Statistical NLP: course notes

Çağrı Çöltekin — SfS / University of Tübingen

2020-04-24

These notes are prepared for the class *Statistical Natural Language Processing* taught in Seminar für Sprachwissenschaft, University of Tübingen.

1 Mathematical preliminaries

Being an interdisciplinary field, it is often difficult to assume that all students of computational linguistics possess a (fresh) knowledge of some of the mathematical topics and notation. This chapter provides a highly coarse overview of some topics in linear algebra and calculus. The aim of this chapter is to provide a refresher or an listing of concepts from basic math that is required in this course. The discussion here is necessarily incomplete and informal. The interested reader should follow the references provided for in-depth treatments of these subjects.

Section 1.1 introduces some topics from linear algebra. We will mainly introduce vectors, matrices and operations on vectors and matrices. These topics and notation will be important particularly in understanding the machine learning methods covered in the class.

Section 1.2, briefly revisits derivatives and integrals. Derivatives are used for finding maxima or minima of functions, which is the basis for many of the machine learning methods. The integrals will also come back in our discussion of some of the machine learning methods, and in discussion of probabilistic learning and inference.

1.1 Linear algebra

In many NLP methods, we make heavy use of *vectors* and *matrices*, which are objects studied in *linear algebra*. Vectors are used for representing *features* in many machine learning methods. Operations on vectors and matrices also have important applications. In this section we will review some of the properties of vectors and matrices, and the operations defined on them. If you had a linear algebra course, of if you know, for example, matrix multiplication, or dot product of vectors, you can safely skip this section.

1.1.1 Vectors

A vector is a mathematical object with a magnitude and a direction. Graphically, we can represent or visualize a (two-dimensional) vector as in Figure 1.1. The 'picture' is useful for getting a better intuition about the objects and operations under study. However, we can only visualize vectors in two and (with some effort) three dimensions. Nevertheless, most of these intuitions generalize neatly to higher dimensional spaces.[1]

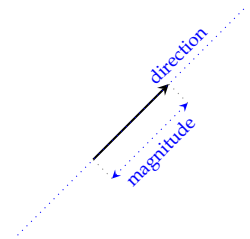

Figure 1.1: A graphical representation of a vector.

[1] Not without exceptions, however. Unexpected or unintuitive behavior of mathematical objects and operations in higher dimensional spaces are often noted under the term *curse of dimensionality*.

More commonly, we represent vectors by an ordered list of number, such as $(1, 0, 1)$. A vector defined with n real numbers is said to be in the vector space $\mathbb{R}^n$. We often write a vector of n real numbers (vectors in $\mathbb{R}^n$) as $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$. Note that the $\boldsymbol{v}$ that stands for the vector is typeset in boldface font. It can alternatively be marked with a arrow over it, like $\vec{v}$. Other notations for vectors of n numbers include $\boldsymbol{v} = \langle v_1, v_2, \ldots, v_n \rangle$ or $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$.

Geometrically, we represent vectors as arrows as in Figure 1.2. The individual numbers on the notation represent their projection to the respective axis. In the example on the right, for example, the green and blue vectors have the same magnitude, but their directions are opposite of each other. If we take the projections of the vector to the x and y axes, they correspond to real first and the second number in our notation respectively.

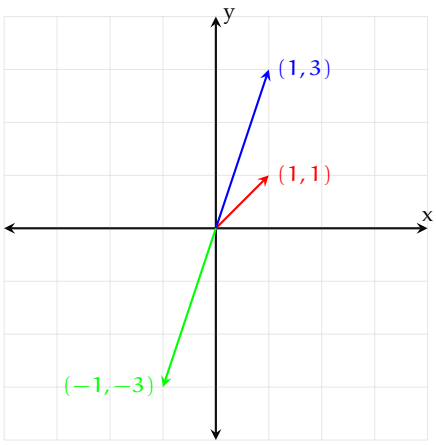

Figure 1.2: Example vectors in 2-dimensional Euclidean space.

Many operations on (real) numbers have analogues forms for vectors, and they are used in machine learning and natural language processing, as well as many other branches of science and engineering.

VECTOR NORMS are a generalization of the magnitude of a vector. A *norm* assigns a non-negative *length* or *size* to a vector. Norms are related to distance metrics which by themselves are useful in comparing objects represented as vectors.[2] The most familiar norm is the Euclidean norm, which is also known as L2 (or L_2) norm. L2 norm of a vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ is

$$\|\boldsymbol{v}\|_2 = \sqrt{v_1^2 + \ldots + v_n^2} \,.$$

[2] The norm of a vector is the distance from its tail to its tip, or the distance between two objects is the norm of the difference between their vector representations.

The subscript 2 in $\|v\|_2$ indicates that the norm is L2 norm. The L2 norm is often taken to be the default. If the subscript is omitted then we mean the L2 norm. Another interesting norm for our purposes is the L1 norm, which is related to the so-called taxi-cab, city-block or Manhattan distance. It is defined as

$$\|\boldsymbol{v}\|_1 = |v_1| + \ldots + |v_n| \,.$$

Figure 1.3 visualizes the L1 and L2 norms in two-dimensional Euclidean space. For the example vector in Figure 1.3, we have

$$\|(3, 3)\|_2 = \sqrt{3^2 + 3^2} = \sqrt{18} \approx 4.24$$

$$\|(3, 3)\|_1 = |3| + |3| = 6 \,.$$

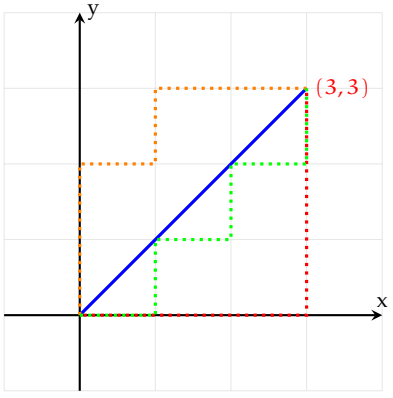

Figure 1.3: Visualizations of L2(solid blue) and example L1 (dotted green, orange and red) norms vector $(3, 3)$.

Like any other vector operation or property, vector norms can be generalized to vectors of any dimension.

The concept of vector norm can also be generalized to any positive integer p the L_p norm for an n-dimensional vector is defined as

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}$$

In this course, we will only work with L1 and L2 norms defined above. You may occasionally see L_0 or L_∞ norms used in some related literature.[3]

[3] with some simplification, L_0 norm of vector is number of non-zero entry of the vector, and L_∞ norm is the largest absolute value among the entries of the vector.

SCALAR MULTIPLICATION is the operation of multiplying a vector with a scalar (for our purposes a scalar is a real number). Given a vector $\boldsymbol{v} = (v_1, \ldots, v_n)$, its multiplication with scalar a is defined as

$$a\boldsymbol{v} = (av_1, \ldots, av_2)$$

Multiplying a vector with a positive scalar, changes its magnitude('scales' it) but does not change its direction. Multiplying a vector with a negative scalar reverses the direction of the original vector.

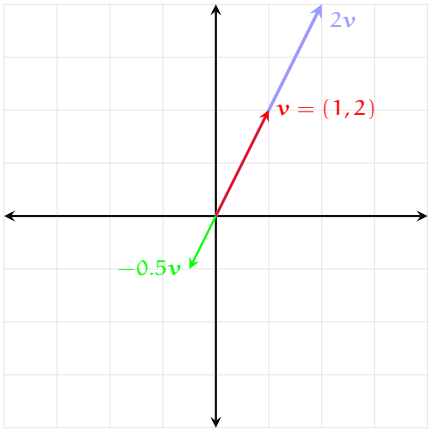

Figure 1.4: Scalar multiplication.

VECTOR ADDITION AND SUBTRACTION are defined on two vectors with the same dimensionality. For n-dimensional vectors $\boldsymbol{v} = (v_1, \ldots, v_n)$ and $\boldsymbol{w} = (w_1, \ldots, v_n)$,

$$\boldsymbol{v} + \boldsymbol{w} = (v_1 + w_1, \ldots, v_n + w_n)$$

The subtraction is simply addition where the second vector is multiplied by -1.

$$\boldsymbol{v} - \boldsymbol{w} = \boldsymbol{v} + (-\boldsymbol{w}) = (v_1 - w_1, \ldots, v_n - w_n)$$

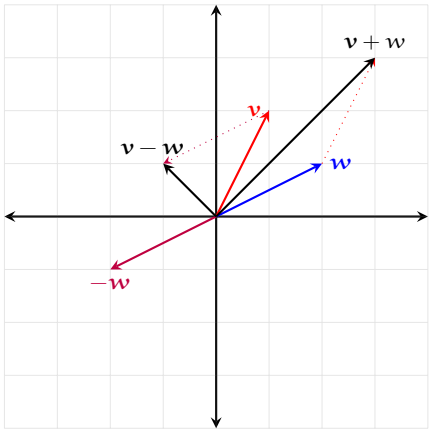

Figure 1.5: Vector addition and subtraction.

DOT PRODUCT is a very important quantity that will come up regularly in this course. Dot product of two vectors, $\boldsymbol{v} = (v_1, \ldots, v_n)$ and $\boldsymbol{w} = (w_1, \ldots, v_w)$, is a scalar defined as:

$$\boldsymbol{v} \cdot \boldsymbol{w} = v_1 \times w_1 + \ldots + v_n \times w_n$$

It should be emphasized that dot product yields a scalar (real number), not a vector. There are other vector product operations: *outer product* that we will discuss below, and *cross product* defined for vectors in $\mathbb{R}^3$. Hence, without 'dot' the notation $\boldsymbol{vw}$ is ambiguous. However, it is common to treat k-dimensional vectors as $k \times 1$ matrices for which multiplication is not ambiguous (we discuss this notation on page 8 below).

There is an alternative way to define the dot product, which also leads to a nice geometric interpretation. We can calculate the dot product as

$$\boldsymbol{v} \cdot \boldsymbol{w} = \|\boldsymbol{v}\| \|\boldsymbol{w}\| \cos \alpha \tag{1.1}$$

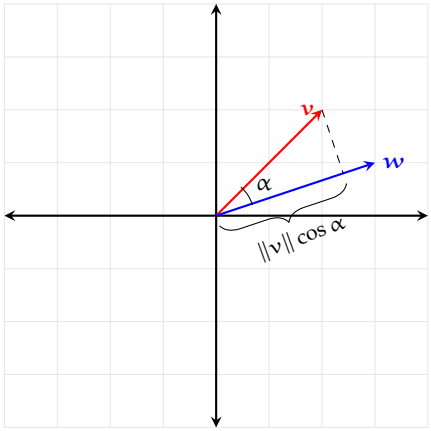

Figure 1.6: Dot product of two vectors.

where α is the angle between the two vectors (see Figure 1.6). This also allows us to interpret the dot product geometrically. The dot product of two vectors is proportional to each vector's magnitude, and also to the cosine of the angle between them. Since the cosine of the angle will be larger for smaller angles, the dot product will be larger for vectors that point to similar directions (keeping the magnitudes constant). The dot product of two orthogonal vectors (vectors with a 90° angle between them) is 0. If the angle is larger than 90°, the dot product is negative. Remember that like the other operations we discuss here, the dot product and its interpretations generalizes to higher dimensional vectors.

COSINE SIMILARITY is a similarity measure related to dot product, which we will often use for measuring similarities between objects of interest, e.g., documents. We can rewrite Equation 1.1, above to calculate the cosine of the angle between two vectors as,

$$\cos \alpha = \frac{\boldsymbol{v}\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|} .$$

The range of the cosine similarity is between -1 and 1. The cosine similarity for vectors that point to the same direction is 1 (regardless of their magnitude) and the vectors that point exact opposite directions have a cosine similarity of -1. Note that by dividing the vectors to their Euclidean (L2) norms, we are scaling them to unit vectors while keeping their directions the same. As a result, cosine similarity ignores the magnitudes of the vectors. This is generally more appropriate when the ratios between the entries of a vector matters more than the magnitude of the vector. For example, if we represent documents with vectors of word counts (we will return to this representation later), the cosine similarity would less sensitive to document length in comparison to the dot product.

1.1.2 Matrices

Matrices are the second type of mathematical objects we often encounter in various NLP methods. A matrix is simply a two-dimensional array of numbers, which is noted as a rectangular placement of scalars. A matrix of n rows and m columns is an $n \times m$ matrix. A real-valued $n \times m$ matrix is said to be in $\mathbb{R}^{n \times m}$. We can think about a matrix as a collection of column or row vectors. We denote matrices with boldface capital letters, like $\mathbf{A}$. While referring to a matrix' elements, we subscript the element first with its row and then its column.

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,m} \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots & a_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,m} \end{bmatrix}$$

We will briefly revisit some of the operations on matrices in this section.

TRANSPOSE OF A MATRIX simply replaces its rows by columns. Transpose of a matrix $\mathbf{A}$ is denoted with $\mathbf{A}^T$.

$$\text{If } \mathbf{A} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}, \mathbf{A}^T = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix}.$$

MULTIPLICATION BY A SCALAR is also defined for matrices. To multiply a matrix with a scalar, each element of the matrix is multiplied by the scalar. For example,

$$2\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 2\times 2 & 2\times 1 \\ 2\times 1 & 2\times 4 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 8 \end{bmatrix}$$

MATRIX ADDITION AND SUBTRACTION require two matrices of same dimensions. To obtain sum (or difference) of two matrices, each element of the second matrix is added to (or subtracted from) the corresponding element of the first matrix. For example:

$$\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2+0 & 1+1 \\ 1+1 & 4+0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

MATRIX MULTIPLICATION is a slightly complicated operation. The matrix multiplication $\mathbf{A}\times\mathbf{B}$ is defined only if $\mathbf{A}$ has the same number of columns as the number of rows in $\mathbf{B}$. Multiplying a $n\times k$ matrix with a $k\times m$ matrix results in a $n\times m$ matrix. Note that both $\mathbf{A}\times\mathbf{B}$ and $\mathbf{B}\times\mathbf{A}$ is defined only for square matrices (of same dimensions).

For an $n\times k$ matrix $\mathbf{A}$ and a $k\times m$ $\mathbf{B}$, if $\mathbf{A}\times\mathbf{B} = \mathbf{C}$, $c_{i,j}$, the element of the resulting matrix $\mathbf{C}$ on row i and column j, is calculated as:

$$c_{ij} = \sum_{\ell=0}^{k} a_{i\ell} b_{\ell j}$$

Figure 1.7 demonstrates the matrix multiplication.

$$\begin{bmatrix} \boxed{a_{11} & a_{12} & \dots & a_{1k}} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{bmatrix} \times \begin{bmatrix} b_{11} & \boxed{b_{12}} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \dots & b_{km} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nm} \end{bmatrix}$$

$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + \dots a_{1k}b_{k2}$$

Figure 1.7: Matrix multiplication. The calculation of the resulting matrix c_{12} is highlighted.

Note that the element c_{ij} is the dot product of i^{th} row vector of $\mathbf{A}$ and j^{th} column vector of $\mathbf{B}$. Hence, we can view dot-product as matrix multiplication of a row vector (on the left) and column vector (on the right). Dot product of two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ is often noted as

$\boldsymbol{v}\boldsymbol{w}^T$.[4] Technically, result of a matrix multiplication of a $1 \times k$ vector with a $k \times 1$ vector is a 1×1 matrix, not a scalar. However, this notation is prevalent in machine learning and NLP literature, and, in general, it is common not to distinguish scalars from vectors and matrices with single items.

For example, $\boldsymbol{w} = (2, 2)$ and $\boldsymbol{v} = (2, -2)$,

$$\boldsymbol{v}^T\boldsymbol{w} = \begin{bmatrix} 2 & 2 \end{bmatrix} \times \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 2 \times 2 + 2 \times -2 = 4 - 4 = 0$$

[4] It is a common convention to assume that vectors are column vectors unless stated otherwise.

What does the result of dot product (0) say about the vectors?

OUTER PRODUCT of two vectors with the same dimensionality, can also be defined as matrix multiplication. This time we put the column vector to the left and the row vector to the right. So, in the notation used above, outer product of two matrices $\boldsymbol{v}$ and $\boldsymbol{w}$ is $\boldsymbol{v}^T\boldsymbol{w}$. Note that result of outer product of two k-dimensional vectors is a $k \times k$ matrix, not a scalar. The following is an example of outer product of two 3-dimensional vectors.

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \times \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 & 5 & 4 \\ 12 & 10 & 8 \\ 18 & 15 & 12 \end{bmatrix}$$

AN IDENTITY MATRIX is a square matrix in which all the elements of the main diagonal are ones, and all other elements are zeros. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$. When there is no ambiguity, we omit the subscript, and simply write $\mathbf{I}$. Multiplying a matrix with a compatible identity matrix does not change the original matrix. For $n \times m$ matrix $\mathbf{A}$,

$$\mathbf{I}_n\mathbf{A} = \mathbf{A}\mathbf{I}_m = \mathbf{A}$$

$$\mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 1.8: The 4×4 identity matrix.

MULTIPLYING A VECTOR WITH A MATRIX (linearly) transforms it to another (possibly a different dimensional) vector. These linear transformations have many applications, and they will also be useful for understanding some of the machine learning concepts. Here are a few interesting transformations in 2-dimensional space:

- Identity transformation has no effect on the vector to be transformed

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- Stretch along x-axis

$$\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

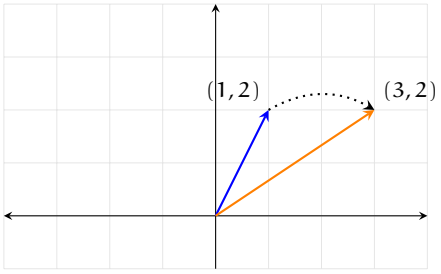

Figure 1.9: Stretch (three times) along x.

- Multiplying a vector with

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

rotates it with θ degrees. For example, for 90-degrees rotation,

$$\begin{bmatrix} \cos 90 & -\sin 90 \\ \sin 90 & \cos 90 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

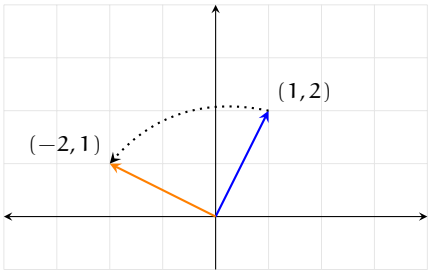

Figure 1.10: Rotate 90 degrees.

These linear operations can be combined (composed) for more complex transformations.

SOLVING A SET OF LINEAR LINEAR EQUATIONS has been one of the main applications of linear algebra. We will not discuss how to solve a linear equations here (since we rarely do this by hand), but we will demonstrate how a set of linear equations are represented using matrices and vectors. We will encounter this in various forms during the course.

The set of equations,

$$\begin{aligned} 2x_1 + x_2 &= 6 \\ x_1 + 4x_2 &= 17 \end{aligned}$$

can be written as:

$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}}_{W} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 6 \\ 17 \end{bmatrix}}_{b} \tag{1.2}$$

which allows finding a solution (if one exists) using a method called *Gaussian elimination*.

For our purposes, important part is to realize that this amounts to the matrix/vector operations operations we have been reviewing so far.

INVERSE OF A MATRIX is defined for square matrices. Inverse of matrix W is denoted by W^{-1}. Multiplying a matrix with its inverse yields the identity matrix.

$$WW^{-1} = W^{-1}W = I$$

Now that we have defined the inverse of a matrix, we can solve a set of linear equations represented with matrices and vectors as in Equation 1.2 easily:

$$\begin{aligned} Wx &= b \\ W^{-1}Wx &= W^{-1}b \\ Ix &= W^{-1}b \\ x &= W^{-1}b \end{aligned}$$

Calculating inverse of a matrix involves using a set of operations, called *elementary row operations*, on the augmented matrix that contains the original matrix and the identity matrix side by side. We will

not cover this here, as we rarely do this by hand. Interested readers should check any of the linear algebra sources listed at the end of the chapter.

THE DETERMINANT OF A MATRIX is a scalar value with some interesting properties and applications, including

- A matrix is invertible if it has a non-zero determinant
- A system of linear equations has a unique solution if the coefficient matrix has a non-zero determinant

We denote the determinant of a matrix with vertical bars around it, determinant of $\mathbf{A}$ is denoted by $|\mathbf{A}|$. Geometric interpretation of determinant is the (signed) changed in the volume of a unit (hyper)cube caused by the transformation defined by the matrix.

The determinant of a 2×2 matrix can be calculated by the formula:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

The above formula generalizes to larger matrices through a recursive definition.

EIGENVALUES AND EIGENVECTORS of a matrix also have important applications. An *eigenvector*, $\boldsymbol{v}$ and corresponding *eigenvalue*, λ, of a matrix $\mathbf{A}$ is defined such that

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v} \, .$$

In (other) words, multiplying a matrix with one of its eigen vectors only changes the magnitude of the vector and does not change its direction.

Eigenvalues an eigenvectors have many applications from communication theory to quantum mechanics. A better known example (and close to home) is Google's PageRank algorithm. We will return to them while discussing PCA and SVD.

1.2 *Derivatives and integrals*

Differentiation and integration are two fundamental concepts in calculus. The reason we review some of the basic calculus here has to do with the fact that these operations are often used in probability theory and machine learning. In many machine learning problems, learning is achieved through minimizing the error or maximizing an objective (e.g., likelihood). A particularly important use of derivatives in machine learning is to find maxima or minima of error or objective function. This section will give a refresher on these topics, and define some notation that we will use throughout the course. You can safely skip this section if you know how to differentiate polynomial functions or what a *gradient* is.

Derivative of a function indicates the rate of change. The familiar example from physics is that the derivative of the velocity of a moving object is its acceleration. For example, the velocity of a car changes proportional to its acceleration or deceleration.

One of the common ways of denoting a function's derivative is using the 'prime notation'. For example derivative of the function function $f(x)$ written as $f'(x)$. Another common notation is $\frac{df}{dx}(x)$. For multi-variate functions, this notation makes it clear that the derivative is taken with respect to the variable x.

If defined, derivative of a function is another function. A well known example is the polynomials, whose derivatives are lower degree polynomials. For example, if $f(x) = x^2 - 2x$ then $f'(x) = 2x - 2$, which means that the rate of change of a quadratic function doubles as x is increased one unit. Note that if a polynomial of degree n is differentiated n times, it becomes a constant. Derivative of a linear function is a constant value, since a linear function changes with the same rate everywhere. On the other hand, derivative of a constant (function) is 0, since there is no change.

When evaluated at a particular x value, the derivative of the function is the slope of the tangent line at that point, which is indication of the direction and the rate of change. Figure 1.11 presents a simple quadratic function, $f(x) = x^2 - 2x$, and its derivatives at three points. The genral formular for the derivative $f'(x) = 2x - 2$, which is -3, 0 and 4 at $x = -0.5$, $x = 1$ and $x = 3$ respectively. Note that the function has a minimum (in fact its only minimum) at $x = 1$, where the its derivative evaluates to 0. The derivative is negative at $x = -0.5$, indicating the function is decreasing at this point, and positive derivative at $x = 3$ indicates that function is increasing. Also note that the solopes of lines in Figure 1.11 indicate the rate of change. The rate of decrease at -0.5 is less steep in comparison to increase at 3.

The derivative of a continuous function is equal to 0 at the maximum and minimum points. And this is the most important reason for all the earlier notes in this brief informal introduction. In many methods we see later on, we are interested in maximizing or minimizing functions, where this will be a handy tool. In general, derivative evaluated at a particular point will be 0 for maxima and minima of functions, it will be a negative value if the function is decreasing (as x increases), and a positive value if the function is increasing with x.

So far, we've considered differentiation of functions of a single variables. In machine learning and NLP, we often deal with multivariate functions, functions of more than one arguments (or vector arguments). One can also differentiate a function of multiple variables with respect to one of its arguments. Total derivative of a function with respect to one of the variables require considering the dependence between the variables should be considered. We will not review how to take (total) derivatives of functions of multiple variables. However, we will introduce *partial derivatives* briefly here. A partial derivative is similar to a total derivative, but we assume that

A quick refresher on polynomial functions: if $f(x) = x^n$,

$$f'(x) = \frac{df}{dx} = nx^{n-1} .$$

For example, for $f(x) = x^3 + 2x^2$,

$$f'(x) = \frac{df}{dx} = 3x^2 + 4x .$$

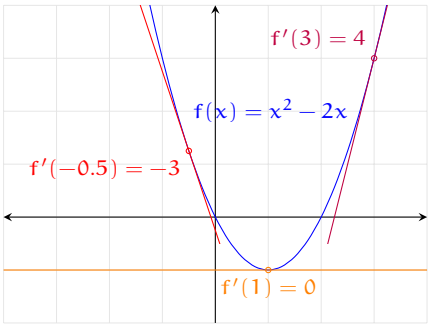

Figure 1.11: The function $f(x) = x^2 - 2x$ and its derivative evaluated at, different x values.

except the variable along which we take the derivative, all other variables are constants. So, when you evaluate the partial derivative of a function at a particular point, it gives you the rate of change along one of the axes.

The partial derivative of a function f with respect variable x is denoted by $\frac{\partial f}{\partial x}$. For example, if $f(x, y) = x^3 + yx$,

$$\frac{\partial f}{\partial x} = 3x^2 + y, \text{ and } \frac{\partial f}{\partial y} = x.$$

The vector formed by all partial derivatives of a function of n-variables is called its *gradient*. Gradient of a function f is denoted by ∇f, or $\vec{\nabla} f$.

$$\nabla f(x_1, \ldots, x_n) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

Similar to derivative of a function of a single variable, gradient points to the direction of the greatest change, and the magnitude of the gradient indicates the steepness of the change. Areas where the gradient is 0 are (local) minima, maxima and saddle points. As a result, it is an important tool in finding minimum and maximum values of (objective) functions.

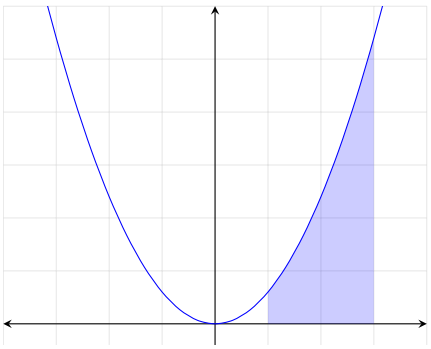

Figure 1.12: Integral of the function $f(x) = 3x^2$ in range $[1, 3]$.

INTEGRATION is the inverse of the derivation. In general, the integral of a function in a given range corresponds to the (signed) area (or volume) under a function in this range. The notation used for integral of a function $f(x)$ is $F(x) = \int f(x)dx$. This is called an indefinite integral. Often we want the integral of a function in an interval $[a, b]$, which can be calculated by

$$\int_a^b f(x)dx = F(b) - F(a).$$

For example, if $f(x) = 3x^2$, we know that $F(x) = x^3$ (since the integral is the antiderivative, and $F'(x) = f(x) = 3x^2$). If we want to know the area under $f(x)$ within range $[1, 3]$, we simply calculate $F(3) - F(1) = 27 - 3 = 24$.

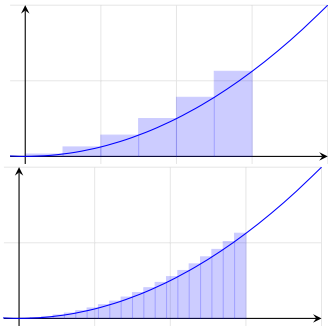

Figure 1.13: Demonstration of numerical approximation to an integral. Note that as the rectangles get smaller (as in the figure below), sum of their areas gets closer to the area under the curve.

Often integrating functions analytically (in closed form) is not easy or possible. In these cases, integrals can be computed with numeric approximation. One way to do this is to sum the areas of rectangles as demonstrated in Figure 1.13. As we decrease the width of the rectangles, or equivalently, increase the number of rectangles in a fixed range, the approximation will be more precise. This also hints at interpreting integrals as a infinite sum. This interpretation will be useful for understanding some concepts we will see later (often in probability theory).

Summary

In this lecture, we reviewed some concepts from linear algebra and calculus. The aim is to provide a refresher for readers who studied

these topics, familiarize the readers with the notation that will be used, and also give a feeling of the what mathematical concepts will be useful for following the rest of the course. This overview here is necessarily informal and incomplete. Below, a number of potential sources are listed if you need a better introduction to these concepts.

For linear algebra, Cherney, Denton, and Waldron (2013) and Beezer (2014) are two textbooks that are freely available online. A classic reference textbook for linear algebra is Strang (2009). For a more practical/geometric approach, see Farin and Hansford (2014) or Shifrin and Adams (2011).

For the concepts we reviewed briefly from calculus, any textbook introduction to calculus should be sufficient. A well-known (also available online) textbook is Strang (1991). For more alternatives on open textbooks on mathematics see `http://www.openculture.com/free-math-textbooks`.

Bibliography

Beezer, Robert A. (2014). *A First Course in Linear Algebra*. version 3.50. Congruent Press. ISBN: 9780984417551. URL: http://linear.ups.edu/.

Cherney, David, Tom Denton, and Andrew Waldron (2013). *Linear algebra*. math.ucdavis.edu. URL: https://www.math.ucdavis.edu/~linear/.

Farin, Gerald E. and Dianne Hansford (2014). *Practical linear algebra: a geometry toolbox*. Third edition. CRC Press. ISBN: 978-1-4665-7958-3.

Shifrin, Theodore and Malcolm R Adams (2011). *Linear Algebra. A Geometric Approach*. 2nd. W. H. Freeman. ISBN: 978-1-4292-1521-3.

Strang, Gilbert (1991). "Calculus". In: *Wellesley-Cambridge press*. URL: https://ocw.mit.edu/resources/res-18-001-calculus-online-textbook-spring-2005/textbook/.

– (2009). *Introduction to Linear Algebra, Fourth Edition*. 4th ed. Wellesley Cambridge Press. ISBN: 9780980232714.