1 Summary of linear algebra and its applications in physics

There are basically two approaches to linear algebra: the pedestrian approach of row and column vectors and matrices, and the more high-falutin' approach of abstract vector spaces and linear mappings and operators. We will start with a quick review of the pedestrian approach, to remind us of the most important concepts, before moving onto the abstract approach. It is important for you to be familiar with both approaches, and to be comfortable moving back forth between them

1.1 Matrices

In this subsection a *vector* is just a row or column of (real or complex) numbers, while a *matrix* is a rectangular array of numbers. (For the time being, we will only deal with finite vectors and matrices.) We write the *entries* (or *elements*, or *components*) of a vector v as v_i, and those of a matrix M as M_{ij}, where the first index represents the row and the second the column. Later on, we will sometimes also put the indices up, as in v^i, M^{ij}, or $M_i{}^j$, for reasons that will be explained.

I assume that you know how to multiply vectors and matrices, but here is a handy notation that we will sometime use: we automatically sum over repeated indices. Thus, the product MN of two matrices has components

$$(MN)_{ik} = M_{ij}N_{jk}\,, \tag{1}$$

which means $\sum_j M_{ij}N_{jk}$. An index that is summed over is sometimes called a *dummy* index.

The *transpose* of a matrix M is the matrix M^T whose components are $(M^T)_{ij} = M_{ji}$. Its *adjoint* or *Hermitian conjugate* $M^\dagger = (M^T)^*$ is the complex conjugate of the transpose. Note that $(MN)^T = N^T M^T$, and $(MN)^\dagger = N^\dagger M^\dagger$.

For the rest of this subsection we deal only with *square* matrices. We denote the unit matrix I; logically, its components should be written I_{ij}, but instead we write δ_{ij}, which is 1 if $i = j$ and 0 otherwise. Two useful properties of a matrix M are its *trace* $\operatorname{tr} M = M_{ii}$ and its *determinant* $\det M$ (whose definition is more complicated, but I assume you know it). The trace is obviously *linear*, $\operatorname{tr}(aM + bN) = a \operatorname{tr} M + b \operatorname{tr} N$ (where a and b are numbers); it also has the *cyclic* property $\operatorname{tr}(MN) = \operatorname{tr}(NM)$. The determinant has the property $\det(MN) = \det M \det N$ (from which the cyclic property follows). They are both invariant under transposition: $\operatorname{tr}(M^T) = \operatorname{tr} M$, $\det M^T = \det M$.

A matrix M is called

- *diagonal* if $M_{ij} = 0$ for $i \neq j$,
- *singular* if $\det M = 0$,
- *symmetric* if $M = M^T$,
- *antisymmetric* if $M = -M^T$,
- *Hermitian* if $M = M^\dagger$,
- *anti-Hermitian* if $M = -M^\dagger$,
- *orthogonal* if $MM^T = I$ (usually used in the context of real matrices),
- *unitary* if $MM^\dagger = I$,
- *normal* if $MM^\dagger = M^\dagger M$.

Any non-singular matrix M is invertible, that is, has a unique *inverse* M^{-1}, such that $MM^{-1} = I$. It also satisfies $M^{-1}M = I$. If M is singular, then it has no inverse. It's easy to see that, if M is invertible, then $(M^{-1})^T = (M^T)^{-1}$.

Two matrices M and N are said to be *similar* if there exists an invertible matrix S such that

$$M = SNS^{-1}. \tag{2}$$

It is easy to see that similar matrices have the same trace and determinant. A matrix is called *diagonalizable* if it is similar to a diagonal matrix, and *defective* if it is not.[1] (A simple example of a defective matrix is $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$.) The following two theorems, which we will not prove, are extremely important:

- Every real symmetric matrix is similar, by a real orthogonal matrix, to a real diagonal matrix (i.e. is *orthogonally diagonalizable*). Furthermore, if two such matrices commute ($M_1M_2 = M_2M_1$), then they can be diagonalized by the *same* real orthogonal matrix (they are *simultaneously diagonalizable*).

[1] In the context of real matrices, the term *diagonalizable* is sometimes applied only when S is real. Many real matrices, such as $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, are diagonalizable only using a complex S.

- Every Hermitian matrix is similar, by a unitary matrix, to a real diagonal matrix (i.e. is *unitarily diagonalizable*). Furthermore, if two such matrices commute, then they are simultaneously diagonalizable.

It follows from the second theorem that every normal matrix is unitarily diagonalizable (but not necessarily to a *real* diagonal matrix).

If, for a matrix M, a non-zero vector v, and a number a, we have

$$Mv = av\,, \tag{3}$$

then a and v are said to be an *eigenvalue* and *eigenvector* respectively of M. It is easy to see that similar matrices have the same eigenvalues, and it follows from the above theorem about Hermitian matrices that the latter have only real eigenvalues. For a given eigenvalue a, the set of vectors v satisfying (3) is called its *eigenspace*, and the dimensionality of that space its *multiplicity*. If the multiplicity is more than one, the eigenvalue is *degenerate*. If M is diagonalizable, $M = SDS^{-1}$, where D is diagonal, then the columns of S are eigenvectors of M, with the diagonal entries of D being the corresponding eigenvalues. It follows that the trace of M is the sum of its eigenvalues, while its determinant is their product (including multiplicities).

The *exponential* of a matrix M is defined as

$$\exp M = \sum_{n=0}^{\infty} \frac{1}{n!} M^n\,. \tag{4}$$

(Other functions can be defined similarly through their Taylor series.) The matrix $\exp M$ has the same eigenvectors as M, with eigenvalues that are the exponentials of those of M. (The same applies to any function of a matrix defined through its Taylor series.) We have $(\exp M)^T = \exp M^T$ and $(\exp M)^{-1} = \exp(-M)$. It follows that the exponential of a symmetric matrix is symmetric, and of an antisymmetric matrix is orthogonal. For example,

$$\exp \begin{pmatrix} 0 & \theta \\ \theta & 0 \end{pmatrix} = \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix}\,. \tag{5}$$

$$\exp \begin{pmatrix} 0 & \theta \\ -\theta & 0 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\,. \tag{6}$$

Similarly, the exponential of a Hermitian matrix is Hermitian, and of an anti-Hermitian matrix is unitary. A useful formula (which applies whether or not M is diagonalizable) is

$$\det \exp M = \exp \operatorname{tr} M\,. \tag{7}$$

1.2 Abstract vector spaces

An abstract vector space V is a set which admits addition and scalar multiplication. Under addition, the set must form an abelian group: there is a unique vector $\mathbf{0}$ for which $\mathbf{0}+\mathbf{v}=\mathbf{v}$ for all vectors $\mathbf{v}$; each vector $\mathbf{v}$ has an inverse $-\mathbf{v}$ such that $\mathbf{v}+(-\mathbf{v})=0$; for all vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, we have $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$ and $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$. For a real vector space, the scalars are real numbers, for a complex vector space complex numbers. Scalar multiplication is associative and distributive: $a(b\mathbf{u})=(ab)\mathbf{u}$, $a(\mathbf{u}+\mathbf{v})=a\mathbf{u}+a\mathbf{v}$, $(a+b)\mathbf{u}=a\mathbf{u}+b\mathbf{u}$.

Examples of real vector spaces include (except where otherwise stated, all numbers, matrices, and functions considered are real):

1. The set with one element $\{\mathbf{0}\}$.
2. The real numbers $\mathbf{R}$.
3. The complex numbers $\mathbf{C}$.
4. $\mathbf{R}^n$, the set of column vectors of length n as discussed in the previous section.
5. The set of column vectors orthogonal to a given row vector.
6. The eigenspace associated with a given eigenvalue of a given real matrix.
7. The set of $n\times n$ matrices.
8. The set of traceless $n\times n$ matrices.
9. The set of Hermitian complex $n\times n$ matrices.
10. The set of cubic (and lower-order) polynomials.
11. The set of functions on $\mathbf{R}$.
12. The set of even functions on $\mathbf{R}$.
13. The set of functions $f(x)$ on $\mathbf{R}$ such that $f(-1)=f(\pi)=0$.
14. The set of infinitely differentiable functions on $\mathbf{R}$.
15. The set of infinitely differentiable real functions on $\mathbf{R}$ satisfying $f''(x)-3x^3f(x)=0$.

16. The set of functions on $\mathbf{R}$ with period 2π.
17. The set of infinitely differentiable functions on the interval $[0, 1]$ satisfying $f(0) = f(1) = 0$.
18. The set of functions on $\mathbf{R}$ such that $\int dx f(x)^2$ is finite.
19. The set of functions f on $\mathbf{R}$ such that $\int dx f(x)g(x) = 0$ for all continuous functions g.
20. The set of functions on $\mathbf{R}$ modulo constants.
21. The set of source-free time-dependent solutions to the Maxwell equations in infinite space.

In sum, vector spaces are all around us. Below we will discuss the reasons they are ubiquitous in physics.

A subset of vector space that is a vector space in its own right is called a *subspace*. (Which are the above examples are subspaces of other examples?)

The most basic distinction among vector spaces is between *finite-dimensional* and *infinite-dimensional* ones. If there exists a finite set of vectors which spans the whole space (i.e. such that every vector in the space is a linear combination of vectors in the set), then the space is finite-dimensional. (Which of the above examples is finite-dimensional?) A *basis* for a finite-dimensional vector space is a set that spans the space and that is linearly independent (no element is a linear combination of the others). A given vector space will admit many different choices of basis, but it can be shown that they all have the same number of elements, which is the *dimensionality* of the space. (What is the dimensionality of each of the finite-dimensional examples above?) The utility of a basis is that it turns considerations of an abstract vector space into considerations about numbers and matrices, as discussed in the previous section. Given a basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots\}$, any vector $\mathbf{v}$ can be uniquely decomposed as a linear combination of basis elements:

$$\mathbf{v} = \mathbf{e}_i v^i \,. \tag{8}$$

The numbers $\{v^i\}$ are called the *components* of $\mathbf{v}$ in the given basis. If we now change to a different basis $\{\mathbf{e}'_1, \mathbf{e}'_2, \ldots\}$, each element of the old basis can be expanded in terms of the new one:

$$\mathbf{e}_i = \mathbf{e}'_j S^j{}_i \,. \tag{9}$$

Similarly the other way around:

$$\mathbf{e}'_j = \mathbf{e}_i S'^i{}_j \,. \tag{10}$$

Inserting the first relation into the second, we see that S and S' are inverses of each other as matrices. The components of $\mathbf{v}$ in the basis $\{\mathbf{e}_i\}$ can be transformed into its components in the basis $\{\mathbf{e}'_j\}$ using the matrix S:

$$v'^j = S^j{}_i v^i. \tag{11}$$

Infinite-dimensional vector spaces are technically more difficult to handle, essentially because to have a useful notion of a basis, you need a useful notion of infinite sums of vectors, in other words a notion of convergence, and for that you need a notion of a norm with certain properties. This chain of reasoning leads to the concept of a *Hilbert space*, which we will discuss below. For now we will stick to the finite-dimensional case, starting with real vector spaces.

1.3 Linear and bilinear functions

A linear function $\hat{\mathbf{v}} : V \to \mathbf{R}$ is called a *dual vector*. The dual vectors themselves form a vector space $\hat{V}$. (What is the dual of each of the finite-dimensional examples above?) Because of its linearity, to know how $\hat{\mathbf{v}}$ acts on any vector it's sufficient to know how it evaluates on all the elements of a basis:

$$\hat{v}_i = \hat{\mathbf{v}}(\mathbf{e}_i)\,. \tag{12}$$

This shows that $\hat{V}$ has the same dimensionality as V. For any vector $\mathbf{v}$, we then have

$$\hat{\mathbf{v}}(\mathbf{v}) = \hat{v}_i v^i\,. \tag{13}$$

The numbers $\hat{v}_i$ are the components of $\hat{\mathbf{v}}$ in a basis $\{\hat{\mathbf{e}}^j\}$ for $\hat{V}$ called the *dual basis*, defined by the property

$$\hat{\mathbf{e}}^j(\mathbf{e}_i) = \delta^j_i\,. \tag{14}$$

Under a change of basis, the components $\hat{v}_j$ transform using S^{-1}:

$$\hat{v}'_j = \hat{v}_i (S^{-1})^i{}_j\,. \tag{15}$$

Note that, in equation (13), under a change of basis the S and the S^{-1} cancel, ensuring (as was necessarily the case) that the result is invariant.

You are now hopefully starting to see the utility of upper and lower indices. If you are consistent in your usage of them, then:

1. When you sum over a repeated index, one copy is always up and the other down. This is an important check.

2. Components with upper indices always transform under a change of basis with S, and components with lower indices always with S^{-1}.

One more thing about dual vector spaces: Let us pick a fixed vector $\mathbf{v} \in V$. The expression $\hat{\mathbf{v}}(\mathbf{v})$ can then be considered a *linear map* from $\hat{V}$ to $\mathbf{R}$. In other words, each element of V is itself a *dual vector* of $\hat{V}$. It's not hard to see that every dual vector of $\hat{V}$ can be written as $\hat{\mathbf{v}}(\mathbf{v})$ for some fixed element $\mathbf{v}$ of V. In summary, $\hat{\hat{V}} = V$. So you don't get more and more vector spaces if you keep dualizing!

Next, let us consider symmetric bilinear functions, or *quadratic forms* on V, i.e. $\mathbf{g} : V \times V \to \mathbf{R}$, where $\mathbf{g}(\mathbf{u}, \mathbf{v}) = \mathbf{g}(\mathbf{v}, \mathbf{u})$. (Example: on a space of matrices, $\mathrm{tr}(MN)$ is a quadratic form.) This is represented in a given basis by a symmetric matrix with two lower indices:

$$g_{ij} = \mathbf{g}(\mathbf{e}_i, \mathbf{e}_j)\,. \tag{16}$$

We then have, for general vectors $\mathbf{u}$, $\mathbf{v}$,

$$(\mathbf{u}, \mathbf{v}) = u^i g_{ij} v^j\,. \tag{17}$$

The components g_{ij} transform with S^{-1}, which acts twice, once on each index:

$$g'_{i'j'} = (S^{-1})^i{}_{i'} g_{ij} (S^{-1})^j{}_{j'}\,. \tag{18}$$

In matrix language, we would write

$$g' = (S^{-1})^T g S^{-1}\,. \tag{19}$$

Since g is symmetric, according to one of the theorems mentioned above, we can find an *orthogonal* matrix S such that g' is diagonal. We can then make another change of basis, using a *diagonal* matrix D, to reduce g' to a form where it only has 1, 0, and -1 on the diagonal:

$$g'' = D g' D = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \ddots & & \\ & & & & -1 & \\ & & & & & \ddots \end{pmatrix} \tag{20}$$

There is no change of basis that will change the numbers of 1's, 0's, and -1's on the diagonal. If the quadratic form has one or more zero eigenvalues (i.e.

0's on the diagonal of g''), then it's *degenerate*. A non-degenerate quadratic form is sometimes called a *metric*. A basis in which a metric takes the form (20) is called *orthonormal*. The *norm* of a vector $\mathbf{v}$ with respect to a metric is $\mathbf{g}(\mathbf{v}, \mathbf{v})$.

Example: In special relativity, displacements in space and time about a given event (spacetime point) form a four-dimensional vector space. This space is equipped with a metric which can be written, in any appropriate basis,

$$g = \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}. \tag{21}$$

Displacements with positive norm are called "spacelike"; those with negative norm "timelike"; those with zero norm "null" or "lightlike".

Given a metric $\mathbf{g}$, there is a natural map from V to $\hat{V}$ defined by $\mathbf{v} \mapsto \hat{\mathbf{v}} = \mathbf{g}(\cdot, \mathbf{v})$. In components, we write

$$v^i \mapsto \hat{v}_j = v^i g_{ij} \,. \tag{22}$$

This is called *lowering the index*. To *raise* the index, we multiply by the inverse of g, whose components are usually written g^{ij}:

$$\hat{v}^i \mapsto v_j = g^{ij} \hat{v}_j \,. \tag{23}$$

This map is an isomorphism. Therefore, if we are working with a *fixed* metric, as in the example of special relativity, then we can forget about the distinction between vectors and dual vectors.

A metric with only positive eigenvalues (i.e. such that $g'' = I$) is sometimes called an *inner product*. (Beware: the terms *metric* and *inner product* are sometimes used interchangeably; their precise definition is context-sensitive.) Often in physical applications a vector space will "arrive with an inner product pre-installed". A vector space with a fixed inner product is called an *inner product space*. The inner product in that case is usually denoted simply $(\mathbf{u}, \mathbf{v})$ (or $\langle \mathbf{u}, \mathbf{v} \rangle$, $\langle \mathbf{u} | \mathbf{v} \rangle$, or $\mathbf{u} \cdot \mathbf{v}$). In an orthonormal basis, by definition, the inner product has components δ_{ij}, and lowering and raising indices has no effect. Therefore, as long as we stick to orthonormal bases, we can forget about upper and lower indices! However, it is not always convenient to restrict ourselves to orthonormal bases.

1.4 Linear maps

We now move on to linear maps between vector spaces, $\mathbf{M} : V \to W$. We usually write $\mathbf{Mv}$, rather than $\mathbf{M}(\mathbf{v})$. Given a basis for V and a basis for W, $\mathbf{M}$ is represented by a matrix M, such that, if $\mathbf{Mv} = \mathbf{w}$, then

$$w^i = M^i{}_j v^j \,. \tag{24}$$

There are two interesting subspaces associated with any linear mapping $\mathbf{M}$: its *image* $\mathbf{M}(V) \subset W$, whose dimension is its *rank*; and its *kernel* (or *null space*), $\mathbf{M}^{-1}(\mathbf{0}) \subset V$, whose dimension is its *nullity.* There is a theorem that says that the rank plus the nullity of $\mathbf{M}$ equals the dimensionality of V (explicitly, $\mathbf{M}$ is a isomorphism between the image of $\mathbf{M}$ and the quotient space $V/\ker(\mathbf{M})$).

A linear map from a space V to itself, $\mathbf{M} : V \to V$, is called a linear *transformation* or *operator.* It is represented by a square matrix M. Note that the trace and determinant of M don't depend on what basis we choose; therefore we can speak of the *trace* and *determinant* of the operator itself, without referring to any matrix. As for matrices, if we have

$$\mathbf{Mv} = \lambda \mathbf{v} \,, \tag{25}$$

for some vector $\mathbf{v}$ and number λ, then these are called an *eigenvector* and *eigenvalue* of $\mathbf{M}$. (Another way to say this is that $\mathbf{v}$ is in the kernel of the operator $\mathbf{M} - \lambda\mathbf{I}$.) This implies that, in any given basis, the column vector v representing $\mathbf{v}$ is an eigenvector with eigenvalue λ of the matrix M representing $\mathbf{M}$; however, the notions of eigenvector and eigenvalue are basis-independent. The eigenvectors of $\mathbf{M}$ span V if and only if M is diagonalizable. The set of all eigenvalues of an operator is called its *spectrum.*

We can define the *exponential* (or other function) of a linear operator in the same way as for a square matrix, and the properties discussed above for matrices apply also to linear operators.

In an inner product space, an operator $\mathbf{M}$ is said to be *orthogonal* if $(\mathbf{Mu}, \mathbf{Mv}) = (\mathbf{u}, \mathbf{v})$, and *symmetric* if $(\mathbf{u}, \mathbf{Mv}) = (\mathbf{Mu}, \mathbf{v})$ (for all $\mathbf{u}, \mathbf{v}$). These are equivalent to the matrix M representing $\mathbf{M}$ *in an orthonormal basis* being orthogonal or symmetric. (Please note that, in the absence of an inner product, it makes no sense to speak of an orthogonal or symmetric operator!) Furthermore, just as we can forget about the distinction between a vector and a dual vector, we can forget about the distinction between a symmetric operator and a quadratic form. To any symmetric operator $\mathbf{M}$ we can associate a unique quadratic form $\mathbf{h}$,

$$\mathbf{h}(\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbf{Mv}) \,, \tag{26}$$

and for any quadratic form $\mathbf{h}$ there is a unique symmetric operator $\mathbf{M}$ satisfying (26). In components, we are just raising and lowering indices:

$$h_{ij} = g_{ii'} M^{i'}{}_j \,, \qquad M^i{}_j = g^{ii'} h_{i'j} \,. \tag{27}$$

Even better, in an orthonormal basis, $g_{ij} = \delta_{ij}$, so $h = M$. Since that matrix is symmetric, it has a complete, orthonormal set of eigenvectors with real eigenvalues. In the basis-independent language, the eigenvector $\mathbf{v}$ with eigenvalue λ obeys $\mathbf{h}(\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v})$ for all $\mathbf{u} \in V$. Furthermore, eigenvectors with different eigenvalues are orthogonal: $(\lambda_1 - \lambda_2)(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{h}(\mathbf{v}_1, \mathbf{v}_2) - \mathbf{h}(\mathbf{v}_2, \mathbf{v}_1) = 0$. (Note that, unless you are in an inner product space, it makes no sense to speak of a quadratic form having eigenvectors and eigenvalues.)

1.5 New vector spaces out of old ones

Above we've mentioned several examples of subspaces of vector spaces. Given such a subspace $W \subset V$, the *quotient space* V/W is simply the set of equivalences classes of vectors in V under the relation $\mathbf{v} \sim \mathbf{v} + \mathbf{w}$ for all $\mathbf{w} \in W$. It's easy to see that $\dim V/W = \dim V - \dim W$.

Given two vector spaces V and W, their Cartesian product $V \times W$ forms a vector space, with the obvious addition and multiplication rules $(\mathbf{v}_1, \mathbf{w}_1) + (\mathbf{v}_2, \mathbf{w}_2) = (\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2)$, $\lambda(\mathbf{v}, \mathbf{w}) = (\lambda\mathbf{v}, \lambda\mathbf{w})$. This space is sometimes called the *direct sum* or *tensor sum*, and is written $V \oplus W$. A basis for $V \oplus W$ is provided by the union of $(\mathbf{e}_i, \mathbf{0}_W)$ and $(\mathbf{0}_V, \mathbf{e}'_j)$, where $\mathbf{e}_i$ is a basis for V and $\mathbf{e}'_j$ is a basis for W. Therefore we have $\dim(V \oplus W) = \dim V + \dim W$.

Slightly trickier is the *direct product* or *tensor product* space $V \otimes W$. This is easiest to construct by starting with bases $\mathbf{e}_i$ and $\mathbf{e}'_j$ for V and W respectively, and considering all linear combinations of

$$\mathbf{e}_i \otimes \mathbf{e}'_j \,. \tag{28}$$

From the definition we obviously have $\dim(V \otimes W) = \dim V \times \dim W$. If you are ever confused about whether you are working with a direct sum or product of two spaces, ask yourself whether the dimensions add or multiply. An example of a tensor product is (these are all considered real vector spaces):

$$\begin{aligned} &\mathbf{C} \otimes \{\text{cubic polynomials in } x \text{ with real coefficients}\} \\ &\qquad = \{\text{cubic polynomials in } x \text{ with complex coefficients}\} \,. \end{aligned} \tag{29}$$

Another example is

$$\begin{aligned}&\{\text{cubic polynomials in } x \text{ with real coefficients}\}\\ &\qquad\otimes\{\text{cubic polynomials in } y \text{ with real coefficients}\}\\ &\qquad=\{\text{cubic polynomials in } x \text{ and } y \text{ with real coefficients}\}\,. \end{aligned} \tag{30}$$

Yet another example:

$$\{\text{linear maps } V \to W\} = W \otimes \hat{V} \tag{31}$$

(think about it).

1.6 Linearization

One of the most common ways that linear structures appear in physics is through the approximation procedure called *linearization.* Suppose we have some continuous space, for example the configuration space of some system. This space need not be a vector space. We'll denote the points in this space $\mathbf{x}$. Suppose that we have a map $\mathbf{f}(\mathbf{x})$ from this space to another one (which, again, need not be a vector space). Two examples:

1. We attach one end of a spring to a wall, and move the other end around. We let $\mathbf{x}$ denote the position of the other end, and $\mathbf{f}(\mathbf{x})$ the force exerted by the spring on our hand.
2. Consider an elastic solid in its equilibrium shape. We now deform the solid by exerting some external forces on it. We describe the new configuration by saying that $\mathbf{f}(\mathbf{x})$ is the new position of the atom that was at the position $\mathbf{x}$ in the equilibrium configuration.

Now, an important statement is that, in any continuous space, given a fixed point $\mathbf{x}_0$, the small displacements $\delta\mathbf{x}$ about $\mathbf{x}_0$ form a vector space, which we will call V. (This rough statement can be made rigorous.) Furthermore, $\mathbf{f}(\mathbf{x})$ defines, through its derivative $\mathbf{f}'(\mathbf{x}_0)$, a linear map from that vector space to the vector space W of small displacements about $\mathbf{f}(\mathbf{x}_0)$:

$$\mathbf{f}(\mathbf{x}_0 + \delta\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{f}'(\mathbf{x}_0)\delta\mathbf{x} + \mathcal{O}(\delta\mathbf{x}^2)\,. \tag{32}$$

In example 1, $\mathbf{f}'(\mathbf{x}_0)$ would be called the spring constant (at least if $\mathbf{x}_0$ is the equilibrium configuration), while in example 2 it (actually its symmetric part) would be called the strain. Now here is the point: essentially every linear law of classical physics you've every been taught (Hooke's law, Ohm's law, $\mathbf{F} = m\mathbf{a}$, "the friction between two objects is proportional to the normal force", . . .) is the result of linearizing some much more complicated, non-linear law. (As we will discuss below, the same does *not* apply to the Schrödinger equation, which is also linear.)

Another common instance of linearization is for a scalar function of a continuous space:

$$f(\mathbf{x}_0 + \delta\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)\delta\mathbf{x} + \mathcal{O}(\delta\mathbf{x}^2)\,. \tag{33}$$

Note that the gradient of f, $\nabla f(\mathbf{x}_0)$, is an element of the dual space to the space of small displacements. If we let $\mathbf{e}_i$ be a basis for the space of small displacements, $\delta\mathbf{x} = \delta x^i \mathbf{e}_i$, then we have

$$(\nabla f(\mathbf{x}_0))_i = \left.\frac{\partial f}{\partial \delta x^i}\right|_{\delta x=0}\,. \tag{34}$$

Here we make a very important notational observation: *an upper index in the denominator of a derivative is effectively a lower index.* Now, if $\nabla f(\mathbf{x}_0) = 0$, then $\mathbf{x}_0$ is an extremum of f, and we may be interested in Taylor expanding to one higher order:

$$f(\mathbf{x}_0 + \delta\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{2}\mathbf{h}(\delta\mathbf{x}, \delta\mathbf{x}) + \mathcal{O}(\delta\mathbf{x}^3)\,. \tag{35}$$

The second derivative $\mathbf{h}$ is a quadratic form. In components, we have

$$h_{ij} = \left.\frac{\partial^2 f}{\partial \delta x^i\, \partial \delta x^j}\right|_{\delta x=0}\,. \tag{36}$$

If $\mathbf{h}$ is positive definite, then $\mathbf{x}_0$ is a local minimum.

Probably the most common application of this technique in physics is when V is the potential energy function on the configuration space of a system. If V admits a local minimum $\mathbf{x}_0$, then to study the dynamics of motions very close to $\mathbf{x}_0$ it is sufficient to consider the quadratic approximation (35). Similarly, we can consider the kinetic energy $K(\dot{\mathbf{x}})$. It is a general principle of physics that $\dot{\mathbf{x}} = 0$ is the (global) minimum of the kinetic energy. (It's interesting to think about what would go wrong if this were not the case.) For sufficiently small velocities (for example, if the system is a particle in special relativity, then for $|\dot{\mathbf{x}}| \ll c$), we can therefore approximate $K(\dot{\mathbf{x}})$ by its second derivative; this defines an *inner product*, which is so fundamental that we'll simply write it $(\cdot, \cdot)$:

$$K(\dot{\mathbf{x}}) = E_0 + \frac{1}{2}(\dot{\mathbf{x}}, \dot{\mathbf{x}}) + \mathcal{O}(\dot{\mathbf{x}}^3)\,. \tag{37}$$

In this approximation, the equation of motion is

$$\ddot{\mathbf{x}} = -\mathbf{M}\mathbf{x}\,, \tag{38}$$

where $\mathbf{M}$ is defined as in (26). $\mathbf{M}$ is a symmetric operator (with respect to the inner product $(\cdot,\cdot)$), and therefore admits an orthonormal and complete set of eigenvectors. Furthermore, it is positive-definite, so all of its eigenvalues are positive. In this context, each eigenvector is a normal mode of the system, and each eigenvalue is the square of the corresponding oscillation frequency.

In the above we started with a small set of assumptions, namely that (1) the kinetic energy is a smooth function of the velocities, whose minimum is at 0, and (2) the potential energy is a smooth function of the configuration, with at least one local minimum somewhere. We derived the fact that, at low energies, the system behaves like a set of harmonic oscillators. The first assumption is, as I said above, more or less a basic postulate of physics. The second assumption is fairly mild—we expect most functions in physics to be smooth, and as long as the potential energy is bounded below, it must have at least one minimum. These assumptions, and the mathematics that followed from them, are the reason—and the only reason—for the ubiquity of the harmonic oscillator in physics. It is also important to point out, however, that there are exceptions to assumption (2). A prominent example is the Coulomb potential, which is not smooth, is unbounded below, and has no local minima. Other examples occur when the minimum occurs on the boundary of the configuration space (this might sound exotic, but it's not: consider an object sitting on the surface of the Earth). More generally, the assumption of smoothness in physics, although very common (and often made without being stated), does *not* hold universally. Many functions in physics are far from smooth; consider the trajectory of a molecule in a gas, or fractals such as the shape of a shoreline. An example from outside of physics is stock prices.

1.7 Complex vector spaces

When dealing with complex vector spaces, most of the concepts discussed for real vector spaces go through unchanged. The biggest change concerns quadratic forms: rather than dealing with *symmetric* quadratic forms, we usually deal with *Hermitian* ones. These are anti-linear in the first argument, $(\lambda\mathbf{u},\mathbf{v}) = \lambda^*\mathbf{g}(\mathbf{u},\mathbf{v})$, and obey $\mathbf{g}(\mathbf{u},\mathbf{v}) = \mathbf{g}(\mathbf{v},\mathbf{u})^*$. This implies, usefully, that the norm of any vector is real. Again, a positive-definite quadratic form is an inner product; the above statements about orthogonal and symmetric operators continue to hold if we replace orthogonal with *unitary* and symmetric with *Hermitian*. (Hermitian operators are also sometimes called *self-adjoint*.)

An example of the statement that in the presence of a fixed inner product we can forget about the distinction between vectors and dual vectors is implicit in the bra(c)ket notation used for quantum mechanical systems. Such a system is represented by a complex vector space equipped with a fixed Hermitian inner product. A vector is written as a ket, $|\psi\rangle$, and a dual vector as a bra, $\langle\psi|$. The inner product provides a on-to-one map between vectors and dual vectors. Thus, the scalar $\langle\psi_1|\psi_2\rangle$ can be thought of either as the dual vector $\langle\psi_1|$ acting on the vector $|\psi_2\rangle$, or equivalently as the inner product of the two vectors $|\psi_1\rangle$ and $|\psi_2\rangle$.

1.8 Infinite-dimensional vector spaces

Many of the vector spaces we deal with in physics, such as function spaces, are infinite-dimensional. For example, consider the space of complex functions on $\mathbf{R}$:

$$\mathcal{F}(\mathbf{R}) = \{f : \mathbf{R} \to \mathbf{C}\}\,. \tag{39}$$

In such a space you often want to consider *infinite* sums of vectors. For example, you might want to be able to write each vector in the space as an infinite linear combination of basis vectors. To define a notion of convergence you need a *norm*, or, better yet, an *inner product*. The most common inner product appearing in physics is called L^2:

$$(f, g) = \int dx\, f(x)^* g(x)\,. \tag{40}$$

(A small variation is $(f, g) = \int dx \rho(x) f(x)^* g(x)$, where ρ is a fixed positive function.) This is obviously Hermitian, and is non-negative definite: $(f, f) \geq 0$. However, two problems present themselves:

1. Many functions in $\mathcal{F}(\mathbf{R})$ (constant ones, for example) have infinite norm. Fortunately, the ones with finite norm, the *square-integrable* functions, still form a vector space. We call this subspace $C^0(\mathbf{R})$:

$$C^0(\mathbf{R}) = \left\{(f : \mathbf{R} \to \mathbf{C}) : \int dx |f(x)|^2 < \infty\right\}\,. \tag{41}$$

2. Many functions in $C^0(\mathbf{R})$ are *degenerate*, meaning that $(f,g)=0$ for all g. An example is

$$f(x)=\begin{cases}1, & x=0\\ 0, & x\neq 0\end{cases}. \tag{42}$$

   Hence $C^0(\mathbf{R})$ is still not an inner product space. Note that such degenerate functions themselves form a vector space $\mathcal{F}_0(\mathbf{R})\subset C^0(\mathbf{R})$:

$$\mathcal{F}_0(\mathbf{R})=\{f\in C^0(\mathbf{R}):\forall g\in\mathcal{F}(\mathbf{R}),(f,g)=0\}. \tag{43}$$

There are two ways to get around the existence of degenerate vectors. The first is to eliminate them. For example, we could restrict ourselves to *continuous* functions, or, even better, to *infinitely differentiable* functions. This subspace of $C^0(\mathbf{R})$, which we will call $C^\infty(\mathbf{R})$, *does* form an inner product space. The differentiability condition is also useful, of course, because we are frequently interested in differentiating functions, i.e. considering *differential operators*. The simplest example is the momentum operator $-i\partial/\partial x$. It is easy to see that this operator is Hermitian. Sadly, however, despite what we might expect from the finite-dimensional case, it doesn't have even a single eigenvector! Its would-be eigenvectors are functions of the form e^{-ikx} for fixed k, which have infinite L^2 norm. This is where physicists start talking about "delta-function normalizable" functions, a notion that we will make slightly more precise (but still not rigorous) below.

However, the inner product space $C^\infty(\mathbf{R})$ has another problem: you can have a infinite sequence of functions that converge nicely, in the sense that they are getting closer and closer to each other according to the norm, yet do not have any limit. This is what happens, for example, in the set of rational numbers: the sequence 3, 3.1, 3.14, 3.141, . . . appears to be converging (technically, it's a *Cauchy sequence*), but it has no limit in the rationals. The rationals are *incomplete*, which makes doing analysis on them essentially impossible; when you complete them you get the reals. Similarly, it's easy to write down a sequence of C^∞ functions that seem to be converging nicely with respect to the L^2 norm but don't have a limit (e.g. $e^{-x^2}\tanh(nx)$). This incompleteness essentially rules out having a useful concept of an infinite basis.

One manifestation of the incompleteness of $C^\infty(\mathbf{R})$ is that, unlike for a finite-dimensional inner product space, its *dual vector space* is much larger than it is. For example, the linear functionals $f\mapsto f(0)$ and $f\mapsto f'(0)$ cannot be written by integrating f against some other function g in $C^\infty(\mathbf{R})$.

To write them as the L^2 inner product of f with another function, the other function must be a *delta function* ($\delta(x)$ and $-\delta'(x)$ respectively). The dual space of C^∞ is called the *Schwartz space* $\mathcal{S}(\mathbf{R})$; this is how mathematicians rigorously define functions spaces that include delta functions.

1.9 Hilbert spaces

Given the incompleteness of the inner product space $C^\infty(\mathbf{R})$, we now turn to a different way of solving problem 2 above, the presence of degenerate vectors in $C^0(\mathbf{R})$. Rather than *removing* them from the vector space, we *quotient* it by them. In other words, we declare that they are equivalent to $\mathbf{0}$. The quotient space $C^0(\mathbf{R})/\mathcal{F}_0(\mathbf{R})$ is called $L^2(\mathbf{R})$, and it can be shown that it is complete. More generally, a *complete inner product space* is called a *Hilbert space*. (A Hilbert space can be finite or infinite-dimensional; in particular, every finite-dimensional inner product space is a Hilbert space.)

A theorem tells us that every[2] Hilbert space admits a *Hilbert basis*; this is a (countably infinite, if the space is infinite-dimensional) set of vectors $\mathbf{e}_i$ that are (1) orthonormal, i.e. $(\mathbf{e}_i, \mathbf{e}_j) = \delta_{ij}$, and (2) complete, i.e. every vector $\mathbf{v}$ can be written uniquely in the form $v_i \mathbf{e}_i$. Note that the squares of the coefficients add up to the norm of $\mathbf{v}$, $v_i^* v_i = (\mathbf{v}, \mathbf{v})$. A Hilbert basis for $L^2(\mathbf{R})$ is given, for example, by the harmonic oscillator wavefunctions $e^{-x^2/2} H_i(x)$, where H_i are the Hermite polynomials.

The above construction works for any space on which we can integrate functions. For example, we can consider functions on $\mathbf{R}$ that are periodic with period 2π, and consider the inner product

$$(f, g) = \int_0^{2\pi} dx\, f(x)^* g(x)\,. \tag{44}$$

There a Hilbert basis is provided by the Fourier modes $(2\pi)^{-1/2} e^{inx}$.

Although it has a lot of nice properties, the Hilbert space $L^2(\mathbf{R})$ also has some strange ones. First, because it includes all kinds of discontinuous functions, derivative operators are not defined on it. Even stranger, a given element of $L^2(\mathbf{R})$, although it can be represented by a function, does not have a definite value at a given point x. For example, for $f \in L^2(\mathbf{R})$, you cannot ask what $f(0)$ is, since you can always add a function of the form (42) without changing the element of $L^2(\mathbf{R})$. This eliminates those delta functions from the dual space; in fact, it can be shown that, just as for

[2] separable (a technicality)

finite-dimensional inner product spaces, the dual space of a Hilbert space is itself.

To get around these problems, it is possible to use the fact that $C^\infty(\mathbf{R})$ is actually a *dense* subspace of $L^2(\mathbf{R})$, in the same sense that the rationals are dense in the reals: any element of $L^2(\mathbf{R})$ can be approximated to arbitrary accuracy by some element of $C^\infty(\mathbf{R})$. This allows one to define various useful concepts, such as the *spectrum* of an operator, even where (as for the momentum operator) it is neither well-defined nor has any eigenvectors. Doing this rigorously entails a lot of real analysis which, although interesting enough for those who like that sort of thing, is rarely relevant in the day-to-day work of physicists. Instead, we will, as von Neumann said of Dirac, "adhere to the fiction that each self-adjoint operator can be put in diagonal form. In the case of those operators for which this is not actually the case, this requires the introduction of 'improper' functions with self-contradictory properties [i.e. delta functions]." (*Mathematical Foundations of Quantum Mechanics*) The basic statement is as follows: A Hermitian operator on a Hilbert space has a spectrum, which is a subset of $\mathbf{R}$ that may include both discrete points and intervals. We call the discrete set D and the continuum part C. To each point $\lambda \in D$ there corresponds an eigenvector $|\lambda\rangle$ which is normalizable. (The eigenvalue may of course be degenerate, meaning it has more than one linearly independent eigenvector. To simplify the notation, we'll ignore this possibility for now.) To each point $\alpha \in C$ there corresponds an eigenvector $|\alpha\rangle$ which is "delta-function normalizable", meaning $\langle\alpha_1|\alpha_2\rangle = \delta(\alpha_1 - \alpha_2)$. Furthermore, these eigenvectors form a complete basis for the Hilbert space, meaning that any vector $|\psi\rangle$ can be (uniquely) expanded as

$$|\psi\rangle = \sum_{\lambda\in D} \psi_\lambda |\lambda\rangle + \int_C d\alpha\, \psi(\alpha)|\alpha\rangle\,, \tag{45}$$

where

$$\psi_\lambda = \langle\lambda|\psi\rangle\,, \qquad \psi(\alpha) = \langle\alpha|\psi\rangle\,. \tag{46}$$

For example, on $L^2(\mathbf{R})$, the momentum operator $-i\partial/\partial x$ (which, as mentioned above, is defined via its action on the dense subspace $C^\infty(\mathbf{R})$), has a spectrum which covers the entire real line. The normalized eigenvectors are the functions $|k\rangle = (2\pi)^{-1/2}e^{ikx}$, which obey $\langle k|k'\rangle = \delta(k-k')$. Any function can be expanded in these eigenfunctions:

$$f(x) = \frac{1}{(2\pi)^{1/2}}\int dk\, \tilde{f}(k)e^{ikx}\,, \qquad \tilde{f}(k) = \frac{1}{(2\pi)^{1/2}}\int dx\, f(x)e^{-ikx}\,. \tag{47}$$

On the other hand, for example, the harmonic oscillator Hamiltonian

$$H = -\frac{1}{2}\frac{\partial^2}{\partial x^2} + \frac{1}{2}x^2 \tag{48}$$

has a purely discrete spectrum, namely $\mathbf{N} + \frac{1}{2}$. The eigenfunctions are the ones mentioned above as a Hilbert basis for $L^2(\mathbf{R})$.

1.10 Quantum mechanics

According to the axioms of quantum mechanics, associated with any quantum system is a state space $\mathcal{H}$, which is a complex Hilbert space that may be finite or infinite dimensional. The distinct physical states are elements of the *projectivization* $P(\mathcal{H})$ of $\mathcal{H}$, that is to say vectors with unit norm $\langle\psi|\psi\rangle = 1$ identified by $|\psi\rangle \sim e^{i\alpha}|\psi\rangle$, where α is real. (Note that $P(\mathcal{H})$ is not itself a vector space.)

Each observable is represented by a Hermitian operator A. The possible experimental values obtained from a measurement of the observable are given by the spectrum of A. Furthermore, if the system is in the state $|\psi\rangle$ before the measurement, then the probability of obtaining the value λ in the discrete part of the spectrum is $|\psi_\lambda|^2$ (in the expansion (45)), while the probability of obtaining a value between α_1 and α_2 within one of the intervals in the continuous part is $\int_{\alpha_1}^{\alpha_2} d\alpha|\psi(\alpha)|^2$. It follows that the expectation value of the measurement is $\langle\psi|A|\psi\rangle$. (Note that, because $\langle\psi|\psi\rangle = 1$, the total probability is 1.) Furthermore, the system immediately after the measurement will be in the state represented by the eigenvector corresponding to the measured eigenvalue (or, if the measured eigenvalue is in the continuum part of the spectrum, then a linear combination of eigenvectors corresponding to nearby eigenvalues, depending on the measuring apparatus).[3]

One observable with a special status is the *Hamiltonian*, which generates the time evolution of the system according to the Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = H|\psi(t)\rangle\,. \tag{49}$$

(H may also be time-dependent.) Because H is Hermitian, the operator relating the state at a time t_1 to that at a time t_2 is unitary:

$$|\psi(t_2)\rangle = U(t_2, t_1)|\psi(t_1)\rangle\,. \tag{50}$$

[3] This set of rules is called the *Copenhagen interpretation*. The truth is slightly more complicated, but the Copenhagen interpretation works in practice.

For example, if H is time-independent, then

$$U(t_2, t_1) = e^{-i(t_2-t_1)H}\,. \tag{51}$$

As far as anyone knows, the appearance of these linear structures in quantum mechanics is *not* due to any process of linearization. Many people have considered the idea that the Schrödinger equation might be just the linear approximation to some more general non-linear equation. It turns out that this idea quickly runs into serious and fundamental problems, and it seems likely that the linear structure of quantum mechanics is an essential and exact feature of the laws of physics.