Mathematics

Laurent Lemmens
GQCG

Friday 18th May, 2018 08:50

Abstract

These notes serve as a summary of mathematics I have thorougly studied.

Contents

1 Functions - relations between sets

1.1 Sets

Modern algebra starts from the notion of a set. A set S is a collection of elements:

$$S = \{s_1, s_2, \ldots, s_n\}\,. \tag{1.1}$$

Some examples of sets are the set of natural numbers $\mathbb{N}$, the set of real numbers $\mathbb{R}$, the set of complex numbers $\mathbb{C}$. We can also have smaller sets, for example

$$S = \{0, 1\}\,, \tag{1.2}$$

being the set of the numbers 0 and 1. Sets don't necessarily have to contain only numbers. We can, for example, collect all invertible $n \times n$-matrices in a set:

$$\mathrm{GL}(n, \mathbb{R}) = \{A \in \mathbb{R}^{n\times n} \mid A \text{ is invertible}\}\,, \tag{1.3}$$

in which the symbol GL has to do with 'general linear', but more on that when we encounter the general linear group.

The previous examples are all concrete (i.e. not abstract) examples of sets. Now say we have a mathematical object, called E (we haven't specified anything about it), we can say that

$$G = \{E\} \tag{1.4}$$

is also a set, but in a more abstract sense than the previous examples. We can enlarge this set by adding the elements C_2, σ_v and σ_v', to end up with

$$G = \{E, C_2, \sigma_v, \sigma_v'\}\,, \tag{1.5}$$

in which we still haven't specified anything about the nature of its elements, but in mathematics that is perfectly fine.

1.2 Functions - introduction

Naturally, if we have two different sets, we would like to be able to define relations between their elements. This is exactly what a function does. A function (or map, mapping, these are all synonyms) is a relation between two sets X and Y:

$$f : X \to Y : x \mapsto y = f(x), \tag{1.6}$$

subject to the important condition that every input $x \in X$ is related to exactly one output $y \in Y$. We would read the definition in equation (1.6) as follows: f is a function from the set X to the set Y, in which every $x \in X$ is related to a $y \in Y$, which we call $f(x)$.

We give special names to the sets X and Y, depending on which role they play in the function. The set X is called the domain of the function: it is the set of all inputs for the function. The set Y is then called the function's codomain: it is the set of values that *could* occur as output values for the function. We can then define another set, Z, which is called the image of the function: it is the set of *actual* values for the outputs. A visual clarification of the terms can be found in Figure 1.1.

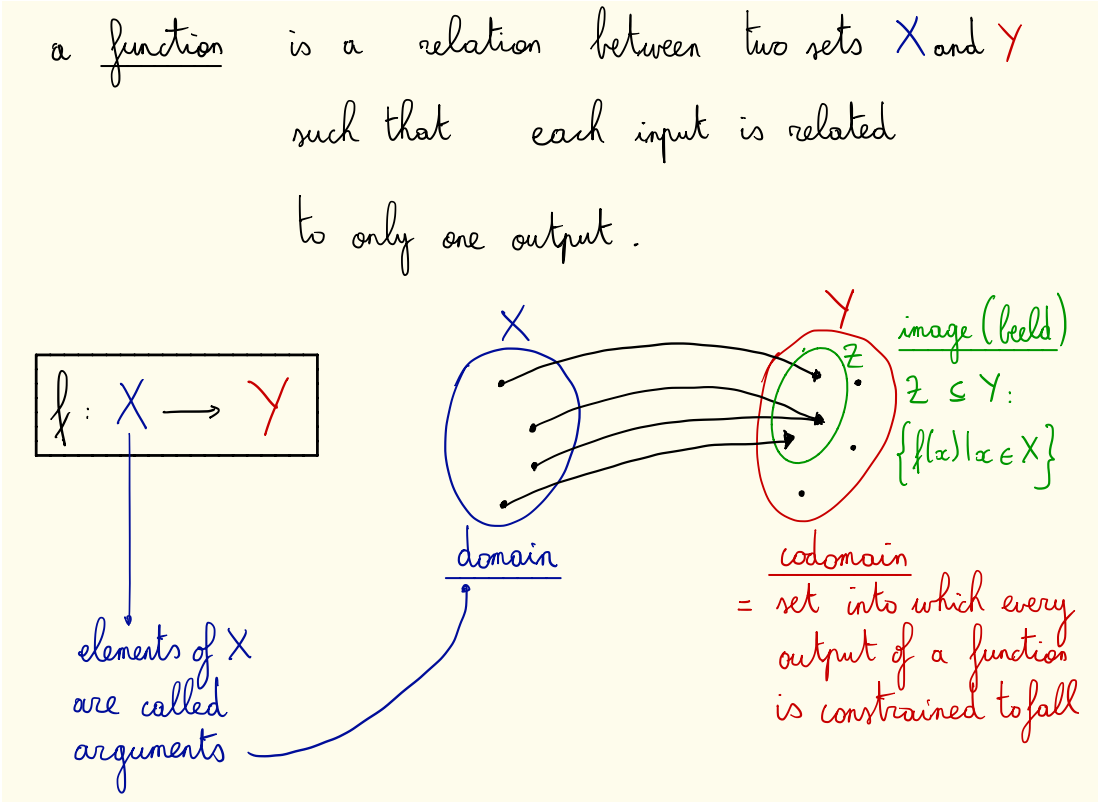

Figure 1.1: The definition of a function, and a visual clarification of the terms domain, codomain and image. The domain is the set of input values, the codomain is the set of possible output values, and the images is the set of actual output values.

An example of a function could be:

$$f : \mathbb{R} \to \mathbb{R} : x \mapsto f(x) = x^2 + 2 \tag{1.7}$$

Its domain is $\mathbb{R}$, and its codomain is also $\mathbb{R}$. Here, we can also see the difference between the codomain and the image. The codomain of this function is defined to be $\mathbb{R}$, but its image $[2, +\infty[$. Another example is shown in Figure 1.2.

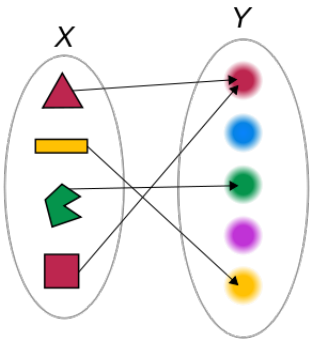

Figure 1.2: An example of a function that maps an object to its color.

The reason why this is a function is because every input element of the set X, being shapes in a certain color, is related to its color, represented as elements of the set Y. In this case, Y, the codomain is the set of colors depicted as elements of Y, and the range is the set of colors red, green and yellow.

We have seen some examples of functions already, but what are some examples relations that are not functions? There are actually two requirements to the definition of a function:

1. Every input (element of the domain) has to be related to an output (element of the codomain)
2. No two outputs (elements of the codomain) may be related to the same input (element of the domain)

With these two criteria in mind, it is possible to come up with many examples of relations between two sets that are not functions, for example those shown in Figure 1.3.

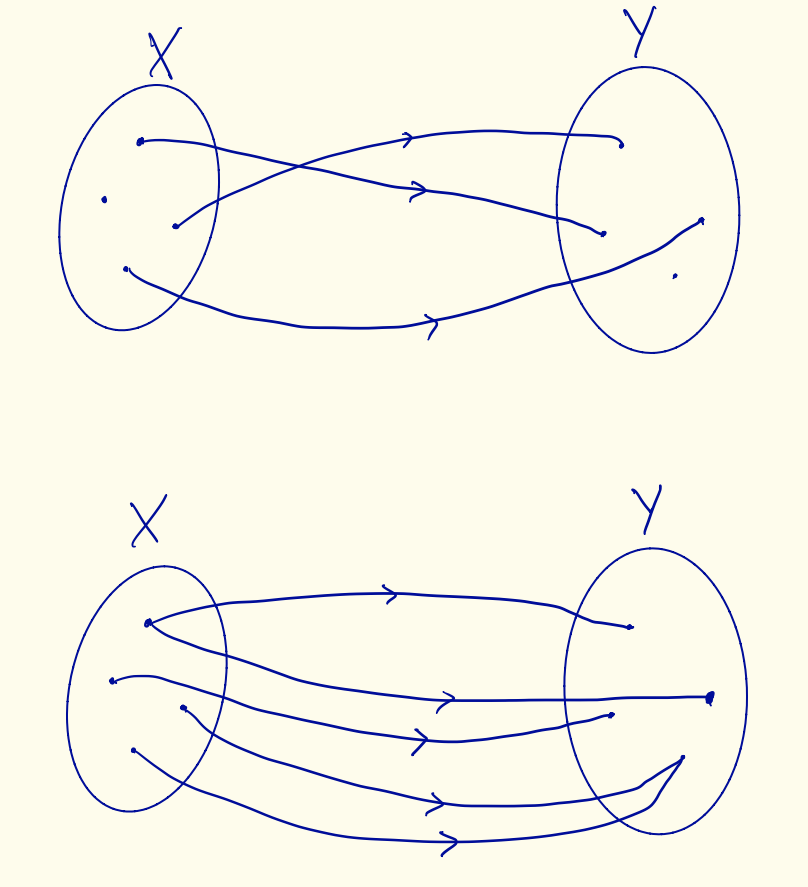

Figure 1.3: Two examples of non-functions. The first relation between X and Y is not a function, because not every element of X is related to an element of Y. The second figure is not a function because there is an element of X related to two elements of Y.

The identity function on a set S is the function:

$$\mathrm{id}_S : S \to S : \mathrm{id}_S(s) = s\,. \tag{1.8}$$

An operation can hardly be distinguished from the definition of a function that was already given. It is a function

$$\omega : Y_1 \times \cdots \times Y_N \to X\,, \tag{1.9}$$

which takes as an input one element of every of its input sets Y_i and relates that combination to the output set X. For me, this definition is exactly equal to the definition that was given previously. However, it seems that the term operation is often used to talk about associativity,

commutativity, etc.

We know many operations, for example number addition, which takes two numbers as input and relates that combination to a number. Number multiplication is another example of a binary operation: we take two numbers, whose product is a number. If we take, for example, set of $n \times n$ matrices, then matrix multiplication is an example of a binary operation. Furthermore, we don't have to take inputs from the same sets! We can define scalar multiplication as an operation that relates a number and a vector to a vector.

1.3 Classification of functions

By now, we have seen some examples of functions. In order to classify functions, we will introduce the terms surjection, injection, and bijection. For a visual overview of the terms, see Figure 1.4.

A function is called surjective (or: onto), if every element of its codomain is mapped to. *Sur* means 'above', which relates to the fact that the function's codomain is completely covered. Its definition can be written as follows: a function f is said to be surjective if

$$f : X \to Y : \forall y \in Y : \exists x \in X : f(x) = y\,. \tag{1.10}$$

A function is called injective (or: one-to-one) if no element of its codomain is mapped to twice. Mathematically, we would write: f is an injective function if

$$f : X \to Y : \forall a, b \in X : f(a) = f(b) \Rightarrow a = b\,. \tag{1.11}$$

A function is called bijective (or: one-to-one and onto, a one-to-one correspondence), if it is both injective and surjective: every element of the codomain is mapped to by exactly one element of the domain.

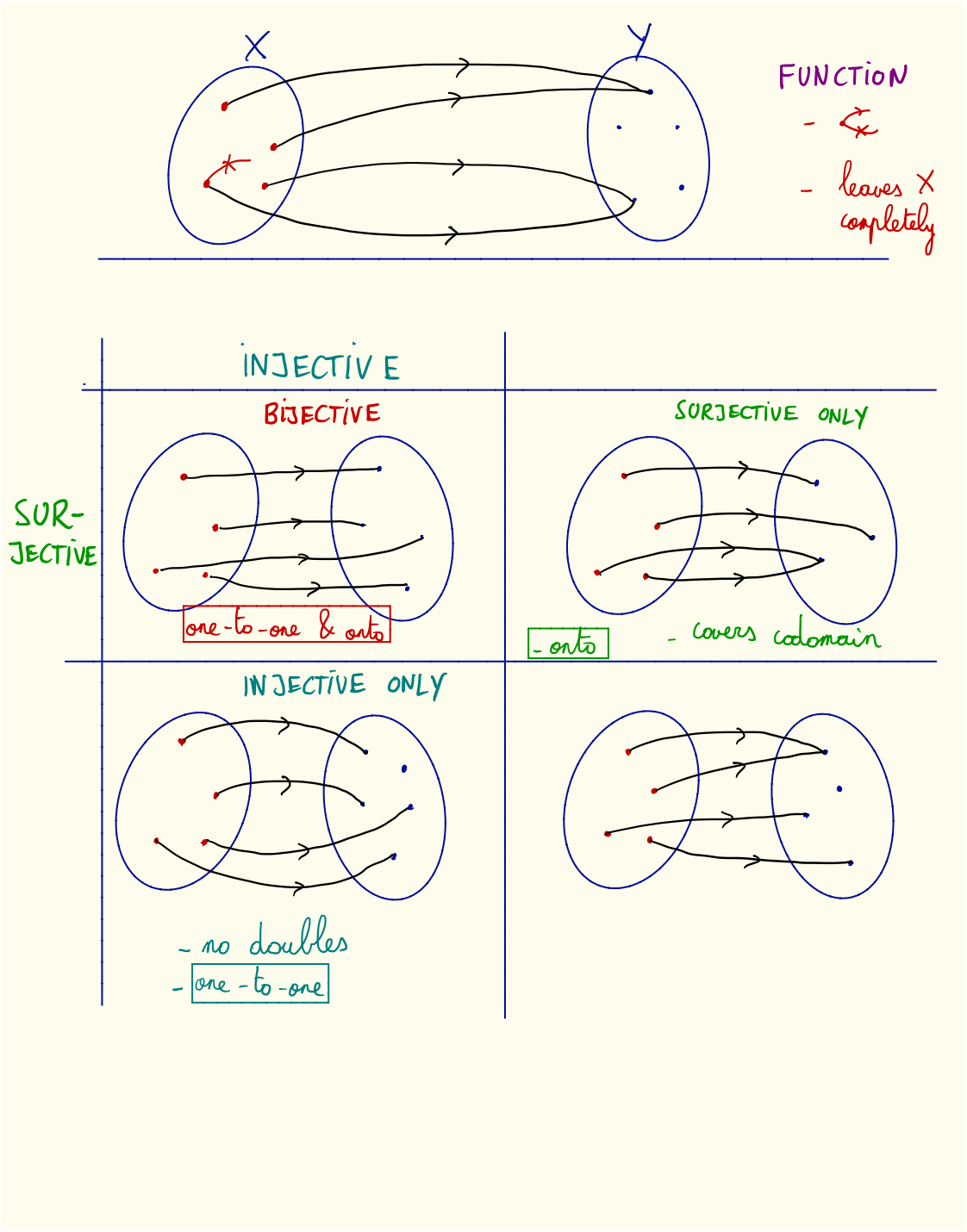

Figure 1.4: A visual scheme of the terms surjective, injective and bijective.

Given Figure 1.4, we can already see examples of surjective and injective functions. In Figure 1.2, we can see an example of a function that is not surjective (not every color is mapped to), nor injective (the color red is mapped to twice).

2 Algebraic structures

Algebraic structures are a combination of a set, together with one or more operations, satisfying a list of axioms.

2.1 The mathematical definition of a group

A group has a mathematical definition. It is a set G[1] with elements $G = \{g_1, g_2, \ldots, g_n\}$, together with an operation $\cdot$ (which is often called the group multiplication), meeting the following axioms:

1. closedness
$$\forall g_1, g_2 \in G : g_1 \cdot g_2 \in G \tag{2.1}$$
2. associativity
$$\forall g_1, g_2, g_3 \in G : g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3 \tag{2.2}$$
3. identity element
$$\exists!\, e \in G : \forall g \in G : e \cdot g = g \cdot e = g \tag{2.3}$$
4. inverses
$$\forall g \in G : \exists!\, g^{-1} \in G : g \cdot g^{-1} = g^{-1} \cdot g = e \tag{2.4}$$

If the axiom

5. commutativity
$$\forall g_1, g_2 \in G : g_1 \cdot g_2 = g_2 \cdot g_1 \tag{2.5}$$

is also met, the group is called Abelian.

2.2 Examples of groups

Since the definition of a group is so abstract, let us try to examine some examples of groups.

As a first, let us consider a group that is familiar to all of us. Let us take the set $\mathbb{R}_0$: the rational numbers excluding 0, together with the operation of multiplication. We can check that every group axiom holds (the identity element is 1, and we know the inverse of every real number),

[1] We often use the same symbol to denote the set of group elements and the actual group. I don't think there is anything wrong with this, as the distinction is most often clear from the context.

even the commutative one. We can therefore say that $\mathbb{R}_0$ with multiplication is an Abelian group.

In section 2.1, we gave a general name to the group operation: group multiplication. This doesn't mean that the group operation can't be addition, for example, as 'group multiplication' is just a name. A perfectly valid example of an Abelian group is the set of integers $\mathbb{Z}$, together with addition. Again, we can check that all group axioms hold (the identity element is 0, and we all know the inverse of integers with respect to addition).

As a slightly more complicated example of a group, we will consider the general linear group over $\mathbb{R}$ of degree n, denoted by $\mathrm{GL}(n, \mathbb{R})$. This is the set of all invertible $n \times n$-matrices with real entries, with the operation of matrix multiplication. The identity element is I_n: the $n \times n$-identity matrix (a diagonal matrix with 1 on the diagonal), and since we have specified the set as being the set of invertible matrices, every matrix has an inverse. We should emphasize that $\mathrm{GL}(n, \mathbb{R})$ is not Abelian, as, in general, matrix multiplication is not commutative. A special case of this group is formed by requiring that the determinant of the invertible $n \times n$-matrices is equal to 1. We call this set of matrices, together with matrix multiplication, the special linear group $\mathrm{SL}(n)$.

Many sets of matrices, together with the operation of multiplication form a group. We have for example $\mathrm{O}(n)$, being the set of $n \times n$ orthogonal ($Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I_n$) matrices under matrix multiplication. A special group that is related to $\mathrm{O}(n)$ is $\mathrm{SO}(n)$, being the set of orthogonal matrices with determinant equal to 1, under matrix multiplication. Furthermore, we also have the group $\mathrm{U}(n)$, being the set of $n \times n$ unitary matrices ($UU^\dagger = U^\dagger U = I_n$), under the group operation of matrix multiplication. Again, a special variant is $\mathrm{SU}(n)$, being the set of $n \times n$ unitary matrices with determinant equal to 1, under matrix multiplication.

The previous examples are examples of matrix Lie groups.

As a first more abstract example, let us take a look at the trivial group. It consists of the set $G = \{e\}$ under group multiplication. We have to specify that e is the element for which equation (2.3) holds: e is the identity element, and consequently its own inverse. With this in mind, we can check that the trivial group is Abelian.

As another abstract example, let's take the set of elements

$$G = \{E, C_2, \sigma_v, \sigma_v'\}\,, \tag{2.6}$$

with the multiplication table given in Table 1.

	E	C_2	σ_v	σ_v'
E	E	C_2	σ_v	σ_v'
C_2	C_2	E	σ_v'	σ_v
σ_v	σ_v	σ_v'	E	C_2
σ_v'	σ_v'	σ_v	C_2	E

Table 1: An example multiplication table

A multiplication table is read as follows. Take an element from the first column (for example E), and take an element of the second column (C_2), and find their product as $C_2 \cdot E = C_2$ (note that we read group multiplication conventionally from right to left). This set G, together with the multiplication $\cdot$ specified in the multiplication table, forms a group as all four group axioms are fulfilled. As commutativity is also fulfilled[2], this group is even Abelian.

We can even introduce bigger sets:

$$G = \{E, C_3, C_3^2, \sigma_v, \sigma_v', \sigma_v''\}\,, \tag{2.7}$$

	E	C_3	C_3^2	σ_v	σ_v'	σ_v''
E	E	C_3	C_3^2	σ_v	σ_v'	σ_v''
C_3	C_3	C_3^2	E	σ_v'	σ_v''	σ_v
C_3^2	C_3^2	E	C_3	σ_v''	σ_v	σ_v'
σ_v	σ_v	σ_v''	σ_v'	E	C_3^2	C_3
σ_v'	σ_v'	σ_v	σ_v''	C_3	E	C_3^2
σ_v''	σ_v''	σ_v'	σ_v	C_3^2	C_3	E

Table 2: Another example of a multiplication table

Given the multiplication table in Table 2, we can verify that the set G, together with the group multiplication forms a non-Abelian group.

2.3 The mathematical definition of a field

A field is a set $\mathbb{F}$ together with two binary operations $+$ and $\cdot$, which fulfills

1. $\mathbb{F}$, together with the operation $+$ is an Abelian group

[2] An easy way to confirm the commutative property, is to verify that the multiplication table is symmetric with respect to its diagonal.

2. $\mathbb{F}\backslash\{0_+\}$[3], together with the operation $\cdot$ is an Abelian group,

3. $\cdot$ is distributive with respect to $+$

The last property, distributivity of $\cdot$ over $+$ means the following:

$$\forall a, b, c \in \mathbb{F} : a \cdot (b + c) = a \cdot b + a \cdot c \tag{2.8}$$

$$(a + b) \cdot c = a \cdot c + b \cdot c \tag{2.9}$$

Some important examples of fields include the field of the real numbers without 0 ($\mathbb{R}_0$) together with multiplication and addition, and the field of the complex numbers without 0 ($\mathbb{C}_0$) with the operations addition and multiplication.

In some sense, this distributive property just means that scalar multiplication is a bilinear operation.

2.4 The mathematical definition of a vector space

A vector space over a field F is a set of vectors ($\mathbf{v} \in V$) together with two binary operations: the vector addition, $+$, and the scalar multiplication with an element of the field, $\cdot$, fulfilling the following axioms:

1. $(V, +)$ is an Abelian group

2. V is closed with respect to scalar multiplication

$$c \cdot \mathbf{v} \in V \tag{2.10}$$

3. V has an identity element for scalar multiplication

$$1 \in F : 1 \cdot \mathbf{v} = \mathbf{v} \tag{2.11}$$

4. scalar multiplication is compatible with field multiplication

$$a \cdot (b \cdot \mathbf{v}) = (ab) \cdot \mathbf{v} \tag{2.12}$$

5. scalar multiplication is distributive over vector addition

$$a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + a \cdot \mathbf{v} \tag{2.13}$$

6. scalar multiplication is distributive over field addition

$$(a + b) \cdot \mathbf{u} = a \cdot \mathbf{u} + b \cdot \mathbf{u} \tag{2.14}$$

[3] The set $\mathbb{F}$ without the identity element of the operation $+$.

2.5 Examples of vector spaces

Every field $\mathbb{F}$ is, in a sense, a vector space over itself, in which scalar multiplication is replaced by the field multiplication, and vector addition is replaced by field addition.

$\mathbb{R}^n$, with elements being column matrices of dimension n, together with matrix addition and scalar multiplication, forms a vector space over $\mathbf{R}$.

$\mathbb{R}^{m \times n}$, with elements being the $(m \times n)$-matrices, together with matrix addition and scalar multiplication, forms a vector space over $\mathbf{R}$.

3 Linear maps

3.1 Linear maps and linear operators

A linear map T is a function (= map, mapping) between two vector spaces V and W over a field $\mathbb{F}$:

$$T : V \to W \,, \tag{3.1}$$

such that $\forall \mathbf{v}_1, \mathbf{v}_2 \in V; \forall a \in \mathbb{F}$:

$$T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2) \qquad T \text{ 'preserves' vector addition} \tag{3.2}$$

$$T(a\mathbf{v}_1) = aT(\mathbf{v}_1) \qquad T \text{ 'preserves' scalar multiplication} \tag{3.3}$$

We will call the set of all linear maps from V to W $\mathcal{L}(V, W)$. If we define the sum of two linear maps S and T and the scalar product of an element $a \in \mathbb{F}$ with a linear map T such that $\forall \mathbf{v} \in V$:

$$(S + T)(\mathbf{v}) = S(\mathbf{v}) + T(\mathbf{v}) \tag{3.4}$$

$$(aS)(\mathbf{v}) = a(S(\mathbf{v}))\,, \tag{3.5}$$

respectively, we can show that $\mathcal{L}(V, W)$ forms a vector space over the field $\mathbb{F}$.

A linear operator is a linear map T from a vector space V to itself:

$$T : V \to V \,. \tag{3.6}$$

Obviously, $\mathcal{L}(V) = \mathcal{L}(V, V)$ also forms a vector space over $\mathbb{F}$.

3.2 Bilinear maps

Let U, V, W be vector spaces over a field $\mathbb{F}$. A bilinear function is a function

$$f : U \times V \to W : (\mathbf{u}, \mathbf{v}) \mapsto f(\mathbf{u}, \mathbf{w}) = \mathbf{w}\,, \tag{3.7}$$

such that f is linear in both of its arguments. This means that $\forall \mathbf{u}_1, \mathbf{u}_2 \in U; \forall \mathbf{v}_1, \mathbf{v}_2 \in V; \forall a, b, c, d \in \mathbb{F}$:

$$f(a\mathbf{u}_1 + b\mathbf{u}_2, c\mathbf{v}_1 + d\mathbf{v}_2) = ac\, f(\mathbf{u}_1, \mathbf{v}_1) + ad\, f(\mathbf{u}_1, \mathbf{v}_2) + bc\, f(\mathbf{u}_2, \mathbf{v}_1) + bd\, f(\mathbf{u}_2, \mathbf{v}_2)\,. \tag{3.8}$$

An example of a bilinear map is general matrix multiplication. In the most general case, matrix multiplication is a bilinear map between $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{m\times p}$.

In the case that $U = V$, and W is the field $\mathbb{F}$ itself, we would talk about a bilinear form.

An example of a bilinear form would be an inner product on V.

4 Algebras

Now that we have defined a bilinear operation, we can continue by adding another, more advanced, algebraic structure called an algebra.

4.1 The mathematical definition of an algebra

If we have a vector space V over a field $\mathbb{F}$, we already have two operations available: vector addition and scalar multiplication. The natural way to extend this concept, is to define a map that combines two vectors into another vector. That is exactly how we end up with an algebra. If now add a bilinear operator $\star$ to a vector space V over $\mathbb{F}$, then we will call V an algebra (with $\star$) over V. Again, there is an unfortunate notation in which both V represents the algebra, as well as the

In some sense we could say that algebras are a generalization of fields in the way that field multiplication is now generalized to the bilinear operation of the algebra. In a sense, we can call the field multiplication a bilinear operation (in which the vector space associated to the bilinear operation is the field over itself).

Let $\{\mathbf{e}_i \,|\, i = 1, \ldots, n\}$ be a basis for the underlying n-dimensional vector space V of the algebra. It is then possible, in much the same way as operators can be represented as matrices in a certain basis, to characterize the the multiplication $\star$ of the algebra as

$$\mathbf{e}_i \star \mathbf{e}_j = \sum_k^n f_{ijk} \mathbf{e}_k \,, \tag{4.1}$$

in which f_{ijk} are called the structure constants of the algebra.

If $\star$ is associative, i.e.

$$\forall \mathbf{u}, \mathbf{v}, \mathbf{w} : \mathbf{u} \star (\mathbf{v} \star \mathbf{w}) = (\mathbf{u} \star \mathbf{v}) \star \mathbf{w} \,, \tag{4.2}$$

then the algebra is called associative.

4.2 Examples of algebras

We all know examples of algebras, with the easiest example being the $(n \times n)$-matrices with matrix multiplication.

4.3 The mathematical definition of a Lie algebra

A Lie algebra is an algebra $\mathfrak{g}$ over the field $\mathbb{F}$, in which the bilinear operation is the Lie bracket. The Lie bracket $[\cdot, \cdot]$ is a bilinear function that further obeys

1. alternativity
$$\forall T_a \in \mathfrak{g} : [T_a, T_a] = 0 \tag{4.3}$$
2. the Jacobi identity
$$\forall T_a, T_b, T_c \in \mathfrak{g} : [T_a, [T_b, T_c]] + [T_c, [T_a, T_b]] + [T_b, [T_c, T_a]] = 0 \tag{4.4}$$

It can be shown that bilinearity and alternativity together imply anticommutativity:

$$\forall T_a, T_b \in \mathfrak{g} : [T_b, T_a] = -[T_a, T_b] \,. \tag{4.5}$$

In the physics community, the elements $T_a, T_b, \cdots$ are called the generators of the algebra if they are a basis for the underlying vector field.

For the structure constants, the Jacobi identity implies

$$\sum_d^n (f_{bcd} f_{ade} + f_{abd} f_{cde} + f_{cad} f_{bde}) = 0 \,. \tag{4.6}$$

It is interesting to note that every associative algebra A over a field $\mathbb{F}$ admits a Lie algebra $L(A)$ over the same field $\mathbb{F}$ (both having the same underlying vector space V), by defining the Lie bracket as the commutator:

$$[T_a, T_b] = T_a T_b - T_b T_a \,. \tag{4.7}$$

The associative algebra A is then called the enveloping algebra of the Lie algebra $L(A)$.

4.4 Examples of Lie algebras

The most important examples of Lie algebras are those that are associated to a matrix Lie group. We have

- $\mathfrak{gl}(n, \mathbb{R})$: the Lie algebra of the $(n \times n)$-matrices with real entries,
- $\mathfrak{sl}(n, \mathbb{R})$: the Lie algebra of the $(n \times n)$-matrices with real entries and trace 0,
- $\mathfrak{o}(n) = \mathfrak{so}(n)$: the Lie algebra of the $(n \times n)$ skew-symmetric matrices with real entries,
- $\mathfrak{u}(n) = \mathfrak{su}(n)$: the Lie algebra of the $(n \times n)$ anti-Hermitian matrices.

The relation between the matrix Lie groups G and their associated Lie algebras $\mathfrak{g}$ is

$$\mathfrak{g} = \{M \in \mathbb{F}^{n \times n} \,|\, \forall t \in \mathbb{R} : \exp(tM) \in G\} \,. \tag{4.8}$$

4.5 The Lie algebra $\mathfrak{su}(2)$

If we have a vector space of operators spanned by $\mathfrak{g} = \{T^x, T^y, T^z\}$ with the commutation relations

$$[T^x, T^y] = iT^z \tag{4.9}$$
$$[T^y, T^z] = iT^x \tag{4.10}$$
$$[T^z, T^x] = iT^y\,, \tag{4.11}$$

the algebra $\mathfrak{g}$ is said to be $\mathfrak{su}(2)$. Another basis for $\mathfrak{su}(2)$ could be given by

$$T^+ = T^x + iT^y \tag{4.12}$$
$$T^- = T^x - iT^y \tag{4.13}$$
$$T^z = T^z\,, \tag{4.14}$$

or, reversely

$$T^x = \frac{1}{2}\big(T^+ + T^-\big) \tag{4.15}$$
$$T^y = \frac{1}{2i}\big(T^+ - T^-\big) \tag{4.16}$$
$$T^z = T^z\,, \tag{4.17}$$

with the commutators

$$\big[T^+, T^-\big] = 2T^z \tag{4.18}$$
$$\big[T^z, T^\pm\big] = \pm T^\pm\,, \tag{4.19}$$

and T^+ being the Hermitian adjoint of T^-

$$T^- = \big(T^+\big)^\dagger\,, \tag{4.20}$$

and T^z being Hermitian:

$$(T^z)^\dagger = T^z\,. \tag{4.21}$$

The (quadratic) Casimir invariant of $\mathfrak{su}(2)$ is given by

$$T^2 = (T^x)^2 + (T^y)^2 + (T^z)^2 \tag{4.22}$$
$$= T^+T^- - T^z + (T^z)^2\,, \tag{4.23}$$

and it commutes with every generator:

$$\left[T^2, T^x\right] = \left[T^2, T^y\right] = \left[T^2, T^z\right] = 0 \tag{4.24}$$

$$\left[T^2, T^+\right] = \left[T^2, T^-\right] = \left[T^2, T^z\right] = 0\,. \tag{4.25}$$

If we were to take N copies of the $\mathfrak{su}(2)$-algebra, the following commutation relations hold:

$$\left[T_i^+, T_j^-\right] = 2\delta_{ij} T_i^z \tag{4.26}$$

$$\left[T_i^z, T_j^\pm\right] = \pm\delta_{ij} T_i^\pm\,. \tag{4.27}$$

4.6 The generalized Gaudin algebra

Let us take a vector space spanned by $\{S_u^x, S_u^y, S_u^z\}$, where $S_u^\kappa \equiv S^\kappa(u)$ and $u \in \mathbb{C}$. The generalized Gaudin algebra is characterized by the commutation relations ($u \neq v$)

$$[S_u^\kappa, S_v^\kappa] = 0 \tag{4.28}$$

$$[S_u^x, S_v^y] = i(Y_{uv} S_u^z - X_{uv} S_v^z) \tag{4.29}$$

$$[S_u^y, S_v^z] = i(Z_{uv} S_u^x - Y_{uv} S_v^x) \tag{4.30}$$

$$[S_u^z, S_v^x] = i(X_{uv} S_u^y - Z_{uv} S_v^y)\,, \tag{4.31}$$

in which $\kappa = x, y, z$, $X_{uv} \equiv X(u,v)$, $Y_{uv} \equiv Y(u,v)$ and $Z_{uv} \equiv Z(u,v)$ are antisymmetric, i.e.

$$X(v,u) = -X(u,v) \tag{4.32}$$

Furthermore, for the commutation relations (4.28) to (4.31) to be consistent with the Jacobi identity $[S_u^x, [S_v^x, S_w^y]]$, the Yang-Baxter equation

$$X(u,v)Y(v,w) + Y(w,u)Z(u,v) + Z(v,w)X(w,u) = 0 \tag{4.33}$$

has to hold. Gaudin [1] found solutions to this equation, and we will focus on the so-called XXX-case, in which

$$X(u,v) = Y(u,v) = Z(u,v) = \frac{1}{u-v}\,, \tag{4.34}$$

such that the XXX-Gaudin algebra reduces to ($u \neq v$)

$$[S_u^\kappa, S_v^\kappa] = 0 \tag{4.35}$$

$$[S_u^x, S_v^y] = \frac{i}{u-v}(S_u^z - S_v^z) \tag{4.36}$$

$$[S_u^y, S_v^z] = \frac{i}{u-v}(S_u^x - S_v^x) \tag{4.37}$$

$$[S_u^z, S_v^x] = \frac{i}{u-v}(S_u^y - S_v^y)\,, \tag{4.38}$$

In this case, it is a good idea to work in the basis

$$S_u^+ = S_u^x + iS_u^y \tag{4.39}$$
$$S_u^- = S_u^x - iS_u^y \tag{4.40}$$
$$S_u^z = S_u^z\,, \tag{4.41}$$

with the commutators becoming ($u \neq v$)

$$\left[S_u^+, S_v^+\right] = \left[S_u^-, S_v^-\right] = [S_u^z, S_v^z] = 0 \tag{4.42}$$
$$\left[S_u^+, S_v^-\right] = \frac{2}{u-v}(S_u^z - S_v^z) \tag{4.43}$$
$$\left[S_u^z, S_v^\pm\right] = \frac{\pm}{u-v}\left(S_u^\pm - S_v^\pm\right). \tag{4.44}$$

So far, the form of the generators of the Gaudin algebra has not been specified. To make a specification, we will use N copies of an $\mathfrak{su}(2)$-algebra, each carrying a real number ε_i, and we will choose in particular

$$S_u^+ = \sum_i^N \frac{T_i^+}{u-\varepsilon_i} \tag{4.45}$$
$$S_u^- = \sum_i^N \frac{T_i^-}{u-\varepsilon_i} \tag{4.46}$$
$$S_u^z = \frac{1}{g} - \sum_i^N \frac{T_i^z}{u-\varepsilon_i}\,, \tag{4.47}$$

in which g is a real number. Using the $\mathfrak{su}(2)$-based realization of the Gaudin algebra, we can indeed show that the commutator relations (4.42) to (4.44) hold. Let us also calculate S_u^2:

$$S_u^2 = \frac{1}{g^2} - \frac{2}{g}\sum_i^N \frac{T_i^z}{u-\varepsilon_i} + \frac{1}{2}\sum_{ij}^N \frac{T_i^+T_j^- + T_i^-T_j^+ + 2T_i^zT_j^z}{(u-\varepsilon_i)(u-\varepsilon_j)}\,. \tag{4.48}$$

We can see that S_u^2 has a single pole when $u \to \varepsilon_k$, in one of the terms, and a double pole when $u \to \varepsilon_k$ in another term. The corresponding residue is calculated as

$$R_k = -\frac{2}{g}T_k^z + \sum_{i\neq k}^N \frac{T_i^+T_k^- + T_i^-T_k^+ + 2T_i^zT_k^z}{\varepsilon_k - \varepsilon_i}\,. \tag{4.49}$$

Paul Johnson [2] suggests rescaling g such that

$$R_k = T_k^z - g \sum_{i \neq k}^{N} \frac{T_i^+ T_k^- + T_i^- T_k^+ + 2 T_i^z T_k^z}{\varepsilon_k - \varepsilon_i} \,. \tag{4.50}$$

A specific linear combination [2] of the residues leading to

$$H_{\mathrm{RG}} = \sum_{i}^{N} \varepsilon_i T_i^z - g \sum_{ij}^{N} T_i^+ T_j^- \,, \tag{4.51}$$

up to a constant, where the subscript RG stands for Richardson-Gaudin.

5 Morphisms

A morphism is a map from one algebraic structure to another, preserving its structure.

5.1 Group homomorphisms and group isomorphisms - mathematical definitions

Say we have a group G with elements $\{a, b, \cdots\}$ and group multiplication $\cdot$. Say we also have another group G' with elements $\{a', b', \cdots\}$ and group multiplication $\cdot'$. A group homomorphism is a function

$$h : G \to G' : g \mapsto g' = h(g) \tag{5.1}$$

such that

$$\forall a, b \in G : h(a \cdot b) = h(a) \cdot' h(b)\,. \tag{5.2}$$

From this definition we can show that the identity e of G is mapped onto the identity e' of G' and that

$$\forall a \in G : h(a^{-1}) = h(a)^{-1} \tag{5.3}$$

such that we can say that the function h (the relation between the two groups) is compatible with the group structure.

Equivalently, we can write for a group homomorphism h:

$$h : G \to G' : \forall a, b, c \in G : \tag{5.4}$$

$$a \cdot b = c \Rightarrow h(a) \cdot' h(b) = h(c)\,. \tag{5.5}$$

A special type of group homomorphism is a group endomorphism. This is a group homomorphism from a set G to itself:

$$h : G \to G \tag{5.6}$$

Another special group homomorphism is a group isomorphism. It is a group homomorphism that is bijective.

5.2 Examples of group homomorphisms and group isomorphisms

As I stumble upon examples, they will be included here.

5.3 Examples of isomorphisms

Let us consider a subset C of the matrices with real entries ($x, y \in \mathbb{R}$), consisting of matrices of the form

$$\begin{pmatrix} x & y \\ -y & x \end{pmatrix} \tag{5.7}$$

Then, defining f as the field isomorphism $f : C \to \mathbb{C}$;

$$\begin{pmatrix} x & y \\ -y & x \end{pmatrix} \mapsto x + yi\,, \tag{5.8}$$

it can be seen that the field $(C, +, \cdot)$ is isomorphic to the field $(\mathbb{C}, +, \cdot)$.

We have seen that $\mathrm{Aut}(V)$ is the automorphism group of V. By introducing a basis in V, we can represent these invertible linear operators as invertible linear matrices and we can say that $\mathrm{Aut}(V)$ is isomorphic (one-to-one correspondence) to $\mathrm{GL}(n)$.

5.4 Examples of vector space morphisms

From its definition, we can see that a linear map can be called a vector space homomorphism.

6 The mathematical definition of a representation

The mathematical branch that connects groups and vector spaces is called representation theory. A representation of a finite group G on a finite-dimensional vector space V is a homomorphism

$$\rho : G \to \mathrm{GL}(V) : g \mapsto \rho(g) \tag{6.1}$$

of the group G to the general linear group $\mathrm{GL}(V)$, such that every group element g is associated to an element of the general linear group. In other words, we associate every group element g with an $n \times n$-matrix $\rho(g)$. The term homomorphism means that group structure is preserved:

$$\forall g_1, g_2 \in G : \rho(g_1 \cdot g_2) = \rho(g_1)\rho(g_2)\,, \tag{6.2}$$

which means that the matrix representation $\rho(g_1 \cdot g_2)$ of the group multiplication of two group elements g_1 and g_2 is the matrix product of their respective matrix representations $\rho(g_1)$ and $\rho(g_2)$.

7 Formulas for commutators and anticommutators

When an addition and a multiplication are both defined for all elements of a set $\{A, B, \ldots\}$, we can check if multiplication is commutative by calculation the commutator:

$$[A, B] = AB - BA\,. \tag{7.1}$$

A and B are said to commute if their commutator is zero. We can analogously define the anticommutator between A and B as

$$[A, B]_+ = AB + BA\,. \tag{7.2}$$

From these definitions, we can easily see that

$$[A, B] = -[B, A] \tag{7.3}$$
$$[A, B]_+ = [B, A]_+\,. \tag{7.4}$$

Letting † stand for the Hermitian adjoint, we can write for operators or A and B:

$$[A, B]^\dagger = \left[B^\dagger, A^\dagger\right] = -\left[A^\dagger, B^\dagger\right] \tag{7.5}$$
$$[A, B]_+^\dagger = \left[A^\dagger, B^\dagger\right]_+ \tag{7.6}$$

If U is a unitary operator or matrix, we can see that

$$\left[U^\dagger A U, U^\dagger B U\right] = U^\dagger [A, B] U\,. \tag{7.7}$$

Using the definitions, we can derive some useful formulas for converting commutators of products to sums of commutators:

$$[A, BC] = B[A, C] + [A, B]C \tag{7.8}$$
$$[AB, C] = A[B, C] + [A, C]B \tag{7.9}$$
$$[AB, CD] = A[B, C]D + AC[B, D] + [A, C]DB + C[A, D]B \tag{7.10}$$
$$[ABC, D] = AB[C, D] + A[B, D]C + [A, D]BC \tag{7.11}$$
$$[A, BCD] = BC[A, D] + B[A, C]D + [A, B]CD \tag{7.12}$$

In general, we can summarize these formulas as

$$[A, B_1 B_2 \cdots B_n] = \left[A, \prod_{k=1}^{n} B_k\right] = \sum_{k=1}^{n} B_1 \cdots B_{k-1} [A, B_k] B_{k+1} \cdots B_n\,. \tag{7.13}$$

Concerning sufficiently well-behaved functions f of B, we can prove that

$$[[A,B],B] = 0 \qquad \Rightarrow \qquad [A,f(B)] = f'(B)[A,B]\,. \tag{7.14}$$

In electronic structure theory, we often want to end up with anticommutators:

$$[A,BC] = [A,B]_+C - B[A,C]_+ \tag{7.15}$$
$$[AB,C] = A[B,C]_+ - [A,C]_+B \tag{7.16}$$

In electronic structure theory, we often end up with anticommutators. In case there are still products inside, we can use the following formulas:

$$[A,BC]_+ = [A,B]C + B[A,C]_+ \tag{7.17}$$
$$[A,BC]_+ = [A,B]_+C - B[A,C] \tag{7.18}$$
$$[AB,C]_+ = A[B,C]_+ - [A,C]B \tag{7.19}$$
$$[AB,C]_+ = [A,C]_+B + A[B,C] \tag{7.20}$$

The elementary BCH (Baker-Campbell-Hausdorff) formula reads

$$\exp(A)\exp(B) = \exp\left(A + B + \frac{1}{2}[A,B] + \cdots\right), \tag{7.21}$$

where higher order nested commutators have been left out. From this, two special consequences can be formulated:

$$\exp(-A)\,B\,\exp(A) = B + [B,A] + \frac{1}{2!}[[B,A],A] + \cdots \tag{7.22}$$
$$= \sum_{n=0}^{+\infty} \frac{1}{n!}\,{}_n[B,A]\,, \tag{7.23}$$

in which ${}_n[B,A]$ is the n-fold nested commutator in which the increased nesting is in the left argument, and

$$\exp(A)\,B\,\exp(-A) = B + [A,B] + \frac{1}{2!}[A,[A,B]] + \cdots \tag{7.24}$$
$$= \sum_{n=0}^{+\infty} \frac{1}{n!}[A,B]_n\,, \tag{7.25}$$

in which $[A,B]_n$ is the n-fold nested commutator in which the increased nesting is in the right argument.

8 Multivariate calculus

In Apostol [3], chapter 8, we can find a very nice mathematical summary of multivariate differential calculus. We'll start with multivariate functions, and subsequently discuss multivariate vector-valued functions.

8.1 Elementary topology

An open n-ball $\mathcal{B}(\mathbf{a}, \mathbf{r})$ is the set

$$\mathcal{B}(\mathbf{a}, \mathbf{r}) = \{\mathbf{x} \in \mathbb{R}^n \mid \ ||\mathbf{x} - \mathbf{a}|| < \mathbf{r}\} \,. \tag{8.1}$$

A point $\mathbf{a}$ is called an interior point of $S \subset \mathbb{R}^n$ if there exists an open n-ball such that $\mathcal{B}(\mathbf{a}, \mathbf{r}) \subset S$.

A set $S \subseteq \mathbb{R}^n$ is called open if every point inside it is an interior point of S. Examples are in 1D an open interval, and in 3D the sphere without boundary.

A neighborhood of a point $\mathbf{a}$ is an open set S in which $\mathbf{a}$ lies.

A point $\mathbf{x}$ is called an exterior point of $S \subset \mathbb{R}^n$ if there exists an open n-ball that does not contain any points of S.

A point $\mathbf{b}$ is called a boundary point of S if it is not interior nor exterior. The set of all boundary points of a set S is called the boundary and is denoted by ∂S.

8.2 Limits and continuity for scalar functions

Let S be an open subset of $\mathbb{R}^n$, and let $\mathbf{f}$ be a function

$$\mathbf{f} : S \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) \,, \tag{8.2}$$

in which we will designate vector-valued functions (i.e. $m \leq 1$) by a bold-face symbol. If we instead want to emphasize a scalar function (i.e. $m = 1$), we will use an italic symbol.

The limit notation, in which $\mathbf{x}$ approaches the interior point $\mathbf{a}$ has two equivalent meanings:

$$\lim_{\mathbf{x} \to \mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{b} \iff \lim_{||\mathbf{x}-\mathbf{a}|| \to \mathbf{0}} ||\mathbf{f}(\mathbf{x}) - \mathbf{b}|| = 0 \,. \tag{8.3}$$

The function $\mathbf{f}$ is called continuous at $\mathbf{a}$ if

$$\lim_{\mathbf{x} \to \mathbf{a}} \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) \,. \tag{8.4}$$

8.3 Directional derivatives for scalar functions

Let f be a scalar function:

$$\mathbf{f} : S \to \mathbb{R} : \mathbf{x} \mapsto f(\mathbf{x}) \,, \tag{8.5}$$

in which S is an open subset of $\mathbb{R}^n$ and $\mathbf{y}$ is a vector in $\mathbb{R}^n$. Let $\mathbf{a}$ be an interior point of S. We then call

$$\lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{y}) - f(\mathbf{a})}{h} \right) = f'(\mathbf{a}; \mathbf{y}) \tag{8.6}$$

the derivative of f with respect to $\mathbf{y}$ at $\mathbf{a}$.

If

$$g(t) = f(\mathbf{a} + t\mathbf{y}) \,, \tag{8.7}$$

then

$$g'(t) = f'(\mathbf{a} + t\mathbf{y}; \mathbf{y}) \,. \tag{8.8}$$

If $f(\mathbf{a} + t\mathbf{y})$ is differentiable for all $t \in [0, 1]$, then by the mean value theorem we have:

$$\exists \theta \in]0, 1[: \qquad f(\mathbf{a} + \mathbf{y}) - f(\mathbf{a}) = f'(\mathbf{a} + \theta\mathbf{y}; \mathbf{y}) \,. \tag{8.9}$$

If $\mathbf{y}$ is a unit vector (i.e. $||\mathbf{y}|| = 1$), then we call $f'(\mathbf{a}; \mathbf{y})$ a special name: the directional derivative of f w.r.t. $\mathbf{y}$ in $\mathbf{a}$ and we assign a new symbol to it:

$$\boldsymbol{\nabla}_{\mathbf{y}} f(\mathbf{a}) = \lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{y}) - f(\mathbf{a})}{h} \right) = f'(\mathbf{a}; \mathbf{y}) \,, \tag{8.10}$$

which is equivalent with

$$\lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{y}) - f(\mathbf{a}) - h\boldsymbol{\nabla}_{\mathbf{y}} f(\mathbf{a})}{h} \right) = 0 \,. \tag{8.11}$$

The i-th partial derivative of f in $\mathbf{a}$ is defined to be the directional derivative along $\mathbf{e}_i$. We define a new symbol

$$\frac{\partial f(\mathbf{a})}{\partial x_i} = \boldsymbol{\nabla}_{\mathbf{e}_i} f(\mathbf{a}) = f'(\mathbf{a}; \mathbf{e}_i) = \lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{e}_i) - f(\mathbf{a})}{h} \right) . \tag{8.12}$$

8.4 The total derivative for scalar functions

Let S be an open subset of $\mathbb{R}^n$ and $\mathbf{a}$, such that we define the scalar function f:

$$\mathbf{f} : S \to \mathbb{R} : \mathbf{x} \mapsto f(\mathbf{x}) \,, \tag{8.13}$$

Let $\mathbf{a}$ be an interior point of S, and let's choose an $\mathbf{r}$ such that the open n-ball is contained by S:

$$\mathcal{B}(\mathbf{a}, \mathbf{r}) \subseteq S \,. \tag{8.14}$$

Now let's choose a $\mathbf{v}$ with $||\mathbf{v}|| < ||\mathbf{r}||$ such that

$$\mathbf{a} + \mathbf{v} \in \mathcal{B}(\mathbf{a}, \mathbf{r}) \,. \tag{8.15}$$

These are all the ingredients we need to define differentiability. We call the function f differentiable at $\mathbf{a}$ if there exists a linear transformation $\mathbf{T_a}$

$$\mathbf{T_a} : \mathbb{R}^n \to \mathbb{R} \tag{8.16}$$

such that f admits a first-order Taylor formula:

$$f(\mathbf{a} + \mathbf{v}) = f(\mathbf{a}) + \mathbf{T_a}(\mathbf{v}) + ||\mathbf{v}|| E(\mathbf{a}, \mathbf{v}) \,, \tag{8.17}$$

in which $E(\mathbf{a}, \mathbf{v})$ is a scalar function with the behaviour that $E(\mathbf{a}, \mathbf{v}) \to 0$ if $||\mathbf{v}|| \to 0$. We call this linear map the total derivative and we can equivalently write:

$$\lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{y}) - f(\mathbf{a}) - h\mathbf{T_a}(\mathbf{y})}{h} \right) = 0 \,. \tag{8.18}$$

The total derivative is related to the directional derivative:

$$\mathbf{T_a}(\mathbf{y}) = f'(\mathbf{a}; \mathbf{y}) = \boldsymbol{\nabla}_{\mathbf{y}} f(\mathbf{a}) = \lim_{h \to 0} \left(\frac{f(\mathbf{a} + h\mathbf{y}) - f(\mathbf{a})}{h} \right) , \tag{8.19}$$

and the following useful formula holds:

$$\mathbf{T_a}(\mathbf{y}) = \boldsymbol{\nabla} f(\mathbf{a}) \cdot \mathbf{y} \,, \tag{8.20}$$

in which we have introduced the gradient of f at the point $\mathbf{a}$. This is the vector of partial derivatives:

$$\Big(\boldsymbol{\nabla} f(\mathbf{x}) \Big)_i \equiv \Big(\frac{\partial f}{\partial \mathbf{x}} \Big)_i = \frac{\partial f(\mathbf{x})}{\partial x_i} \,. \tag{8.21}$$

The directional derivative of a function f along a vector $\mathbf{a}$ is defined as

$$\boldsymbol{\nabla}_{\mathbf{a}} f(\mathbf{x}) = \lim_{h \to 0} \left(\frac{f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x})}{h} \right) \tag{8.22}$$

and can be calculated using

$$\boldsymbol{\nabla}_{\mathbf{a}} f(\mathbf{x}) = \boldsymbol{\nabla} f(\mathbf{x}) \cdot \mathbf{a} \tag{8.23}$$

for functions that are differentiable at $\mathbf{x}$.

Some useful formulas concerning the derivative of scalar fields with respect to a vector are:

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x} \cdot \mathbf{y}) = \mathbf{y} \tag{8.24}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x} \cdot \mathbf{x}) = 2\mathbf{x} \tag{8.25}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x} \cdot (\mathbf{A}\mathbf{y})) = \mathbf{A}\mathbf{y} \tag{8.26}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{y} \cdot (\mathbf{A}\mathbf{x})) = \mathbf{A}^{\mathrm{T}}\mathbf{y} \tag{8.27}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x} \cdot (\mathbf{A}\mathbf{x})) = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\mathbf{x} \,. \tag{8.28}$$

8.5 The Hessian for scalar functions

We can also calculate second-order (and subsequently higher-order) derivatives of the scalar function with respect to the components of $\mathbf{x}$. This second-order derivative is a symmetric matrix for twice-differentiable functions and is called the Hessian:

$$\mathbf{H}(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \,. \tag{8.29}$$

Using the previously defined gradient and Hessian, we can write the Taylor expansion of the function f around the point $\mathbf{x}_0$ as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \boldsymbol{\nabla} f(\mathbf{x}_0) + \frac{1}{2!} (\mathbf{x} - \mathbf{x}_0) \cdot (\mathbf{H}(\mathbf{x}) (\mathbf{x} - \mathbf{x}_0)) + \cdots \tag{8.30}$$

$$= f(\mathbf{x}_0) + \Delta\mathbf{x} \cdot \boldsymbol{\nabla} f(\mathbf{x}_0) + \frac{1}{2!} \Delta\mathbf{x}^{\mathrm{T}} \cdot (\mathbf{H}(\mathbf{x}_0) \, \Delta\mathbf{x}) + \cdots \,, \tag{8.31}$$

in which

$$\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0 \,. \tag{8.32}$$

Equation (8.30) is actually just short-hand notation for the following:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \sum_i^n \left.\frac{\partial f(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0} (x_i - x_{0,i}) + \frac{1}{2!}\sum_{ij}^n \left.\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\mathbf{x}_0} (x_i - x_{0,i})(x_j - x_{0,j}) + \cdots . \tag{8.33}$$

Often, we would like to separate the n variables contained in $\mathbf{x}$ in say m variables contained in $\mathbf{y}$ and l variables contained in $\mathbf{z}$. Then, f is the function

$$f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} : (\mathbf{y}, \mathbf{z}) \mapsto f(\mathbf{y}, \mathbf{z}) . \tag{8.34}$$

The gradient of f is then a blocked vector:

$$\boldsymbol{\nabla} f(\mathbf{y}, \mathbf{z}) = \begin{pmatrix} \frac{\partial f(\mathbf{y},\mathbf{z})}{\partial \mathbf{y}} \\ \frac{\partial f(\mathbf{y},\mathbf{z})}{\partial \mathbf{z}} \end{pmatrix} , \tag{8.35}$$

and the Hessian is a blocked matrix:

$$\mathbf{H}(\mathbf{y}, \mathbf{z}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{y},\mathbf{z})}{\partial \mathbf{y}^2} & \frac{\partial^2 f(\mathbf{y},\mathbf{z})}{\partial \mathbf{y} \partial \mathbf{z}} \\ \frac{\partial^2 f(\mathbf{y},\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{y}} & \frac{\partial^2 f(\mathbf{y},\mathbf{z})}{\partial \mathbf{z}^2} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{\mathbf{yy}}(\mathbf{y}, \mathbf{z}) & \mathbf{H}_{\mathbf{yz}}(\mathbf{y}, \mathbf{z}) \\ \mathbf{H}_{\mathbf{zy}}(\mathbf{y}, \mathbf{z}) & \mathbf{H}_{\mathbf{zz}}(\mathbf{y}, \mathbf{z}) \end{pmatrix} . \tag{8.36}$$

This means that an expression for the Taylor expansion of f around $(\mathbf{y}_0, \mathbf{z}_0)$ becomes

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}) =& f(\mathbf{y}_0, \mathbf{z}_0) + \Delta\mathbf{y} \cdot \frac{\partial}{\partial \mathbf{y}} f(\mathbf{y}_0, \mathbf{z}_0) + \Delta\mathbf{z} \cdot \frac{\partial}{\partial \mathbf{z}} f(\mathbf{y}_0, \mathbf{z}_0) \\ &+ \frac{1}{2!}\Delta\mathbf{y} \cdot (\mathbf{H}_{\mathbf{yy}}(\mathbf{y}_0, \mathbf{z}_0)\Delta\mathbf{y}) + \frac{1}{2!}\Delta\mathbf{y} \cdot (\mathbf{H}_{\mathbf{yz}}(\mathbf{y}_0, \mathbf{z}_0)\Delta\mathbf{z}) \\ &+ \frac{1}{2!}\Delta\mathbf{z} \cdot (\mathbf{H}_{\mathbf{zy}}(\mathbf{y}_0, \mathbf{z}_0)\Delta\mathbf{y}) + \frac{1}{2!}\Delta\mathbf{z} \cdot (\mathbf{H}_{\mathbf{zz}}(\mathbf{y}_0, \mathbf{z}_0)\Delta\mathbf{z}) + \cdots . \end{aligned} \tag{8.37}$$

8.6 Vector fields - multivariate vector functions

Let $\mathbf{f}(\mathbf{x})$ be a vector-valued function, i.e. a vector field:

$$\mathbf{f} : S \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) , \tag{8.38}$$

in which S is an open subset of $\mathbb{R}^n$. We will also write:

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})) \tag{8.39}$$

$$\forall f_i : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto f_i(\mathbf{x}) , \tag{8.40}$$

in which the functions f_i are sometimes called coordinate functions [4]. In a sense, the vector field associates to every vector $\mathbf{x} \in \mathbb{R}^n$ a vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$. Since this is a generaliation from $\mathbb{R}$ to $\mathbb{R}^n$, we can easily generalize the previous formulas for scalar functions to vector functions.

We can now write

$$\mathbf{f}(\mathbf{a};\mathbf{y}) = \lim_{h\to 0}\left(\frac{\mathbf{f}(\mathbf{a}+h\mathbf{y})-\mathbf{f}(\mathbf{a})}{h}\right). \tag{8.41}$$

The vector function $\mathbf{f}$ is now called differentiable at a point $\mathbf{a}$ is there exists a linear map called the total derivative of f at $\mathbf{a}$

$$\mathbf{T}_\mathbf{a} : \mathbb{R}^n \to \mathbb{R}^n\,, \tag{8.42}$$

such that $\mathbf{f}$ admits a Taylor formula:

$$\mathbf{f}(\mathbf{a}+\mathbf{v}) + \mathbf{T}_\mathbf{a}(\mathbf{v}) + ||\mathbf{v}||\mathbf{E}(\mathbf{a},\mathbf{v})\,, \tag{8.43}$$

in which $\mathbf{E}(\mathbf{a},\mathbf{v}) \to \mathbf{0}$ as $||\mathbf{v}|| \to \mathbf{0}$.

We can write this total derivative also as

$$\mathbf{T}_\mathbf{a}(\mathbf{y}) = \mathbf{f}'(\mathbf{a},\mathbf{y}) = \sum_i^n \boldsymbol{\nabla} f_i(\mathbf{a})\cdot \mathbf{y}\mathbf{e}_i = \mathbf{J}(\mathbf{a})\mathbf{y}\,, \tag{8.44}$$

which leads to the conclusion that the matrix $\mathbf{J}$, which we will call the Jacobian matrix, is the matrix representation of the total derivative for vector fields.

We can also say that first-order derivative of a vector field is the matrix

$$\mathbf{J} \equiv \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\,, \tag{8.45}$$

which is called the Jacobian and has entries

$$\mathbf{J}(\mathbf{x})_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}\,. \tag{8.46}$$

Since the gradient of a scalar function is also a vector, we can take the Jacobian of this gradient, leading to

$$\frac{\partial}{\partial \mathbf{x}}\Big(\boldsymbol{\nabla} f(\mathbf{x})\Big) = \mathbf{H}(\mathbf{x})^{\mathrm{T}}\,, \tag{8.47}$$

which means that the Jacobian of the gradient is the transpose of the Hessian. So, for twice differentiable functions the Hessian is equal to the Jacobian of the gradient.

A useful formula concerning the derivative of vector fields with respect to a vector is

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}\,. \tag{8.48}$$

9 Solving systems of equations

The goal of solving systems of equations is to solve equations of the type

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}\,, \tag{9.1}$$

in which $\mathbf{f}$ is a vector field as in equation (8.39). In order to solve equation (9.1), we use a linear approximation of $\mathbf{f}$ at a specified $\mathbf{x}_0$:

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\,, \tag{9.2}$$

in which $\mathbf{J}(\mathbf{x}_0)$ is the Jacobian (cfr. equation (8.46)) of $\mathbf{f}$, calculated at the point $\mathbf{x}_0$, which immediately signifies the meaning of the Jacobian: it is a generalization of the concept of derivative.

9.1 Newton's method

Combining equations (9.1) and (9.2), we get

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \approx \mathbf{0} \tag{9.3}$$

$$\mathbf{J}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \approx -\mathbf{f}(\mathbf{x}_0)\,, \tag{9.4}$$

which means that if we have an initial $\mathbf{x}_0$ (a guess for the solution), we can in principle calculate an improved guess $\mathbf{x}$. Newton's method thus goes as follows [4]:

1. Choose an initial guess $\mathbf{x}_0$
2. In the $(i+1)$-th iteration, solve (9.4) for $\Delta\mathbf{x}_{i+1} = \mathbf{x}_{i+1} - \mathbf{x}_i$.
3. Update
$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\mathbf{x}_{i+1} \tag{9.5}$$
4. Recalculate the Jacobian and the vector field at $\mathbf{x}_{i+1}$.
5. Repeat steps 2 to 4 until convergence is achieved:
$$||\Delta\mathbf{x}_{i+1}|| < \epsilon\,. \tag{9.6}$$

9.2 Broyden's method

Broyden's method [4, 5] is a quasi-Newton method that eliminates the need to recompute the Jacobian matrix at every iteration step. In the $i+1$-th step of the iteration, we use an approximation for the Jacobian matrix, which we shall denote by $\mathbf{A}_i$:

$$\mathbf{A}_i(\mathbf{x}_{i+1} - \mathbf{x}_i) = \mathbf{f}(\mathbf{x}_{i+1}) - \mathbf{f}(\mathbf{x}_i)\,, \tag{9.7}$$

or using a simplified notation:

$$\mathbf{A}_i \Delta\mathbf{x}_{i+1} = \Delta\mathbf{f}_{i+1}\,. \tag{9.8}$$

These equations show the action of $\mathbf{A}_{i+1}$ on $\Delta\mathbf{x}_{i+1}$. However, to fully characterize $\mathbf{A}_{i+1}$, we must also know its action on a vector $\mathbf{z}_i$ that is in the orthogonal complement of $\Delta\mathbf{x}_{i+1}$. Since we have no information about this, we specify that there should be no change made in this direction, i.e.

$$\forall \mathbf{z}_i : \mathbf{z}_i \cdot \Delta\mathbf{x}_i = 0 : \qquad \mathbf{A}_{i+1}\mathbf{z}_i = \mathbf{A}_i\mathbf{z}_i \tag{9.9}$$

We can show that the (unique) solution to equations (9.8) and (9.9) is

$$\mathbf{A}_{i+1} = \mathbf{A}_i + \frac{\Delta\mathbf{f}_{i+1} - \mathbf{A}_i\Delta\mathbf{x}_{i+1}}{||\Delta\mathbf{x}_{i+1}||^2}\Delta\mathbf{x}_{i+1}^T\,, \tag{9.10}$$

which is of the form $(\mathbf{A} + \mathbf{x}\mathbf{y}^T)$, for $\mathbf{A}$ nonsingular and $\mathbf{y}^T\mathbf{A}^{-1}\mathbf{x} \neq -1$, such that we can use the Sherman-Morrison formula [4]

$$(\mathbf{A} + \mathbf{x}\mathbf{y}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{y}^T\mathbf{A}^{-1}}{1 + \mathbf{y}^T\mathbf{A}^{-1}\mathbf{x}}\,, \tag{9.11}$$

which leads to

$$\mathbf{A}_{i+1}^{-1} = \mathbf{A}_i^{-1} + \frac{(\Delta\mathbf{x}_{i+1} - \mathbf{A}_i^{-1}\Delta\mathbf{f}_{i+1})\Delta\mathbf{x}_{i+1}^T\mathbf{A}_i^{-1}}{\Delta\mathbf{x}_{i+1}^T\mathbf{A}_i^{-1}\Delta\mathbf{f}_{i+1}}\,. \tag{9.12}$$

All in all, this makes the solution to equation (9.8) when requiring $\mathbf{f}(\mathbf{x}_{i+1}) = \mathbf{0}$ easier to compute:

$$\Delta\mathbf{x}_{i+1} = -\mathbf{A}_i^{-1}\mathbf{f}(\mathbf{x}_i)\,, \tag{9.13}$$

which means that the (possibly) time-consuming step of calculating the Jacobian over and over again in Newton' method, is avoided.

Broyden's method goes as follows:

1. Choose an initial guess $\mathbf{x}_0$ and calculate $\mathbf{A}_0^{-1} = \mathbf{J}(\mathbf{x}_0)^{-1}$, with $\mathbf{J}$ the exact Jacobian
2. Solve equation (9.13)

3. Update the guess

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta\mathbf{x}_{i+1} \tag{9.14}$$

4. Update the inverse of the approximate Jacobian matrix $\mathbf{A}_{i+1}$ through equation (9.12)

5. Repeat steps 2 through 4 until convergence is achieved:

$$||\Delta\mathbf{x}_{i+1}|| < \epsilon \tag{9.15}$$

9.3 DIIS

DIIS [6] (direct inversion of the iterative subspace) is a technique to accelerate convergence in solving linear equations. Let's say we have already found n estimates to the problem $\{\mathbf{x}_i\}$. We can construct a new estimate $\mathbf{x}^*$ as a linear combination of the previous ones:

$$\mathbf{x}^* = \sum_i^n c_i \mathbf{x}_i \,. \tag{9.16}$$

Let's say we also have a way to construct an error vector $\mathbf{e}_i$ for every estimate, such that our goal will be to minimize the error function in a least squares sense:

$$\left\| \sum_i^n c_i \mathbf{e}_i \right\|^2 , \tag{9.17}$$

under the constraint that

$$\sum_i^n c_i = 1 \,. \tag{9.18}$$

We can then construct the Lagriangian for this problem as

$$\mathcal{L}(\{c_i\}, \lambda) = \left\| \sum_i^n c_i \mathbf{e}_i \right\|^2 - \lambda \left(\sum_i^n c_i - 1 \right) , \tag{9.19}$$

and by introducing a suitable scalar product between the error vectors

$$B_{ij} = (\mathbf{e}_i, \mathbf{e}_j) \,, \tag{9.20}$$

we can write the Lagrangian as

$$\mathcal{L}(\{c_i\}, \lambda) = \sum_{ij}^n c_i c_j B_{ij} - \lambda \left(\sum_i^n c_i - 1 \right) . \tag{9.21}$$

Differentiating with respect to λ gives us back the original constraint, and differentiating with respect to every coefficient leads to

$$\frac{\partial \mathcal{L}}{\partial c_k} = 0 = 2\sum_i^n c_i B_{ik} - \lambda \,, \tag{9.22}$$

as the inner product matrix is symmetric. Absorbing the factor 2 in the Lagrange multiplier, we find the following system of equations

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1n} & -1 \\ B_{21} & B_{22} & \cdots & B_{2n} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} & -1 \\ -1 & -1 & \cdots & -1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} . \tag{9.23}$$

We can recognize a linear system of equations of the form $\mathbf{B}'\mathbf{y} = \mathbf{b}$, which can be easily solved using various decomposition techniques, leading to the coefficients $\{c_i\}$ which are then used in constructing the new estimate as in equation (9.16).

What remains is to construct a way to define the error vectors associated to every estimate $\mathbf{x}_i$, and a standard way would be to define residual vectors as

$$\mathbf{e}_i = \mathbf{x}_{i+1} - \mathbf{x}_i \,. \tag{9.24}$$

As an example, after 3 standard iterations, we end up with a set $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. We can then calculate

$$\mathbf{e}_1 = \mathbf{x}_2 - \mathbf{x}_1 \tag{9.25}$$

$$\mathbf{e}_2 = \mathbf{x}_3 - \mathbf{x}_2 \,, \tag{9.26}$$

We can then set up the DIIS equations to find the coefficients c_1 and c_2, in order to construct a new guess vector $\mathbf{x}_3$, which is in a sense 'overwritten'. This DIIS-improved guess vector is then used as a starting point for the next iteration.

10 The variation method

Suppose we have introduced a basis set of L vectors $\{|i\rangle\}$, which are not necessarily orthonormal:

$$S_{ij} = \langle i|j\rangle \neq \delta_{ij}, \tag{10.1}$$

such that we can linearly expand a state vector $|\mathbf{c}\rangle$ in this L-dimensional basis as

$$|\mathbf{c}\rangle = \sum_{i}^{L} c_i \, |i\rangle \ . \tag{10.2}$$

The energy of a system characterized by this wave function and the Hamiltonian $\hat{\mathcal{H}}$ is then

$$E = \frac{\langle \mathbf{c}|\hat{\mathcal{H}}|\mathbf{c}\rangle}{\langle \mathbf{c}|\mathbf{c}\rangle} \, , \tag{10.3}$$

which is a real[4]-valued function of the complex parameters $\{c_i\}$ and their complex conjugates $\{c_i^*\}$. For notational convenience, we will collect both sets in complex-valued vectors $\mathbf{c}$ and $\mathbf{c}^*$. We will then rewrite equation (10.3) as

$$E(\mathbf{c}, \mathbf{c}^*) \, \langle \mathbf{c}|\mathbf{c}\rangle = \ \langle \mathbf{c}|\hat{\mathcal{H}}|\mathbf{c}\rangle \ . \tag{10.4}$$

and by deriving equation (10.4) with respect to c_i and c_i^*, we obtain

$$\frac{\partial E(\mathbf{c}, \mathbf{c}^*)}{\partial c_i} \, \langle \mathbf{c}|\mathbf{c}\rangle + E(\mathbf{c}, \mathbf{c}^*) \sum_{j}^{L} c_j^* S_{ji} = \sum_{j}^{L} c_j^* H_{ji} \tag{10.5}$$

and

$$\frac{\partial E(\mathbf{c}, \mathbf{c}^*)}{\partial c_i^*} \, \langle \mathbf{c}|\mathbf{c}\rangle + E(\mathbf{c}, \mathbf{c}^*) \sum_{j}^{L} c_j S_{ij} = \sum_{j}^{L} c_j H_{ij} \, , \tag{10.6}$$

in which we have introduced the Hamiltonian matrix:

$$H_{ij} = \ \langle i|\hat{\mathcal{H}}|j\rangle \ . \tag{10.7}$$

In obtaining these two equations, we have used the properties [7] ($\forall z, z \in \mathbb{C}$)

$$\frac{\partial z}{\partial z} = 1 \qquad\qquad \frac{\partial z}{\partial z^*} = 0 \tag{10.8}$$

$$\frac{\partial z^*}{\partial z} = 0 \qquad\qquad \frac{\partial z^*}{\partial z^*} = 1 \, , \tag{10.9}$$

[4] because of the Hermiticity of the Hamiltonian

which lead to

$$\frac{\partial c_j}{\partial c_i} = \delta_{ij} \tag{10.10}$$

and

$$\frac{\partial c_j^*}{\partial c_i} = 0\,. \tag{10.11}$$

At a minimum of the energy $E(\mathbf{c}, \mathbf{c}^*)$, its partial derivatives with respect to c_i and c_i^* must vanish, leading to

$$\sum_j^L (H_{ji} - ES_{ji})c_j^* = 0 \tag{10.12}$$

and

$$\sum_j^L (H_{ij} - ES_{ij})c_j = 0\,, \tag{10.13}$$

which are two equations that are each other's complex conjugate. This means that these equations are not independent, and therefore coincide. We will keep the last equation, which is equivalent with the generalized eigenvalue problem

$$\mathbf{Hc} = E\mathbf{Sc}\,. \tag{10.14}$$

In the special case that the initially introduced L-dimensional basis set $\{|i\rangle\}$ is orthonormal, we have

$$S_{ij} = \langle i|j\rangle = \delta_{ij}, \tag{10.15}$$

such that we can recognize the regular eigenvalue problem

$$\mathbf{Hc} = E\mathbf{c}\,. \tag{10.16}$$

11 The symmetric eigenvalue problem

In the following, we will be looking for the solutions of the symmetric eigenvalue problem [8]:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}\,, \tag{11.1}$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{A}$ symmetric:

$$\mathbf{A}^{\mathrm{T}} = \mathbf{A}\,. \tag{11.2}$$

11.1 Ritz pairs

On a symmetric matrix $\mathbf{A}$, we can perform a *symmetric Schur decomposition*. This states that there exists a real, orthogonal matrix $\mathbf{Q}$

$$\mathbf{Q} = \begin{pmatrix}\mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n\end{pmatrix} \qquad\qquad \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_{(n)} \tag{11.3}$$

such that

$$\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)\,. \tag{11.4}$$

Moreover,

$$\mathbf{A}\mathbf{q}_k = \lambda_k\mathbf{q}_k\,. \tag{11.5}$$

Let $\mathbf{Q}_1 \in \mathbb{R}^{n\times r}$ (first r rows of an orthogonal matrix) have independent columns $\mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1 = \mathbf{I}_{(r)}$. For some $\mathbf{S} \in \mathbb{R}^{r\times r}$, we call

$$\mathbf{A}\mathbf{Q}_1 - \mathbf{Q}_1\mathbf{S} \tag{11.6}$$

the *residual matrix*. Then there exist $\mu_1, \ldots, \mu_r \in \lambda(\mathbf{A})$:

$$\forall k = 1, \ldots, r : |\mu_k - \lambda_k(\mathbf{S})| \leq \sqrt{2}\,||\mathbf{A}\mathbf{Q}_1 - \mathbf{Q}_1\mathbf{S}||_F\,, \tag{11.7}$$

in which $\lambda_k(S)$ is the k-th largest eigenvalue of $\mathbf{S}$ and $||.||_F$ represents the matrix 2-norm (or Frobenius norm). This results says that the eigenvalues of $\mathbf{S}$ tend to the eigenvalues of $\mathbf{A}$, with the better approximation for the smaller Frobenius norm of the residual matrix.

It is then natural to look at which $\mathbf{S}$ minimizes the right-hand side of equation (11.7). It can be shown that

$$\min_{\mathbf{S}\in\mathbb{R}^{r\times r}} ||\mathbf{A}\mathbf{Q}_1 - \mathbf{Q}_1\mathbf{S}||_F = ||(\mathbf{I}_{(n)} - \mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}})\mathbf{A}\mathbf{Q}_1||_2\,, \tag{11.8}$$

with the minimizer being

$$\mathbf{S} = \mathbf{Q}_1^{\mathrm{T}}\mathbf{A}\mathbf{Q}_1\,. \tag{11.9}$$

This means that we can associate any r-dimensional subspace $C(\mathbf{Q}_1)$ with a set of r "optimal" eigenvalue-eigenvector approximates. Let us Schur-decompose the minimizer (equation (11.9)) to yield

$$\mathbf{Z}^{\mathrm{T}}\mathbf{S}\mathbf{Z} = \mathrm{diag}(\theta_1, \ldots, \theta_r)\,, \tag{11.10}$$

and

$$\mathbf{Q}_1\mathbf{Z} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_r \end{pmatrix}\,. \tag{11.11}$$

The tuple $(\theta_k, \mathbf{y}_k)$ is then called a *Ritz pair*, with θ_k being the *Ritz value* and $\mathbf{y}_k$ begin the *Ritz vector*, which are the optimal eigensystem approximates of the symmetric matrix $\mathbf{A}$ in the r-dimensional subspace $C(\mathbf{Q}_1)$.

11.2 Approximate eigenpairs and corrections

Let's say we have a current (subscript c) approximation $\{\lambda_c, \mathbf{x}_c\}$ to the symmetric eigenvalue problem. Denoting the correction to the eigenvalue by $\delta\lambda$ and the correction to the eigenvector by $\boldsymbol{\delta}\mathbf{x}$, we then want

$$\mathbf{A}(\mathbf{x}_c + \boldsymbol{\delta}\mathbf{x}) = (\lambda_c + \delta\lambda)(\mathbf{x}_c + \boldsymbol{\delta}\mathbf{x}) \tag{11.12}$$

to hold. By introducing the current *residual vector*

$$\mathbf{r}_c = \mathbf{A}\mathbf{x}_c - \lambda_c\mathbf{x}_c\,, \tag{11.13}$$

we can rewrite equation (11.12) as

$$\left(\mathbf{A} - \lambda_c\mathbf{I}_{(n)}\right)\boldsymbol{\delta}\mathbf{x} - (\delta\lambda)\mathbf{x}_c = -\mathbf{r}_c\,. \tag{11.14}$$

Unfortunately, this is an underdetermined system of equations.

11.3 The Davidson diagonalization method

As proposed initially by Davidson in 1975 [9], his diagonalization method applied to any symmetry, diagonally dominant matrix of dimension N. It is a method to solve the associated eigenvalue problem for this matrix (which can have insanely large dimensions - large enough to be stored in the modern RAM-memory of a computer), finding its lowest eigenvalue and associated eigenvector.

In summary, the algorithm takes an initial guess for the lowest-eigenvalue eigenvector and produces new estimates by solving the diagonalization in an ever increasing subspace of previous estimates. Starting from that initial guess, the algorithm goes as follows.

1. Let $\mathbf{t}$ be an initial guess vector. Calculate $\mathbf{v}_1$, being a new subspace vector, as

$$\mathbf{v}_1 = \frac{\mathbf{t}}{||\mathbf{t}||} \,. \tag{11.15}$$

2. Calculate the (expensive) matrix-vector product:

$$\mathbf{v}_1^{\mathbf{A}} = \mathbf{A}\mathbf{v}_1 \,. \tag{11.16}$$

3. Expand the subspace $\mathbf{V}$ (initially it is empty) to

$$\mathbf{V}_1 = \left(\mathbf{v}_1\right) \tag{11.17}$$

and expand (again: initially $\mathbf{V}^{\mathbf{A}}$ is empty)

$$\mathbf{V}_1^{\mathbf{A}} = \left(\mathbf{v}_1^{\mathbf{A}}\right) \,. \tag{11.18}$$

4. Calculate the Rayleigh quotient:

$$\theta = \mathbf{v}_1^{\mathrm{T}}\mathbf{A}\mathbf{v}_1 \tag{11.19}$$

$$= \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1^{\mathbf{A}} \,. \tag{11.20}$$

5. Calculate the the associated residual vector:

$$\mathbf{r}_1 = \mathbf{v}_1^{\mathbf{A}} - \theta\,\mathbf{v}_1 \,. \tag{11.21}$$

6. If the norm of the residual vector is lower than a specified threshold, the algorithm has already converged (but let's not get our hopes up: this doesn't happen) and we have found the lowest eigenpair of $\mathbf{A}$, being $(\theta, \mathbf{v}_1)$.

7. Approximately solve the residue correction equation. This is where the 'Davidson' algorithm got his name. By using a diagonal approximation $\mathbf{A}'$ of the matrix $\mathbf{A}$, we get a clear and easy expression for a new $\mathbf{t}$-vector

$$\mathbf{t} = (\theta\,\mathbf{I} - \mathbf{A}')^{-1}\,\mathbf{r} \,, \tag{11.22}$$

or, written coefficient-wise:

$$t_i = \frac{r_i}{\theta - A_{ii}} \,. \tag{11.23}$$

Davidson originally suggested that if $|\theta - A_{ii}|$ is smaller than a threshold, the components t_i are set to zero.

We are now in the position to phrase the algorithm for all iterations $k > 1$.

1. Project $\mathbf{t}$ onto the orthogonal complement of $\mathbf{V}_{k-1}$:

$$\mathbf{t}^{\perp} = (\mathbf{I} - \mathbf{V}_{k-1}\mathbf{V}_{k-1}^{\mathrm{T}})\mathbf{t} \tag{11.24}$$

$$= \mathbf{t} - \sum_{i}^{\dim V_{k-1}} (\mathbf{v}_i^{\mathrm{T}}\mathbf{t})\mathbf{v}_i \tag{11.25}$$

2. Calculate the new subspace vector:

$$\mathbf{v}_k = \frac{\mathbf{t}^{\perp}}{||\mathbf{t}^{\perp}||} \,. \tag{11.26}$$

3. Calculate the (expensive) matrix-vector product:

$$\mathbf{v}_k^{\mathbf{A}} = \mathbf{A}\mathbf{v}_k \,. \tag{11.27}$$

4. Expand the subspace $\mathbf{V}$ to

$$\mathbf{V}_k = \begin{pmatrix} \mathbf{V}_{k-1} & \mathbf{v}_k \end{pmatrix} \tag{11.28}$$

and expand

$$\mathbf{V}_k^{\mathbf{A}} = \begin{pmatrix} \mathbf{V}_{k-1}^{\mathbf{A}} & \mathbf{v}_k \end{pmatrix} \,. \tag{11.29}$$

5. Calculate the subspace matrix:

$$\mathbf{M}_k = \mathbf{V}_k^{\mathrm{T}}\mathbf{A}\mathbf{V}_k \,. \tag{11.30}$$

Since in iteration k, we have already calculated the upper-left $\dim V_{k-1} \times \dim V_{k-1}$ block of $\mathbf{M}_k$, we actually only require to calculate

$$M_{ik} = \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_k^{\mathbf{A}} = M_{ki} \tag{11.31}$$

for $i = 1, \ldots, k$, which is the same as calculating the k-th row of $\mathbf{M}$:

$$\mathbf{m}_k = \mathbf{V}^{\mathrm{T}}(\mathbf{A}\mathbf{v}_k) \tag{11.32}$$

$$= \mathbf{V}^{\mathrm{T}}\mathbf{v}_k^{\mathbf{A}} \,. \tag{11.33}$$

6. Solve the subspace eigenvalue problem:

$$\mathbf{M}_k\mathbf{s} = \theta\,\mathbf{s} \,. \tag{11.34}$$

7. Calculate the new residual vector by first calculating

$$\mathbf{u} = \mathbf{V}_k \mathbf{s} \tag{11.35}$$

and

$$\mathbf{u}^{\mathbf{A}} = \mathbf{V}_k^{\mathbf{A}} \mathbf{s}\,, \tag{11.36}$$

and subsequently calculation the actual residual vector

$$\mathbf{r} = (\mathbf{A} - \theta\,\mathbf{I})\mathbf{V}_k \mathbf{s} \tag{11.37}$$

$$= \mathbf{u}^{\mathbf{A}} - \theta\,\mathbf{u}\,. \tag{11.38}$$

8. If the norm of the residual vector is lower than a specified threshold, the algorithm has converged and we have found the lowest eigenpair of $\mathbf{A}$, being $(\theta, \mathbf{u}_k)$.

9. Approximately solve the residue correction equation:

$$\mathbf{t} = (\theta\,\mathbf{I} - \mathbf{A}')^{-1}\mathbf{r}\,, \tag{11.39}$$

or, written coefficient-wise:

$$t_i = \frac{r_i}{\theta - A_{ii}}\,. \tag{11.40}$$

Davidson originally suggested that if $|\theta - A_{ii}|$ is smaller than a threshold, the components t_i are set to zero.

In case the dimension of the subspace matrix is getting too big (bigger than a predetermined maximum subspace dimension D of 10-15), the algorithm should restart:

- Pulay wrote that taking the two latest ones is enough to prevent convergence instabilities.
- We can therefore take 2 linear combinations

$$\mathbf{v}^{'(j)} = \mathbf{V}_k \mathbf{s}^{(j)}\,, \tag{11.41}$$

in which $\mathbf{V}_k$ is the current subspace, (j) is 1 or 2 and $\mathbf{s}^{(j)}$ is either the lowest or second lowest eigenvector of the subspace matrix $\mathbf{M}$.

12 Miscellaneous

Splitting up sums with respect to one of the allowed index numbers is not that hard to do. Let's start by a single sum:

$$\sum_{i}^{N} C_i = \sum_{i \neq a} C_i + C_a \,, \tag{12.1}$$

and double sums are just a little more involved:

$$\sum_{ij}^{N} C_{ij} = C_{aa} + \sum_{i \neq a}^{N} (C_{ia} + C_{ai}) + \sum_{\substack{i \neq a \\ j \neq a}} C_{ij} \,. \tag{12.2}$$

Borchardt's theorem states that for a Cauchy matrix $\mathbf{C}$:

$$|\mathbf{C}|_{+} = \frac{|\mathbf{C} \circ \mathbf{C}|}{|\mathbf{C}|} \,, \tag{12.3}$$

in which $\circ$ stands for the Hadamard (i.e. element-wise) product.

References

[1] M. Gaudin. Diagonalisation d'une classe d'hamiltoniens de spin. *Journal de Physique*, 37(10):1087–1098, 1976.

[2] Paul Andrew Johnson. *Model Wavefunction Forms to Describe Strong Correlation in Quantum Chemistry*. PhD thesis, McMaster University, 2014.

[3] Tom M. Apostol. *CALCULUS - Volume II, Multi Variable Calculus and Linear Algebra, with Applications to Differential Equations and Probability*, volume 2. John Wiley & Sons, 1969.

[4] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Brooks/Cole, ninth edition, 2011.

[5] C. G. Broyden. A Class of Methods for Solving Nonlinear Simultaneous Equations. *Mathematics of Computation*, 19(92):577, 1965.

[6] Péter Pulay. Convergence Acceleration of Iterative Sequences. The Case of SCF iIeration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[7] R Remmert. *Theory of Complex Functions*. 2010.

[8] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 2013.

[9] Ernest R. Davidson. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *Journal of Computational Physics*, 17(1):87–94, 1975.