma derivative. The notation $\nabla \cdot A$, which is a generalization of the symbol commonly used to denote the divergence of a vector function in three dimensions, is sometimes also used.

The covariant derivative of a *covariant tensor* A_a is

$$A_{a\,;b} = \frac{\partial A_a}{\partial x^k} - \Gamma^a_{bk} A_k \,,$$

25.2. **Some Derivatives.** Consider a scalar field $f = f\big[t, x^k\big]$ (a scalar-valued function of the position vector). Then, the total time derivative of the scalar field, denoted $\dot{f}$, is defined by

$$\dot{f} = \partial_t f + \partial_i f \dot{x}^i \,.$$

Partial time derivative operator:

$$\partial_t = \frac{\partial}{\partial t} \,.$$

Partial spatial (coordinate) derivative operator:

$$\partial_k = \partial_{x^k} = \frac{\partial}{\partial x^k}.$$

Absolute derivative of a tensor field T upon the parameter t:

$$\mathrm{D}_t T = \dot{\overline{T}} = \frac{\mathrm{D}T}{\mathrm{d}t} = \frac{\nabla T}{\mathrm{d}t} \,.$$

25.3. **Continuity Equation.** Recall that the most important equation in fluid dynamics, as well as in general continuum mechanics, is the celebrated equation of continuity, (we explain the symbols in the following text)

$$\partial_t \rho + \operatorname{div}(\rho u) = 0 \,.$$

As a warm-up for turbulence, we will derive the continuity equation, starting from the mass conservation principle. Let $\mathrm{d}m$ denote an infinitesimal mass of a fluid particle. Then, using the absolute time derivative operator $\dot{} = \mathrm{D}t$, the mass conservation principle reads

$$\dot{\overline{\mathrm{d}m}} = 0 \,.$$

If we further introduce the fluid density $\rho = \mathrm{d}m/\mathrm{d}v$, where $\mathrm{d}v$ is an infinitesimal volume of a fluid particle, then the mass conservation principle can be rewritten as

$$\dot{\overline{\rho \mathrm{d}v}} = 0 \,,$$

which is the absolute derivative of a product, and therefore expands into

$$\dot{\rho}\mathrm{d}v + \rho\dot{\overline{\mathrm{d}v}} = 0 \,.$$

Now, as the fluid density $\rho = \rho\big[x^k, t\big]$ is a function of both time t and spatial coordinates x^k, for $k = 1, 2, 3$, that is, a scalar-field, its total time derivative $\dot{\rho}$ is defined by

$$\dot{\rho} = \partial_t \rho + \partial_{x^k} \rho \partial_t x^k = \partial_t \rho + \rho_{;k} u^k \,,$$

or, in vector form,

$$\dot{\rho} = \partial_t \rho + \operatorname{grad} \rho \cdot u \,,$$

where $u = u^k = u^k\big[x^k, t\big]$ is the velocity vector-field of the fluid.

Regarding $\overline{\dot{\mathrm{d}v}}$, the other term figuring in the absolute derivative of a product, we start by expanding an elementary volume $\mathrm{d}v$ along the sides $\left\{\mathrm{d}x^k_{(p)}, \mathrm{d}x^k_{(q)}, \mathrm{d}x^k_{(r)}\right\}$ of an elementary parallelepiped, as

$$\mathrm{d}v = \frac{1}{3!}\delta^{pqr}_{ijk}\,\mathrm{d}x^k_{(p)}\mathrm{d}x^k_{(q)}\mathrm{d}x^k_{(r)}\,, \qquad [i,j,k,p,q,r=1,2,3]$$

so that its absolute derivative becomes [maths here :)] which finally simplifies into

$$\overline{\dot{\mathrm{d}v}} = u^k{}_{;k}\mathrm{d}v = \mathrm{div}\,(u)\mathrm{d}v\,,$$

Substituting the products into the continuity equation gives

$$\overline{\dot{\rho\mathrm{d}v}} = \left(\partial_t\rho + \rho_{;k}u^k\right)\,\mathrm{d}v + \rho u^k{}_{;k}\mathrm{d}v = 0$$

As we are dealing with arbitrary fluid particles, then $\mathrm{d}v \neq 0$, so from the last equation follows

$$\partial_t\rho + \rho_{;k}u^k + \rho u^k{}_{;k} = \partial_t\rho + \left(\rho u^k\right)_{;k} = 0\,,$$

The last equation is the covariant form of the continuity equation, which in standard vector notation becomes

$$\partial_t\rho + \mathrm{div}(\rho u) = 0\,.$$

25.4. **Differential Forms.** Consider a set of coordinates $\left\{x^1,\ldots,x^n\right\}$ on a manifold $\mathcal{M}$ and consider the set

$$[\gamma_1,\ldots,\gamma_n] = \left[\frac{\partial}{\partial x^1},\ldots,\frac{\partial}{\partial x^n}\right]$$

be the basis for $T_m\mathcal{M}$.

Consider the set

$$\left[\gamma^1,\ldots,\gamma^n\right] = \left[\mathrm{d}x^1,\ldots,\mathrm{d}x^n\right]$$

be the dual basis for $T^*_m\mathcal{M}$.

At each point $m \in \mathcal{M}$, we can write a 2-form as

$$\Omega_m[v,w] = \Omega_{ij}[m]\,v^iw^j\,,$$

where

$$\Omega_{ij}[m] = \left[\frac{\partial}{\partial x^i},\frac{\partial}{\partial x^j}\right]\,.$$

25.5. **Algebra of Differential Forms.** Consider the forms $\left\{\mathrm{d}x^i,\mathrm{d}x^j\right\}$ and the scalar $\alpha \in \mathcal{R}$, then

- addition is commutative: $\mathrm{d}x^i + \mathrm{d}x^j = \mathrm{d}x^j + \mathrm{d}x^i$;
- multiplication by a scalar is commutative: $\alpha\mathrm{d}x^i = \mathrm{d}x^i\alpha$;
- subtraction: $\mathrm{d}x^i - \mathrm{d}x^j = \mathrm{d}x^i + (-1)\mathrm{d}x^j$;
- multiplication: $\mathrm{d}x^i\mathrm{d}x^j = 0$, when $i = j$, and $\mathrm{d}x^i\mathrm{d}x^j = -\mathrm{d}x^j\mathrm{d}x^i$, when $i \neq j$. Both conditions can be summarized by

  $$\left[\mathrm{d}x^i,\mathrm{d}x^j\right]_+ = 0\,;$$

  or, expanding the anti-commutator brackets:

  $$\mathrm{d}x^i\mathrm{d}x^j + \mathrm{d}x^j\mathrm{d}x^i = 0\,.$$

26. Heat Transfer

26.1. **Relation of heat transfer and thermodynamics.**

26.1.1. *First law for closed systems.* The result of applying the first law of thermodynamics – conservation of energy – to a closed system is

$$\dot{h} = \dot{w} + \dot{u}\,,$$

wherein $\dot{h}$ is the heat transfer rate, $\dot{w}$ the work transfer rate and $\dot{u}$ the internal energy change rate. All the terms in the last equation have dimensions of energy flow, $[E/T]$. The sign convention adopted is that $\dot{h}$ is positive ($\dot{h} > 0$) when heat is added to the system, $\dot{w} > 0$ when energy is taken away from the system and $\dot{u} > 0$ when the system's energy increases. There's also, however, another sign convention, IUPAC's and Planck's: all net energy transfers to the system are taken as positive, all net energy transfers from the system are negative. Finally, the most suitable way of finding the correct signs of the energy transfer terms is via the conservation of energy statement – all in rate change:

$$+ \text{ accumulation} = + \text{ input - output } + \text{ generation, or,}$$

$$\dot{u} = \dot{h} - \dot{w}\,.$$

The most important measurable quantity in heat analysis is temperature. Then, we need a way to relate the internal energy of a system to its temperature. For homogeneous bodies, based on experimental results, it is regularly assumed that the internal energy of a macroscopic body is proportional to its average thermodynamic temperature, T; *i.e.*,

$$u \propto T \implies u = CT\,,$$

where C is a material's property called *heat capacity* (the capacity that a body has to store heat :). So defined, heat capacity is an *extensive property* (a property that depends on the body's mass), $\dim C = [E/\Theta]$. It is more convenient, though, to work with the *intensive property* c called *specific heat capacity*; that is, the capacity that a body has to store heat per unit mass, $\dim c = [E/M\Theta]$. Then, the internal energy of a body can be calculated as

$$u = mcT\,,$$

where m is the body's mass.

If $p\mathrm{d}v$ work is the only work that occurs, and the body is assumed to be homogeneous (so $\dot{c} = 0$), then

$$\dot{h} = p\dot{v} + \dot{u}\,.$$

The last equation has two important cases:

$$\dot{h} = \begin{cases} \dot{u} = mc_v\dot{T}\,, & \text{constant volume process; } i.e.,\ \mathrm{d}v = 0, \\ \dot{H} = mc_p\dot{T}\,, & \text{constant pressure process; } i.e.,\ \mathrm{d}p = 0, \end{cases}$$

where $H = u + pv$ is the enthalpy, c_v and c_p are the specific heat capacities at constant volume and constant pressure and mass is assumed to be constant, $\dot{m} = 0$.

26.1.2. *Thermodynamic relations and definition of heat capacities.* The internal energy of a closed system changes either by adding "heat" to the system of by the system performing work. Mathematically,

$$\mathrm{d}u = \delta h + \delta w\,.$$

For work as a result of an increase of the system volume, then we have

$$\mathrm{d}u = \delta h - p\mathrm{d}v\,.$$

If now heat is added at constant volume, then the second term of the last equation vanishes and we have

$$\left(\frac{\partial u}{\partial T}\right)_v = \left(\frac{\partial h}{\partial T}\right)_v = c_v\,.$$

This defines the heat capacity at constant volume, c_v, $\dim c_v = [E/\Theta]$.

Another useful quantity is the heat capacity at constant pressure, c_p. With the enthalpy of the system given by

$$H = u + pv\,,$$

then, the equation for du changes to

$$\mathrm{d}H = \delta h + v\mathrm{d}p\,,$$

and, therefore, at constant pressure, we have

$$\left(\frac{\partial H}{\partial T}\right)_p = \left(\frac{\partial h}{\partial T}\right)_p = c_p\,.$$

26.1.3. *Relation between heat capacities.* Measuring the heat capacity at constant volume can be difficult for liquids and solids, since small temperature changes typically require large pressures to keep a solid or liquid at constant volume, implying that the containing vessel must be nearly rigid or at least very strong. Instead it's easier to measure the heat capacity at constant volume – allowing the material to expand or contract freely – and solve for the heat capacity at constant volume using maths derived from the basic thermodynamic laws. Starting from the fundamental thermodynamic relation, we can show

$$C_p - C_v = T\left(\frac{\partial p}{\partial T}\right)_{v,n} T\left(\frac{\partial v}{\partial T}\right)_{p,n}\,,$$

where the partial derivatives are taken at constant volume and constant number of particles, and constant pressure and constant number of particles.

This can also be written as

$$C_p - C_v = vT\frac{\alpha^2}{\beta_T}\,,$$

where α is the *coefficient of thermal expansion* and β_T is the isothermal compresibility.

For an ideal gas, evaluation of the partial derivatives according to the equation of state $pv = nrT$, where r is the gas constant, gives

$$C_p - C_v = r\,.$$

26.1.4. *First law of thermodynamics.* This is a version of the law of conservation of energy specialized for thermodynamic systems: the energy of an isolated system is constant.

When a system expands in a quasistatic process, the work done by the system on the environment is $p\mathrm{d}v$, whereas the work done on the system is $-p\mathrm{d}v$. Using either convention sign for work gives the same relation for the change in internal energy:

$$\mathrm{d}u = \delta h - p\mathrm{d}v\,.$$

Work and heat are expressions of actual physical processes which supply or remove energy, while internal energy is a math abstraction that keeps account of the exchanges of energy that befall the system. Thus, the term δh means that amount of energy added or removed by heat conduction or radiation, rather than referring to a form of energy within the system. Likewise, work energy for δw meas that amount of energy gained or lost as result of work. Internal energy is a system property, whereas work done and heat supplied are not.

Note. Thermodynamics is independent on the underlying atomic theory of matter. It only deals with macroscopic properties: pressure, temperature, *etc*. This is thermodynamics major strength, for any microscopic theory must submit its results to thermodynamics. Paradoxically, this very strength might be seen as a weakness in physical explanation: it is desirable to have a microscopically based mechanical explanation for heat transfer phenomena.

Note. Thermodynamics should be called *thermostatics*, because it only describes systems in thermal equilibrium. Thus, since thermodynamics doesn't take into account t, the problem is how to determine temperatures, because the internal energy as a function of time, $u[t]$, cannot be predicted a priori. Therefore, some principles must be added to predict u, h and T. These principles are called *transport laws* and are not a part of thermodynamics. They include Fourier's law of heat conduction, Newton's law of cooling and Stefan-Boltzmann law for thermal radiation. These are experimental laws.

26.1.5. *Heat capacity for solids and liquids.* When the substances undergoing the process in *incompressible*, then $\mathrm{d}v = 0$ for any pressure variation. Therefore, the two specific heats are equal: $c_v = c_p = c$, implying that

$$\dot{h} = \dot{u} = mc\dot{T}\,.$$

Since solids and liquids con often be approximated as being incompressible, we shall frequently use the last equation.

26.2. **Modes of heat transfer.** The basic modes of heat transfer are

- heat conduction;
- heat convection and
- heat radiation.

26.2.1. *Heat conduction.* Fourier's law (empirical law): the local heat *flux* j resulting from thermal conduction is proportional to the magnitude of the temperature gradient and opposite to it in sign; *i.e.*,

$$j = -k\nabla T\,,$$

where $k > 0$ is the proportional constant that depends on the material called the *thermal conductivity*. It's a constant if the material is homogeneous or isotropic. The dimensions of the quantities in the last equation are $\dim j = \left[E/L^2T\right]$, $\dim k = [E/TL\Theta]$, $\dim \nabla = [1/T]$ and $\dim T = [\Theta]$.

The integral form of Fourier's law is obtained by integrating the differential form over the material's total surface S:

$$\dot{h} = -k \oint_S \nabla T \cdot \mathrm{d}A\,,$$

where $\dot{h}$ is the amount of heat transferred per unit time, $\dim \dot{h} = [E/T]$, and $\mathrm{d}A$ the oriented surface element (remember: $\mathrm{d}A = n\mathrm{d}a$, where n is a unit vector normal to $\mathrm{d}A$).

Thermal conductivity values: because of how molecules are arranged, solids will have generally higher thermal conduction than gases. Thus, the process of heat transfer is more efficient in solids than in gases. In a gas k is proportional to the molecular speed and molar specific heat and inversely proportional to the cross-sectional area of molecules. Values for k can be found in tables, but it's desirable to have an idea of the k order of magnitude.

26.2.2. *Heat convection.* Consider a typical convection cooling situation: cool gas flows past a warm body. The fluid immediately adjacent to the body forms a slowed-dense region called *boundary layer*. Heat is conducted into this layer, which sweeps it away and, farther downstream, mixes it into the streams. We call such process of carrying heat away from a moving fluid *convection*. Newton considered the convection process and suggested that the cooling would be such that

$$\frac{\mathrm{d}T_{\mathrm{body}}}{\mathrm{d}t} \propto T_{\mathrm{body}} - T_\infty\,,$$

where T_∞ is the temperature of the incoming fluid. This statement suggests that energy is flowing from the body. But, if energy is constantly replenished, then the body temperature need not change. Thus, with $h = mc\dot{T}$, we get

$$h \propto T_{\mathrm{body}} - T_\infty\,.$$

This equation can then be rephrased in therms of $j = h/a$, where a is the body outer surface area, as

$$j = \bar{h}\,(T_{\mathrm{body}} - T_\infty)\,,$$

This is the steady-state of Newton's law of cooling. The constant h is the *film coefficient* or *heat transfer coefficient*. The bar over h indicates that's an average over the surface of the body. Without the bar, h denotes the "local" value of the heat transfer coefficient at a point on the surface. The dimensions of h and $\bar{h}$ are $\dim h = \dim \bar{h} = \left[E/TL^2\Theta\right]$.

It turns out that Newton oversimplified the process description when he made his conjecture. Heat convection is complicated and $\bar{h}$ can depend on the temperature difference $(T_{\mathrm{body}} - T_\infty) = \Delta T$:

- h is really independent of ΔT when the fluid is forced past a body and ΔT is not too large. This is called *forced convection*.

- When fluid buoys up from a hot body or down from a cold one, h varies as some weak power of ΔT – typically as $\Delta T^{1/4}$ or $\Delta T^{1/3}$. This is called *free* or *natural convection*. If the body is hot enough to boil a liquid surrounding it, h will typically vary as ΔT^2.

Typical values of h are presented in tables.

Lumped-capacity solution: the problem now is to predict the transient (time dependent) cooling of a convectively cooled object. Apply the first law statement (accumulation = -out energy; out energy: energy that goes from the system into the surrounding fluid) to have:

$$\dot{u} = -j \implies \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho c v \left(T - T_{\mathrm{ref}}\right)\right) = -\bar{h} a \left(T - T_\infty\right) ,$$

where a and v are the surface area and volume of the body, T is the temperature of the body, $T = T[t]$, and T_{ref} is an arbitrary temperature at which u is defined to equal zero. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(T - T_\infty\right) = -\frac{\bar{h} a}{\rho c v}\left(T - T_\infty\right) .$$

The general solution to this equation is

$$\ln\left(T - T_\infty\right) = -\frac{t}{\tau} + C ,$$

where the group $\tau = \rho c v/\bar{h} a$ is the time constant. If the initial temperature is $T[t = 0] = T_{\mathrm{i}}$, then $C = \ln\left(T_{\mathrm{i}} - T_\infty\right)$ and the cooling of the body is given by

$$\frac{T - T\infty}{T_{\mathrm{i}} - T\infty} = \exp[-t/\tau] .$$

All the physical parameters in the problem have now been 'lumped" into the time constant. It represents the time required for a body to cool to $1/e$ or 37 % of its initial temperature above (or below) $T\infty$. The ratio t/τ can also be interpreted as

$$\frac{t}{\tau} = \frac{\bar{h} a t}{\rho c v} = \frac{\text{capacity for convection from surface}}{\text{heat capacity of the body}} .$$

Note. Thermal conductivity is missing from the last equations. The reason is that we have assumed that the temperature of the body is nearly uniform, and thus means that internal conduction is not important. If $L/(k_{\mathrm{b}}/\bar{h}) \ll 1$, then the temperature of the body, T_{b}, is almost constant within the body at any time. Thus,

$$\frac{\bar{h} L}{k_{\mathrm{b}}} \ll 1 \implies T_{\mathrm{b}}[x, t] \sim T[t] \sim T_{\mathrm{surface}}$$

and the thermal conductivity k_{b} becomes irrelevant to the cooling process. This condition must be satisfied if the lumped solution is to be accurate.

We call the group

$$\frac{\bar{h} L}{k_{\mathrm{b}}} = \Pi_{\mathrm{bi}}$$

Biot number. If Π_{bi} were large, the situation would be reversed. In this case, $\Pi_{\mathrm{bi}} \gg 1$ and the convection process offers little resistance to heat transfer. We could solve the heat diffusion equation:

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{\alpha}\frac{\partial T}{\partial t} ,$$

subject to the simple boundary condition $T[x, t] = T\infty$, when $x = L$ to determine the temperature in the body and its rate of cooling in this case.

Biot number will therefore be the basis for determining what sort of problem we have to solve.

To calculate the rate of entropy production in a lumped-capacity system, we note that the entropy always in the universe is the sum of the entropy decrease of the body and the more rapid entropy increase of the surroundings. The source of irreversibility is heat flow through the boundary layer. Accordingly, we unite the time rate of change of entropy of the universe as

$$\dot{S}_{\mathrm{un}} = \dot{S}_{\mathrm{b}} + \dot{S}_{\mathrm{fl}} = -\frac{h_{\mathrm{rev}}}{T_{\mathrm{b}}} + \frac{h_{\mathrm{rev}}}{T_\infty} = -\rho c v \frac{\mathrm{d}T_{\mathrm{b}}}{\mathrm{d}t}\left(\frac{1}{T_\infty} - \frac{1}{T_{\mathrm{b}}}\right) .$$

26.2.3. *Heat radiation – thermal radiation.* Electromagnetic radiation is generated by the thermal motion of charged particles in matter. All matter with temperature greater than the absolute zero emits thermal radiation. Examples of thermal radiation are the visible light and infrared light emitted by an incandescent light bulb, the infrared radiation emitted by animals and detectable with an infrared camera. Thus, thermal radiation can be seen as a conversion of thermal energy into electromagnetic energy.

If a radiation-emitting object meets the physical characteristics of a black body in thermodynamic equilibrium, the radiation is called *black body radiation*. Planck's law describes the *spectrum of black-body radiation*, which depends only on the object's temperature. Wien's displacement law determines the most likely *frequency of the emitted radiation* and Stephan-Boltzmann law given the *radiation intensity*.

Heat transfer by thermal radiation: all bodies constantly emit energy by a process of em radiation. The intensity of such energy flux depends upon the temperature of the body. Most of the heat that reaches you when you sit in front of a fire is radiant energy. Radiant energy warms you when you walk in the sun.

Objects that are cooler that the fire or the sun emit much less energy because the energy emission varies as the fourth power of absolute temperature. Very often, the emission of energy, or radiant heat transfer, from cooler bodies can be neglected in comparison with convection and conduction – approximate analyses, order of magnitude analyses and limiting (extreme cases) analyses can be helpful here! But heat transfer processes that occur at high temperature or with conduction or convection suppressed by evacuated insulators normally involve a significant fraction of radiation.

The em spectrum: thermal radiation occurs in a range of the em spectrum of energy emission. Accordingly, it inhabits the same wavelike properties as light or radio waves. Each quantum of radiant energy has a wavelength λ and a frequency ν associated with it.

The full spectrum includes an enormous range of energy-bearing waves, of which heat is only a small part. Tables list the various forms over a range of wavelengths that spams 17 orders of magnitude. Heat radiation, whose main component is normally the spectrum of infrared radiation, passes through a three-order-of-magnitude window in λ or ν.

Black bodies: the model for the perfect thermal radiator is the so-called *black body*. This is a body that absorbs all energy that reaches it and reflects nothing. The term is a bit confusing, since they *emit* energy. Thus, under infrared vision, a black body would glow with "color" appropriate to its temperature. Perfect radiators *are* "black" in the sense that they absorb all visible light (and all other radiation) that reaches them.

To model a black body a "Hohlraum" is used. What are the important features of a thermally black body? First consider a distinction between heat and infrared radiation: *infrared radiation* refers o a particular range of wavelengths, while *heat* refers to the whole range of radiant energy flowing from one body to another. Suppose that a radiant heat flux j falls upon a translucent plate that's not black. A fraction α of the total incident energy, called the *absorptance*, is absorbed by the body; a fraction ρ, called *reflectance*, is reflected from it and a fraction τ, called *transmittance*, passes through. Thus,

$$1 = \alpha + \rho + \tau \,.$$

This relation can also be written for the energy carried by each wavelength in the distribution of wavelengths that makes up *heat* from a source at any temperature

$$1 = \alpha_\lambda + \rho_\lambda + \tau_\lambda \,.$$

All radiant energy incident on a black body is absorbed, so that α_{b} or $\alpha_{\lambda_{\mathrm{b}}} = 1$ and $\rho_{\mathrm{b}} = \tau_{\mathrm{b}} = 0$. Furthermore, the energy emitted by a black body reaches a theoretical maximum given by Stephan-Boltzmann law.

Stephan-Boltzmann law: the energy flux radiating from a body is commonly designated by $e[t]$, $\dim e = E/L^2T$. The symbol $e_\lambda[\lambda, T]$ designates the distribution function of radiative flux in λ, or the *monochromatic emission power*:

$$e_\lambda[\lambda, T] = \frac{\mathrm{d}e}{\mathrm{d}t}[\lambda, T] \qquad \text{or} \qquad e[\lambda, T] = \int_0^\lambda e_\lambda[\lambda, T]\ \mathrm{d}\lambda \,.$$

Thus,

$$e[T] = E[\infty, T] = \int_0^\infty e_\lambda[\lambda, T]\ \mathrm{d}\lambda\,.$$

The dependence of $e[T]$ on T for a black body was found experimentally by Stefan... The Stephan-Boltzmann law is

$$e_\mathrm{b} = \sigma T^4\,,$$

where the Stephan-Boltzmann constant σ is $5.670\,373(21) \times 10^{-8}\,\mathrm{W/m^2K^4}$:

$$\sigma = \frac{2\pi^5}{15}\frac{k_\mathrm{b}^4}{h^3c^2} \qquad \text{and} \qquad \dim\sigma = \frac{E}{TL^2\Theta^4}\,,$$

or, in terms of the gas constant r,

$$\sigma = \frac{2\pi^5}{15}\frac{r^4}{h^3c^2n_\mathrm{a}^2}\,,$$

where n_a is the Avogadro's number.

A useful mnemonic for σ is 5-6-7-8: $\sigma \sim 5.67 \times 10^{-8}\,\mathrm{W/m^2K^4}$.

e_λ *vs* λ: nature requires that, at a given temperature, a body will emit a unique distribution of energy in wavelength. Thus, when you heat a poker in the fire, it first glows a dull red – emitting most of its energy at long wavelengths and just a bit in the visible regime. When it's white-hot, the energy distribution has been both greatly increased and shifted towards he shorter-wavelength visible range. At each temperature, a black body yields the highest value of e_λ that a black body can attain.

Measurements of the black body systems are shown... The locus of maxima of the curves is plotted. It obys a relation called Wien's law:

$$(\lambda T)_{e_\lambda = max} = 2989\,\mu\mathrm{mK}\,.$$

About 3/4 of the radiant energy of a black body lies to the right of this line. Notice that, while the locus of maxima leans towards the visible range at higher temperatures, only is a small fraction of the radiation visible at the highest temperature.

Predicting how the monochromatic emission power of a black body depends on λ was solved by Planck. He made the prediction and set the basis for quantum mechanics. He found that

$$e_{\lambda_\mathrm{b}} = 2\pi\frac{hc_0^2}{\lambda^5\left(\exp[hc_0/k_\mathrm{b}T\lambda] - 1\right)}\,,$$

where c_0 is the speed of light in vacuum, h is the Planck's constant, k_b is Boltzmann constant.

Radiant heat exchange: suppose that a heated object (1) radiates only to some other object (2) and that both objects are thermally black. All heat leaving object (1) arrives at object (2) and all heat arriving at object (1) comes from object (2). Thus, the net heat transferred from object (1) to object (2), q_net, is the difference. Let now $q_{1\text{ to }2} = a_1 e_\mathrm{b}[T_1]$ and $q_{2\text{ to }1} = a_1 e_\mathrm{b}[T_2]$:

$$q_\mathrm{net} = a_1\sigma\left(T_1^4 - T_2^4\right)\,.$$

We have seen that non-black bodies absorb less radiation than black bodies, which are perfect absorbers. Likewise, non-black bodies emit less radiation that black bodies, which also happens to be perfect emitters. We can characterize the emissive power of a non-black body using a property known as *emittance* ϵ:

$$e_\text{non-black} = \epsilon e_\mathrm{b} = \epsilon\sigma T^4\,,$$

where $0 < \epsilon \leq 1$. When radiation is exchanged between two bodies that are not-black, we have

$$q_\mathrm{net} = a_1 f_{1\text{-}2}\sigma\left(T_1^4 - T_2^4\right)\,,$$

wherein the *transfer factor* $f_{1\text{-}2}$ depends on the emittance of both bodies as well as the geometrical "view".

26.3. **A look ahead.** To solve actual problems, three tasks must be completed:

(1) heat diffusion equation must be solved subject to appropriate boundary and initial conditions;
(2) the convective heat transfer coefficient h must be determined if convection is relevant;
(3) the factor $f_{1\text{-}2}$ must be determined to calculate radiative heat transfer.

There are three types of heat transfer problems:

(1) theoretical: a systematic statement of principles; a formulation of apparent relationships or underlying principles of certain observed phenomena;
(2) analysis: the solving by means of equations; the breaking up of any whole into its parts so as to find out their nature, function, relationship and so forth;
(3) practice: the doing of something as an application of knowledge.

27. Problem Solving

> Before developing the necessary mathematics, survey the crucial physics.
>
> — JOHN F. LINDNER, Electromagnetism with Spacetime Algebra, 2011

> Too much mathematical rigor teaches rigor mortis: the fear of making an unjustified leap even when it lands on a correct result. Instead of paralysis, have courage – shoot first and ask questions later. Although unwise as public policy, it is a valuable problem-solving philosophy.
>
> — SANJOY MAHAJAN, Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving, 2010

27.1. **Dimensional Analysis.**

27.1.1. *Fundamental Constants as Conversion Factors.* Follow Michael Duff's ideas:

> the laws of physics are inherently dimensionless and fundamental constants as c, $\hbar$ or G, in the fundamental equations of physics, must be seen as mere conversion factors to convert mass, time and length into each other or to scale between the macroscopic and the microscopic world.

To see this, consider c, the speed of light in vacuum. This constant is not only a fundamental constant but also universal; *i.e.*, it holds in the whole Universe! So, using the definition of velocity, $c = x/t$, we find that

$$x = ct\,,$$

which is to say, we can measure distances in meters or in seconds or, conversely, time in seconds or in meters. The latter fact is used in the Theory of Relativity and Astronomy: the distance from Earth to our neighbor galaxy, Canus Major Dwarf, is 25 000 light−yr; yeap! Distance measured in years!

Another example is given by Boltzmann constant k_b and the ideal gas law: $pv = nRT = Nk_\mathrm{b}T$. Here, $\dim k_\mathrm{b} = [E/\Theta]$, where $[\Theta]$ represents the dimension of temperature. Thus k_b converts T, a macroscopic property, to energy, so that the product can be coupled with N, the number of particles, a microscopic property. In other words, k_b provides a bridge to move from the macro-world – pressure, forces, temperature, volume, *etc.*, to the micro-world!

27.1.2. *Energy of an Ideal Gas under Pressure.* Say you have a piston acting on a cylinder of cross section a containing a fluid of volume v. Say you apply a force f perpendicular to the cross area. Then, the pressure p exerted by the force is $p = f/a$. This pressure compresses the fluid, reducing its volume. See that pressure manifests macroscopically as a surface phenomenon. The question now is what changes in the fluid? How does the fluid react, internally, to f? To develop an approximate answer, use dimensional analysis.

First, determine the dimensions of pressure:

$$\dim p = \dim f/a = [ML] \,/\, \left[T^2L^2\right]\,.$$

Multiply the RHS of the last equation by a factor $[L/L]$ to find

$$\dim p = \left[ML^2/T^2L^3\right] = \left[E/L^3\right] = [E/V] \sim e/v\,,$$

where e represents the energy imparted by the piston to the fluid. (We have used the tilde notation $\sim$ because dimensions match but dimensionless quantities are hidden to dimensional analysis.) Thus, we get that

$$pv \sim e\,.$$

On the other hand, say that the fluid is a gas, an ideal gas. Then, according to the ideal gas law, we have

$$pv = NRT\,,$$

where p is the gas pressure, v the gas volume, N the amount of gas, R the ideal gas constant and T the gas thermodynamic temperature. Since the external energy imparted by the piston must equal the energy changed in the fluid, we have

$$e \sim NRT = \Pi NRT\,,$$

where Π is a dimensionless quantity.

Therefore, aided by dimensional analysis, we found that the ideal gas internal energy is proportional NRT and that it changes upon the action of p.

In general, *i.e.*, not only for ideal gases, the relationship shows that e increases when p does, then if a compression force acts of a fluid, we would expect that the fluid energy increases, manifested as a change in the fluid temperature.

Incidentally, for ideal gases, Π is a gas property called the *dimensionless heat capacity*, denoted c. It's related to the *heat capacity*, C, by the relation

$$C = cNR\,,$$

where $\dim C = \dim NR = [E/\Theta]$ and $c = 3/2$. Therefore, the energy of an ideal gas can be expressed as

$$e = CT = \frac{3}{2}T\,,$$

which says that the internal energy of an ideal gas depends only on temperature.

27.1.3. *Ratio between Electric and Gravitational Forces.* Find the ratio of the electric force to the gravitational force between the nuclei of two hydrogen atoms in a rest frame.

Solution. Consider both atoms to be ^{1}H. Model the electrostatic force between them $|f_\mathrm{e}|$ by Coulomb's force law: $|f_\mathrm{e}| = k_\mathrm{e}e^2/r^2$, where k_e represents Coulomb's constant, e the proton's electric charge – the elementary charge – and r the separation between the centers of the nuclei. Express next Coulomb's constant as a function of fundamental constants via the relation $k_\mathrm{e} = \alpha c_0 h/2\pi e^2$, where α represents the fine-structure constant, c_0 the speed of light in vacuum and h Planck's constant. Replace then the last equation into Coulomb's force law to have

$$|f_\mathrm{e}| \approx \frac{\alpha}{2\pi} c_0 h\,.$$

On the other hand, model the gravitational force between the nuclei $|f_\mathrm{g}|$ by Newton's force law of universal gravitation: $|f_\mathrm{g}| = Gm_\mathrm{p}^2/r^2$, where G stands for the Newtonian constant of gravitation and m_p for a proton's mass. Use the proton-electron mass ratio β to express m_p as a function of the electron's mass m_e, $m_\mathrm{p} = \beta m_\mathrm{e}$, then express m_e as a function of fundamental constants: $m_\mathrm{e} = 2hR_\infty/\alpha^2 c_0$, where R_∞ stands for Rydberg constant, and thus rewrite Newton's force law:

$$|f_\mathrm{g}| \approx \frac{4\beta^2}{\alpha^4}\frac{h^2 G R_\infty^2}{c_0^2}\,.$$

Finally, find the $|f_\mathrm{e}|$ to $|f_\mathrm{g}|$ ratio:

$$\frac{|f_\mathrm{e}|}{|f_\mathrm{g}|} = \frac{\alpha^5}{8\pi\beta^2}\frac{c_0^3}{hGR_\infty^2}\,. \qquad \square$$

Dim. Analysis. Verify the dimensional homogeneity of the last equation by performing dimensional analysis on it – note that the left hand side and the first term of the right hand side are manifestly dimensionless, so must the second term be:

$$\dim \frac{c_0^3}{hGR_\infty^2} = \left[\frac{L^3}{T^3}\right]\left[\frac{MT^2}{L^3}\right]\left[\frac{T}{ML^2}\right]\left[\frac{L^2}{1}\right] = 1\,. \qquad \square$$

Approx. Solution. Find the order of magnitude of the ratio $|f_\mathrm{e}|/|f_\mathrm{g}|$ given by the expression

$$\frac{|f_\mathrm{e}|}{|f_\mathrm{g}|} = \frac{\alpha^5}{8\pi\beta^2}\frac{c_0^3}{hGR_\infty^2}\,.$$

Use the values provided by the NIST [26]:

- $\alpha = 7.297\,352\,569\,8 \times 10^{-3}$;
- $\pi \sim 3.141\,592\,653\,59$;
- $\beta = 1836.152\,672\,45$;
- $c_0 = 299\,792\,458\,\mathrm{m/s}$;
- $h = 6.626\,069\,57 \times 10^{-34}\,\mathrm{J/s}$;
- $G = 6.673\,84 \times 10^{-11}\,\mathrm{m^3/kg\,s^2}$;
- $R_\infty = 10\,973\,731.568\,539\,\mathrm{m^{-1}}$.

[26] List of frequently used constants: `http://physics.nist.gov/cuu/Constants/`.

Use the "back-of-the-envelope" technique:

- Approximate the values of the constants to 1, 3 or 10 for the "small part": $\alpha \sim 1 \times 10^{-2}$, $8 \sim 10$, $\pi \sim 3$, $\beta \sim 1 \times 10^{3}$, $c_0 \sim 3 \times 10^{8}$, $h \sim 1 \times 10^{-33}$, $G \sim 1 \times 10^{-10}$ and $R_\infty \sim 1 \times 10^{7}$;
- Replace the approximate values into the equation:
$$\frac{\alpha^5}{8\pi\beta^2}\frac{c_0^3}{hGR_\infty^2} \sim \frac{(1\times 10^{-2})^5}{10\times 3\times (1\times 10^{3})^2}\frac{(3\times 10^{8})^3}{1\times 10^{-33}\times 1\times 10^{-10}\times (1\times 10^{7})^2} .$$
- Calculate the "big part", the powers of ten: 1×10^{35}, and calculate then the "small part", $3^3/3$, to finally have
$$\frac{|f_{\mathrm{e}}|}{|f_{\mathrm{g}}|} \sim 9 \times 10^{35} \sim 1 \times 10^{36} \sim O\big[10^{36}\big] .$$

This is, the electric force is *ca.* 36 orders of magnitude greater than the gravitational force.

27.2. **From Approximate Solutions to Formal Analytic Solutions.** To illustrate various problem solving techniques, we will analyze the motion of a charged particle using Newtonian Physics. We will do so by showing various math and physics methods in different levels of sophistication: guessing, dimensional analysis, approximations and analytic techniques. Finally, we present a final wrapped-up solution.
Example. Consider a particle of constant electric charge q and constant mass m moving with velocity v due to an interaction with a constant electromagnetic field. Assuming Newtonian physics, find the rate at which the particle's kinetic energy k changes in time t.

27.2.1. *Guessing the Solution.* As a first approximation to the solution, instead of working with the general case, we go to an specific example by considering the moving particle to be an electron and the electric field to be originated by a proton. The dynamics is described by Lorentz force law.

Let's first analyze the electron-electric field interaction. The proton creates an electric field due to its charge q_{p}. Lorentz force states that the proton's field strength $|e_{\mathrm{p}}|$ is given by $|e_{\mathrm{p}}| \propto |q_{\mathrm{p}}|/r^2$, where r is the distance from the proton's center. Geometrically, this means that $|e_{\mathrm{p}}|$ creates concentric surfaces of equal electric potential in $\mathcal{E}^3$, called *isoelectric surfaces*, just like a static "heat" source forms concentric isothermal surfaces around its center. When something moves towards the proton, it will "pierce" such surfaces. Note that the field strength scales *inversely* with the *squared* distance: for instance, if the distance is *halved*, the field strengthens by a factor of *four*. In other words, the closer to the proton's center, the stronger the interaction with its field becomes. On the other hand, when an electron, with charge $|q_{\mathrm{e}}| < 0$, enters the field, it is "attracted" to the proton's center as the force between them, $|f_{\text{p-e}}| \propto -|q_{\mathrm{e}}||q_{\mathrm{p}}|/r^2$, increases with decreasing distance. In turn, the electron's velocity v_{e} increases and so does its kinetic energy $k_{\mathrm{e}} \propto v_{\mathrm{e}}^2$. Therefore, we expect $\dot{k}_{\mathrm{e}} \sim -q_{\mathrm{e}}e_{\mathrm{p}}v_{\mathrm{e}}$. (Notice the negative sign in the expression. It says that the electron looses energy as it falls into the proton! Also, see that $\dot{k}_{\mathrm{e}}$ does not depend on the electron's mass.)

Now, let's analyze the electron-magnetic field interaction. An electron moving in a magnetic field experiences a *sideways* force f_{m} proportional to (i) the strength of the magnetic "field" $|b|$, (ii) the component of the velocity perpendicular to such field v_{e} and (iii) the charge of the electron q_{e}; *i.e.*, the second term of the Lorentz force: $f_{\mathrm{m}} = q_{\mathrm{e}}v_{\mathrm{e}} \times b$. Note that f_{m} is always *perpendicular* to both the v_{e} and the b that created it, mathematically expressed by the (cross) product $v_{\mathrm{e}} \times b$. Then, when the electron moves in the field, it traces an helical path in which the helix axis is parallel to the field and in which v_{e} remains constant. Because the magnetic force is always perpendicular to the motion, the b can do *no* work. It can only do work *indirectly*, via the electric field generated by a changing b. This means that, if no work is directly created by the magnetic field, then the change rate of the electron's kinetic energy should not depend directly on it, but rather indirectly, via the electron's velocity: $k_{\mathrm{e}} \propto v_{\mathrm{e}}^2 \implies \dot{k}_{\mathrm{e}} \propto v_{\mathrm{e}}$, which has the same dependence as the equation obtained in the electron-electric field analysis.

Finally, because an electron moving towards a proton is an example of a more general case, expect the *form* of the electron-proton case to work for *any* moving charged particle under a constant electromagnetic field. This means that, physically, the change of the particle's kinetic energy $\dot{k}$ should directly depend only on the electric field (and not on the

magnetic induction), the particle's charge and its velocity: $\dot{k} \sim qev$. Mathematically, see that, since e and v are both vectors, the product ev must be a product between vectors. The only suitable product is the inner product, *aka* scalar product, because it is the only one to return a scalar; this would agree with the scalar nature of $\dot{k}$. This means, therefore,

$$\dot{k} \sim qe \cdot v \,.$$

We expect this guessed equation to be obtained by formal methods.

27.2.2. *Dimensional Analysis.* For the next solution, we will use dimensional analysis to determine the *functional* form of the model to the phenomenon.

To find the *functional form of the physical model* by means of dimensional analysis follow the steps:

(1) Instead of using the SI fundamental dimensions, use the set $\{[F], [L], [T], [Q]\}$ of *four* dimensionally independent quantities, where $[F]$ represents the dimension of force, $[L]$ length, $[T]$ time and $[Q]$ electric charge.
(2) In the chosen set, the dimensions of the *six* physical quantities that model the phenomenon are $\dim k = [FL]$, $\dim t = [T]$, $\dim q = [Q]$, $\dim e = \left[FQ^{-1}\right]$, $\dim v = \left[LT^{-1}\right]$ and $\dim b = \left[FTQ^{-1}L^{-1}\right]$.
(3) According to the Buckinham's theorem, *aka* Π theorem, there are $6 - 4 = 2$ dimensionless quantities Π. The first one is $\Pi_1 = k/(tevq)$ and the second $\Pi_2 = bv/e$.
(4) Finally, the model should have the form:

$$g[\Pi_1, \Pi_2] = g\left[\frac{k}{tevq}, \frac{bv}{e}\right] = 0 \implies \frac{k}{t} = qev\, h\left[\frac{bv}{e}\right] ,$$

where h is a function of (bv/e).

In the last equation, the precise form of the function h must be determined by experimentation or by analytic means. However, dimensional analysis confirms our suspicion: $\dot{k} \sim k/t \sim qev$; *i.e.*, the product qev "lives upstairs" in the equation. The second term, the function h, should be equal to a dimensionless parameter Π if our guess is to be correct. We will keep h, nevertheless, for it may be that our guess is not correct.

27.2.3. *Approximate Methods.* For a second approximation, we will use actual equations and will apply to them approximate methods to find the *form* of the physical model. This helps to better understand the physics behind the process by avoiding the distractions of unnecessary constants, numeric factors and complicated notation. Additionally, it helps, as a sketch, to develop and to present the analytic solution.

First, write the complete set of equations modeling the phenomenon:

$$k = \frac{1}{2}mv^2 \,, \qquad \text{[kinetic energy]}$$
$$f = q(e + v \times b) \,, \qquad \text{[Lorentz force law]}$$
$$f = ma = m\dot{v} \,, \qquad \text{[Newton's second law of motion]}$$

where the variables were already defined during guessing and dimensional analysis.

Then, drop unnecessary constants and numeric factors, use the secant method to approximate derivatives [27] and treat vectors as scalars [28] to find

$$k \sim mv^2 \,, \qquad \text{[approx. kinetic energy]}$$
$$\dot{k} \sim k/t \sim mv^2/t \sim (mv/t)v \,, \qquad \text{[approx. kinetic energy time rate change]}$$
$$f \sim q(e + vb) \,, \qquad \text{[approx. Lorentz force law]}$$
$$f \sim mv/t \,. \qquad \text{[approx. Newton's second law of motion]}$$

Find the equation of motion by equating Newton's law to Lorentz law: $(mv/t) \sim q(e+vb)$. Plug this equation into the one for k/t, via the factor (mv/t):

$$k/t \sim q(e + vb)v \sim qev + qvbv \,.$$

[27] In the *secant method*, tangents (derivatives) are replaced by secants (quotients); *i.e.*, if $f = f[x]$, then $\mathrm{d}f/\mathrm{d}x \sim f/x$.

[28] This means to replace vectors by scalars and to replace products between vectors by multiplications between scalars.

In the last equation, the term $(qvbv)$ is likely to vanish, because v is to enter $(v \times b)$ as $(v \times b) \cdot v$, for v comes from $k \sim mv^2 \sim mv \cdot v$ and thus $(qv \times b) \cdot v \sim (qvbv) = 0$, since $v \times b$ is orthogonal to v. Then, the expression would be $k/t \sim qev$ with some product of vectors between e and v – the scalar product. The model could thus be written as $k/t \sim qe \cdot v$. Finally, remembering that $k/t \sim \dot{k}$, then

$$\dot{k} \sim qe \cdot v \,. \qquad \square$$

The equation found by approximate means agrees with our guess and, partially, with dimensional analysis. This increases our confidence in understanding the phenomenon! Besides, all the previous methods have cleared the derivation plan: i) find $\dot{k}$ from k; ii) find the equation of motion by using the definition of linear momentum, by equating Newton's law to Lorentz law and by leaving mv on one side and iii) finally, plug in the equation of motion onto $\dot{k}$ and play with products between vectors to arrive to the final solution.

27.2.4. *Wordy Derivation.* We solve the problem now by presenting a "wordy-version" of the analytic solution: we describe the math derivation in detail.

The particle kinetic energy is $2k = mv^2$. This could be rewritten as

$$2k = mv \cdot v \,,$$

since v is colinear to itself; *i.e.*, its outer product is zero; *viz.*, $v^2 = vv = v \cdot v + v \wedge v = v \cdot v$.

Then, calculate the kinetic energy change rate with time by

$$2k = mv \cdot v \implies 2\dot{k} = m(\dot{v} \cdot v + v \cdot \dot{v}) = m(\dot{v} \cdot v + \dot{v} \cdot v) = 2m\dot{v} \cdot v \,,$$

where the product rule for the differentiation of the inner product $[(f \cdot g)' = f' \cdot g + f \cdot g'$, for vector-valued functions f and $g]$, the commutativity property of the inner product [for vectors a and b, $a \cdot b = b \cdot a$] and the dot notation $[\dot{k} := \mathrm{d}k/\mathrm{d}t]$ were used.

Next, one cancels out the numerical factor 2 in both sides of the equality to find that

$$\dot{k} = m\dot{v} \cdot v \,.$$

On the other hand, the particle's motion can be modeled by equating Newton's second law of motion with Lorentz force, since the particle interacts with an electromagnetic field. Thus, we find that

$$\dot{p} = q(e + v \times b) \,,$$

where p is the particle's linear momentum. By definition, $p = mv$, so $\dot{p} = \dot{m}v + m\dot{v} = m\dot{v}$, because mass is constant, $\dot{m} = 0$, then we have that

$$m\dot{v} = q(e + v \times b) \,.$$

Plug in the last equation (equation of motion) into the $\dot{k}$ expression:

$$\dot{k} = qe \cdot v + q(v \times b) \cdot v \,.$$

For vectors x, y, z, the product $(x \times y) \cdot z$ is called the *scalar triple product*. This product equals zero whenever $x = z$. In our case, we have that $x = z = v$, or, more precisely, $(v \times b) \cdot v = 0$. Therefore, one finally finds

$$\dot{k} = qe \cdot v \,,$$

the rate at which the particle's kinetic energy changes with respect to time.

This (analytic) solution confirms our guessed model and the approximate solutions. Then, it creates confidence, not only on our intuition, but also on the efficacy of approximate methods.

27.2.5. *Formal Derivation.* Finally, we present a more formal solution, suitable for publishing.

Agree on the given hypotheses and on the symbols and notation previously established.

To begin, model the motion of the charged particle by equating Newton's second law of motion with Lorentz force law to find the particle's equation of motion:

$$m\dot{v} = q(e + v \times b) \,. \tag{27.1}$$

On the other hand, write the particle's kinetic energy as $2k = mv^2 = mv \cdot v$. Then, calculate the change rate of kinetic energy with respect to time $\dot{k}$:

$$\dot{k} = m\dot{v} \cdot v \,. \tag{27.2}$$

Plug eq. (27.1) into eq. (27.2) to find that $\dot{k} = qe \cdot v + q(v \times b) \cdot v$. Since the scalar triple product vanishes this gives, finally,

$$\dot{k} = qe \cdot v \,. \qquad \square$$

The formal solution was obtained from the derivation of the wordy solution. They only differ in presentation. In the formal solution,

- the presentation is brief, concise, straight to the point, but not incomplete. It only leaves "obvious details" to be filled in – for instance, nowhere it is written that $\dot{p} = \dot{m}v + m\dot{v} = m\dot{v}$, because under hypotheses, m is constant, so it is "well-known" that $f = ma$ in such a case;
- equations are referred to by proper, technical names (Newton's second law of motion, scalar triple product and so on);
- only "important" equations, derivations and results are displayed, whereas small equations, non-trivial, but small, derivations and partial results are presented in-line – with the running text;
- verbs changed to the imperative to avoid the use of personal grammar forms – we, us, one and so on – and of the passive voice.

27.3. **Newton's, Lagrange's and Hamilton's Formalism of Classical Mechanics.** Consider a simple harmonic oscillator: a mass m attached to a spring of constant κ object to a force $f[x] = -\kappa x$, where x is the mass position and where frictional forces are neglected. Find the equation of motion for the oscillator.

27.3.1. *Newton's.* The equation of motion reads, directly from Newton's second law of motion:

$$f = m\ddot{x} \implies m\ddot{x} + \kappa x = 0 \,.$$

Notice that $f \neq 0$, thus linear momentum is not conserved. Additionally, the force is central, so no angular momentum defined.

27.3.2. *Lagrange's.* Since the force depends only on position $f = f[x]$, then it is conservative and therefore arises from a potential v given by $2v = \kappa x^2$. This means that Lagrange's formalism can be applied.

The kinetic energy of the mass k is $2k = m\dot{x}^2$. Then, the Lagrangian L for the system is

$$L = k - v = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}\kappa x^2 \,.$$

The generalized force acting on the system is then $L_{,x} = -\kappa x$. Since this is not zero, the generalized momentum is not conserved. Additionally, there are no cyclic quantities. Moreover, since L is t independent, then the total energy of the system is conserved.

On the other hand, the generalized momentum of the system is $L_{,\dot{x}} = m\dot{x}$ and thus $\mathrm{d}(L_{,\dot{x}})/\mathrm{d}t = m\ddot{x}$. Therefore, the equation of motion can be read from Euler-Lagrange's equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = m\ddot{x} + \kappa x = 0 \,.$$

27.3.3. *Hamilton's.* Consider the Lagrangian found in the previous section. By applying Legendre's transform, replace the generalized velocity $\dot{x}$ by the generalized momentum p in L:

- the Lagrangian is well behaved for all x. The Lagrangian first derivative $L_{,\dot{x}}$ is also well behaved for all x. And the second derivative $L_{,\dot{x}\dot{x}} = m$ is always positive, since $m > 0$. Therefore, Legendre's transform can be applied to L.
- Define $p = L_{,\dot{x}} = m\dot{x}$. This implies that $\dot{x} = p/m$.
- Define the Hamiltonian of the system by $H = p\dot{x} - L$. Replace the correspondent quantities to find

$$H = p\dot{x} - L = \frac{p^2}{m} - \left(\frac{1}{2}m\dot{x}^2 - \frac{1}{2}\kappa x^2\right) = \frac{1}{2m}\left(p^2 + \kappa x^2\right) \,.$$

In the last equation, note that H is t independent, so the total energy of the system is conserved.

Finally, find the equations of motion from Hamilton's equations:

$$\dot{x} = [p, H]_{\text{pb}} = \frac{\partial H}{\partial p} = \frac{p}{m}\,,$$
$$\dot{p} = [x, H]_{\text{pb}} = -\frac{\partial H}{\partial x} = -\kappa x\,.$$

27.3.4. *Comparison.* From Newton's formalism, we know that f arises from a potential k, that linear momentum is not conserved, that angular momentum is not defined and that the equation of motion reads $m\ddot{x} + \kappa x = 0$. From Lagrange's, we learn that momentum is not conserved, but total energy is, and that the equation of motion is the same as the one found by Newton's. Finally, from Hamilton's, we find that the momentum is not conserved, but total energy is, that the total energy is given by $2em = p^2 + \kappa x^2$ and that the Hamilton's equations of motion are $\dot{x} = p/m$ and $\dot{p} = -\kappa x$.

27.4. **Nondimensionalization.**

27.4.1. *Damped Oscillator.* Consider a mass m attached to a spring with stiffness k that is set into motion from a equilibrium position x_0 at time t_0. Consider the mass is object to a frictional force proportional to the mass velocity. Then, find the non-dim. equation of motion of the system.

Use Hooke's law to model the restoring force f_{r} of the spring: $f_{\text{r}} = -kx$, where x represents the position of the mass from its equilibrium position.

Then, find the frictional force as $f_{\text{f}} = -c\dot{x}$, where c represents the viscous damping coefficient.

Model the equation of motion by applying Newton's second law of motion:

$$m\ddot{x} + c\dot{x} + kx = 0\,.$$

Before non-dim. the system equation of motion, verify its dimensional homogeneity:

$$\dim m\ddot{x} = \left[ML/T^2\right] = [F]\ ,\ \dim c\dot{x} = [M/T \cdot L/T] = [F] \text{ and } \dim kx = \left[M/T^2 \cdot L\right] = [F]\ .$$

Since all the terms have dimensions of force, $[F]$, the model equation is homogeneous, thus, nondim. can proceed.

To begin with non-dim., rewrite the equation of motion as

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + c\frac{\mathrm{d}x}{\mathrm{d}t} + kx = 0\,.$$

In this 2nd-order ordinary differential equation, the independent variable is t, the dependent one x and the parameters m, c and k.

Scale time $\bar{t}$ and position $\bar{x}$ by finding characteristic quantities t_c and x_c satisfying $\bar{t} = t/t_c$ and $\bar{x} = x/x_c$. With these replacements, find

$$x = x_c\bar{x} \implies \mathrm{d}x = x_c\mathrm{d}\bar{x} \implies \mathrm{d}^2x = x_c\mathrm{d}^2\bar{x}\,,$$
$$t = t_c\bar{t} \implies \mathrm{d}t = t_c\mathrm{d}\bar{t} \implies \mathrm{d}t^2 = t_c^2\mathrm{d}\bar{t}^2\,.$$

Replace the characteristic and scaled quantities in the equation of motion to have

$$\frac{mx_c}{t_c^2}\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + \frac{cx_c}{t_c}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + kx_c\bar{x} = 0\,.$$

Divide the last equation through the coefficient of the highest order term; *i.e.*, mx_c/t_c^2, to get

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + \frac{ct_c}{m}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \frac{kt_c^2}{m}\bar{x} = 0\,.$$

Since the last equation has only one characteristic quantity, do

$$\frac{kt_c^2}{m} = 1 \implies t_c^2 = \frac{m}{k} \implies t_c = \sqrt{\frac{m}{k}}\,.$$

Replace these quantities in the equation of motion:

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + \frac{c}{m}\sqrt{\frac{m}{k}}\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \bar{x} = 0\,.$$

Define the quantity $2\zeta = c/m\sqrt{m/k} = c/\sqrt{mk}$ and replace it in the previous equation

$$\frac{\mathrm{d}^2\bar{x}}{\mathrm{d}\bar{t}^2} + 2\zeta\frac{\mathrm{d}\bar{x}}{\mathrm{d}\bar{t}} + \bar{x} = 0\,.$$

The quantity ζ physically represents the *damping ratio*, whereas the inverse of the characteristic time, denoted ω_0, the *undamped angular frequency of the oscillator*. ω_0 is thus given by

$$\omega_0 = \frac{1}{t_c} = \sqrt{\frac{k}{m}}\,.$$

The value of the damping ratio ζ critically determines the behavior of the system. A damped harmonic oscillator can be:

- Overdamped ($\zeta > 1$): The system returns (exponentially decays) to steady state without oscillating. Larger values of the damping ratio return to equilibrium slower.
- Critically damped ($\zeta = 1$): The system returns to steady state as quickly as possible without oscillating. This is often desired for the damping of systems such as doors.
- Underdamped ($\zeta < 1$): The system oscillates (with a slightly different frequency than the undamped case) with the amplitude gradually decreasing to zero. The angular frequency of the underdamped harmonic oscillator is given by

$$\omega_1 = \omega_0\sqrt{1-\zeta^2}\,.$$

The Q factor of a damped oscillator is defined as

$$Q = 2\pi\frac{\text{Energy stored}}{\text{Energy lost per cycle}}\,.$$

Q is related to the damping ratio by the equation

$$Q = \frac{1}{2\zeta}\,.$$

Finally, the nondimensionalized equation of motion is called the *universal oscillation equation*, since all second order linear oscillatory systems can be reduced to this form.

27.5. **Think Physically.** Maths methods are fine, since they provide a formal way to arrive to and present results in a mechanical way. However, this is the downside from a physical viewpoint: sometimes, physical arguments are forgotten.

The idea in this section is to revert this pattern; *i.e.*, to use physical reasoning supported by maths methods. Let's explain it with examples.

27.5.1. *Circular Motion.* Consider a particle tracing a 2-dim. circular shape while it moves during time t. The particle's position $x[t]$ can be tracked by means of two Cartesian coordinates, say $x[t] = [x[t]\,, y[t]]$. Now, in this case, maths are better expressed by using polar coordinates; *i.e.*, by tracking the particle's position by $x[t] = [r[t]\,, \theta[t]]$, where $r[t]$ is the distance from the circle's center, $\mathcal{O}$, to $x[t]$ and $\theta[t]$ the angle between the vector from $\mathcal{O}$ to $x[t]$ and a reference line: chosen to be the x-axis in Cartesian coordinates.

From a mathematical perspective, the next step is to express $[x[t]\,, y[t]]$ as functions of $[r, \theta]$, then write $x[t]$ and finally find all the physical parameters: velocity, acceleration, forces, equations of motion and so on. This is the path that we will not follow. We will use physical reasoning to find $x[t]$ and other parameters.

The particle position $x[t]$ can be tracked by a vector with two coordinates: a radial component $r[t]$ measuring only radial displacements (as if our circle would be expanding or contracting) and an arc component $s[t]$ measuring only the arclength the particle traces in the circle. In this coordinates, $x[t] = [r[t]\,, s[t]]$. The particle velocity, then, becomes

$$\dot{x}[t] = [\dot{x}[t]\,, \dot{s}[t]]\ .$$

Now we relate $\dot{s}[t]$ to $\theta[t]$ via the definition of radians: $\theta = s/r$. So $\dot{s} = r\dot{\theta}[t]$ and we call $\dot{\theta}[t]$ the angular velocity. Therefore, the equation of $\dot{x}[t]$ becomes

$$\dot{x}[t] = \left[\dot{r}, r\dot{\theta}\right]\ .$$

The first component, $\dot{r}$, called radial velocity, measures the rate at which the particle's position moves towards or apart the center of the circle, $\mathcal{O}$, whereas The angular velocity measures how the angle changes in time. Note that $r\dot{\theta}$ is tangent to the curve traced by the particle.

28. Lists

List of derivatives: https://en.wikipedia.org/wiki/List_of_derivatives.

List of derivatives: https://en.wikipedia.org/wiki/Derivative.

List of integrals: https://en.wikipedia.org/wiki/List_of_integrals.

List of Taylor series: https://en.wikipedia.org/wiki/Taylor_series.

List of vector identities: https://en.wikipedia.org/wiki/Vector_identities.

List of vector calculus identities: https://en.wikipedia.org/wiki/Vector_calculus_identities.

List of vector algebra identities: https://en.wikipedia.org/wiki/Vector_algebra_relations.

Cylindrical and polar coordinates: https://en.wikipedia.org/wiki/Del_in_cylindrical_and_spherical_coordinates.

Orthogonal coordinates: https://en.wikipedia.org/wiki/Orthogonal_coordinates.

Curvilinear coordinates: https://en.wikipedia.org/wiki/Curvilinear_coordinates.

Vector algebra *vs*. Geometric algebra: https://en.wikipedia.org/wiki/Comparison_of_vector_algebra_and_geometric_algebra.

List of mathematical functions: https://en.wikipedia.org/wiki/List_of_mathematical_functions.

List of sums: https://en.wikipedia.org/wiki/Sums.

List of trig. indentities: https://en.wikipedia.org/wiki/Trigonometric_identity.

Differential equations of mathematical physics: https://en.wikipedia.org/wiki/Differential_equations_of_mathematical_physics.

Harmonic oscillator: https://en.wikipedia.org/wiki/Harmonic_oscillator.

Dimensional analysis: https://en.wikipedia.org/wiki/Dimensional_analysis.

List of dimensionless numbers: https://en.wikipedia.org/wiki/List_of_dimensionless_numbers#List_of_dimensionless_quantities.

List of orders of magnitude: https://en.wikipedia.org/wiki/Orders_of_magnitude_(numbers).

List of physical quantities: https://en.wikipedia.org/wiki/List_of_physical_quantities.

Physical property: https://en.wikipedia.org/wiki/Physical_property.

Physical quantity: https://en.wikipedia.org/wiki/Physical_quantity.

29. Notation

29.1. General Commands.

- to be defined by: a defby b: $a \coloneqq b$.
- difference operator: diff a: Δa.
- text in equations: eqtxt.

29.2. Sets.

- set: set A: $\mathcal{A}$.
- elements of a set: elset(a,b,c): $\{a, b, c\}$.
- set with a property: setprop(x)(x>0): $\{x : x > 0\}$.
- Cartesian (set) product: set A sprod set B: $\mathcal{A} \otimes \mathcal{B}$.
- Cartesian power: nset An: $\mathcal{A}^n$.
- union of sets: set A union set B: $\mathcal{A} \cup \mathcal{B}$.
- intersection of sets: set A inter set B: $\mathcal{A} \cap \mathcal{B}$.
- Dim-grade space (2 is the dimension and 3 is the grade): dgspace V23: V_3^2.
- n-dim Euclidean space: espace n: $\mathcal{E}^n$.
- n-dim Minkowski space: mkspace n: $\mathcal{M}^n$.
- geometric algebra: ga: $\mathcal{G}$.
- geometric algebra on a n-dim. linear space $\mathcal{V}^n$: nga n: $\mathcal{G}^n$.
- dimension grade geometric algebra (2 is the dimension and 3 is the grade): dgga 23: $\mathcal{G}_3^2$.
- tuple: tuple(1,2,3): $[1, 2, 3]$.

29.3. Probability.

- event A: A: A.
- not event A: lnot A: $\neg A$.
- probability of event A occurring: p vat A: $p[A]$.
- probability of event A not occurring: p vat(lnot A): $p[\neg A]$.
- A and B: A land B: $A \wedge B$.
- A or B: A lor B: $A \vee B$.
- A given B (provided B): A given B: $A \,|\, B$.

29.4. Functions.

- function definition: fdef(f)(set A cartprod set B)(set R): $f : \mathcal{A} \otimes \mathcal{B} \to \mathcal{R}$.
- function mapping: fmap(f)(x)(x**2): $f : x \mapsto x^2$.
- maps to: x mapsto x**2: $x \mapsto x^2$.
- function class (calculus): class k: C^k.
- value at: f vat(x): $f[x]$.
- function composition: f fcomp g: $f \circ g$.
- a binary operation: a bprod b: $a * b$.
- derivative operator (on functions): fder f: Df.
- partial derivative operator (on functions): derivative with respect to x of f: fpder xf: $\partial_x f$.

29.5. Sequences and Series.

- sequence: seq ak()(): $\{a\}_k$.
- sequence with limits: seq(a)(k)(k=1)(10): $\{a\}_k^{k=1}\, 10$.
- series: serie(ak)(k=1)(n): $\sum_{a_k}^{k=1} n$.
- Fibonacci numbers: fib vat 10: fib[1] 0.

29.6. Geometric Algebra.

- point: point P: $\mathcal{P}$.
- curve: curve C: $\mathcal{C}$.
- surface: surf S: $\mathcal{S}$.
- region: region R: $\mathcal{R}$.
- bound: bound region A: $\partial\mathcal{A}$.
- vector: vec a: $\vec{a}$.
- normal (unit) vector: nvec(a): $\hat{a}$.
- omitted vector from a product: ovec(a): $\breve{a}$.
- clifs or multivectors: (use capitals) A: A.
- pseudoscalar: pscl: i.
- better, less typing, use lower-case for vectors a and upper-case for other objects (bivectors, trivectors,...): A.
- use i for the pscl: i.
- orthogonal: ortho: $a \perp b$.
- parallel: parallel: $a \parallel b$.
- to be perpendicular to: perto(a)(b): $a_{\perp b}$.
- to be parallel to: parto(a)(b): $a_{\parallel b}$.
- to be orthogonal to: ortto(a)(b): $a_{\perp b}$.
- to be colinear to: colto(a)(b): $a_{\parallel b}$.
- projection of p onto q: projon pq: $p_{\parallel q}$.
- rejection of p onto q: rejon pq: $p_{\perp q}$.
- magnitude: magn(a): $|a|$.
- inverse: inv(a): a^{-1}.
- reverse: rev(a): $a^\dagger$.
- hodge dual: hdual(a): ${}^*a$.
- anticommutator: acom(a)(b): $[a, b]_+$.
- commutator: com(a)(b): $[a, b]_-$.
- expanded anticommutator: xacom(a)(b): $ab + ba$.

- expanded commutator: xcom(a)(b): $ab - ba$.
- step: step(A1): $\langle A \rangle_1$.
- scalar step: sstep(A): $\langle A \rangle_0$.
- grade operator: Grade A: grade A.
- grade: grade A2: $\langle A \rangle_2$.
- scalar grade: sgrade A: $\langle A \rangle_0$.
- clif with step: sclif Ak: $A_{\bar{k}}$.
- even part: even(A): A_+.
- odd part: odd(A): A_-.
- gorm (geometric norm?): gorm A: gorm A.
- expanded gorm: xgorm A: $\langle A^\dagger A \rangle_0$.
- metric: metric: g.
- Kronecker delta: kron: δ.
- signature: diag a: diag a.
- signature: sign a: sig a.
- inner product: iprod: $a \cdot b$.
- outer product: oprod: $a \wedge b$.
- cross product: cprod: $a \times b$.
- cannonical decomposition of the geometric product: cgprod ab: $a \cdot b + a \wedge b$.

29.7. **Geometric Calculus.**

- ordinary one-dim. derivative: dx x: $\mathrm{d}x$.
- ordinary time derivative (dot derivative): dt x: $\dot{x}$.
- ordinary second time derivative (dot-dot derivative): ddt x: $\ddot{x}$.
- expanded ordinary derivative: xod Hq: $\dfrac{\mathrm{d}H}{\mathrm{d}q}$.
- expanded partial derivative: xpd Hq: $\dfrac{\partial H}{\partial q}$.
- expanded material derivative: xmd phi t: $\dfrac{\partial \phi}{\partial t}$.
- expanded n-order ordinary derivative: nxod 3xt: $\dfrac{\mathrm{d}^3 x}{\mathrm{d}t^3}$.
- expanded n-order partial derivative: nxpd 3xt: $\dfrac{\partial^3 x}{\partial t^3}$.
- comma derivative: cder phi k: $\phi_{,k}$.
- semi-colon: covariant derivative: coder(cntens Aa)(k): $A^a{}_{;k}$.
- material derivative: mder phi t: $\mathrm{D}_t \phi$.
- absolute time derivative: abstder a: $\dot{\bar{a}}$.
- Christoffel symbol: chris abc: Γ^a_{bc}.
- geometric derivative: gder(a): ∇a.
- directional derivative: dder(F)(a): $\nabla_a F$.
- Laplace derivative: lder(a): $\nabla^2 a$.
- Laplace operator: lap a: lap a.
- D'Alembert operator: dalder(phi): $\Box \phi$.
- gradient: grad(phi): grad ϕ.
- divergence: div(phi): div ϕ.
- curl: curl(phi): curl ϕ.
- rotational (curl): rot(phi): rot ϕ.

29.8. **Tensors.**

- spatial coordinates: scord k: x^k.
- spatial coordinates time derivative: dtscoord k: $\dot{x}^k$.
- tensor: tens T: T.
- (empty) slot: tuple(slot, a, slot): $[-, a, -]$.
- tensor product: a tprod b: $a \otimes b$.
- tensor contraction: tcont(a tprod b): cont$(a \otimes b)$.
- indexed tensor contraction: itcont(1,2)(a tprod b tprod c): $\mathrm{cont}_{1,2}(a \otimes b \otimes c)$.
- tensor components: tcomp T: comp T.
- covariant tensor components: cotens T(ij): T_{ij}.
- contravariant tensor components: cntens T(ij): T^{ij}.
- Levi-Civita tensor: lct: ϵ.
- covariant tensor time derivative: dtcotens ak: $\dot{a}_k$.
- contravariant tensor time derivative: dtcntens ak: $\dot{a}^k$.

29.9. **Index Notation.**

- frame element, vector: fvec: γ.
- frame: frm(k): $\{\gamma_k\}$.
- indexed frame: ifrm(k)(0)(n): $\{\gamma_k; 0 \ldots n\}$.
- reciprocal frame: rfrm k: $\{\gamma^k\}$.
- indexed frame vector: ifvec k: γ_k.
- indexed reciprocal frame vector: rfvec k: γ^k.
- components of vector in frame: comp vk: v^k.
- components of vector in reciprocal frame: rcomp vk: v_k.
- metric coefficients in frame: imet kl: g_{kl}.
- metric coefficients in reciprocal frame: rmet kl: g^{kl}.
- mixed metric coefficients: mmet kl: g^k_l.
- kronecker delta coefficients in frame: ikron kl: δ_{kl}.
- kronecker delta coefficients in reciprocal frame: rkron kl: δ^{kl}.
- mixed kronecker coefficients: mkron kl: δ^k_l.

- indexed geometric derivative (in reciprocal frame): igder k: ∂_k.
- indexed geometric derivative (in frame): rgder k: ∂^k.

29.10. **Dimensional Analysis.**

- dimension: dim k: $\dim k$.
- dimension and system of units (use underscore as lim): sdim(FLT)k: $\dim_{FLT} p = \left[F/L^2\right]$.
- unit: unit k: $\mathrm{unit}\, k$.
- physical dimension: phdim k: $[k]$.
- dimensionless quantity: kdim: Π.
- characteristic physical quantity: chpq a: a_c.
- scaled physical quantity: scpq x: $\bar{x}$.
- reynolds number: rey: Π_{re}.
- biot number: biot: Π_{bi}.

29.11. **Mechanics.**

- position vector: pvec vat t: $x[t]$.
- value at time and position vector (for functions): vattpvec: $[t, x[t]]$.
- separation vector between points: svec: s.
- linear momentum: lmom: p.
- kinetic energy: ken vat t: $k[t]$.
- potential energy: pen vat t: $v[t]$.
- Action functional: action: A.
- Lagrange function: lag: L.
- Hamilton function: ham: H.
- Hamilton Kinetic energy: hken: H_{kin}.
- Hamilton potential energy: hpen: H_{pot}.
- Euler-Lagrange Equation: eleqn(q)(i): $\dfrac{\mathrm{d}}{\mathrm{d}t}\left(\dfrac{\partial L}{\partial \dot{q}^i}\right) - \dfrac{\partial L}{\partial q^i}$.
- contra-variant vector: cnvec pi: p^i.
- (contra-variant) indexed vector: ivec pi: p^i.
- covariant vector (covector): covec pi: p_i.
- basis vector: bvec: γ.
- natural basis vector: nbvec i: γ_i.
- dual basis vector: dbvec i: γ^i.
- generalized position vector: gpvec: q.
- indexed generalized position: gpos i: q^i.
- indexed generalized velocity: gvel i: $\dot{q}^i$.
- indexed generalized momentum: gmom i: p_i.
- indexed generalized force: gfor i: f_i.

29.12. **Transport Phenomena.**

- thermodynamic temperature: temp: θ.
- substance: subs A: A.
- flux: flux: j.
- mass flux of substance A: mflux A: j_{A}.
- concentration of substance A: conc A: c_{A}.
- bracket concentration of substance A: bconc Aa: $[\mathrm{A}]^a$.
- chemical amount of substance A: amount A: n_{A}.
- reaction rate of substance A: rrate A: r_{A}.
- time derivative of conc.: dtconc A: $\dot{c}_{\mathrm{A}}$.
- time derivative of chem. amount: dtamount A: $\dot{n}_{\mathrm{A}}$.

29.13. **Various.**

- Iverson brackets: iverson(k=l): $[k = l]_{\mathrm{iv}}$.
- Poisson brackets: poisson(f,g): $[f, g]_{\mathrm{pb}}$.
- matrix representation: mtrx metric: $[g]$.
- Taylor series generated by f at the point a: tseries(f)(x)(a): $T_\infty f[x\,;a]$.
- Taylor polynomial of degree n generated by f at the point a: ntpol(n)(f)(x)(a): $T_n f[x\,;a]$.
- Fourier series generated by f at the point x: fseries fx: $F_\infty f[x]$.
- partial sums of the Fourier series generated by f at the point x: nfsum(f)(n)(x): $F_n f[x]$.
- Legendre transform of a function: ltrans f: $f_\star$.
- conjugate variable (under Legendre transf.): cvar: $x_\star$.
- average quantity: avg a: $\langle a \rangle$.

29.14. **Constants.**

- Boltzmann constant: boltz: k_{b}.
- speed of ligth in vacuum: ligth: c.
- Avogadro's number: avog: n_{a}.

29.15. **Alphabet.**

- Latin minuscules:

$abcdefghijklmnopqrstuvwxyz$.

- Latin majuscules:

$ABCDEFGHIJKLMNOPQRSTUVWXYZ$.

- Greek:

$\alpha\beta\gamma\delta\epsilon\varepsilon\zeta\eta\theta\vartheta\iota\kappa\lambda\mu\nu\xi\pi\varpi$

- Greek:

$\rho\varrho\sigma\varsigma\tau\upsilon\phi\varphi\chi\psi\omega\Gamma\Delta\Theta\Lambda\Xi\Pi\Sigma\Upsilon\Phi\Psi\Omega$