C Linear Algebra

C.1 Introduction

In this appendix chapter we briefly review basic ideas from linear algebra that are fundamental to understanding machine learning. These include vector and matrix arithmetic, vector and matrix norms, and eigenvalue decomposition. The reader is strongly encouraged to ensure familiarity with all concepts mentioned in this chapter before proceeding with the rest of the text.

C.2 Vectors and Vector Operations

We begin by reviewing the fundamental notion of a vector, as well as vector arithmetic.

C.2.1 The vector

A *vector* is another word for an ordered listing of numbers. For example,

$$[-3\ 4\ 1] \tag{C.1}$$

is a vector of three *elements* or *entries*, also referred to as a vector of *size* or *dimension* three. In general, a vector can have an arbitrary number of elements, and can contain numbers, variables, or both. For example,

$$[x_1\ x_2\ x_3\ x_4] \tag{C.2}$$

is a vector of four variables. When numbers or variables are listed out horizontally (or in a row) we call the resulting vector a *row* vector. However, we can list them vertically (or in a column) just as well, in which case we refer to the resulting vector as a *column* vector. For instance,

$$\begin{bmatrix} -3 \\ 4 \\ 1 \end{bmatrix} \tag{C.3}$$

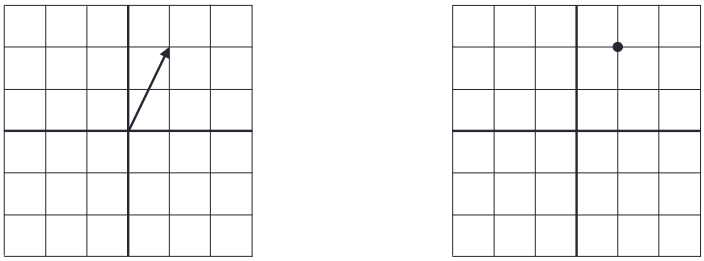

Figure C.1 A two-dimensional vector visualized as an *arrow* stemming from the origin (left panel), or equivalently as a single *point* in a two-dimensional plane (right panel).

is now a column vector of size three. We can swap back and forth between a row and column version of a vector by *transposing* each. Transposition is usually denoted by a superscript T placed just to the right and above a vector, and simply turns a row vector into an equivalent column vector and vice versa. For example, we have

$$\begin{bmatrix} -3 \\ 4 \\ 1 \end{bmatrix}^T = [-3\ 4\ 1] \qquad \text{and} \qquad [-3\ 4\ 1]^T = \begin{bmatrix} -3 \\ 4 \\ 1 \end{bmatrix}. \tag{C.4}$$

To discuss vectors more generally we use algebraic notation, typically a bold lowercase (often Roman) letter, e.g., $\mathbf{x}$. The transpose of $\mathbf{x}$ is then denoted as $\mathbf{x}^T$. This notation does not denote whether or not the vector is a row or column, or how many elements it contains. Such information must therefore be given explicitly. Throughout the text, unless stated otherwise, we assume all vectors are column vectors by default.

Vectors of length two (or three) are easy to intuit since they live in two- (or three-) dimensional spaces that are familiar to our human senses. For example, the two-dimensional vector

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{C.5}$$

can be drawn in a two-dimensional plane as an *arrow* stemming from the origin and ending at the point whose horizontal and vertical coordinates are 1 and 2, respectively, as illustrated in the left panel of Figure C.1. However, as shown in the right panel of the figure, $\mathbf{x}$ can alternatively be drawn (and thought of) as a single *point*, i.e., the arrow's endpoint. When plotting a low-dimensional machine learning dataset (that is simply a collection of vectors) we often employ the latter visual style.

C.2.2 Vector addition

We add (and subtract) two vectors element-wise, noting that, in order to be able to do so, the two vectors must have the same number of elements (or dimension), and both must be row or column vectors. For example, vectors

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{C.6}$$

are added element-wise to form

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{bmatrix}. \tag{C.7}$$

Subtraction of $\mathbf{y}$ from $\mathbf{x}$ is defined similarly as

$$\mathbf{x} - \mathbf{y} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \vdots \\ x_N - y_N \end{bmatrix}. \tag{C.8}$$

Thinking of vectors as arrows stemming from the origin, the addition of two vectors is equal to the vector representing the far corner of the parallelogram formed by the two vectors in the sum. This is typically called the *parallelogram law*, and is illustrated in Figure C.2 for two input vectors colored black, with their sum shown in red. The dashed lines here are merely visual guides helping to outline the parallelogram underlying the sum.

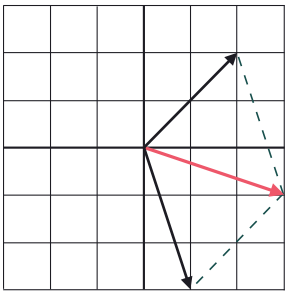

Figure C.2 The parallelogram law illustrated.

C.2.3 Vector multiplication

Unlike addition, there is more than one way to define vector multiplication. In what follows we review multiplication of a vector by a scalar, element-wise multiplication of two vectors, as well as inner- and outer-product of two vectors.

Multiplication of a vector by a scalar

We can multiply any vector $\mathbf{x}$ by a scalar c, by treating the multiplication element-wise as

$$c\,\mathbf{x} = \begin{bmatrix} c\,x_1 \\ c\,x_2 \\ \vdots \\ c\,x_N \end{bmatrix}. \tag{C.9}$$

Element-wise product of two vectors

The element-wise product, sometimes called the Hadamard product, works precisely how it sounds: we multiply two vectors element by element. Note that, just like addition, we need both vectors to have the same dimension in order to make this work. Notationally, the element-wise product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is written as

$$\mathbf{x} \circ \mathbf{y} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_N y_N \end{bmatrix}. \tag{C.10}$$

Inner-product of two vectors

The inner-product (also referred to as the dot product) is another way to multiply two vectors of the same dimension. Unlike the element-wise product, the inner-product of two vectors produces a *scalar* output. To take the inner-product of two vectors we first multiply them together element-wise, and then simply add up the elements in the resulting vector. The inner-product of vectors $\mathbf{x}$ and $\mathbf{y}$ is written as

$$\mathbf{x}^T\mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_N y_N = \sum_{n=1}^{N} x_n y_n. \tag{C.11}$$

Vector length or magnitude

The well-known Pythagorean theorem provides a useful way to measure the length of a vector in two dimensions. Using the Pythagorean theorem we can treat the general two-dimensional vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{C.12}$$

as the hypotenuse of a right triangle, and write

$$\text{length of } \mathbf{x} = \sqrt{x_1^2 + x_2^2}. \tag{C.13}$$

Notice, we can also express the length of $\mathbf{x}$ in terms of the inner-product of $\mathbf{x}$ with itself, as

$$\text{length of } \mathbf{x} = \sqrt{\mathbf{x}^T\mathbf{x}}, \tag{C.14}$$

and this generalizes to vectors of any dimension. Using the notation $\|\mathbf{x}\|_2$ to denote the length of an N-dimensional vector $\mathbf{x}$, we have

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{n=1}^{N} x_n^2}. \tag{C.15}$$

Geometric interpretation of the inner-product

The inner-product of two vectors $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{x}^T\mathbf{y} = \sum_{n=1}^{N} x_n y_n \tag{C.16}$$

can be expressed in terms of the lengths of $\mathbf{x}$ and $\mathbf{y}$, via the so-called *inner-product rule*, as

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2 \cos(\theta) \tag{C.17}$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. This rule is perhaps best intuited after a slight rearrangement of its terms, as

$$\left(\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right)^T \left(\frac{\mathbf{y}}{\|\mathbf{y}\|_2}\right) = \cos(\theta) \tag{C.18}$$

where vectors $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ and $\frac{\mathbf{y}}{\|\mathbf{y}\|_2}$ still point in the same direction as $\mathbf{x}$ and $\mathbf{y}$, respectively, but both have been normalized to have unit length, since

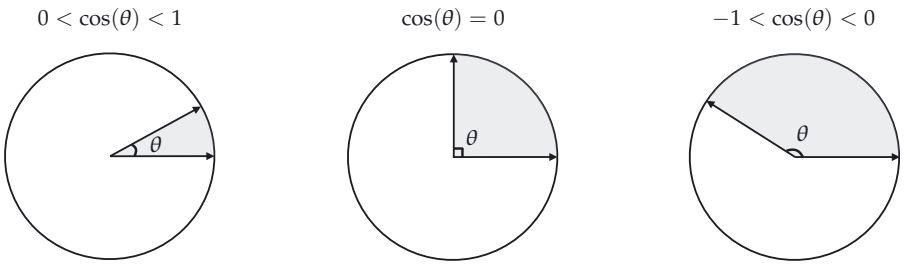

Figure C.3 The inner-product of two unit-length vectors is equal to the cosine of the angle θ created between them.

$$\|\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\|_2 = \|\frac{\mathbf{y}}{\|\mathbf{y}\|_2}\|_2 = 1. \tag{C.19}$$

Note that because cosine always lies between -1 and 1, so too does the inner product of any two unit-length vectors. When they point in the exact same direction, $\theta = 0$, and their inner-product is maximal (i.e., 1). As the two vector start to point away from each other, θ increases, and the inner-product starts to shrink. When the two vectors are *perpendicular* to each other, their inner-product is equal to zero. The inner-product reaches its minimal value (i.e., -1) when $\theta = \pi$, and the two vectors point in completely opposite directions (see Figure C.3).

Outer-product of two vectors

The outer-product is another way to define multiplication between two vectors. With two column vectors (of not necessarily the same dimension)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \tag{C.20}$$

their outer-product is written as $\mathbf{x}\mathbf{y}^T$, and defined as

$$\mathbf{x}\mathbf{y}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_M \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_M \\ x_2y_1 & x_2y_2 & \cdots & x_2y_M \\ \vdots & \vdots & \ddots & \vdots \\ x_Ny_1 & x_Ny_2 & \cdots & x_Ny_M \end{bmatrix}. \tag{C.21}$$

The result is an $N \times M$ matrix, which can be thought of as a collection of M column vectors of length N stacked side by side (or likewise, as a collection of N

row vectors of length M stacked on top of each other). We will return to matrices and discuss them further in the next section.

C.2.4 Linear combination of vectors

A linear combination is an operation that generalizes simple addition of two vectors by combining addition and scalar multiplication. Given two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ of the same dimension, their linear combination is formed by multiplying each with a scalar first and then adding up the result, as

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 \tag{C.22}$$

where α_1 and α_2 are real numbers. Notice that for a given pair of values (α_1, α_2) the linear combination is a vector itself with the same dimension as $\mathbf{x}_1$ and $\mathbf{x}_2$. In Figure C.4 we show the linear combination of vectors

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{x}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{C.23}$$

for three distinct settings of (α_1, α_2). The set of all such vectors created by taking a linear combination of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is referred to as the *span* of $\mathbf{x}_1$ and $\mathbf{x}_2$, and written as

$$\text{span of } \mathbf{x}_1 \text{ and } \mathbf{x}_2 = \left\{\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 \,|\, (\alpha_1, \alpha_2) \in \mathbb{R}^2\right\}. \tag{C.24}$$

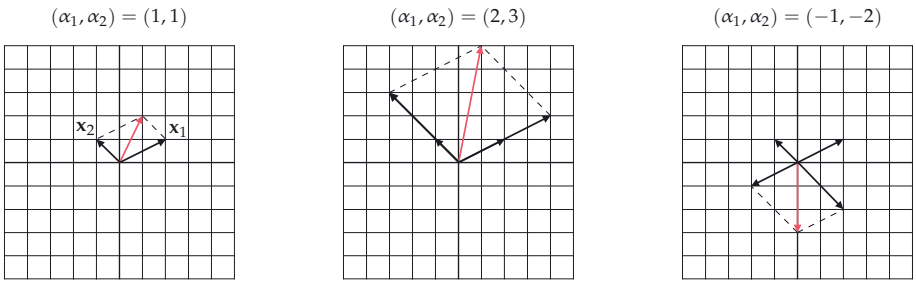

Figure C.4 The linear combination (in red) of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ defined in Equation (C.24) for three different settings of (α_1, α_2). As you can see by changing the values of α_1 and α_2 in $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$ we get a new vector each time. The set of all such vectors is referred to as the span of $\mathbf{x}_1$ and $\mathbf{x}_2$, which in this case is the entire two-dimensional plane.

For the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in Equation (C.23) the span is the entire two-dimensional plane. But this is not necessarily always the case for any pair of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. Take

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{x}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \tag{C.25}$$

for instance. Because these two vectors point at the same direction (one is a scalar multiple of the other), any linear combination of the two will have the same direction. In this case the span of $\mathbf{x}_1$ and $\mathbf{x}_2$ is no longer the entire two-dimensional plane, but a one-dimensional line that can be traced out using scalar multiples of any of the two vectors. In other words, given either one of $\mathbf{x}_1$ or $\mathbf{x}_2$ the other one becomes redundant (in terms of finding their span). In linear algebra terms such vectors are called *linearly dependent*.

The notion of linear combination of vectors can be extended in general to a set of k vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ (all of the same dimension), taking the form

$$\sum_{i=1}^{k} \alpha_i \mathbf{x}_i = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_k \mathbf{x}_k. \tag{C.26}$$

If these vectors span a k-dimensional space they are called *linearly independent*. Otherwise, there is at least one vector in the set that can be written as a linear combination of the rest.

C.3 Matrices and Matrix Operations

In this section we review the concept of a matrix as well as the basic operations one can perform on a single matrix or pairs of matrices. These completely mirror those of the vector in the previous section, including the transpose operation, addition/subtraction, and several multiplication operations. Because of the close similarity to vectors this section is much more terse than the previous section.

C.3.1 The matrix

If we take a set of N row vectors, each of dimension M

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} x_{21} & x_{22} & \cdots & x_{2M} \end{bmatrix}$$

$$\vdots$$

$$\mathbf{x}_N = \begin{bmatrix} x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix}$$

and stack them one by one on top of each other we form an object called a *matrix*

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix} \tag{C.27}$$

of dimension $N \times M$, where the first number N is the number of rows in the matrix, with the second number M denoting the number of columns. The notation we use to describe a matrix in the text is a bold uppercase letter, e.g., $\mathbf{X}$. Like the vector notation nothing about the dimensions of the matrix is detailed by its notation and they must be explicitly stated.

The transpose operation we originally saw for vectors is defined by extension for matrices. When performed on a matrix, the transpose operation flips the entire matrix around: every column is turned into a row, and then these rows are stacked one on top of the other, forming an $M \times N$ matrix

$$\mathbf{X}^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \cdots & x_{NM} \end{bmatrix}. \tag{C.28}$$

C.3.2 Matrix addition

As with vectors, addition (and subtraction) is performed element-wise on matrices of the same dimensions. For example, with two $N \times M$ matrices

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix} \tag{C.29}$$

their sum is defined as

$$\mathbf{X} + \mathbf{Y} = \begin{bmatrix} x_{11} + y_{11} & x_{12} + y_{12} & \cdots & x_{1M} + y_{1M} \\ x_{21} + y_{21} & x_{22} + y_{22} & \cdots & x_{2M} + y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} + y_{N1} & x_{N2} + y_{N2} & \cdots & x_{NM} + y_{NM} \end{bmatrix}. \tag{C.30}$$

C.3.3 Matrix multiplication

As with vectors, there are a variety of ways to define matrix multiplication which we review here.

Multiplication of a matrix by a scalar

We can multiply any matrix $\mathbf{X}$ by a scalar c, and this operation is performed element by element as

$$c\,\mathbf{X} = \begin{bmatrix} c\,x_{11} & c\,x_{12} & \cdots & c\,x_{1M} \\ c\,x_{21} & c\,x_{22} & \cdots & c\,x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ c\,x_{N1} & c\,x_{N2} & \cdots & c\,x_{NM} \end{bmatrix}. \tag{C.31}$$

Multiplication of a matrix by a vector

Generally speaking, there are two ways to multiply an $N \times M$ matrix $\mathbf{X}$ by a vector $\mathbf{y}$. The first, referred to as *left multiplication*, involves multiplication by an N-dimensional row vector $\mathbf{y}$. This operation, written as $\mathbf{yX}$, results in a row vector of dimension M whose mth element is the inner-product of $\mathbf{y}$ with the mth column of $\mathbf{X}$

$$\mathbf{yX} = \begin{bmatrix} \sum_{n=1}^{N} y_n\,x_{n1} & \sum_{n=1}^{N} y_n\,x_{n2} & \cdots & \sum_{n=1}^{N} y_n\,x_{nM} \end{bmatrix}. \tag{C.32}$$

Likewise, *right multiplication* is defined by multiplying $\mathbf{X}$ on the right by an M-dimensional column vector $\mathbf{y}$. Written as $\mathbf{Xy}$, right multiplication results in an N-dimensional column vector whose nth element is the inner-product of $\mathbf{y}$ with the nth row of $\mathbf{X}$

$$\mathbf{Xy} = \begin{bmatrix} \sum_{m=1}^{M} y_m\,x_{1m} \\ \sum_{m=1}^{M} y_m\,x_{2m} \\ \vdots \\ \sum_{m=1}^{M} y_m\,x_{Nm} \end{bmatrix}. \tag{C.33}$$

Element-wise multiplication of two matrices

As with vectors, we can define element-wise multiplication on two matrices of the same dimensions. The element-wise product of two $N \times M$ matrices $\mathbf{X}$ and $\mathbf{Y}$ is written as

$$\mathbf{X} \circ \mathbf{Y} = \begin{bmatrix} x_{11}\,y_{11} & x_{12}\,y_{12} & \cdots & x_{1M}\,y_{1M} \\ x_{21}\,y_{21} & x_{22}\,y_{22} & \cdots & x_{2M}\,y_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1}\,y_{N1} & x_{N2}\,y_{N2} & \cdots & x_{NM}\,y_{NM} \end{bmatrix}. \tag{C.34}$$

General multiplication of two matrices

The general product (or simply product) of two matrices $\mathbf{X}$ and $\mathbf{Y}$ can be defined based on the vector outer-product operation, provided that the number of columns in $\mathbf{X}$ matches the number of rows in $\mathbf{Y}$. That is, we must have $\mathbf{X}$ and $\mathbf{Y}$ of dimensions $N \times M$ and $M \times P$ respectively, for the matrix product to be defined as

$$\mathbf{XY} = \sum_{m=1}^{M} \mathbf{x}_m \mathbf{y}_m^T \tag{C.35}$$

where $\mathbf{x}_m$ is the mth column of $\mathbf{X}$, and $\mathbf{y}_m^T$ is the transpose of the mth column of $\mathbf{Y}^T$ (or equivalently, the mth row of $\mathbf{Y}$). Note that each summand in Equation (C.35) is itself a matrix of dimension $N \times P$, and so too is the final matrix $\mathbf{XY}$.

General matrix multiplication can also be defined element-wise, using vector inner-products, where the entry in the nth row and pth column of $\mathbf{XY}$ is found as the inner-product of (transpose of) the nth row in $\mathbf{X}$ and the pth column in $\mathbf{Y}$.

C.4 Eigenvalues and Eigenvectors

In this section we review general linear functions and their relationship to matrices. We particularly focus on the special case of the square matrix, for which we discuss the important topics of eigenvectors and eigenvalues.

C.4.1 Linear functions and matrix multiplication

As we discussed in the previous section, the product of an $N \times M$ matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{bmatrix} \tag{C.36}$$

by an M-dimensional column vector

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} \tag{C.37}$$

is an N-dimensional column vector written as $\mathbf{Xw}$. Treating the vector $\mathbf{w}$ as input, $\mathbf{Xw}$ defines a function g written formally as

$$g(\mathbf{w}) = \mathbf{Xw}. \tag{C.38}$$

Writing $g(\mathbf{w})$ explicitly as

$$g(\mathbf{w}) = \begin{bmatrix} x_{11}w_1 + x_{12}w_2 + \cdots + x_{1M}w_M \\ x_{21}w_1 + x_{22}w_2 + \cdots + x_{2M}w_M \\ \vdots \\ x_{N1}w_1 + x_{N2}w_2 + \cdots + x_{NM}w_M \end{bmatrix} \tag{C.39}$$

it is clear that each of its elements is a linear function in variables w_1 through w_M, and hence g itself is called a *linear* function.

C.4.2 Linear functions and square matrices

When the number of rows in a matrix $\mathbf{X}$ is identical to the number of columns in it, i.e., $N = M$, the matrix is called a *square* matrix. When $N = M = 2$ we can visually examine the effect of a linear function $g(\mathbf{w}) = \mathbf{X}\mathbf{w}$ by viewing the way two-dimensional points $\mathbf{w}$ are transformed via g.

In Figure C.5 we provide just such a visualization using the 2×2 matrix $\mathbf{X}$ whose entries were set at random as

$$\mathbf{X} = \begin{bmatrix} 0.726 & -1.059 \\ -0.200 & -0.947 \end{bmatrix}. \tag{C.40}$$

In the left panel of the figure we show a coarse set of grid lines. The point of this grid is to help visualize how each point constituting the grid lines (and thus the entire space itself) is transformed using the matrix $\mathbf{X}$ in Equation (C.40). For visualization purposes, a circle of radius 2 is drawn on top of the grid, and is transformed along with it. In the right panel of the figure we illustrate how the space shown in the left panel is warped by the function $g(\mathbf{w}) = \mathbf{X}\mathbf{w}$.

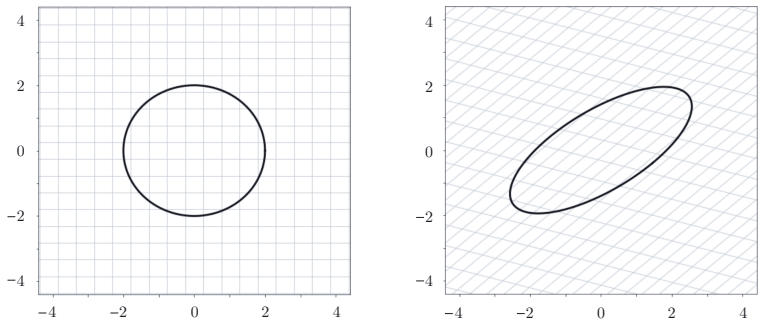

Figure C.5 The input (left panel) and output (right panel) spaces of the linear function $g(\mathbf{w}) = \mathbf{X}\mathbf{w}$ with the matrix $\mathbf{X}$ defined in Equation (C.40).

C.4.3 Eigenvalues and eigenvectors

The previous visualization in Figure C.5 is interesting in its own right, but if examined closely can also be used to provoke the notion of what are called *eigenvectors*. These are the handful of directions that, unlike most others that are warped and twisted by multiplication with the given matrix, are only *scaled* by the function. In other words, eigenvectors are those special vectors in the input space that retain their direction, after having gone through the linear transformation g. In Figure C.6 we again show the transformation provided by the random matrix $\mathbf{X}$ in Equation (C.40). This time, however, we also highlight two such eigenvectors as black arrows. Comparing the left and right panels of the figure, notice how neither direction gets twisted or warped by the transformation: they are only scaled.

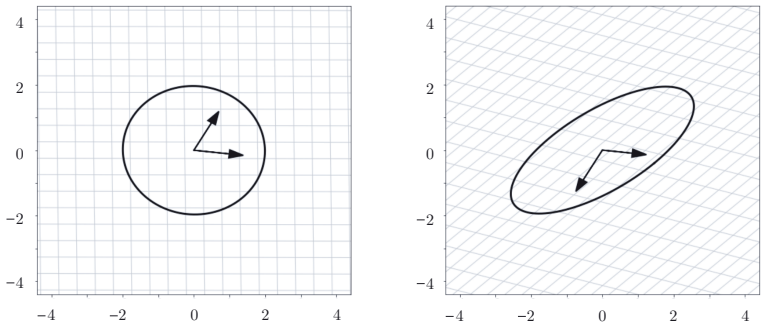

Figure C.6 A redrawing of Figure C.5 with the two eigenvectors of $\mathbf{X}$ added as two black arrows in both input and output spaces.

What we saw with linear functions based on 2×2 square matrices holds more generally for higher dimensions as well: a linear function based on an $N \times N$ matrix affects *at most* N linearly independent directions by simply scaling them. For an $N \times N$ matrix $\mathbf{X}$ each such direction $\mathbf{v} \neq \mathbf{0}_{N\times 1}$ satisfying

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{C.41}$$

is called an *eigenvector*. Here the value λ is precisely the amount by which $\mathbf{X}$ scales $\mathbf{v}$, and is called an *eigenvalue*. In general, λ can take on real or complex values.

C.4.4 The special case of the symmetric matrix

A *symmetric* matrix, that is a square matrix $\mathbf{X}$ where $\mathbf{X} = \mathbf{X}^T$, is an important special case of a square matrix that arises in a wide range of contexts (e.g., Hessian matrices, covariance matrices, etc.). One of the main advantages such matrices have over merely square ones is the following: their eigenvectors are

always perpendicular to each other, and their eigenvalues are *always* real numbers [74, 75, 76]. This fact has significant repercussions in the analysis of such matrices as we can *diagonalize* them as follows.

Stacking all of the eigenvectors of $\mathbf{X}$ column-wise into a matrix $\mathbf{V}$, and placing the corresponding eigenvalues along the diagonal of a matrix $\mathbf{D}$, we can write the Equation (C.41) simultaneously for all eigenvectors/values, as

$$\mathbf{XV} = \mathbf{VD}. \tag{C.42}$$

When the eigenvectors are all perpendicular to each other, $\mathbf{V}$ is an orthonormal matrix[1] and we have $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. Thus multiplying both sides of Equation (C.42) by $\mathbf{V}^T$ (on the right) we can express $\mathbf{X}$ completely in terms of its eigenvectors/values as

$$\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^T. \tag{C.43}$$

C.5 Vector and Matrix Norms

In this section we discuss popular vector and matrix norms that will arise frequently in our study of machine learning, particularly when discussing regularization. A norm is a kind of function that measures the length of real vectors and matrices. The notion of length is extremely useful as it enables us to define distance (or similarity) between any two vectors (or matrices) living in the same space.

C.5.1 Vector norms

The ℓ_2 norm

We begin with the most widely used vector norm in machine learning, the ℓ_2 norm, defined for an N-dimensional vector $\mathbf{x}$ as

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{n=1}^{N} x_n^2}. \tag{C.44}$$

Using the ℓ_2 norm we can measure the distance between any two points $\mathbf{x}$ and $\mathbf{y}$ via $\left\|\mathbf{x} - \mathbf{y}\right\|_2$, which is simply the length of the vector connecting $\mathbf{x}$ and $\mathbf{y}$. For example, the distance between

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 9 \\ 8 \end{bmatrix} \tag{C.45}$$

[1] Here we have assumed every eigenvector $\mathbf{v}$ that satisfies Equation (C.41) has unit length, i.e., $\|\mathbf{v}\|_2 = 1$. If not, we can always replace $\mathbf{v}$ with $\frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ and Equation (C.41) will still hold.

is calculated as $\sqrt{(1-9)^2 + (2-8)^2} = 10$, as shown pictorially (in red) in Figure C.7.

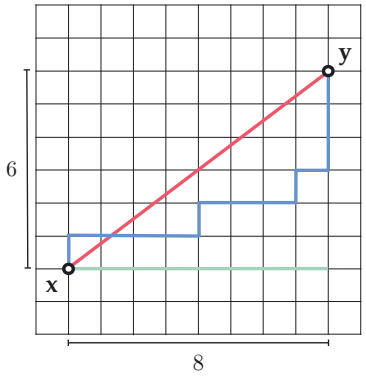

Figure C.7 The ℓ_1 (blue), ℓ_2 (red), and ℓ_∞ (green) based distances between the points $\mathbf{x}$ and $\mathbf{y}$ defined in Equation (C.45).

The ℓ_1 norm

The ℓ_1 norm of a vector $\mathbf{x}$ is another way to measure its length, defined as the sum of the absolute values of its entries

$$\|\mathbf{x}\|_1 = \sum_{n=1}^{N} |x_n| . \tag{C.46}$$

In terms of the ℓ_1 norm the distance between $\mathbf{x}$ and $\mathbf{y}$ is given by $\|\mathbf{x} - \mathbf{y}\|_1$, which provides a measurement of distance different from the ℓ_2 norm. As illustrated in Figure C.7 the distance defined by the ℓ_1 norm is the length of a path consisting of perpendicular pieces (shown in blue). Because these paths are somewhat akin to how an automobile might travel from $\mathbf{x}$ to $\mathbf{y}$ if they were two locations in a gridded city, having to traverse perpendicular city blocks one after the other, the ℓ_1 norm is sometimes referred to as the *taxicab norm*, and the distance measured via the ℓ_1 norm, the *Manhattan distance*. For $\mathbf{x}$ and $\mathbf{y}$ in Equation (C.45) the Manhattan distance is calculated as $|1-9| + |2-8| = 14$.

The ℓ_∞ norm

The ℓ_∞ norm of a vector $\mathbf{x}$ is equal to its largest entry (in terms of absolute value), defined mathematically as

$$\|\mathbf{x}\|_\infty = \max_n |x_n| . \tag{C.47}$$

For example, the distance between $\mathbf{x}$ and $\mathbf{y}$ in Equation (C.45) in terms of the ℓ_∞ norm is found as $\max(|1-9|, |2-8|) = 8$, as illustrated in Figure C.7 (in green).

C.5.2 Common properties of vector norms

The ℓ_2, ℓ_1, and ℓ_∞ norms share a number of useful properties that we detail below. Since these properties hold in general for *any* vector norm, we momentarily drop the subscript and represent the generic norm of $\mathbf{x}$ simply by $\|\mathbf{x}\|$.

1. Norms are always nonnegative, that is, $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x}$. Furthermore, the equality holds if and only if $\mathbf{x} = \mathbf{0}$, implying that the norm of any nonzero vector is always greater than zero.

2. The norm of $\alpha\mathbf{x}$, that is a scalar multiple of $\mathbf{x}$, can be written in terms of the norm of $\mathbf{x}$ as $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$. With $\alpha = -1$ for example, we have that $\|-\mathbf{x}\| = \|\mathbf{x}\|$.

3. Norms also satisfy the so-called *triangle inequality* where for any three vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ we have $\|\mathbf{x}-\mathbf{z}\| + \|\mathbf{z}-\mathbf{y}\| \geq \|\mathbf{x}-\mathbf{y}\|$. As illustrated in Figure C.8 for the ℓ_2 norm (left panel), the ℓ_1 norm (middle panel), and the ℓ_∞ norm (right panel), the triangle inequality simply states that the distance between $\mathbf{x}$ and $\mathbf{y}$ is always smaller than (or equal to) the distance between $\mathbf{x}$ and $\mathbf{z}$, and the distance between $\mathbf{z}$ and $\mathbf{y}$, combined. In other words, if one wanted to travel from a given point $\mathbf{x}$ to a given point $\mathbf{y}$, it would be always better to travel directly from $\mathbf{x}$ to $\mathbf{y}$ than to travel first to a third point z, and then to y. With the change of variables $\mathbf{u} = \mathbf{x} - \mathbf{z}$ and $\mathbf{v} = \mathbf{z} - \mathbf{y}$, the triangle inequality is sometimes written in the simpler form of $\|\mathbf{u}\| + \|\mathbf{v}\| \geq \|\mathbf{u} + \mathbf{v}\|$ for all vectors $\mathbf{u}$ and $\mathbf{v}$.

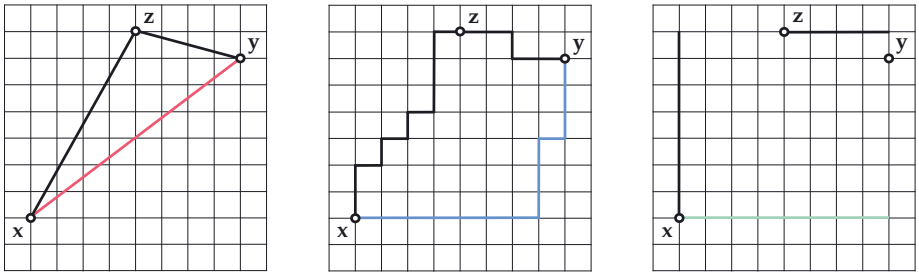

Figure C.8 The triangle inequality illustrated for the ℓ_2 norm (left panel), ℓ_1 norm (middle panel), and ℓ_∞ norm (right panel).

In addition to the general properties mentioned above and held by any norm, the ℓ_2, ℓ_1, and ℓ_∞ norms share a stronger bond that ties them together: they are all members of the ℓ_p norm family. The ℓ_p norm is generally defined as

$$\|\mathbf{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{\frac{1}{p}} \tag{C.48}$$

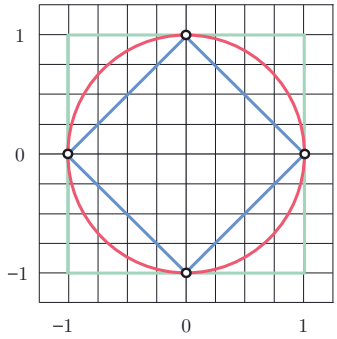

Figure C.9 Illustration of the ℓ_1 (blue), ℓ_2 (red), and ℓ_∞ (green) unit balls.

for $p \geq 1$. One can easily verify that with $p = 1$, $p = 2$, and as $p \longrightarrow \infty$, the ℓ_p norm reduces to the ℓ_1, ℓ_2, and ℓ_∞ norm, respectively.

The ℓ_p norm balls

A norm ball is the set of all vectors $\mathbf{x}$ with same norm value, that is, all $\mathbf{x}$ such that $\|\mathbf{x}\| = c$ for some constant $c > 0$. When $c = 1$, this set is called the unit norm ball, or simply the unit ball. The ℓ_1, ℓ_2, and ℓ_∞ unit balls are plotted in Figure C.9.

The ℓ_0 norm

The ℓ_0 norm is yet another way of defining a vector's length as

$$\|\mathbf{x}\|_0 = \text{number of nonzero entries of } \mathbf{x}. \tag{C.49}$$

Calling the ℓ_0 norm a *norm* is technically a misnomer as it does not hold the scalability property held by all vector norms. That is, $\|\alpha\mathbf{x}\|_0$ is generally not equal to $|\alpha|\,\|\mathbf{x}\|_0$. Nevertheless, the ℓ_0 norm arises frequently when modeling vectors with a large number of zeros (also called *sparse* vectors).

C.5.3 Matrix norms

The Frobenius norm

Recall that the ℓ_2 norm of a vector is defined as the square root of the sum of the squares of its elements. The Frobenius norm is the intuitive extension of the ℓ_2 norm for vectors to matrices, defined similarly as the square root of the sum of the squares of all the elements in the matrix, and written for an $N \times M$ matrix $\mathbf{X}$ as

$$\|\mathbf{X}\|_F = \sqrt{\sum_{n=1}^{N}\sum_{m=1}^{M} x_{nm}^2}. \tag{C.50}$$

For example, the Frobenius norm of the matrix $\mathbf{X} = \begin{bmatrix} -1 & 2 \\ 0 & 5 \end{bmatrix}$ is calculated as $\sqrt{(-1)^2 + 2^2 + 0^2 + 5^2} = \sqrt{30}$.
The connection between the ℓ_2 norm and the Frobenius norm goes further: collecting all singular values of $\mathbf{X}$ in the vector $\mathbf{s}$ we have

$$\|\mathbf{X}\|_F = \|\mathbf{s}\|_2 . \tag{C.51}$$

The spectral and nuclear norms

The observation that the ℓ_2 norm of the vector of singular values of a matrix is identical to its Frobenius norm motivates the use of other ℓ_p norms on the vector $\mathbf{s}$. In particular, the ℓ_1 norm of $\mathbf{s}$ defines the nuclear norm of $\mathbf{X}$ denoted by $\|\mathbf{X}\|_*$

$$\|\mathbf{X}\|_* = \|\mathbf{s}\|_1 \tag{C.52}$$

and the ℓ_∞ norm of $\mathbf{s}$ defines the spectral norm of $\mathbf{X}$ denoted by $\|\mathbf{X}\|_2$

$$\|\mathbf{X}\|_2 = \|\mathbf{s}\|_\infty . \tag{C.53}$$

Because the singular values of real matrices are always nonnegative, the spectral norm and the nuclear norm of such a matrix are simply its largest and the sum of all its singular values, respectively.