A primer on linear algebra and probability

STCS Vigyan Vidushi 2024

1 Linear algebra

Linear algebra deals with “vectors”, “vector spaces”, “matrices”, “eigenvalues”, “determinants” etc. In this primer we give you a brief overview of some basic concepts in linear algebra.

1.1 Definition and examples

What are vector spaces (informally)?

A collection of “vectors” that can be *added*, *subtracted*, and *scaled*.

Most of the times, we tend to think of “vectors” as a *tuple* of coordinates (for example: $(1, 2, -1) \in \mathbb{R}^3$, or $a + \imath b \in \mathbb{C}$ but thought as the tuple $(a, b) \in \mathbb{R}^2$). But the only things that a vector space needs to have is the notion of “addition”, “subtraction” and “scaling”.

Definition 1.1 (Vector spaces)**.** *A* vector space V *over a* field $\mathbb{F}$ *(such as the real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$) is a set of elements, and three binary operations corresponding to “addition”, “subtraction” and “scaling” with the following properties:*

The *zero* vector: *The set of vectors contains a special* zero *vector (often denoted by just $\mathbf{0}$).*

Addition and subtraction[1]**:** *There is an associative binary operations $+ : V \times V \to V$ that is*

- *commutative (i.e., $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$; order of addition does not matter),*
- *associative (i.e., $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$; order of bracketing does not matter), and*
- *satisfies $\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u}$ (adding the* zero *vector does nothing).*
- *every vector $\mathbf{u} \in V$ has a unique vector $-\mathbf{u} \in V$ that is its negative in the sense that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$. Using this, subtraction of two vectors* u *and $\mathbf{v}$ is just defined as $\mathbf{u} + (-\mathbf{v})$ (we will use $\mathbf{u} - \mathbf{v}$ as shorthand for $\mathbf{u} + (-\mathbf{v})$).*

Scaling: *There is a binary operation $\cdot : \mathbb{F} \times V \to V$ that “scales” a vector by a field constant, such that*

- $1 \cdot \mathbf{v} = \mathbf{v}$ *(i.e., scaling a vector by 1 does nothing),*
- $\alpha \cdot (\beta \cdot \mathbf{v}) = (\alpha\beta) \cdot \mathbf{v}$ *(i.e., successive scalings compose naturally), and*
- $\alpha \cdot (\mathbf{u} + \mathbf{v}) = (\alpha \cdot \mathbf{u}) + (\alpha \cdot \mathbf{v})$ *(i.e., scaling distributes over addition, or the scale of a sum of vectors is the sum of the scaled vectors).* ◊

Base version: (2024-07-09 11:40:19 +0530) , b1c8621

The most standard examples are vector spaces such as $\mathbb{R}^3 = \{(x, y, z) \ : \ x, y, z \in \mathbb{R}\}$, or the set of complex number $\mathbb{C} = \{a + \imath b \ : \ a, b \in \mathbb{R}\}$ (as a vector space over $\mathbb{R}$), or polynomials of degree at most 3 ($\{f_0 + f_1 x + f_2 x^2 + f_3 x^3 \ : \ f_1, f_2, f_3, f_4 \in \mathbb{R}\}$), etc.

All of these are examples where there inherently appear to be "coordinates". However, this need not always be the case. There are some non-standard examples to keep in mind

A non-standard example: Functions from $\mathbb{R}$ to $\mathbb{R}$

If V is the set of all possible functions from $\mathbb{R}$ to $\mathbb{R}$ (i.e., $V = \{f : \mathbb{R} \to \mathbb{R}\}$) then we can certainly add/subtract/scale functions — for example, $f + g$ is the function that maps any x to $f(x) + g(x)$, and $3 \cdot f$ is the function that maps x to $3 \cdot f(x)$:

$$\begin{aligned} f + g &: x \mapsto f(x) + g(x) \\ 3 \cdot f &: x \mapsto 3 \cdot f(x). \end{aligned}$$

Although there does not appear to be any "coordinates" here, the above is still an example of a legitimate vector space.

In most situations that we will deal with, the vector spaces have finite *dimension*, and coordinates then become more meaningful.

1.2 Linear combination, linear span, linear dependence, basis and dimension

We say that a vector $\mathbf{w}$ is a *linear combination* of a sequence of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ if there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that

$$\mathbf{w} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_k \mathbf{v}_k.$$

We say that a sequence of vectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ is *linearly dependent* if there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$, at least one of them non-zero, such that

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_k \mathbf{v}_k = \mathbf{0}.$$

If the sequence is not linearly dependent, then we call it linearly independent. A *set* S (instead of a sequence) of vectors is linearly independent if every *finite* sequence of distinct vectors from S is linearly independent.

Examples

The sequence $\left([0,1,-1]^T, [-1,0,-1]^T, [1,1,0]^T\right)$ of vectors from $\mathbb{R}^3$ is linearly dependent as

$$1 \cdot [0,1,-1]^T + (-1) \cdot [-1,0,-1]^T + (-1) \cdot [1,1,0]^T = [0,0,0]^T.$$

On the other hand, the sequence $\left([0,1,-1]^T, [-1,0,-1]^T\right)$ is linearly independent, because if

$$\alpha \cdot \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} + \beta \cdot \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

then both α and β must be zero.

Example

Suppose $S = \left\{[0,1,-1]^T, [-1,0,-1]^T, [1,1,0]^T\right\}$, then

$$\mathrm{span}(S) = \left\{ \begin{bmatrix} \beta - \alpha \\ \beta \\ -\alpha \end{bmatrix} : \alpha, \beta \in \mathbb{R} \right\}.$$

Consider the following collection of vectors all from $\mathbb{R}^4$:

$$\{(1,-1,0,3), (1,1,1,1), (2,0,1,4), (0,-2,-1,2), (3,1,2,5), (4,2,3,6), (4,0,2,8)\}.$$

Let us say that each Vigyan Vidushi participant is asked to choose a *maximal* linearly independent set. That is, each participant chooses some set S of vectors that happens to be linearly independent, and it is *maximal* in the sense that they cannot extend the set S by another vector from the above collection and still keep it linearly independent. In general, a set S being maximal according to some property does not mean that this is the largest possible among all sets; it only means that there are no ways of extending S by additional elements and still maintaining the property.
Hence, it should be surprising to learn that each Vigyan Vidushi participant will have picked a set of size two. That is, all maximal linearly independent subsets all have the same size! This is a non-trivial fact about linear independence.

Let S be a set of vectors from some vector space V. Some other vector can be obtained as a linear of combination of (finite sequences of) vectors from S. The collection of all vectors that can be obtained as a linear combination of vectors in S will be called $\mathrm{span}(S)$, that is,

$$\mathrm{span}(S) = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_k\mathbf{v}_k\},$$

Note that $\mathrm{span}(S)$ may not include all vectors in V, but it will be vector space under the same operations of vector addition and scalar multiplication. We call a set S a *spanning* set if $\mathrm{span}(S) = V$. If a vector space V has a finite spanning set, then we say that V is finite-dimensional. We will deal exclusively with finite-dimensional vector spaces.

Linearly independent sets and spanning sets are of opposing flavours. A subset of a linearly independent set is, of course, linearly independent. Similarly, a superset of a spanning set is a spanning set. We will soon see that every vector space has a set that is both linearly independent and spanning.

Theorem 1.2.

- *Every maximal linearly independent set is a spanning set.* (*Not hard.*)
- *Every minimal spanning set is a linearly independent set.* (*Not hard.*)
- *If L is a linearly independent set and S is a spanning set, then $|L| \leq |S|$.* (*Deep fact about vector spaces.*)

The above observations immediately imply that in a finite-dimensional vector space, all linearly independent sets are finite. Furthermore, a maximal independent set (it exists, why) is spanning. A sequence of vectors that is both linearly independent and spanning (e.g., the elements of a maximal linearly independent set ordered in some way) is called a *basis* of the vector space. Do you now see why the third observation implies that all such bases must have the same cardinality[2]? This cardinality is the *dimension* of the vector space. Furthermore, note that every every minimal spanning set is also a basis.

Exercise (optional)

Every subspace of a finite-dimensional vector space is finite-dimensional.

Bases allow us to assign coordinates to vectors. Let us fix a basis $\mathbf{B} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ for a vector space V.

(a) Because $\mathbf{B}$ is spanning, every vector w in V can be written as a linear combination of vectors in B, say,

$$\mathbf{w} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots \alpha_k\mathbf{v}_k,$$

[2] Using cardinality now because we might need size for bit-lengths later.

which we will sometimes write as

$$\mathbf{w} = \left[\mathbf{v}_1,\, \mathbf{v}_2,\, \ldots,\, \mathbf{v}_k\right] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}$$

Here we think of $\left[\mathbf{v}_1,\, \mathbf{v}_2,\, \ldots,\, \mathbf{v}_k\right]$ as a row whose entries are abstract vectors. In this way we think of the column vector $[\alpha_1, \alpha_2, \ldots, \alpha_k]^T$ as a representation of $\mathbf{w}$ with respect to the basis $\mathbf{B}$.

(b) Next, because $\mathbf{B}$ is a basis, this representation is unique. For suppose $[\alpha'_1, \alpha'_2, \ldots, \alpha'_k]^T$ is another representation. Then, we have $\mathbf{w} = \mathbf{B}[\alpha_1, \alpha_2, \ldots, \alpha_k]^T$ and $\mathbf{w} = \mathbf{B}[\alpha'_1, \alpha'_2, \ldots, \alpha'_k]^T$. Then, $0 = B[\ldots, \ldots, \ldots]$. Complete this proof as an exercise.

A very convenient basis for vector spaces such as $\mathbb{R}^3$ is what is called *the standard basis* consisting of vectors $\mathbf{e}_1 = [1,0,0]^T$, $\mathbf{e}_2 = [0,1,0]^T$ and $\mathbf{e}_3 = [0,0,1]^T$. In fact, do have this basis in mind all the time, because the representation of the vector $\mathbf{w} = [2,3,5]^T$ in the basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is $[2,3,5]^T$, that is $\mathbf{w}$ itself, because $\mathbf{w} = 2\mathbf{e}_1 + 3\mathbf{e}_2 + 5\mathbf{e}_3$.

1.3 Linear transformations, and matrices

Linear transformations

Suppose V and W are two vector spaces over the same field $\mathbb{F}$. A linear transformation is a map $\varphi\colon V \to W$ "that behaves linearly". That is, it must satisfy properties such as

$$\varphi(\alpha \cdot \mathbf{u} + \beta \cdot \mathbf{v}) = \alpha \cdot \varphi(\mathbf{u}) + \beta \cdot \varphi(\mathbf{v}) \qquad \text{for any } \alpha, \beta \in \mathbb{F}\ ,\ \mathbf{u},\ \in V.$$

Notice that the scaling operation and the addition operation on the LHS are in the vector space V, and on the RHS are in the vector space W. Hence, the map φ and structure of V, W "works well together" in the sense that it doesn't matter if you apply the operations and then apply φ, or if you apply φ and then apply the operations.

Suppose we have two vector space V and W over the same field, how do we specify such a linear transformation? Suppose you had a basis $\{\mathbf{v_1}, \ldots, \mathbf{v_r}\}$ for V, then knowing $\varphi(\mathbf{v_i})$ for $i = 1, \ldots, r$ is sufficient to figure out $\varphi(\mathbf{v})$ for any other $\mathbf{v} \in V$ (because we know any $\mathbf{v} \in V$ can be expressed as a linear combination in terms of the basis, and the above property of linear transformations will let us work out $\varphi(\mathbf{v})$ should be). If we also have a basis $\{\mathbf{w_1}, \ldots, \mathbf{w_s}\}$ for the space W, then we can express $\varphi(\mathbf{v_i})$ in terms of this basis. Hence, suppose

$$\varphi(\mathbf{v_i}) = a_{1i}\mathbf{w_1} + \cdots + a_{si}\mathbf{w_s}, \ \text{for all } i = 1, \ldots, r,$$

we can express this "data" as an $s \times r$ matrix

$$M = \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{sr} \end{bmatrix}$$

where the i-th column is just $\varphi(\mathbf{v_i})$ expressed as the linear combination of $\{\mathbf{w_1}, \dots, \mathbf{w_s}\}$. Therefore, if $\mathbf{v} = c_1\mathbf{v_1} + \cdots + c_r\mathbf{v_r}$, then $\varphi(\mathbf{v})$ when expressed in terms of $\{\mathbf{w_1}, \dots, \mathbf{w_s}\}$ is simply the matrix vector product

$$\begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots \\ a_{s1} & \cdots & a_{sr} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_r \end{bmatrix}.$$

For any linear transformation $\varphi : V \to W$, once we fix a basis for V and a basis for W, we can represent the linear transformation using an $\dim(W) \times \dim(V)$ matrix $M_\varphi = (m_{ij} : i = 1, 2 \dots, \dim(W), j = 1, 2, \dots, \dim(V))$. Indeed, we can fix these bases so that matrix has very special *diagonal* form: the only non-zero entries appear along the principal diagonal of the matrix, that is, $m_{ij} = 0$ whenever $i \neq j$. If coordinates are assigned to vectors using bases wrt which the matrix has this diagonal form, then computation becomes straight forward; to apply the transformation, we need to scale the coordinates appropriately, drop some coordinates, or pad some with zeros.
When the domain and co-domain of the linear transformation are the same (such linear transformations are called linear operators), then it natural to use the same basis for representing vectors in the domain and co-domain. Note that in this case, the transformation is represented by a square matrix. Motivated by the remarks above, we can ask if one can, by choosing an appropriate basis, represent such linear operators by diagonal matrices.

1.4 Diagonalization, eigenvalues and eigenvectors

Suppose $\varphi : V \to V$ is a linear transformation. Let $\mathbf{B} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ be a basis for which φ is represented as a diagonal matrix with scalars $(\lambda_1, \lambda_2, \dots, \lambda_n)$ along the diagonal. Then,

$$\varphi(\mathbf{v}_i) = \lambda_i \mathbf{v}_i, \text{ for } i = 1, 2, \dots, n.$$

That is, the action of φ on the i-th basis vector amounts to scaling it by a factor λ_i. In general, we say that a vector $\mathbf{v} \neq 0$ is an *eigenvector* of the linear operator φ, if

$$\varphi(\mathbf{v}) = \lambda \mathbf{v}.$$

The scalar (for it scales!) λ is the *eigenvalue* associated with the eigenvector $\mathbf{v}$. Clearly, repeatedly applying φ to the eigenvector $\mathbf{v}$ repeatedly scales $\mathbf{v}$:

$$\varphi^{(r)}(\mathbf{v}) = \lambda^r \mathbf{v}.$$

Note that this equality holds for $r = 0$, if we regard $\varphi^{(0)}$ as the identity linear operator. (What if φ is invertible and r is negative?) In general, suppose $p(X)$ is a polynomial: $p_r X^r + p_{r-1} X^{r-1} + \cdots + p_0$, the we may substitute φ for X, and regard $p(\varphi)$ as the linear operator

$$p_r \varphi^{(r)} + p_{r-1} \varphi^{(r-1)} + \cdots + p_0 \varphi^{(0)}.$$

Suppose $\mathbf{v}$ is an eigenvector of φ with eigenvalue λ. What is $p(\varphi)\mathbf{v}$?

Exercise: Suppose $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are eigenvectors with distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Show that $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ is linearly independent. Hint: construct a polynomial $p_i(X)$ such that $p_i(\lambda_i) = 1$ and $p_i(\lambda_j) = 0$ for $j \neq i$. Conclude that φ can have at most $\dim(V)$ distinct eigenvalues.

Exercise: Give an example of a linear operator φ on $\mathbb{R}^2$ that has (i) two linearly independent eigenvectors, (ii) no eigenvector, (iii) has an eigenvector but does not have two linearly independent eigenvectors.

The linear operator $\varphi : V \to V$ has a representation as diagonal matrix iff V has a basis consisting of eigenvectors of φ.

1.5 Determinants and volume

Consider a linear operator $\varphi : \mathbb{R}^d \to \mathbb{R}^d$. Under this operator shapes get deformed. In general, a cube becomes a parallelepiped, a ball becomes an ellipsoid, etc. However, it is a remarkable property of linear operators that the ratio of the volumes of the original shape and its image under φ is a fixed constant independent of the shape. This quantity can be computed using the matrix M_φ of φ (with respect to some basis). Given such an $n \times n$ matrix $M = (m_{ij})$, we define its determinant to be

$$\det(M) = \sum_{\sigma \in \mathcal{S}_n} \prod_{i=1}^{n} \operatorname{sign}(\sigma) \cdot m_{i\sigma(i)},$$

where σ ranges over the set $\mathcal{S}_n$ of all permutations of $\{1, 2, \ldots, n\}$. Remarkably, $\det(M_\varphi)$ does not depend on the basis with respect to which it is written; so we may refer this quantity as $\det(\varphi)$. The determinant, which is a polynomial function of the entries of M is a very important quantity. It has many important properties. For example, suppose M is diagonalizable, under the basis of eigenvectors $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ with corresponding eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then $\det(M)$ is the product $\lambda_1 \lambda_2 \cdots \lambda_n$. The determinant can be computed efficiently, for example, by Gaussian elimination, even for matrices that are not diagonalizable; interestingly, the claim that Gaussian elimination is efficient can be established by considering certain determinants that arise in the course of Gaussian elimination.

Returning to our discussion on volumes, the of a body A is transformed to a body B under the transformation φ, then

$$\mathrm{vol}(B) = |\det(\varphi)|\, \mathrm{vol}(A).$$

2 Fields

Most of the time, we would be dealing with vectors where the coefficients come from familiar "fields" such as the real numbers ($\mathbb{R}$), rational numbers ($\mathbb{Q}$) or complex numbers ($\mathbb{C}$).

What is a field?

A field $\mathbb{F}$ is just a set (of "numbers") where we can meaningfully add, subtract, multiply and divide (by any "nonzero" element). In other words, these are sets of elements that *behave* like the familiar fields like $\mathbb{R}, \mathbb{Q}, \mathbb{C}$ etc.

Formally, a field is defined as follows.

Definition 2.1 (Field). *A field is specified by a set of elements $\mathbb{F}$ and two binary operations $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ and $\times : \ : \mathbb{F} \times F \to \mathbb{F}$ that satisfy the following properties.*

Addition and multiplication are commutative: *For any pair of elements $a, b \in \mathbb{F}$, we have $a + b = b + a$ and $a \times b = b \times a$.*

Contains 'zero' and 'one': *There is a unique element in $\mathbb{F}$, called the 'zero' element, denoted by 0 that satisfies*

$$a + 0 = 0 + a = a \quad \textit{for all } a \in \mathbb{F}.$$

The field also contains a unique element, called the 'one' element denoted by 1, that satisfies

$$a \times 1 = 1 \times a = a \quad \textit{for all } a \in \mathbb{F}.$$

Distributivity: *For any $a, b, c \in \mathbb{F}$, we have $a \times (b + c) = (a \times b) + (a \times c)$.*

Subtraction and division: *For every element $a \in \mathbb{F}$, there is a unique element called $(-a) \in \mathbb{F}$ such that $a + (-a) = 0$. Similarly, for every $0 \neq a \in \mathbb{F}$, there is a unique element called $a^{-1} \in \mathbb{F}$ such that $a \times a^{-1} = 1$. We will use $a - b$ and a/b as shorthand for $a + (-b)$ and $a \times b^{-1}$ respectively.*

◊

Of course, rationals ($\mathbb{Q}$), reals ($\mathbb{R}$) and complex numbers ($\mathbb{C}$) form fields. But here is an example of an *unusual* field:

$$\mathbb{Q}[\sqrt{2}] = \left\{ a + b\sqrt{2} \ : \ a, b \in \mathbb{Q} \right\}.$$

Of course, it is quite clear that the set is closed under the usual addition and multiplication:

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2},$$
$$(a + b\sqrt{2}) \times (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}.$$

The only nontrivial fact is that $1/(a + b\sqrt{2})$ can also be expressed as $c + d\sqrt{2}$ for some rational numbers c and d. And this true because

$$\frac{1}{a + b\sqrt{2}} = \frac{1}{a + b\sqrt{2}} \cdot \frac{a - b\sqrt{2}}{a - b\sqrt{2}} = \left(\frac{a}{a^2 - 2b^2}\right) - \left(\frac{b}{a^2 - 2b^2}\right)\sqrt{2}.$$

All of the fields we have discussed so far have infinitely many elements in them. But turns out, there *are* fields that have finitely many elements in them and these are called *finite fields*.

2.1 Finite fields

Here is an example with just five elements that we will refer to as $\{0, 1, 2, 3, 4\}$:

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\times$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

The above field is called $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$, and in fact we can create a similar field $\mathbb{F}_p = \{0, 1, 2, \ldots, p - 1\}$ for any prime p. These fields are created using *modular arithmetic* and can be succinctly described as follows.

Prime fields

To add two elements of $\mathbb{F}_p := \{0, 1, 2, \ldots, p - 1\}$, just add the elements as integers and output the remainder of the sum when divided by p.
Similarly, to multiply elements of $\mathbb{F}_p := \{0, 1, 2, \ldots, p - 1\}$, just multiply the elements and output the remainder when divided by p.

Often, the notation $a \equiv b \bmod p$ is used to denote that "a and b leave the same remainder when divided by p", which is equivalent to stating that p divides $a - b$. And $(a \bmod p)$ is sometimes used to refer to remainder when divided by p, which is also the unique element $b \in \{0, 1, \ldots, p - 1\}$ such that $a \equiv b \bmod p$.

It can be easily seen that $(-a)$ is nothing but the element $(p - a) \bmod p \in \mathbb{F}_p$. However, it is perhaps surprising that every $0 \neq a \in \mathbb{F}_p$ has a multiplicative inverse $a^{-1} \in \mathbb{F}_p$ (the fact that p was a prime is important here).

Fact 2.2. *Let p be any prime. For any $a \in \{1, 2, \ldots, p-1\}$, there is a unique $b \in \{1, 2, \ldots, p-1\}$ such that $ab \equiv 1 \bmod p$.*

Now that we have these *prime fields*, we can also talk about vector spaces where the coefficients are from $\mathbb{F}_p$ instead since we can now meaningfully add, subtract and scale vectors.

Other finite fields

It turns out that for any prime p, the above field $\mathbb{F}_p$ is the *only* field of size p (up to calling elements by different names (also known as ’isomorphisms’)). Also, for any prime p and a positive integer $r > 0$, there is also a *unique* field $\mathbb{F}_{p^r}$ consisting of exactly p^r elements (and this is NOT just addition and multiplication modulo p^r). Not only that, these are the only finite fields possible. We’ll just write down the addition and multiplication table for $\mathbb{F}_4 = \{0, 1, a, b\}$ as an example:

+	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

×	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

3 Probability

Let us start with some basic definitions: given a finite set Ω, a *probability distribution* on Ω is an assignment of non-negative “weights” to the elements of Ω, so that the weights sum up to 1. Formally, we can write this is a function $\mu : \Omega \to [0, 1]$, such that $\mu(x) \geq 0$ for each $x \in \Omega$, and such that $\sum_{x \in \Omega} p(x) = 1$. A particularly important special case is of the *uniform distribution*, where $p(x)$ is the same for each $x \in \Omega$.

Given a probability distribution, we can think of any function on Ω as a *random variable*. The simplest and most important case is when the function is the identity function: $x \mapsto x$ for each $x \in X$. In that case we say that the random variable “X is sampled according to μ” or has “probability distribution μ” if it takes the value $x \in \Omega$ with probability $\mu(x)$. This is denoted as $X \sim \mu$ and also as

$$\Pr_{X \sim \mu}[X = x] = \mu(x) \text{ for all } x \in \Omega. \tag{3.1}$$

More generally,

$$\Pr_{X \sim \mu}[X \in A] = \mu(A) \text{ for all } A \subseteq \Omega, \tag{3.2}$$

where

$$\mu(A) := \sum_{x \in A} \mu(x). \tag{3.3}$$

Consider now a random variable Y taking values in $\mathbb{R}$: it can then be seen as a function from Ω to $\mathbb{R}$. We then define

$$\Pr_{\mu}[Y = r] := \mu\left(\{x \in \Omega \mid Y(x) = r\}\right). \tag{3.4}$$

The *expectation* of *mean* of such a random variable is then defined as

$$\mathop{\mathbb{E}}_{\mu}[Y] := \sum_{x \in X} \mu(x) Y(x). \tag{3.5}$$

The μ in the subscript in the definitions in eqs. (3.3) and (3.5) denotes the underlying probability distribution, and we will usually drop it when the distribution under discussion is clear from the context.

Finally, we define the *conditional probability*, the probability that "X is in set A, *conditioned on* or *given* the fact that it is known to be in set B" as

$$\Pr_{\mu}[X \in A \mid X \in B] := \frac{\Pr_{\mu}[X \in A \cap B]}{\Pr_{\mu}[X \in B]}. \tag{3.6}$$

Another important quantity is the *variance*, which can be a seen as a measure of how much a random variable "varies" from its mean. It is defined as

$$\mathop{\mathrm{Var}}_{\mu}[X] := \mathop{\mathbb{E}}_{\mu}[(X - M)]^2] = \mathop{\mathbb{E}}_{\mu}[X^2] - M^2, \tag{3.7}$$

where $M := \mathbb{E}_{\mu}[X]$.

3.1 Continuous distributions

We will also need to deal with cases where Ω is not a finite set. This might seem straightforward, but it easy to run into logical problems if one is not careful. To see why, try the following: what should it mean to sample a real number from the "uniform distribution" over the real numbers?

For our purposes, we can avoid such logical pitfalls by restricting to a special but important class of probability distributions. Let Ω be $\mathbb{R}^d$, where d is a positive integer. By a *probability density*, we will mean a non-negative function f "which we can integrate", and whose integral over all of Ω is 1:

$$\int_{x \in \mathbb{R}^d} f(x)\, dx = 1. \tag{3.8}$$

Formalizing the phrase "which we can integrate" can be a bit tricky, and we will again avoid some of the technical apparatus needed for that formalization by restricting our attention to the case when f is a continuous function.

With this we can define probabilities and expectations with respect to f, as in eqs. (3.2) and (3.5). For any set A, the indicator function I_A is defined as $I_x(x) := 1$ when $x \in A$ and $I_A(x) = 0$ when $x \notin A$. We then have

$$\Pr_f[A] = \int\limits_{x \in \mathbb{R}^d} I_A(x) f(x)\, dx, \text{ and} \tag{3.9}$$

$$\mathop{\mathbb{E}}_f[Y] = \int\limits_{x \in \mathbb{R}^d} Y(x) f(x)\, dx, \text{ and} \tag{3.10}$$

(Again, this cannot be done for "all" sets A and all functions Y. However we will only be dealing with sets where the integrations in eqs. (3.9) and (3.10) are well defined.)

The definition of conditional probabilities given in eq. (3.6) carries over to this setting as it is, after replacing the definition of probability in eq. (3.2) by the one in eq. (3.9).

Bounded densities An obvious but often useful manipulation one can do with probability densities is to use lower and upper bounds on densities to upper and lower bound probabilities. As an exercise for this, try the following estimate.

Bounding probabilities with densities

Let X be a real valued random variable with density f. Assume that $f(x) \geq \frac{1}{10}$ for all x satisfying $1 \leq x \leq 3$, and that $f(x) \leq \frac{2}{10}$ for all x satisfying $2 \leq x \leq 5$. Show that

1. $\Pr[X > 3] \leq \frac{8}{10}$.
2. $\frac{1}{10} \leq \Pr[2 \leq X \leq 3] \leq \frac{2}{10}$.
3. $\Pr[X > 3 \mid X \geq 2] \leq \frac{8}{9}$.

3.2 Normal distribution

An important continuous distribution is the *Gaussian* or *Normal* distribution. The *standard normal distribution* distribution over $\mathbb{R}$ is given by the density

$$\nu(x) := \frac{1}{\sqrt{2\pi}} \exp(-x^2/2). \tag{3.11}$$

It can be checked that ν integrates to 1 (this does require a trick, but you should try it if you have not seen it before and are familiar with change of variables while integrating in high dimensions). Also, it can be checked that if X has density ν then

$$\mathop{\mathbb{E}}_\nu[X] = 0 \text{ and } \mathop{\mathrm{Var}}_\nu[X] = 1. \tag{3.12}$$

In general, given real numbers M and $\sigma > 0$, we say that $X \sim \mathcal{N}(M, \sigma^2)$ to mean that X has the density

$$\nu_{M,\sigma^2}(x) := \frac{1}{\sigma}\nu\left(\frac{x-M}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-M)^2}{2\sigma^2}\right). \tag{3.13}$$

It can be checked that if $X \sim \mathcal{N}(a, \sigma^2)$ then

$$\mathbb{E}[X] = M \text{ and } \operatorname{Var}[X] = \sigma^2, \tag{3.14}$$

and also that $Z := (X - M)/\sigma$ satisfies $Z \sim \mathcal{N}(0, 1)$, i.e., Z is a standard normal random variable.

3.3 Some resources

For an introduction to probability and its applications, please see the textbook by Grinstead and Snell, made available for free by the authors through the Chance Project here. A classic textbook is *An introduction to probability theory and its applications*, by William Feller, which has a relatively affordable student edition available.

For more background on measure theoretic foundations, see the book *Probability: Theory and Examples* by Rick Durrett: a draft version is available for free from the author's webpage. For several algorithmic applications, see the textbook *Foundations of Data Science* by Blum, Hopcroft and Kannan. A draft is available from the webpage of one of the authors, and a relatively affordable edition has been published by Hindustan Book Agency.