Notes and Comments for Linear Algebra and Its Applications (642:55)

December 11, 2023

Chapter 1

Linear Algebra and Optimization: An Introduction

For us this chapter is a quick review of some basic concepts and operations in linear algebra—perhaps a few concepts were not introduced or emphasized in a first introductory course. This chapter also has an overview flavor with some of the concepts used not defined in detail.

1.2 Scalars, Vectors, and Matrices

Vectors and matrices are arrays of scalars. The entries of a vector $\bar{x}$ are arranged in a linear order; its ith entry is labeled as x_i. The entries of a matrix A are arranged in a rectangular form; its entry in the ith row and jth column is labeled as A_{ij}, or $A[i, j]$, or a_{ij}. We work with a matrix when we work with a list of data each having a fixed list of attributes. For example an $n \times d$ matrix can be used to represent a list of n data each with d attributes. Thus a vector is a special kind of a matrix.

1.2.1 Basic Operations with Scalars and Vectors

Concepts to be introduced: *sum of two vectors, scalar multiplication of a vector, dot product between two vectors, norm (also called length) of a vector, unit vector and normalizing a vector, Cauchy-Schwarz and triangle inequality, angle between two vectors, orthogonal vectors, a set of orthonormal vectors, L_p-norm of a vector.*

Notation convention: The textbook uses a bar over a letter, such as $\bar{x}, \bar{y}$, to represent a vector, and writes out the entries of a vector in a square bracket, separating

the entries by a comma (sometimes using a space to separate the entries). When a vector is represented as a column vector, one either writes the entries in a column form or as the *transpose* of a row vector.

Similar notations are used in programming languages. In `Matlab`, `a = [1 2 3 4]` is an example of a row vector, `b=[1; 2; 3; 4]` is an example of a column vector, while `A=[1 2 3 4; 5, 6, 7, 8]` is an example of a 2×4 matrix. The ith entry of `a` is called by `a(i)`, and the entry of a matrix `A` in in row i and column j is called by `A(i, j)`; `Matlab` also makes it easy to call a range of entries of a vector or a matrix. For example, the ith row of matrix A is called by `A(i, :)`.

In `Mathematica`, `a={1, 2, 3, 4}` is an example of a row vector, while `b={{1}, {2}, {3},{4}}` is an example of a column vector. Also, `A={{1, 2, 3, 4 }, {5, 6, 7, 8}}` is an example of a 2×4 matrix, and its entry in row i and column j is called by `A[[i,j]]`, while its ith row is called by `A[[i]]` but its jth column is called by `A[[All, j]]`.

When the textbook refers to a d-dimensional vector, it is referring to the vector having d coordinates (or components); in other words, a vector in $\mathcal{R}^d$. The concept of *dimension of a vector space* has a more precise definition, to be given in chapter 2.

Remark 1.2.1

The Euclidean inner product and norm is used when the different coordinates of a vector have the same dimensions or have been non-dimensionalized (or "weights adjusted"), as it may not be reasonable to treat difference of one unit in one component (e.g., one pound difference in weight) the same as the difference of one unit in another component (e.g., one foot difference in height). In such situations one needs to scale different components of the "raw data" in vector $\bar{y} = [y_1, \cdots, y_d]$ by appropriate weights $w_j > 0$ to produce a scaled vector $\bar{x} = [y_1/w_1, \cdots, y_d/w_d]$. Then the Euclidean norm $\|\bar{x}\| = \left(\sum_{j=1}^{d} |x_j|^2\right)^{1/2}$ of $\bar{x}$ is really some weighted norm $\|\bar{y}\|_w := \left(\sum_{j=1}^{d} |y_j|^2/|w_j|^2\right)^{1/2}$ of $\bar{y}$. There is also a corresponding notion of weighted dot product. Both the Cauchy-Schwarz and triangle inequalities continue to hold for the weighted dot product and weighted norm.

When complex entries are used, the definition of dot product between two vectors $\bar{w} = [w_1, \cdots, w_d]$ and $\bar{z} = [z_1, \cdots, z_d]$ is modified to be

$$\bar{w} \cdot \bar{z} = \sum_{j=1}^{d} w_j z_j^*,$$

where z_j^* is the complex conjugate *of z_j. This modification makes it possible to keep the notion of norm of a vector and orthogonal relation between vectors.*

The only change to the properties of the dot product is that it is no longer complex-linear in the second variable but complex-conjugate-linear:

$$\bar{w} \cdot (c_1\bar{z}_1 + c_2\bar{z}_2) = c_1^* \, \bar{w} \cdot \bar{z}_1 + c_2^* \, \bar{w} \cdot \bar{z}_2.$$

1.2.2 Basic Operations with Vectors and Matrices

A matrix in this course is often viewed in one of two ways: (i). as a representation for some data and one is interested in finding certain features of the data, or (ii). as a representation for some linear transformation and one in interested in finding features of the linear transformation often by factoring the matrix as a product of simpler matrices.
The transpose operation transforms a column vector $\bar{x}$ with d entries into a row vector $\bar{x}^{\mathrm{T}}$ and vice versa; it transforms an $n \times d$ matrix A to an $d \times n$ matrix A^{T} making the ith row of A into the ith column of A^{T}.
Two matrices of the same size can be treated as two vectors of the same size in defining scalar multiplication and matrix addition, and share the same properties (commutative and associative).
To define the matrix multiplication AB, the number of columns of A must match the number of rows of B.
In the case that $\bar{x}$ is a column vector with d rows, $[x_1, \cdots, x_d]^{\mathrm{T}}$, A must have d columns, $\bar{a}_1, \cdots, \bar{a}_d$, then

$$A\bar{x} = x_1\bar{a}_1 + \cdots + x_d\bar{a}_d.$$

Even when both AB and BA are defined and have the same size, AB may not be equal BA.
One often used interpretation of this this operation $A\bar{x}$ is that it gives a "weighted sum" of the columns of A; another often used interpretation is that it transforms $\bar{x}$ to $A\bar{x}$ as a linear transformation in the sense that

$$A(a\bar{x} + b\bar{y}) = aA\bar{x} + bA\bar{y} \quad \text{for any scalars } a, b \text{ and vectors } \bar{x}, \bar{y}.$$

In the latter interpretation, the jth column of A is $A\bar{e}_j$, the image of $\bar{e}_j$ after applying multiplication by A, where $\bar{e}_j = [0, \cdots, 1, \cdots, 0]^{\mathrm{T}}$, where the only non-zero entry 1 appears in the jth position.
When B is a $d \times l$ matrix with $\bar{b}_j$ as its jth column and A has d columns, then AB has l columns, with its jth column defined by $A\bar{b}_j$. In short notation

$$A[\bar{b}_1, \cdots, \bar{b}_l] = [A\bar{b}_1, \cdots, A\bar{b}_l].$$

In `Matlab` notation, this can be expressed as: $C = AB$, then $C(:, j) = AB(:, j)$. In terms of the row vectors we have $C(i, :) = A(i, :)B$.
The motivation for defining matrix multiplication is dictated by the role played by matrices in representing linear transformations, as will be explained in chapter 2.
If both $\bar{x}$ and $\bar{y}$ are column vectors with d components, then treating them as $d \times 1$ matrices, $\bar{x}^{\mathrm{T}}\bar{y}$ is a 1×1 matrix and equal to $\bar{x} \cdot \bar{y}$, while $\bar{x}\bar{y}^{\mathrm{T}}$ is a $d \times d$ matrix, whose jth column is simply $y_j\bar{x}$. We give this product a special name, called the *outer product* of $\bar{x}$ and $\bar{y}$: $\bar{x} \otimes \bar{y} := \bar{x}\bar{y}^{\mathrm{T}}$. Note that this notion is defined even if $\bar{y}$ has a different number of components from d, say, it has k components, then $\bar{x}\bar{y}^{\mathrm{T}}$ is a $d \times k$ matrix. Note also that $\bar{x} \otimes \bar{y} \neq \bar{y} \otimes \bar{x}$ in general.
The standard algorithm for computing matrix multiplication AB focuses on a row $A(i, :)$ of A and a column $B(:, j)$ of B and uses the product of a row vector with a column vector $A(i, :)B(:, j)$ for the (i, j)th entry of AB. Using outer product, one has an alternative view of computing matrix multiplication AB: *compute outer product between the kth column of A and kth row of B, $A(:, k)B(k, :)$, for each k and sum up the result to get AB* (see **Lemma 1.2.1**).
Note that it is possible to have two non-zero matrices A and B to produce AB equal the zero matrix! As a consequence, the relations $AB = CB$ and B not zero does not necessarily imply $A = C$.

1.2.3 Special Classes of Matrices

Several special classes of matrices arise in applications: *zero matrix and identity matrix, symmetric matrices, diagonal matrices, block diagonal matrices, rectangular diagonal matrices, upper/lower triangular matrices, strictly upper/lower triangular matrices, sparse matrices.*
Note that the textbook defines a rectangular diagonal matrix as *an $n \times d$ matrix in which each entry (i, j) has a non-zero value if and only if $i = j$*. This definition is a bit unusual in requiring that all entries for $i = j$ be non-zero. The author may have meant to define a rectangular diagonal matrix as an n × d matrix in which the non-zero entries can only occur on a position (i, j) with $i = j$.

1.2.4 Matrix Powers, Polynomials, and the Inverse

If A is a square matrix, then A^k is defined for any positive integer k (we may define $A^0 = I$, the identity matrix, as we do for a scalar $a^0 = 1$ when $a \neq 0$). This then allows us to define a polynomial in A.
A square matrix A has an inverse if there is a square matrix B such that $AB = I$ and $BA = I$. In such a situation we say A is invertible or nonsingular.
Below are a few properties of this notion. Assume A and B are $d \times d$ matrices.

(i). If $AB = I$, then for any $\bar{c} \in \mathcal{R}^d$ the system $A\bar{x} = \bar{c}$ has a solution given by $\bar{x} = B\bar{c}$.

(ii). If the system $A\bar{x} = \bar{c}$ has a solution for any $\bar{c} \in \mathcal{R}^d$, then there exists a $d \times d$ matrix B satisfying $AB = I$, as this condition is equivalent to $AB(:, j) = \bar{e}_j$ for $j = 1, \cdots, d$

(iii). If $AB = I$, then the jth column of B, $B(:, j)$, is orthogonal to all but the jth row of A.

(iv). If $BA = I$, then the only solution to $A\bar{x} = \bar{0}$ is $\bar{x} = \bar{0}$.

(v). If A is invertible, then there can be only one matrix B having the above two properties: If C also satisfies the above two properties, then

$$B = BI = B(AC) = (BA)C = IC = C.$$

This unique matrix B is called the inverse of A and denoted as A^{-1}.

(vi). $AB = I$ is equivalent to $BA = I$, so one of them would imply both (**Lemma 1.2.3**).

(vii). If both A and B are invertible, then so is AB and

$$(AB)^{-1} = B^{-1}A^{-1}.$$

(viii). If A is invertible, then so is A^{T} and

$$(A^{\mathrm{T}})^{-1} = (A^{-1})^{\mathrm{T}}$$

If A is a non-square $n \times d$ matrix, then it may be possible to find a $d \times n$ matrix B such that $AB = I_{n\times n}$ (a necessary condition is $n < d$*), but such a B can't also satisfy $BA = I_{d\times d}$.

1.2.5 The Matrix Inversion Lemma: Inverting the Sum of Matrices

For a scalar a, if $|a| < 1$, then $(1 - a)^{-1} = 1 + a + a^2 + \cdots$. A similar property holds for a square matrix A which is "small" in suitable sense: for such a matrix A, $I - A$ is invertible and

$$(I - A)^{-1} = I + A + A^2 + \cdots .$$

*We will discuss this and more detailed criteria for an invertible matrix later on.

This would involve an infinite series of matrices and a full discussion will be given later. But when A takes the special form $\bar{u}\bar{v}^{\mathrm{T}}$, in the infinite series

$$I + \bar{u}\bar{v}^{\mathrm{T}} + (\bar{u}\bar{v}^{\mathrm{T}})^2 + \cdots = I + \bar{u}\bar{v}^{\mathrm{T}} + \bar{u}(\bar{v}^{\mathrm{T}}\bar{u})\bar{v}^{\mathrm{T}} + \bar{u}(\bar{v}^{\mathrm{T}}\bar{u})^2\bar{v}^{\mathrm{T}} + \cdots$$

$\bar{v}^{\mathrm{T}}\bar{u}$ and its powers are all 1×1 matrices, so $\bar{u}(\bar{v}^{\mathrm{T}}\bar{u})^k\bar{v}^{\mathrm{T}} = (\bar{v}^{\mathrm{T}}\bar{u})^k\bar{u}\bar{v}^{\mathrm{T}}$ for any k, which leads to

$$I+\bar{u}\bar{v}^{\mathrm{T}}+\bar{u}(\bar{v}^{\mathrm{T}}\bar{u})\bar{v}^{\mathrm{T}}+\bar{u}(\bar{v}^{\mathrm{T}}\bar{u})^2\bar{v}^{\mathrm{T}}+\cdots = I+\Big(1 + (\bar{v}^{\mathrm{T}}\bar{u}) + (\bar{v}^{\mathrm{T}}\bar{u})^2 + \cdots\Big)\,\bar{u}\bar{v}^{\mathrm{T}} = I+\frac{\bar{u}\bar{v}^{\mathrm{T}}}{1-\bar{v}^{\mathrm{T}}\bar{u}},$$

provided that $1 - \bar{v}^{\mathrm{T}}\bar{u} \neq 0$. This derivation uses some formal argument involving an infinite series, but one can now verify directly that

$$(I - \bar{u}\bar{v}^{\mathrm{T}})\left(I + \frac{\bar{u}\bar{v}^{\mathrm{T}}}{1-\bar{v}^{\mathrm{T}}\bar{u}}\right) = I + \frac{\bar{u}\bar{v}^{\mathrm{T}}}{1-\bar{v}^{\mathrm{T}}\bar{u}} - \bar{u}\bar{v}^{\mathrm{T}} - \frac{\bar{u}\bar{v}^{\mathrm{T}}\bar{u}\bar{v}^{\mathrm{T}}}{1-\bar{v}^{\mathrm{T}}\bar{u}},$$

then using $\bar{v}^{\mathrm{T}}\bar{u}$ as a 1×1 matrix, one gets $\bar{u}\bar{v}^{\mathrm{T}}\bar{u}\bar{v}^{\mathrm{T}} = (\bar{v}^{\mathrm{T}}\bar{u})\bar{u}\bar{v}^{\mathrm{T}}$, thus

$$(I - \bar{u}\bar{v}^{\mathrm{T}})\left(I + \frac{\bar{u}\bar{v}^{\mathrm{T}}}{1-\bar{v}^{\mathrm{T}}\bar{u}}\right) = I + \left(\frac{1}{1-\bar{v}^{\mathrm{T}}\bar{u}} - 1 - \frac{\bar{v}^{\mathrm{T}}\bar{u}}{1-\bar{v}^{\mathrm{T}}\bar{u}}\right)\bar{u}\bar{v}^{\mathrm{T}} = I.$$

More generally we have

Lemma: (Matrix Inversion Lemma)

Let A be an invertible $d \times d$ matrix, and $\bar{u}$ and $\bar{v}$ be d-dimensional column vectors. Then, $A + \bar{u}\bar{v}^T$ is invertible if and only if $\bar{v}^T A^{-1}\bar{u} \neq -1$. In such a case, the inverse is computed as follows:

$$\left(A + \bar{u}\bar{v}^T\right)^{-1} = A^{-1} - \frac{A^{-1}\bar{u}\bar{v}^T A^{-1}}{1 + \bar{v}^T A^{-1}\bar{u}}.$$

Proof. Note that $A + \bar{u}\bar{v}^{\mathrm{T}} = A\left(I + A^{-1}\bar{u}\bar{v}^{\mathrm{T}}\right)$ and that $\left(I + A^{-1}\bar{u}\bar{v}^{\mathrm{T}}\right)$ is invertible when $1 + \bar{v}^{\mathrm{T}}A^{-1}\bar{u} \neq 0$ (with $-A^{-1}\bar{u}$ playing the role of $\bar{u}$ in the above argument for computing $(I - \bar{u}\bar{v}^{\mathrm{T}})$). Then

$$\left[A\left(I + A^{-1}\bar{u}\bar{v}^{\mathrm{T}}\right)\right]^{-1} = \left(I + A^{-1}\bar{u}\bar{v}^{\mathrm{T}}\right)^{-1}A^{-1} = \left[I + \frac{-A^{-1}\bar{u}\bar{v}^{\mathrm{T}}}{1 + \bar{v}^{\mathrm{T}}A^{-1}\bar{u}}\right]A^{-1}$$

□

Problem 1.2.5: *Let D be an $n \times d$ matrix for which each column sums to 0. Let A be an arbitrary $d \times d$ matrix. Show that the sum of each column of DA is also zero.*

Problem 1.2.11 (Inverting Triangular Matrices) *A $d \times d$ triangular matrix L with non-zero diagonal entries can be expressed in the form $(\Delta + A)$, where Δ is an invertible diagonal matrix and A is a strictly triangular matrix. Show how to compute the inverse of L using only diagonal matrix inversions and matrix multiplicatons/additions. Note that strictly triangular matrices of size $d \times d$ are always nilpotent and satisfy $A^d = 0$.*

Problem 1.2.13 (Push-Through Identity) *If U and V are two $n \times d$ matrices, show the following result:*

$$U^T(I_n + VU^T)^{-1} = (I_d + U^TV)^{-1}U^T.$$

Use the above result to show the following for any $n \times d$ matrix D and scalar $\lambda > 0$:

$$D^T(\lambda I_n + DD^T)^{-1} = (\lambda I_d + D^TD)^{-1}D^T.$$

Question: Does the existence of inverse of one of $(\lambda I_n + VU^{\mathrm{T}})$ and $(\lambda I_d + U^{\mathrm{T}}V)$ imply the existence of inverse of the other?

1.2.6 Frobenius Norm, Trace, and Energy

For the rectangular $n \times d$ matrix A with (i, j)th entry denoted by a_{ij}, its Frobenius norm is defined as follows:

$$\|A\|_{\mathcal{F}} = \|A^{\mathrm{T}}\|_{\mathcal{F}} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} a_{ij}^2}.$$

This is like treating A as a vector in $\mathcal{R}^{nd}$. Likewise we can introduce the dot product between two $n \times d$ matrices A and B as

$$A \cdot B = \sum_{i=1}^{n} \sum_{j=1}^{d} A(i,j)B(i,j).$$

The Cauchy-Schwarz and triangle inequalities continue to hold in this context.
For any square $n \times n$ matrix C, we define its trace as

$$\text{tr}(C) = \sum_{i=1}^{n} C(i,i).$$

Note that

$$\text{tr}(C) = \text{tr}(C^{\text{T}}) \quad \text{and } \text{tr}(AB) = \text{tr}(BA) \text{ if } A \text{ is } n \times d \text{ and } B \text{ is } d \times n.$$

Furthermore,

$$\|A\|_{\mathcal{F}} = \sqrt{\text{tr}(AA^{\text{T}})} \quad \text{and } A \cdot B = \text{tr}(AB^{\text{T}}) \text{ for } n \times d \ A \text{ and } B.$$

Lemma 1.2.1: (Sub-additive Frobenius Norm)

For any pair of matrices A and B of the same size, the triangle inequality $\|A + B\|_{\mathcal{F}} \leq \|A\|_{\mathcal{F}} + \|B\|_{\mathcal{F}}$ is satisfied.

Lemma 1.2.2: (Sub-multiplicative Frobenius Norm)

For any pair of matrices A and B of sizes $n \times k$ and $k \times d$, respectively, the sub-multiplicative property $\|AB\|_{\mathcal{F}} \leq \|A\|_{\mathcal{F}}\|B\|_{\mathcal{F}}$ is satisfied.

This is really a form of the Cauchy-Schwarz inequality. One can write out $\|AB\|_{\mathcal{F}}$ in terms of the entries of A and B, which would involve the dot product between the rows of A and columns of B, and apply the usual Cauchy-Schwarz inequality. As a special case we have $\|A\bar{x}\| \leq \|A\|_{\mathcal{F}}\|\bar{x}\|$ for any $\bar{x} \in \mathcal{R}^k$.

1.3 Matrix Multiplication as a Decomposable Operator

This section has the flavor of a general survey about decomposing a given matrix as the product of a number of "simple" matrices, without giving details of how this is done at

this stage and what would constitute a simple matrix. Examples involving geometric transformations are illustrated with *rotation matrices, reflection matrices*, and *scaling matrices*, while examples involving transformations on matrices are illustrated with *elementary row operator matrices.*
Note that the inverse of a rotation matrix, a reflection matrix, a scaling matrix (with non-zero scaling factors), and an elementary row operator matrix can be constructed easily.

Problem 1.3.1: *Define a* 4×4 *operator matrix so that pre-multiplying any matrix* X *with this matrix will result in addition of* c_i *times the* i*th row of* X *to the* 2*nd row of* X *for each* $i \in \{1, 2, 3, 4\}$ *in one shot. Show that this matrix can be expressed as the product of three elementary addition matrices and a single elementary multiplication matrix.*

Problem 1.3.3: *Let* A *and* B *be two matrices. Let* A_{ij} *be the matrix obtained by exchanging the* i*th and* j*th columns of* A*, and* B_{ij} *be the matrix obtained by exchanging the* i*th and* j*th rows of* B*. Write each of* A_{ij} *and* B_{ij} *in terms of* A *or* B*, and an elementary matrix. Now explain why* $A_{ij}B_{ij} = AB$*.*

Problem 1.3.4: *Let* A *and* B *be two matrices. Let matrix* A' *be created by adding* c *times the* j*th column of* A *to its* i*th column, and matrix* B' *be created by subtracting* c *times the* i*th row of* B *from its* j*th row. Explain using the concept of elementary matrices why the matrices* AB *and* $A'B'$ *are the same.*

Problem 1.3.6: *Suppose that you are told that any invertible square matrix* A *can be expressed as a product of elementary rotation/reflection/scaling matrices as* $A = R_1 R_2 \cdots R_k$*. Express the inverse of* A *in terms of the easily computable inverses of* $R_1, R_2, \cdots, R_k$*.*

1.8 Exercises

Problem 7. Show that any matrix product AB can be expressed in the form $A'\Delta B'$, where A' is a matrix in which the sum of the squares of the entries in each column is 1, B' is a matrix in which the sum of the squares of the entries in each row is 1, and Δ is an appropriately chosen diagonal matrix with nonnegative entries on the diagonal.

Problem 14. **Parallelogram law**: The parallelogram law states that the sum of the squares of the sides of a parallelogram is equal to the sum of the squares of its diagonals. Write this law as a vector identity in terms of vectors $\bar{A}$ and $\bar{B}$ of Figure 1.1. Now use vector algebra to show why this vector identity must hold

Problem 19. Consider the perturbed $d \times d$ matrix $A_\epsilon = A + \epsilon B$, where the value of ϵ is small. Show the following useful approximation for approximating A_ϵ^{-1} from A^{-1}:

$$A_\epsilon^{-1} \approx A^{-1} - \epsilon A^{-1} B A^{-1}.$$

Chapter 2

Linear Transformations and Linear Systems

2.1 Introduction

This chapter introduces/reviews some of the most basic concepts of linear algebra.

2.1.1 What Is a Linear Transform?

Definition 2.1.1: (Linear Transform)

A vector-to-vector function $f(\bar{x})$ defines a linear transform of $\bar{x}$, if the following conditions are satisfied for any scalar c:

$$f(c\bar{x}) = cf(\bar{x}), \forall\, \bar{x} \text{ in the domain of } f(\cdot),$$

$$f(\bar{x} + \bar{y}) = f(\bar{x}) + f(\bar{y}), \forall\, \bar{x}, \bar{y} \text{ in the domain of } f(\cdot).$$

What's left unsaid is what could constitute the domain of a linear transform. Implicit in the domain of a linear transform is the property that, whenever $\bar{x}, \bar{y}$ are in the domain, then so are $c\bar{x}$ for any scalar c and $\bar{x} + \bar{y}$. This is the definition for a *vector space*[*].

A typical example of the domain of a linear transformation is some $\mathcal{R}^d$, but other types of domain appear in applications often. One common domain is a *subspace* of a vector space. A subset set W of a vector space V is called a subspace if, whenever $\bar{x}, \bar{y} \in W$, then $c\bar{x} \in W$ for any scalar c and $\bar{x} + \bar{y} \in W$.

[*]A detailed discussion of this topic is given in the next section

Example 2.1.1

The set of symmetric $n \times n$ matrices $\mathcal{S}_{n\times n}$ is a subspace of the vector space of $n \times n$ matrices $\mathcal{M}_{n\times n}$. So is the set of upper triangular $n \times n$ matrices.
If A is an $n \times d$ matrix, then the set of null-vectors A, defined as $\{\bar{x} \in \mathcal{R}^d : A\bar{x} = \bar{0}\}$, is a subspace of $\mathcal{R}^d$, while if $\bar{b} \neq \bar{0}$, then the set $\{\bar{x} \in \mathcal{R}^d : A\bar{x} = \bar{b}\}$ is not a vector subspace of $\mathcal{R}^d$.
The set of $n \times n$ matrices with trace 0 is a subspace of $\mathcal{M}_{n\times n}$, while set of $n \times n$ matrices with trace 1 is not a subspace of $\mathcal{M}_{n\times n}$
The set of even functions on $\mathcal{R}$ is a subspace of the vector space of functions on $\mathcal{R}$.

An often used operator, the *translation* operator is not a linear transform. Consider the translation of the d-dimensional vector $\bar{x}$ with the vector $\bar{b} = [b_1, \cdots, b_d]^{\mathrm{T}}$ as follows:

$$t(\bar{x}) = \bar{x} + \bar{b}.$$

A translation operator can be composed with a linear transform to produce what is called an *affine* transform. Suppose that $f(\bar{x}) = A\bar{x}$, where A is some $n \times d$ matrix, then $f \circ t(\bar{x}) = A\left(\bar{x} + \bar{b}\right) = A\bar{x} + A\bar{b}$ is an example of an affine transform. If $n = d$, we can also do $t \circ f(\bar{x}) = f(\bar{x}) + \bar{b} = A\bar{x} + \bar{b}$. In generał $f \circ t \neq t \circ f$, aș expected.

Definition 2.1.2: Affine Transform

A vector-to-vector function $f(\bar{x})$ defines an affine transform of $\bar{x}$, if the following condition is satisfied for any scalar λ:

$$f(\lambda\bar{x} + [1 - \lambda]y) = \lambda f(\bar{x}) + [1 - \lambda]f(\bar{y}), \forall\, \bar{x}, \bar{y} \text{ in the domain of } f(\cdot).$$

What's left unsaid is that the domain of such a transform must have the property that $\lambda\bar{x} + [1 - \lambda]\bar{y}$ remains in the domain of $f(\cdot)$ whenever $\bar{x}, \bar{y}$ are in the domain. The set of points on a hyperplane of $\mathcal{R}^d$ defined by $\{\bar{x} \in \mathcal{R}^d : A\bar{x} = \bar{b}\}$ satisfies this property. A subset of a vector space having this property is called an affine subset of the vector space. Geometrically, for any $\bar{x}, \bar{y}$, $\{\lambda\bar{x} + [1 - \lambda]\bar{y} : \lambda \in \mathcal{R}\}$ is the set of points on the straight-line through $\bar{x}$ and $\bar{y}$, and an affine transform f maps points on the straight-line through $\bar{x}$ and $\bar{y}$ to points on the straight-line through $f(\bar{x})$ and $f(\bar{y})$.
Here is a basic
Fact: If W is an affine subset of a vector space, and $\bar{w}_0 \in W$, then $V := \{\bar{v} : \bar{v} = \bar{w} - \bar{w}_0, \bar{w} \in W\}$ is a subspace of the vector space. As a result, $W = \{\bar{w} : \bar{w}_0 + \bar{v}, \bar{v} \in V\}$; in other words, W is the translation of the subspace V.

Furthermore, if f is an affine transform defined on W, then $g(\bar{v}) = f(\bar{w}_0 + \bar{v}) - f(\bar{w}_0)$ is a linear transform on V. This allows us to express $f(\bar{w}) = g(\bar{w} - \bar{w}_0) + f(\bar{w}_0)$, a composition of translations and a linear transform.
This is seen as follows. First, for any scaler c, we have

$$\begin{aligned} g(c\bar{v}) =& f(\bar{w}_0 + c\bar{v}) - f(\bar{w}_0) \\ =& f((1-c)\bar{w}_0 + c[\bar{w}_0 + \bar{v}]) - f(\bar{w}_0) \\ =& (1-c)f(\bar{w}_0) + cf(\bar{w}_0 + \bar{v}) - f(\bar{w}_0) \\ =& cf(\bar{w}_0 + \bar{v}) - cf(\bar{w}_0) \\ =& cg(\bar{v}) \end{aligned}$$

Next, for any scalers c_1, c_2 and vectors $\bar{x}, \bar{y}, \bar{z}$, we first establish, when $c_1 + c_2 \neq 0$,

$$\begin{aligned} & f((1-c_1-c_2)\bar{z} + c_1\bar{x} + c_2\bar{y}) \\ =& f((1-c_1-c_2)\bar{z} + (c_1+c_2)[\frac{c_1}{c_1+c_2}\bar{x} + \frac{c_2}{c_1+c_2}\bar{y}]) \\ =& (1-c_1-c_2)f(\bar{z}) + (c_1+c_2)f(\frac{c_1}{c_1+c_2}\bar{x} + \frac{c_2}{c_1+c_2}\bar{y}) \quad (\text{treating } \lambda = 1-c_1-c_2) \\ =& (1-c_1-c_2)f(\bar{z}) + c_1 f(\bar{x}) + c_2 f(\bar{y}) \quad (\text{treating } \lambda = \frac{c_1}{c_1+c_2}); \end{aligned}$$

but when $c_1 + c_2 = 0$, then either $1 - c_1 \neq 0$ or $1 - c_2 \neq 0$, so we could switch the role between $\bar{x}$ and $\bar{y}$ or $\bar{x}$ to get the same conclusion. Finally we have

$$\begin{aligned} g(c_1\bar{v}_1 + c_2\bar{v}_2) =& f(\bar{w}_0 + c_1\bar{v}_1 + c_2\bar{v}_2) - f(\bar{w}_0) \\ =& f((1-c_1-c_2)\bar{w}_0 + c_1[\bar{w}_0 + \bar{v}_1] + c_2[\bar{w}_0 + \bar{v}_2]) - f(\bar{w}_0) \\ =& (1-c_1-c_2)f(\bar{w}_0) + c_1 f(\bar{w}_0 + \bar{v}_1) + c_2 f(\bar{w}_0 + \bar{v}_2) - f(\bar{w}_0) \\ =& c_1[f(\bar{w}_0 + \bar{v}) - f(\bar{w}_0)] + c_2[f(\bar{w}_0 + \bar{v}_2) - f(\bar{w}_0)] \\ =& c_1 g(\bar{v}_1) + c_2 g(\bar{v}_2) \end{aligned}$$

2.2 The Geometry of Matrix Multiplication

We first show that any linear transform from $\mathcal{R}^d$ to $\mathcal{R}^n$ can be represented as the multiplication of a d-dimensional vector with an $n \times d$ matrix.

Lemma 2.2.1: (Linear Transformation Is Matrix Multiplication)

Any linear mapping $f(\bar{x})$ from d-dimensional vectors to n-dimensional vectors can be represented as the matrix-to-vector product $A\bar{x}$ by constructing A as

follows. The columns of the $n \times d$ matrix A are $f(\bar{e}_1), \cdots, f(\bar{e}_d)$, where $\bar{e}_i$ is the ith column of the $d \times d$ identity matrix.

When writing the vectors as row-vectors, the representation can be written as $\bar{x}A^{\mathrm{T}}$.

Proof. Any $\bar{x} \in \mathcal{R}^d$ can be written as $\bar{x} = \sum_{j=1}^{d} x_j \bar{e}_j$. By the defining property of f, we have

$$f(\bar{x}) = \sum_{j=1}^{d} x_j f(\bar{e}_j) = A[x_1, \cdots, x_d]^{\mathrm{T}}.$$

The last equality uses the multiplication rule of matrices. □

The following property dictates how matrix multiplication is defined:

> Suppose that $f : \mathcal{R}^d \mapsto \mathcal{R}^n$ is a linear transform represented by $A\bar{v}$ for some $n \times d$ matrix A, $\bar{v} \in \mathcal{R}^d$, and $g : \mathcal{R}^k \mapsto \mathcal{R}^d$ is a linear transform represented by $B\bar{u}$ for some $d \times k$ matrix B, $\bar{u} \in \mathcal{R}^k$, then $f \circ g : \mathcal{R}^k \mapsto \mathcal{R}^n$ is a linear transform represented by $(AB)\bar{u}$.

Other commonly used linear transforms include the derivative operator on the vector space of differentiable functions on an open domain, linear differential operators associated with a linear differential equation (For example, $\mathcal{D}^2 + 2\mathcal{D} + x$ is associated with the ODE $u''(x) + 2u'(x) + xu(x) = f(x)$), and Fourier transform on appropriate function spaces.

The rest of this section focuses on the concepts of *orthogonal matrices, rotation matrices*, and *reflection matrices*. The latter two types are particular examples of orthogonal matrices. More explicit examples of rotation matrices are *Givens matrices*, while a general reflection matrix can be constructed as a *Householder reflection matrix*.

Definition 2.2.1

An $n \times n$ matrix A is called an orthogonal matrix if its columns are *orthonormal*. This is equivalent to $A^{\mathrm{T}}A = I_n$.

From the definition we see that if A is an orthogonal matrix, then it is invertible and $A^{-1} = A^{\mathrm{T}}$. Furthermore, $AA^{\mathrm{T}} = AA^{-1} = I_n$, which implies that the columns of A^{T}, namely, the rows of A, are also orthonormal in $\mathcal{R}^n$.

A linear transform defined by an orthogonal matrix is called an *orthogonal transformation*. An orthogonal transformation preserves the dot product between vectors, therefore preserves the length of vectors and angle between vectors. For this reason, an orthogonal transformation is a *rigid transformation*. This is seen by noting that

$$A\bar{x} \cdot A\bar{y} = (A\bar{x})^{\mathrm{T}} A\bar{y} = \bar{x}^{\mathrm{T}} A^{\mathrm{T}} A\bar{y} = \bar{x}^{\mathrm{T}} \bar{y} = \bar{x} \cdot \bar{y}.$$

Another important property is the *closure property under multiplication: The product of any number of orthogonal matrices is always an orthogonal matrix.*
Take two orthogonal matrices A, B, we see that $(AB)^{\mathrm{T}}(AB) = B^{\mathrm{T}}(A^{\mathrm{T}}A)B = B^{\mathrm{T}}I_nB = I_n$, proving that AB continues to be an orthogonal matrix.

Definition 2.2.2

An $n \times n$ matrix A is called rotation matrix if there exist a set of n orthonormal vectors $\{\bar{\eta}_1, \cdots, \bar{\eta}_n\}$ of $\mathcal{R}^n$ and an angle α such that

$$A\bar{\eta}_1 = \cos\alpha\,\bar{\eta}_1 + \sin\alpha\,\bar{\eta}_2, A\bar{\eta}_2 = -\sin\alpha\,\bar{\eta}_1 + \cos\alpha\,\bar{\eta}_2, A\bar{\eta}_k = \bar{\eta}_k \text{ for } k > 2.$$

When $\{\bar{\eta}_1, \bar{\eta}_2\}$ is $\{\bar{e}_j, \bar{e}_k\}$ for some $j \neq k$, such a rotation matrix is called a Givens rotation matrix.
An $n \times n$ matrix A is called reflection matrix if there exist a set of n orthonormal vectors $\{\bar{\eta}_1, \cdots, \bar{\eta}_n\}$ of $\mathcal{R}^n$ such that

$$A\bar{\eta}_1 = -\bar{\eta}_1, A\bar{\eta}_k = \bar{\eta}_k \text{ for } k \geq 2.$$

It may not be self-evident from the definition that a rotation matrix or a reflection matrix is an orthogonal matrix. But if A is a rotation matrix as in the definition, then the relations give

$$A[\bar{\eta}_1, \bar{\eta}_2, \cdots, \bar{\eta}_n] = [\bar{\eta}_1, \bar{\eta}_2, \cdots, \bar{\eta}_n] \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 & \cdots & 0 \\ \sin\alpha & \cos\alpha & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Set $[\bar{\eta}_1, \bar{\eta}_2, \cdots, \bar{\eta}_n]$ as B, then B is an orthogonal matrix, and

$$A = B \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 & \cdots & 0 \\ \sin\alpha & \cos\alpha & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} B^{-1}$$

is the product of three orthogonal matrices, therefore is an orthogonal matrix.
When A is the Givens rotation matrix corresponding to $\{\bar{\eta}_1, \bar{\eta}_2\}$ is $\{\bar{e}_j, \bar{e}_k\}$ for some $j \neq k$, A can be worked out explicitly as given in the textbook—this follows by using the relations of $A\bar{e}_i$ but the ordering of $\{\bar{e}_1, \cdots, \bar{e}_d\}$ of the basis vectors.
The argument for a reflection matrix is similar. In fact we can reconstruct the reflection matrix A as a *Householder matrix* using the $\bar{\eta}_1$ in its definition.

For any vector $\bar{x} \in \mathcal{R}^n$, $\bar{x}_{\|} := (\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1$ is the *orthogonal projection* of $\bar{x}$ along $\bar{\eta}_1$ in the sense that $\bar{x}_{\perp} := \bar{x} - (\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1$ is orthogonal to $\bar{\eta}_1$. Then

$$\bar{x} = (\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1 + (\bar{x} - (\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1) = \bar{x}_{\|} + \bar{x}_{\perp}$$

and

$$-\bar{x}_{\|} + \bar{x}_{\perp} = -(\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1 + (\bar{x} - (\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1) = \bar{x} - 2(\bar{x} \cdot \bar{\eta}_1)\bar{\eta}_1$$

is the orthogonal reflection of $\bar{x}$ with respect to the hyperplane through the origin with $\bar{\eta}_1$ as its unit normal. This vector can be expressed as

$$\left(I_n - 2\bar{\eta}_1 \left(\bar{\eta}_1\right)^{\mathrm{T}}\right) \bar{x},$$

which identifies $I_n - 2\bar{\eta}_1 \left(\bar{\eta}_1\right)^{\mathrm{T}}$ as the Householder reflection matrix determined by $\bar{\eta}_1$.

Problem Verify that this matrix A satisfies $A\bar{\eta}_1 = -\bar{\eta}_1, A\bar{\eta}_k = \bar{\eta}_k$ for $k \geq 2$.

It turns out that product of any two reflection matrices is a rotation matrix. Furthermore

Lemma 2.2.4: (Householder Geometric Decomposition)

Any orthogonal matrix of size $d \times d$ can be expressed as the product of at most d Householder reflection matrices.

Proof. Let A be an $d \times d$ orthogonal matrix. If the last column of A is equal to $\bar{e}_d$, then the last component of each of the first $(d-1)$ columns of A must be 0 and can be treated as a vector in $\mathcal{R}^{d-1}$; they remain orthonormal in $\mathcal{R}^{d-1}$. We can use an induction argument to claim that the $(d-1) \times (d-1)$ orthogonal matrix consisting of the first $(d-1)$ columns of A with their last 0 component stripped can be expressed as the product $H_k \cdots H_1$ of at most $(d-1)$ $(d-1) \times (d-1)$ Householder reflection matrices. Let $\tilde{H}_i$ be the $d \times d$ matrix whose first $(d-1)$ columns are simply the respective column of H_i adjoining 0 as its last component as a vector in $\mathcal{R}^d$, and whose dth column is $\bar{e}_d$, then we have $A = \tilde{H}_k \cdots \tilde{H}_1$.
If the last column of A is not equal to $\bar{e}_d$, then there exists a Householder reflection matrix H_0 which maps that vector to $\bar{e}_d$. Then $H_0 A$ is an $d \times d$ orthogonal matrix whose last column is $\bar{e}_d$. The argument of the last paragraph shows that $H_0 A$ can be written as the product of at most $(d-1)$ Householder reflection matrices. Left-multiplying by $H_0^{-1} = H_0$, we conclude that A can be written as the product of at most d Householder reflection matrices. □

Problem Let A be the Householder reflection matrix defined by the unit vector $\bar{\eta}_1$ and B be the Householder reflection matrix defined by the unit vector $\bar{\eta}_2$. Show that any vector $\bar{v}$ which is orthogonal to both $\bar{\eta}_1$ and $\bar{\eta}_2$ satisfies $AB\bar{v} = \bar{v}$. Use this to show that, if $\bar{\eta}_1 \neq \pm\bar{\eta}_2$, then AB is a rotation.

A theorem to anticipate:

Theorem: (General Factorization of a Linear Transformation)

All linear transformations defined by matrix multiplication can be expressed as a sequence of rotations/reflections, together with a single anisotropic scaling.

Problem 2.2.1 Show that you can express a $d \times d$ elementary row interchange matrix as the product of a 90° Givens rotation and an elementary reflection.

Problem 2.2.4 (Generalized Householder) Show that a sequence of k mutually orthogonal Householder transformations can be expressed as $I - 2QQ^{\mathrm{T}}$ for a $d \times k$ matrix Q containing orthonormal columns. Which $(d-k)$-dimensional plane is this a reflection in?

Exercises 2.14

Exercise 32. Show that any 2×2 Givens rotation matrix is a product of at most two Householder reflection matrices. Think geometrically before wading into the algebra. Now generalize the proof to $d \times d$ matrices.

Exercise 36. Show how any vector $\bar{v} \in \mathcal{R}^n$ can be transformed to $\bar{w} \in \mathcal{R}^n$ as $\bar{w} = cH\bar{v}$,where c is a scalar and H is an $n \times n$ Householder reflection matrix. Think geometrically to solve this exercise. NOTE: It's implicitly assumed that $\bar{v} \neq \bar{0}$ here.

2.3 Vector Spaces and Their Geometry

We already defined vector space and vector subspace, and will focus on other related important concepts: *linear dependence* and *linear independence*, *basis*, *dimension*, *coordinates in a basis* and *coordinate transformations between bases*.

Definition 2.3.4: (Linear Dependence and Independence)

A set of vectors $\bar{a}_1, \cdots, \bar{a}_d$ is linearly dependent, if a set of d scalars $x_1, \cdots, x_d$ can be found so that at least some of the scalars are non-zero, and the following condition is satisfied:

$$\sum_{i=1}^{d} x_i \bar{a}_i = \bar{0}.$$

When no such set of non-zero scalars can be found, the resulting set of vectors is said to be *linearly independent*.

From the definition, $\bar{a}_1, \cdots, \bar{a}_d$ is linearly dependent if one of them is a linear combination of the rest; also $\bar{a}_1, \cdots, \bar{a}_d$ is linearly independent if the *only solution* to $\sum_{i=1}^{d} x_i \bar{a}_i = \bar{0}$ is $x_1 = x_2 = \cdots = x_d = 0$. Using this we see that *a set of non-zero vectors $\bar{a}_1, \cdots, \bar{a}_d$ that are mutually orthogonal must be linearly independent.* For, if there are scalars $x_1, \cdots, x_d$ such that $\sum_{i=1}^{d} x_i \bar{a}_i = \bar{0}$, then for any $j, 1 \leq j \leq d$, taking the dot product with $\bar{a}_j$ on both sides and using the mutual orthogonality relations, we get $x_j \bar{a}_j \cdot \bar{a}_j = 0$. Since $\bar{a}_j \cdot \bar{a}_j \neq 0$, we conclude that $x_j = 0$.

When $\bar{a}_1, \cdots, \bar{a}_d$ are in $\mathcal{R}^n$, whether the set $\{\bar{a}_1, \cdots, \bar{a}_d\}$ is linearly dependent or linearly independent can be answered by knowing whether the linear system of equations

$$[\bar{a}_1, \cdots, \bar{a}_d][x_1, \cdots, x_d]^{\mathrm{T}} = \bar{0}$$

has a non-zero solution or only zero as a solution.

Definition 2.3.5: (Basis)

A basis (or basis set) of a vector space $V \subset \mathcal{R}^n$ is a minimal set of vectors $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\} \subset V$, so that all vectors in V can be expressed as linear combinations of $\bar{a}_1, \cdots, \bar{a}_d$. In other words, for any vector $\bar{v} \in V$, we can find scalars $x_1, \cdots, x_d$ so that $\bar{v} = \sum_{i=1}^{d} x_i \bar{a}_i$, and one cannot do this for any proper subset of $\mathcal{B}$.

The definition implies that a basis must be linear independent. For, if $\{\bar{a}_1, \cdots, \bar{a}_d\}$ is a basis and there are scalars $x_1, \cdots, x_d$, not all zero, such that $\sum_{i=1}^{d} x_i \bar{a}_i = \bar{0}$, then, if $x_j \neq 0$, then we can solve for $\bar{a}_j$ in terms of the rest to express $\bar{a}_j$ as a linear combination of the rest of the vectors in $\{\bar{a}_1, \cdots, \bar{a}_d\}$. This would imply that, after removing $\bar{a}_j$ from this set, all vectors in V can be expressed as linear combinations of the remaining vectors, which would violate the minimality in the definition of a basis.

Definition 2.3.6: (Basis: Alternative Definition)

A basis (or basis set) of a vector space V is a **maximal** set of linearly independent vectors in it.

Theorem 2.3.1: (Dimension Theorem for Vector Spaces)

The number of members in every possible basis set of a vector space V is always the same. This value is referred to as the **dimensionality** (or dimension) of the vector space.

Corollary: (Criteria for a basis if dimension is known)

Let V be a d-dimensional vector space.
Suppose that $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ is a set of linearly independent vectors in V, then it must be a basis of V.
Suppose that every vector $\bar{v}$ in V can be written as a linear combination of $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$, then $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ must be linearly independent and forms a basis of V.

Proof. Suppose that $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ is a set of linearly independent vectors in V, to prove that it is a basis of V we need to show that every vector $\bar{v}$ in V can be written as a linear combination of $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$. If that is not the case, then there exists some vector $\bar{a}_{d+1}$ which can't be written as a linear combination of $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$, then $\{\bar{a}_1, \cdots, \bar{a}_d, \bar{a}_{d+1}\}$ must be linearly independent and can be expanded into a basis of V. But that would mean that V has a basis containing at least $d+1$ elements, violating the Dimension Theorem for Vector Spaces.
Suppose that every vector $\bar{v}$ in V can be written as a linear combination of $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$, and if $\mathcal{B}$ is not a minimal set of vectors having this property, then one could select a proper subset of $\mathcal{B}$ having this property and is a minimal set having this property. But this would mean that V would have a basis with fewer than d vectors, violating the Dimension Theorem for Vector Spaces. □

Corollary: (Expansion into a basis)

Let V be a d-dimensional vector space and $\{\bar{a}_1, \cdots, \bar{a}_k\}$ be a set of linearly independent vectors in V. If $k < d$, then one can find $\{\bar{a}_{k+1}, \cdots, \bar{a}_d\}$ in V such that $\{\bar{a}_1, \cdots, \bar{a}_k, \bar{a}_{k+1}, \cdots, \bar{a}_d\}$ forms a basis of V.

Lemma 2.3.4: (Matrix Invertibility and Linear Independence)

An $n \times n$ square matrix A has linearly independent columns/rows if and only if it is invertible.

Proof. Suppose that A is invertible, then there exists an $n \times n$ matrix B such that $AB = BA = I_n$. To prove that A has linearly independent columns, consider any solution $\bar{x} \in \mathcal{R}^n$ to $A\bar{x} = \bar{0}$, we deduce that $BA\bar{x} = B\bar{0} = \bar{0}$, from which we conclude that $\bar{x} = BA\bar{x} = \bar{0}$, proving that A has linearly independent columns. To prove that A has linearly independent rows, we consider row vector solutions to $\bar{x}^{\mathrm{T}}A = \bar{0}$ in a similar fashion.
Conversely suppose that A has linearly independent columns, then our corollary above implies that the columns of A form a basis of $\mathcal{R}^n$. Therefore we can construct $\bar{b}_i$ solving $A\bar{b}_i = \bar{e}_i$ for each $i = 1, \cdots, n$. Let $B = [\bar{b}_1, \cdots, \bar{b}_n]$, then $AB = I_n$. This property implies that the columns of B are linearly independent and therefore forms a basis of $\mathcal{R}^n$ according to the above corollary. Then there must be some $n \times n$ matrix C such that $BC = I_n$. It then follows that $A = A(BC) = (AB)C = I_nC = C$, so $BA = I_n$ and B is the inverse of A.
An alternative proof uses elementary row reductions to conclude that A must have a pivot in each row, therefore its reduced row echelon form must be I_n. This means that there must be elementary matrices $E_1, \cdots, E_k$ such that $E_k \cdots E_1 A = I_n$. But each E_i is invertible so we obtain $A = E_1^{-1} \cdots E_k^{-1}$ is a product of invertible matrices and conclude that it is invertible.
The situation when A has linearly independent rows is done in a similar fashion. □

2.3.1 Coordinates in a Basis System

Let $V \subset \mathcal{R}^n$ be a d-dimensional subspace of $\mathcal{R}^n$ and $d < n$. For any basis $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ of V, any vector $\bar{v} \in V$ can be expressed as $\bar{v} = \sum_{i=1}^{d} x_i \bar{a}_i$ for some scalars $x_1, \cdots, x_d$. Then $[x_1, \cdots, x_d]^{\mathrm{T}}$ are the coordinates of $\bar{v}$ in the basis $\mathcal{B}$ and $[x_1, \cdots, x_d]^{\mathrm{T}}$ is called the coordinate vector of $\bar{v}$ in the basis $\mathcal{B}$.

Lemma 2.3.5: (Uniqueness of Coordinates)

The coordinates $x = [x_1, \cdots, x_d]^T$ of any vector $\bar{v} \in V$ in terms of a basis set $\mathcal{B} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ are always unique.

This is a consequence of the linear independence of any basis.
The question of finding these coordinates is equivalent to solving the system of n linear equations $A[x_1, \cdots, x_d]^{\mathrm{T}} = \bar{v}$ in the d unknowns $x_1, \cdots, x_d$. The solvability of

such a system can be studied using the Gaussian elimination method.

Example 2.3.1

Consider the null space of the matrix

$$A = \begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 6 & 12 & 2 & 0 & 28 \\ 3 & 6 & 2 & 1 & 23 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix},$$

namely, those vectors $\bar{x} \in \mathcal{R}^5$ such that $A\bar{x} = \bar{0}$. One finds that the reduced row echelon form of A is

$$\text{rref}(A) = \begin{bmatrix} 1 & 2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

thus $\bar{x}$ needs to solve

$$\begin{bmatrix} 1 & 2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

One finds that x_2, x_5 are free variables and $x_4 = -4x_5, x_3 = -5x_5, x_1 = -2x_2 - 3x_5$. In vector form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = x_2 \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} -3 \\ 0 \\ -5 \\ -4 \\ 1 \end{bmatrix}.$$

Thus

$$\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ -5 \\ -4 \\ 1 \end{bmatrix} \right\}$$

forms a basis of the null space of A, and any vector $\bar{x}$ in the null space of A has coordinates (x_2, x_5) in this basis.

When the columns of A are orthonomal, there is an easier solution: for any $i = 1, \cdots, d$, one simply left-multiply both sides of the system by the transpose of the ith column of A to get $A(:,i)^{\mathrm{T}} A[x_1, \cdots, x_d]^{\mathrm{T}} = A(:,i)^{\mathrm{T}} \bar{v}$, and uses the information that $A(:,i)^{\mathrm{T}} A(:,j) = 0$ unless $i = j$ and $A(:,i)^{\mathrm{T}} A(:,i) = 1$ to conclude that $x_i = A(:,i)^{\mathrm{T}} \bar{v}$.

Question: Do the $x_i = A(:,i)^{\mathrm{T}} \bar{v}$'s still solve $A[x_1, \cdots, x_d]^{\mathrm{T}} = \bar{v}$ when $\bar{v} \notin V$?

In the general case of $d < n$, one can't solve $A[x_1, \cdots, x_d]^{\mathrm{T}} = \bar{v}$ for every $\bar{v} \in \mathcal{R}^n$ as $A[x_1, \cdots, x_d]^{\mathrm{T}} \in V$, a proper subspace of $\mathcal{R}^n$. The best one can do is to choose $\bar{x} = [x_1, \cdots, x_d]^{\mathrm{T}}$ to make $\|A\bar{x} - \bar{v}\|$ as small as possible. One can prove that such an $\bar{x}$ exists, and would make $A\bar{x} - \bar{v}$ orthogonal to every vector in V. As a consequence, $A\bar{x} - \bar{v}$ is orthogonal to every column of A, resulting in the equation

$$A^{\mathrm{T}} (A\bar{x} - \bar{v}) = \bar{0}.$$

Then we need to solve the system of d linear equations $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}} \bar{v}$ in the d unknowns $x_1, \cdots, x_d$.
Under the condition that the columns of A are linearly independent, the $d \times d$ square matrix $A^{\mathrm{T}} A$ has the property that the only solution to $A^{\mathrm{T}} A\bar{x} = \bar{0}$ is $\bar{x} = \bar{0}$ (Prove this!). Thus the columns of $A^{\mathrm{T}} A$ are also linearly independent, making $A^{\mathrm{T}} A$ invertible. We can now solve $\bar{x} = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} \bar{v}$ as the coordinate of $\bar{v}$ in the basis consisting of the columns of A. Again $A\bar{x} = \bar{v}$ only when $\bar{v} \in V$; in general we only have $A\bar{x} - \bar{v} \perp A(:,j)$ for each $j = 1, \cdots, d$.
Note that $\left[(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}\right] A = I_d$, but $A\left[(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}\right] \neq I_n$ when $d < n$. This is seen by setting $\left[(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}\right] \bar{y} = \bar{x}$, then $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}} \bar{y}$ but $A\bar{x} \neq \bar{y}$ in general. $(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}$ is called the *left inverse* of A.

2.3.2 Coordinate Transformations Between Basis Sets

The set of vectors $\left\{ \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} -4 \\ 3 \end{bmatrix} \right\}$ is linearly independent in $\mathcal{R}^2$ so forms a basis of $\mathcal{R}^2$, so does $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}$. Any vector $\bar{v} \in \mathcal{R}^2$ has a set of coordinates in each basis:

$$\bar{e} = x_1 \begin{bmatrix} 3 \\ 4 \end{bmatrix} + x_2 \begin{bmatrix} -4 \\ 3 \end{bmatrix} = y_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + y_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

This relation can be written in matrix form

$$\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

We are interested in finding the relations between $[x_1, x_2]^{\mathrm{T}}$ and $[y_1, y_2]^{\mathrm{T}}$. In the case here it is straightforward to see that

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The general case is formulated and solved in the same way. Suppose that $\{\bar{a}_1, \cdots, \bar{a}_d\}$ and $\{\bar{b}_1, \cdots, \bar{b}_d\}$ are sets of basis of a vector subspace V of $\mathcal{R}^n$. Any vector $\bar{v} \in V$ has two sets of coordinates: let $\bar{x}_a$ denotes its coordinates in the basis $\{\bar{a}_1, \cdots, \bar{a}_d\}$ and $\bar{x}_b$ denotes its coordinates in the basis $\{\bar{b}_1, \cdots, \bar{b}_d\}$. Our task is to find the relation between $\bar{x}_a$ and $\bar{x}_b$.
First, we note that

$$\bar{v} = A\bar{x}_a = B\bar{x}_b$$

where $A = [\bar{a}_1, \cdots, \bar{a}_d]$ and $B = [\bar{b}_1, \cdots, \bar{b}_d]$. If $d = n$, then both A and B are invertible so one can get $\bar{x}_a = A^{-1}B\bar{x}_b$.
For the case of $d < n$, each column of A must be a linear combination of the columns of B, so there exists a $d \times d$ matrix, called it $P_{a\to b}$, such that $A = BP_{a\to b}$. This then leads to

$$BP_{a\to b}\bar{x}_a = B\bar{x}_b.$$

Since the columns of B are linearly independent, we conclude that $P_{a\to b}\bar{x}_a = \bar{x}_b$.
What remains to have an algorithm to determine $P_{a\to b}$. Each column of $P_{a\to b}$, say $P_{a\to b}(:, j)$, solves $BP_{a\to b}(:, j) = A(:, j)$. We use our algorithm developed in the last subsection: $B^{\mathrm{T}}BP_{a\to b}(:, j) = B^{\mathrm{T}}A(:, j)$ and $B^{\mathrm{T}}B$ must be invertible to get $P_{a\to b}(:, j) = (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}A(:, j)$. In summary

$$P_{a\to b} = (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}A.$$

Example 2.3.2

Consider the subspace $\mathcal{V}$ in $\mathcal{R}^2$ spanned by $\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} \right\}$. Note that

$$\begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

lie in $\mathcal{V}$ and are linearly independent, so also form a basis of $\mathcal{V}$.

Any vector $\bar{v} \in \mathcal{V}$ has coordinates in each basis

$$\bar{v} = x_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = y_1 \begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix} + y_2 \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}.$$

This can be written in matrix form

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & -3 \\ 2 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

To find the relations between $[x_1, x_2]^{\mathrm{T}}$ and $[y_1, y_2]^{\mathrm{T}}$ we use the relation

$$\begin{bmatrix} -1 & -3 \\ 2 & 0 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad (\dagger)$$

to deduce that

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

Since the two columns of the matrix

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

are linearly independent, the coefficients in the above equality must match, resulting in

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

In the absence of knowing (†), it can be found by left-multiplying both sides by the transpose of

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

to get

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -1 & -3 \\ 2 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix}$$

is an invertible square matrix, from which we obtain

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -1 & -3 \\ 2 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

2.3.3 Span of a Set of Vectors

Definition 2.3.7: (Span)

The span of a finite set of vectors $\mathcal{A} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ is the vector space defined by all possible linear combinations of the vectors in $\mathcal{A}$:

$$\mathrm{Span}(\mathcal{A}) = \left\{ \bar{v} = \sum_{i=1}^{d} x_i \bar{a}_i : \forall\, x_i \in \mathcal{R} \right\}.$$

$\mathrm{Span}(\mathcal{A})$ is always a subspace, although its dimension may be smaller than d. What one can say is

Corollary 2.3.2: (Basis selection)

A basis of $\mathrm{Span}(\{\bar{a}_1, \cdots, \bar{a}_d\})$ can be chosen from among $\{\bar{a}_1, \cdots, \bar{a}_d\}$. One just chooses a maximal subset of $\{\bar{a}_1, \cdots, \bar{a}_d\}$ which is linearly independent.

Corollary 2.3.3

Any subset of more than d vectors in $\mathrm{Span}(\{\bar{a}_1, \cdots, \bar{a}_d\})$ must be linearly dependent.

Example 2.3.3

Consider the span of the column vectors of

$$A = \begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 6 & 12 & 2 & 0 & 28 \\ 3 & 6 & 2 & 1 & 23 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix}.$$

Since the reduced row echelon form of A is

$$\text{rref}(A) = \begin{bmatrix} 1 & 2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

we know that columns $1, 3, 4$ are pivot columns, and the remaining columns are linear combination of these pivot columns. More specifically, setting $x_2 = 1$ and $x_5 = 0$ leads to

$$2\begin{bmatrix} 3 \\ 6 \\ 3 \\ 2 \end{bmatrix} + \begin{bmatrix} 6 \\ 12 \\ 6 \\ 4 \end{bmatrix} = \bar{0},$$

and setting $x_2 = 0$ and $x_5 = 1$ leads to

$$-3\begin{bmatrix} 3 \\ 6 \\ 3 \\ 2 \end{bmatrix} - 5\begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \end{bmatrix} - 4\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 14 \\ 28 \\ 23 \\ 20 \end{bmatrix} = \bar{0}.$$

From this we conclude that $\mathcal{V}$ is also the span of

$$\left\{ \begin{bmatrix} 3 \\ 6 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\}.$$

These three vectors are linearly independent so form a basis of $\mathcal{V}$.

2.3.4 Machine Learning Example: Discrete Wavelet Transform

This subsection illustrates a commonly used orthogonal basis in Discrete Wavelet Transform. Let B be the matrix whose columns are these orthogonal vectors and B' be the matrix whose columns are the normalized vectors of these orthogonal vectors. Then $B = B'\text{diag}[l_1, \cdots, l_n]$ with $l_i = \|B(:, i)\| > 0$. For a given vector $\bar{s} \in \mathcal{R}^n$ in its Euclidean coordinate, its coordinates $\bar{s}_n$ in the orthonormal basis $\{B'(:, i) : i = 1, \cdots, n\}$ satisfy $\bar{s} = B'\bar{s}_n$. B' is an orthogonal matrix, so $\bar{s}_n = (B')^{\mathrm{T}}\bar{s}$.
The textbook explains that in practice one does not work out B' explicitly; rather, using $B^{\mathrm{T}} = \text{diag}[l_1, \cdots, l_n]^{\mathrm{T}}(B')^{\mathrm{T}}$, one sees that $B^{\mathrm{T}}\bar{s} = \text{diag}[l_1, \cdots, l_n]^{\mathrm{T}}\bar{s}_n$. One names $B^{\mathrm{T}}\bar{s}$ the *unnormalized coordinates* $\bar{s}_u$ of $\bar{s}$ in this orthogonal wavelet basis, and determines $\bar{s}_n$ through $\bar{s}_u = \text{diag}[l_1, \cdots, l_n]^{\mathrm{T}}\bar{s}_n$.
Note that in the example given in the textbook, a vector $\bar{s}$ which changes very little over adjacent moments will have small components in the basis vectors corresponding to difference between eighths.

2.3.5 Relationships Among Subspaces of a Vector Space

The intersection (but not the union) of any two subspaces of a vector space is still a subspace.

Definition 2.3.8: (Disjoint Vector Spaces)

Two vector spaces $\mathcal{U} \subset \mathcal{R}^n$ and $\mathcal{W} \subset \mathcal{R}^n$ are disjoint if and only if the two spaces do not contain any vector in common other than the zero vector, in other words, $\mathcal{U} \cap \mathcal{W} = \{\bar{0}\}$.

If $\mathcal{U}$ and $\mathcal{W}$ are disjoint with basis sets $\mathcal{B}_u$ and $\mathcal{B}_w$, the union $\mathcal{B} = \mathcal{B}_u \cup \mathcal{B}_w$ of these basis sets is a linearly independent set.
Suppose that $\mathcal{U} \subset \mathcal{R}^n$ and $\mathcal{W} \subset \mathcal{R}^n$ are two subspaces of $\mathcal{R}^n$. Then $\mathcal{U}+\mathcal{W} := \{\bar{u}+\bar{w} : \bar{u} \in \mathcal{U}, \bar{w} \in \mathcal{W}\}$ is a subspace of $\mathcal{R}^n$.

Proposition 2.3.4

Suppose that $\mathcal{U} \subset \mathcal{R}^n$ and $\mathcal{W} \subset \mathcal{R}^n$ are two subspaces of $\mathcal{R}^n$. Then

$$\dim(\mathcal{U} + \mathcal{W}) = \dim(\mathcal{U}) + \dim(\mathcal{W}) - \dim(\mathcal{U} \cap \mathcal{W}).$$

Proof. Let $\{\bar{z}_1, \cdots, \bar{z}_k\}$ be a basis of $\mathcal{U} \cap \mathcal{W}$. Then there exists $\{\bar{x}_1, \cdots, \bar{x}_l\}$ of $\mathcal{U}$ such that $\{\bar{x}_1, \cdots, \bar{x}_l; \bar{z}_1, \cdots, \bar{z}_k\}$ form a basis of $\mathcal{U}$. Likewise, there exists $\{\bar{y}_1, \cdots, \bar{y}_m\}$ of

$\mathcal{W}$ such that $\{\bar{y}_1, \cdots, \bar{y}_m; \bar{z}_1, \cdots, \bar{z}_k\}$ form a basis of $\mathcal{W}$. We claim that

$$\{\bar{x}_1, \cdots, \bar{x}_l; \bar{z}_1, \cdots, \bar{z}_k; \bar{y}_1, \cdots, \bar{y}_m\}$$

form a basis of $\mathcal{U} + \mathcal{W}$. Then the Proposition follows directly.
It's clear that every vector of $\mathcal{U} + \mathcal{W}$ is a linear combination of vectors in

$$\{\bar{x}_1, \cdots, \bar{x}_l; \bar{z}_1, \cdots, \bar{z}_k; \bar{y}_1, \cdots, \bar{y}_m\}.$$

It remains to prove that this set is linearly independent. Suppose there are coefficients $a_1, \cdots, a_l; b_1, \cdots, b_m; c_1, \cdots, c_k$ such that

$$a_1\bar{x}_1 + \cdots + a_l\bar{x}_l + c_1\bar{z}_1 + \cdots + c_k\bar{z}_k + b_1\bar{y}_1 + \cdots + b_m\bar{y}_m = \bar{0}.$$

This implies that

$$a_1\bar{x}_1 + \cdots + a_l\bar{x}_l = -b_1\bar{y}_1 - \cdots - b_m\bar{y}_m - c_1\bar{z}_1 - \cdots - c_k\bar{z}_k \in \mathcal{W}.$$

But $a_1\bar{x}_1 + \cdots + a_l\bar{x}_l \in \mathcal{U}$, so it must be an element of $\mathcal{U} \cap \mathcal{W}$, therefore, must be a linear combination of $\{\bar{z}_1, \cdots, \bar{z}_k\}$. But $\{\bar{x}_1, \cdots, \bar{x}_l; \bar{z}_1, \cdots, \bar{z}_k\}$ is linearly independent, so we conclude that $a_1 = \cdots = a_l = 0$. Similarly, we also have $b_1 = \cdots = b_m = c_1 = \cdots = c_k = 0$, proving that $\{\bar{x}_1, \cdots, \bar{x}_l; \bar{z}_1, \cdots, \bar{z}_k; \bar{y}_1, \cdots, \bar{y}_m\}$ is linearly independent. □

Definition 2.3.9: (Orthogonal Vector Spaces)

Two vector spaces $\mathcal{U} \subset \mathcal{R}^n$ and $\mathcal{W} \subset \mathcal{R}^n$ are orthogonal if and only if for any pair of vectors $\bar{u} \in \mathcal{U}$ and $\bar{w} \in \mathcal{W}$, the dot product of the two vectors is 0.

If $\bar{u} \in \mathcal{U} \cap \mathcal{W}$, then $\bar{u} \cdot \bar{u} = 0$ implies that $\bar{u} = \bar{0}$. Thus two orthogonal vector subspaces must be disjoint.
Examples of orthogonal vector spaces in $\mathcal{R}^3$ include two orthogonal lines through the origin, or a line and a plane through the origin that are orthogonal to each other.

Definition 2.3.10: (Orthogonal Complementary Subspace)

Let $\mathcal{U}$ be a subspace of $\mathcal{R}^n$. Then, $\mathcal{W}$ is an orthogonal complementary subspace of $\mathcal{U}$ if and only if it satisfies the following properties:

- The spaces $\mathcal{U}$ and $\mathcal{W}$ are orthogonal (and therefore disjoint).
- The union of the basis sets of $\mathcal{U}$ and $\mathcal{W}$ forms a basis for $\mathcal{R}^n$.

Examples of orthogonal vector spaces in $\mathcal{R}^3$ include two orthogonal lines through the origin, or a line and a plane through the origin that are orthogonal to each other. In the latter case the orthogonal line and plane are orthogonal complementary subspaces. Since the union of the basis sets of $\mathcal{U}$ and $\mathcal{W}$ is linearly independent, the condition here implies that it must also span $\mathcal{R}^n$. As a consequence, any vector $\bar{v} \in \mathcal{R}^n$ can be written as $\bar{u} + \bar{w}$, with $\bar{u} \in \mathcal{U}$, $\bar{w} \in \mathcal{W}$, $\bar{u} \cdot \bar{w} = 0$; furthermore, this decomposition is unique.
In fact any subspace $\mathcal{U}$ has a uniquely defined orthogonal complement defined as $\mathcal{U}^\perp = \{\bar{x} : \bar{x} \cdot \bar{u} = 0 \text{ for any } \bar{u} \in \mathcal{U}\}$. $\mathcal{U}^\perp$ is certainly a subspace orthogonal to $\mathcal{U}$. It remains to show that any vector $\bar{v} \in \mathcal{R}^n$ can be written as $\bar{u} + \bar{w}$, with $\bar{u} \in \mathcal{U}$, $\bar{w} \in \mathcal{U}^\perp$. This assertion is a consequence of the least square solution: use a basis of $\mathcal{U}$ as columns to form matrix A, then $\bar{u} = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\bar{v} \in \mathcal{U}$ for any $\bar{v} \in \mathcal{R}^n$ and $\bar{w} = \bar{v} - \bar{u} \perp \mathcal{U}$, for the columns of A are a basis of $\mathcal{U}$, and

$$A^{\mathrm{T}}\bar{w} = A^{\mathrm{T}}\left(\bar{v} - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\bar{v}\right) = A^{\mathrm{T}}\bar{v} - A^{\mathrm{T}}\bar{v} = \bar{0},$$

showing that $\bar{w} \in \mathcal{U}^\perp$.
We summarize this as

Proposition 2.3.5: Orthogonal Complement

For any vector space $\mathcal{U}$ of $\mathcal{R}^n$, $\mathcal{U}$ and $\mathcal{U}^\perp$ are orthogonal complement. As a consequence, any vector $\bar{v} \in \mathcal{R}^n$ can be written uniquely as $\bar{u} + \bar{w}$, with $\bar{u} \in \mathcal{U}$, $\bar{w} \in \mathcal{U}^\perp$; if $\bar{v} \perp \mathcal{U}^\perp$, then $\bar{v} \in \mathcal{U}$. The latter property is encoded in $(\mathcal{U}^\perp)^\perp = \mathcal{U}$.

Problem 2.3.2 Consider two disjoint vector spaces in $\mathcal{R}^3$ with basis sets $\{[1,1,1]^{\mathrm{T}}\}$ and $\{[1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}\}$, respectively. Express the vector $[0,1,1]^{\mathrm{T}}$ as the sum of two vectors, such that each of them belongs to one of the two spaces. Note that you will have to solve a system of three linear equations to solve this problem.

Problem 2.3.3 Let $\mathcal{U} \subset \mathcal{R}^3$ be defined by the basis set $\{[1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}\}$. State the basis sets of two possible complementary subspaces of $\mathcal{U}$. In each case, provide a decomposition of the vector $[1,1,1]^{\mathrm{T}}$ as a sum of vectors from these complementary subspaces.

Problem 2.3.4 Let $\mathcal{U} \subset \mathcal{R}^3$ be defined by the basis set $\mathcal{B} = \{[1,1,1]^{\mathrm{T}}, [1,-1,1]^{\mathrm{T}}\}$. Formulate a system of equations to find the orthogonal complementary subspace $\mathcal{W}$

of $\mathcal{U}$. Use the orthogonality of $\mathcal{U}$ and $\mathcal{W}$ to propose a fast method to express the vector $[2, 2, 1]^{\mathrm{T}}$ as a sum of vectors from these complementary subspaces.

2.4 The Linear Algebra of Matrix Rows and Columns

Definition 2.4.1: (Row Space and Column Space)

For an $n \times d$ matrix A, its column space is defined as the vector space spanned by its columns, and it is a subspace of $\mathcal{R}^n$. The row space of A is defined as the vector space spanned by the columns of A^{T} (which are simply the transposed rows of A). The row space of A is a subspace of $\mathcal{R}^d$.

We note the row space of A by $\mathrm{Row}(A)$ and the column space of A by $\mathrm{Col}(A)$.

Definition 2.4.2: (Null Space)

The null space of an $n \times d$ matrix A is the subspace of $\mathcal{R}^d$ containing all column vectors $\bar{x} \in \mathcal{R}^d$, such that $A\bar{x} = \bar{0}$. We denote this subspace by $\mathrm{Null}(A)$.

Note that $\mathrm{Null}(A)$ is the orthogonal complement of $\mathrm{Row}(A)$ in $\mathcal{R}^d$, so

$$\dim(\mathrm{Null}(A)) + \dim(\mathrm{Row}(A)) = d.$$

A fundamental result in linear algebra is that the dimensionality of the row space (also referred to as row rank) and that of the column space (also referred to as column rank) of any $n \times d$ matrix A is the same.
This would follow if one can show that $\dim(\mathrm{Null}(A)) + \dim(\mathrm{Col}(A)) = d$. This is seen by noting that $\dim(\mathrm{Null}(A))$ equals the number of free variables in solving $A\bar{x} = \bar{0}$ and $\dim(\mathrm{Col}(A))$ equals the number of pivot columns of A, as those columns form a basis of $\mathrm{Col}(A)$.

Definition 2.4.3: (Left Null Space)

The left null space of an $n \times d$ matrix A is the subspace of R^n containing all column vectors $\bar{x} \in \mathcal{R}^n$, such that $\bar{x}^{\mathrm{T}}A = \bar{0}$, equivalently $A^{\mathrm{T}}\bar{x} = \bar{0}$. The left null space of A is the orthogonal complementary subspace of the column space of A.

The left null space of A is really $\text{Null}(A^{\text{T}})$ and $\text{Row}(A^{\text{T}}) = \text{Col}(A)$. Then we have

$$\dim(\text{Null}(A^{\text{T}})) + \dim(\text{Row}(A^{\text{T}})) = n$$

The row space, column space, the right null space, and the left null space are referred to as the *four fundamental subspaces of linear algebra.*

Problem 2.4.1 Find a basis for each of the right and left null spaces of matrix

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Problem 2.4.2 For any $n \times d$ matrix A, show why the matrices $P = A^{\text{T}}A + \lambda I_d$ and $Q = AA^{\text{T}} + \lambda I_n$ always have an empty null space for any $\lambda > 0$.

2.5 The Row Echelon Form of a Matrix

Definition 2.5.1: (Row and Column Equivalence)

Two matrices are said to be row equivalent, if one matrix is obtained from the other by a sequence of elementary row operations such as row interchange, row addition, or multiplication of a row with a non-zero scalar. Similarly, two matrices are said to be column equivalent, if one matrix is obtained from the other with a sequence of elementary column operations.

Lemma 2.5.1

Elementary row operations do not change the vector space spanned by the rows, whereas elementary column operations do not change the vector space spanned by the columns.

Definition 2.5.2: (Row Echelon Form)

An $n \times d$ matrix A is said to be in row echelon form, if and only if (i) the leftmost non-zero entry in each non-zero row is 1, (ii) the column index of the

leftmost non-zero entry in each row increases with row index, and (iii) all zero rows (if any) occur at the bottom of the matrix.

The increasing column index of the leading non-zero entry ensures that *non-zero rows in echelon form are always linearly independent.*
Every elementary row operation on an $n \times d$ matrix A is accomplished by left multiplication of A by an $n \times n$ elementary matrix E.
If we reduce a given $n \times d$ matrix A to an echelon form R, we can express their relation as $E_k \cdots E_1 A = R$, where E_i is the elementary matrix representing the elementary row operation in step i.
Here is a useful
Observation: If matrix B is obtained from A through one or more steps of elementary row operations, then $\text{Row}(A) = \text{Row}(B)$ and $\text{Null}(A) = \text{Null}(B)$, though $\text{Col}(A)$ may not be the same as $\text{Col}(B)$—they do have the same dimensions.

Example 2.5.1

$$\text{rref}(A) = \begin{bmatrix} 1 & 2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

is the reduced row echelon form of

$$A = \begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 6 & 12 & 2 & 0 & 28 \\ 3 & 6 & 2 & 1 & 23 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix}.$$

Thus $\text{Row}(\text{rref}(A)) = \text{Row}(A)$. Evidently $\{[1,2,0,0,3],[0,0,1,0,5],[0,0,0,1,4]\}$ is a basis for $\text{Row}(\text{rref}(A))$. This does not mean that the first three rows of A form a basis of $\text{Row}(A)$—the second row is 2-times the first row of A so the first three rows of A are not linearly independent. If we insist on picking a basis of $\text{Row}(A)$ from the rows of A, since we know the dimension of $\text{Row}(A)$ is 3, then rows $1, 3, 4$ of A must form a basis.
Note also that $\text{Col}(A) \neq \text{Col}(\text{rref}(A))$. However, A and $\text{rref}(A))$ have identical pivot columns. We thus conclude that columns $1, 3, 4$ of A are pivot columns of A and form a basis for $\text{Col}(A)$. $\text{Null}(A) = \text{Null}(\text{rref}(A))$ and a basis can be found by finding the general solution of $\text{rref}(A)\bar{x} = \bar{0}$.

To find the elementary matrices that reduce A to $\text{rref}(A)$, note that

$$\begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 6 & 12 & 2 & 0 & 28 \\ 3 & 6 & 2 & 1 & 23 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix} \xrightarrow{R_2 - 2R_1 \mapsto R_2, R_3 - R_1 \mapsto R_3}$$

$$\begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 9 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix} \xrightarrow{R_4 - 2R_1/3 \mapsto R_4}$$

$$\begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 9 \\ 0 & 0 & 4/3 & 1 & 32/3 \end{bmatrix} \xrightarrow{R_4 - 3R_3/4 \mapsto R_4}$$

$$\begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 9 \\ 0 & 0 & 0 & -1/3 & -4/3 \end{bmatrix} \xrightarrow{R_2 \longleftrightarrow R_4, R_2 \longleftrightarrow R_3, -3R_3 \mapsto R_3}$$

$$\begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 0 & 0 & 1 & 1 & 9 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

This results in an upper triangular matrix. These 7 steps of elementary row operations can be recorded as left multiplications by 7 elementary matrices:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -3/4 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -2/3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 14 \\ 6 & 12 & 2 & 0 & 28 \\ 3 & 6 & 2 & 1 & 23 \\ 2 & 4 & 2 & 1 & 20 \end{bmatrix}.$$

An elementary matrix corresponding to a lower row adding a multiple of a higher row is a lower triangular matrix. The above operations involve two that are not of this type. If one pre-left-multiply A by the last two permutation

matrices, then one can use only lower triangular elementary matrices to left-multiply the resulting matrix to reduce it to a row echelon form, which is an upper triangular matrix.

$$\begin{bmatrix}1&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&0&0&1\\0&0&1&0\\0&1&0&0\end{bmatrix}\begin{bmatrix}3&6&1&0&14\\6&12&2&0&28\\3&6&2&1&23\\2&4&2&1&20\end{bmatrix}$$
$$=\begin{bmatrix}3&6&1&0&14\\3&6&2&1&23\\2&4&2&1&20\\6&12&2&0&28\end{bmatrix}$$

Then

$$\begin{bmatrix}3&6&1&0&14\\3&6&2&1&23\\2&4&2&1&20\\6&12&2&0&28\end{bmatrix}\xrightarrow{R_2-R_1\mapsto R_2,\,R_3-2R_1/3\mapsto R_3,\,R_4-2R_1\mapsto R_4}$$
$$\begin{bmatrix}3&6&1&0&14\\0&0&1&1&9\\0&0&4/3&1&32/3\\0&0&0&0&0\end{bmatrix}\xrightarrow{R_3-4R_2/3\mapsto R_3}$$
$$\begin{bmatrix}3&6&1&0&14\\0&0&1&1&9\\0&0&0&-1/3&-4/3\\0&0&0&0&0\end{bmatrix}=:U.$$

This can be recorded as

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&-4/3&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\-2&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\-2/3&0&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\-1&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}3&6&1&0&14\\3&6&2&1&23\\2&4&2&1&20\\6&12&2&0&28\end{bmatrix}$$
$$=\begin{bmatrix}3&6&1&0&14\\0&0&1&1&9\\0&0&0&-1/3&-4/3\\0&0&0&0&0\end{bmatrix}$$

If we need to obtain its reduced row echelon form, we need to perform further elementary row operations to make the entries above each pivot 0 and make each pivot equal 1. But the elementary matrices representing these operations are now upper triangular matrices.
Note that

$$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & -4/3 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ -2 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ -2/3 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ -1 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}$$

is a lower triangular matrix. Call it L_1. We thus have $L_1PA = U$, where

$$P = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}$$

is a permutation matrix — PA simply permutes the rows of A.

2.5.1 LU Decomposition

When it is possible to perform a sequence of elementary row operations using only row addition operations from a higher row to a lower row and row scaling operations, then each E_i in the relation $E_k \cdots E_1 A = R$ is a lower triangular matrix (why?). It follows that E_i^{-1} and $(E_k \cdots E_1)^{-1} = E_1^{-1} \cdots E_k^{-1}$ are also lower triangular matrices. We thus conclude that $A = (E_k \cdots E_1)^{-1} R$ has the form of the product of a lower triangular matrix $L = (E_k \cdots E_1)^{-1}$ and an upper triangular matrix $U = R$. This is called an *LU decomposition* of A.

Note that L is a square invertible matrix while U may not be square or invertible.

In Example 2.5.1, we have $L_1PA = U$, from which we obtain $PA = LU$, with

$L = L_1^{-1}$:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2/3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -4/3 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2/3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 4/3 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2/3 & 4/3 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}$$

Note that the entry at (i, j) of L, $i > j$, is simply the opposite of the multiplier in the elementary operation that makes the entry at (i, j) equal 0 using the pivot in row j of PA. A common application of an LU decomposition of $A = LU$ is in solving $A\bar{x} = \bar{b}$: we set $\bar{y} = U\bar{x}$, then $L\bar{y} = \bar{b}$ can be solved by *forward substitution* starting from the first equation, and with $\bar{y}$ solved, we can find $\bar{x}$ through $U\bar{x} = \bar{y}$ by *backward substitution* starting from the bottom non-zero row of U—if this row has non-zero elements behind the first non-zero element (which is called a pivot), then this system will have some *free variables*; however, $U\bar{x} = \bar{y}$ may not have a solution if U has some rows of zero vectors for which the component(s) of $\bar{y} \neq 0$.

However, it is not always possible to create an LU decomposition of a matrix without permuting its rows first, as sometimes row interchange operations are unavoidable in applying Gauss elimination. In such cases, it is possible to pre-permute some rows of A so that the resulting matrix has an LU decomposition. This is expressed by identifying a permutation matrix P such that $PA = LU$. The discussion in the textbook using $PL_m \cdots L_1 A = U$ is faulty, at least calling for clarify, as one may need to apply some permutation first to carry out the Gauss elimination.

Suppose we need to solve $A\bar{x} = \bar{b}$, with A the matrix in Example 2.5.1. $PA = LU$ has a LU-decomposition, so we should solve $LU\bar{x} = P\bar{b}$. Set $\bar{y} = U\bar{x}$, then $L\bar{y} = P\bar{b}$ is simply

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2/3 & 4/3 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_3 \\ b_4 \\ b_2 \end{bmatrix}$$

which can be easily solved by forward substitution. Once we have solved for $\bar{y}$, we can solve for $\bar{x}$ by backward substitution.

2.5.2 Application: Finding a Basis Set

Here, the textbook gives a discussion on how to find a basis for Span$\{\bar{a}_1, \cdots, \bar{a}_n\}$, where each $\bar{a}_i \in \mathcal{R}^d$, using Gaussian elimination applied to the matrix A whose rows are the $\bar{a}_i$'s. Note that the vectors in the basis are the non-zero row vectors of the row echelon form, and are not necessarily a subset of $\{\bar{a}_1, \cdots, \bar{a}_n\}$.
To select a subset of $\{\bar{a}_1, \cdots, \bar{a}_n\}$ as a basis for Span$\{\bar{a}_1, \cdots, \bar{a}_n\}$, one should use these vectors as columns to form a $d \times n$ matrix B and apply elementary row operations to reduce it to a row echelon form R. Using the property that $B\bar{x} = \bar{0}$ if and only if $R\bar{x} = \bar{0}$ we see that columns of B corresponding to the pivot columns of R form a basis of Span$\{\bar{a}_1, \cdots, \bar{a}_n\}$—even though Col$(R)$ may not be the same as Col$(B) =$ Span$\{\bar{a}_1, \cdots, \bar{a}_n\}$.
The beginning part of Example 2.5.1 illustrates how this is worked out.

2.5.3 Application: Matrix Inversion

When A is an $n \times n$ invertible matrix, it will have a pivot in each row, so there must be a sequence of elementary matrices $E_1, \cdots, E_k$ such that

$$E_k \cdots E_1 A = I_n.$$

This shows that $E_k \cdots E_1$ must be the inverse of A. But $E_k \cdots E_1$ is also the outcome of applying the same sequence of row operations on I_n. This explains the standard procedure for finding A^{-1} by reducing the *augmented* matrix $[A, I_n]$ via elementary row operations until the first $n \times n$ block becomes I_n, then the remaining $n \times n$ block would be A^{-1}.

2.5.4 Application: Solving a System of Linear Equations

This subsection summarizes what could happen in trying to solve a linear system $A\bar{x} = \bar{b}$.
Depending on A and $\bar{b}$, three cases arise:

1. If the vector $\bar{b}$ does not occur in the column space of A, then no solution exists to this system of linear equations although best fits are possible. This case is studied in detail in Section 2.8.

2. If the vector $\bar{b}$ occurs in the column space of A, and A has linearly independent columns (which implies that the columns form the basis of a d-dimensional subspace of R^n), the solution is unique. This result is based on the uniqueness of coordinates (cf. Lemma 2.3.2). In the special case that A is square, the solution is simply $\bar{x} = A^{-1}\bar{b}$.

3. If the vector $\bar{b}$ occurs in the column space of A and the columns of A are linearly dependent, then an infinite number of solutions exists to $A\bar{x} = \bar{b}$, as there will be free parameters in solving $A\bar{z} = \bar{0}$, which produces an infinite number of solutions to $A\bar{z} = \bar{0}$. Note that if $\bar{x}_0$ is a solution of $A\bar{x}_0 = \bar{b}$ and $\bar{z}$ is a solution of $A\bar{z} = \bar{0}$, then $\bar{x} = \bar{x}_0 + \bar{z}$ is also a solution of $A\bar{x} = \bar{b}$; in fact any solution of $A\bar{x} = \bar{b}$ can be written in the form of $\bar{x} = \bar{x}_0 + \bar{z}$ for some $\bar{z}$ solving $A\bar{z} = \bar{0}$.

Problem 2.5.1 Suppose that no solution exists to the system of equations $A\bar{x} = \bar{b}$, where A is an $n \times d$ matrix and $\bar{b}$ is an n-dimensional column vector. Show that an n-dimensional column vector $\bar{z}$ must exist that satisfies $\bar{z}^{\mathrm{T}}A = \bar{0}$ and $\bar{z}^{\mathrm{T}}\bar{b} \neq \bar{0}$.

Using Proposition 2.3.5 we can formulate a criterion for the solvability of $A\bar{x} = \bar{b}$.

Proposition 2.5.1: Solvability criterion of $A\bar{x} = \bar{b}$

$A\bar{x} = \bar{b}$ has a solution if and only if $\bar{b} \in (\mathrm{Null}(A^{\mathrm{T}}))^{\perp}$, namely, $\bar{b} \perp \bar{v}$ for all $\bar{v}$ such that $A^{\mathrm{T}}\bar{v} = \bar{0}$.

Proof. $A\bar{x} = \bar{b}$ has a solution if and only if $\bar{b} \in \mathrm{Col}(A)$. But $(\mathrm{Col}(A))^{\perp} = \mathrm{Null}(A^{\mathrm{T}})$ so $\mathrm{Col}(A) = \left(\mathrm{Null}(A^{\mathrm{T}})\right)^{\perp}$. □

2.6 The Notion of Matrix Rank

Any $n \times d$ matrix A has a well defined row space and column space and their respective dimensions. A fundamental result of linear algebra says

Lemma 2.6.1

$$\dim(\mathit{Row}(A)) = \dim(\mathit{Col}(A)).$$

This number is called the rank of A.

Proof. Let B be a row echelon form of A, then there exists an invertible matrix R such that $RA = B$. This relation means that each row of B is a linear combination of rows of A. Since $A = R^{-1}B$, we see that each row of A is a linear combination of rows of B, thus $\mathrm{Row}(A) = \mathrm{Row}(B)$. On the other hand $A\bar{x} = \bar{0}$ if and only if $B\bar{x} = \bar{0}$. Thus $\mathrm{Null}(A) = \mathrm{Null}(B)$. Earlier we have identified $\mathrm{Null}(A) = (\mathrm{Row}(A))^{\perp}$, so

$$\dim(\mathrm{Row}(A)) + \dim(\mathrm{Null}(A)) = \dim(\mathrm{Row}(B)) + \dim(\mathrm{Null}(B)) = d.$$

But $\dim(\text{Null}(B))$ equals the number of free variables in solving $B\bar{x} = \bar{0}$, and for each free variable, say x_j, we can set $x_j = 1$ and the rest of the free variables to 0, which gives column j of B as a linear combination of the pivot columns of B, but the same solution also gives column j of A as a linear combination of the pivot columns of A. This shows that the pivot columns of A forms a basis of $\text{Col}(A)$ and the pivot columns of B forms a basis of $\text{Col}(B)$; furthermore, $\dim(\text{Col}(B)) = \dim(\text{Row}(B))$ equals the number of pivots of B. Since A and B have equal number of pivot columns, we conclude that $\dim \text{Col}(A) = \dim \text{Col}(B) = \dim \text{Row}(B) = \dim(\text{Row}(A))$. □

Remark 2.6.1

One ingredient of the proof is that whenever $B = RA$ for an invertible R, then $Row(A) = Row(B)$, therefore, A and B have equal rank. Likewise whenever $B = AC$ for an invertible C, then A and B have the same column spaces, therefore A and B have equal rank.
Thus left-multiplying and right-multiplying a matrix by an invertible matrix does not change its rank. The textbook shows that any matrix A can be reduced to a diagonal matrix Δ through a sequence of elementary row and column operations. This results in $RAC = \Delta$ for some invertible R and C. It follows that

$$rank(A) = rank(RA) = rank(RAC) = rank(\Delta).$$

Corollary 2.6.1

The rank of an $n \times d$ matrix is at most $\min\{n, d\}$.

Corollary 2.6.2

Consider an $n \times d$ matrix A with rank $k \leq \min\{n, d\}$. Then the dimension of the null space of A is $d - k$ and the dimension of the left null space of A is $n - k$.

Corollary 2.6.3: (Matrix Multiplication Upper Bound)

Let A and B be two matrices with ranks a and b, respectively. Then, the rank of AB is at most $\min\{a, b\}$.

Proof. We note that $\text{Col}(AB)$ is a subspace of $\text{Col}(A)$, so $\text{rank}(AB) \leq \text{rank}(A)$. Similarly, $\text{Row}(AB)$ is a subspace of $\text{Row}(B)$, so $\text{rank}(AB) \leq \text{rank}(B)$. □

2.6.1 Effect of Matrix Operations on Rank

Lemma 2.6.2: (Matrix Addition Upper Bound)

Let A and B be two matrices with ranks a and b, respectively. Then, the rank of $A + B$ is at most $a + b$.

Lemma 2.6.3: (Matrix Addition Lower Bound)

et A and B be two matrices with ranks a and b, respectively. Then, the rank of $A + B$ is at least $|a - b|$.

Lemma 2.6.4: (Matrix Multiplication Lower Bound)

Let A and B be $n \times d$ and $d \times k$ matrices of ranks a and b, respectively. Then, the rank of AB is at least $a + b - d$.

Corollary 2.6.4

Multiplying a matrix A with a square matrix B of full rank does not change the rank of matrix A.

Corollary 2.6.5

Let A and B be two square matrices. Then AB is non-singular if and only if A and B are both non-singular.

Lemma 2.6.5: (Linear Independence and Gram Matrix)

The matrix $A^T A$ is said to be the Gram matrix of the column space of an $n \times d$ matrix A. The columns of the matrix A are linearly independent if and only if $A^T A$ is invertible.

Proof. Consider the case where $A^{\mathrm{T}}A$ is invertible. This means that the rank of $A^{\mathrm{T}}A$ is d, and therefore the rank of each of the factors of $A^{\mathrm{T}}A$ must also be at least d. This means that A must have rank at least d, which is possible only when the d columns of A are linearly independent. Note that part of the consequence of A having rank at least d is that $n \geq d$.
Now suppose that A has linearly independent columns. We prove that the only solution $\bar{x}$ of $A^{\mathrm{T}}A\bar{x} = \bar{0}$ is $\bar{x} = \bar{0}$, thereby showing that its rank is d and it is invertible.

Suppose that $A^{\mathrm{T}}A\bar{x} = \bar{0}$. Then $\bar{x}^{\mathrm{T}}A^{\mathrm{T}}A\bar{x} = 0$, which implies that $(A\bar{x})^{\mathrm{T}}(A\bar{x}) = 0$. Thus $A\bar{x} = \mathbf{z}$. But A has linearly independent columns, so $\bar{x} = \bar{0}$.

□

Remark 2.6.2

Buried in the argument above is the following useful fact for a general matrix A:

$$Null(A^T A) = Null(A).$$

2.7 Generating Orthogonal Basis Sets

2.7.1 Gram-Schmidt Orthogonalization and QR Decomposition

We aim to find an orthonormal basis set of the span of $\mathcal{A} = \{\bar{a}_1, \cdots, \bar{a}_d\}$ in $\mathcal{R}^n$. We will use the *Gram-Schmidt* orthogonalization. The basic idea of Gram-Schmidt orthogonalization is to successively remove the projections of previously generated vectors from a vector belonging to $\mathcal{A}$ to iteratively create orthogonal vectors.
The case of starting with a basis of two vectors is the simplest: if $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a basis for a subspace V, and $\mathbf{w}$ is the orthogonal projection of $\mathbf{u}_2$ along the line of $\mathbf{u}_1$, then $\mathbf{z} = \mathbf{u}_2 - \mathbf{w} \perp \mathbf{u}_1$ and $\neq \mathbf{0}$, for, if $\mathbf{z} = \mathbf{0}$, it would mean $\mathbf{u}_2 = \mathbf{w}$, which is a multiple of $\mathbf{u}_1$, and this would contradict the assumption that $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a basis for a subspace V. So if we define $\{\mathbf{v}_1 = \mathbf{u}_1, \mathbf{v}_2 = \mathbf{z} = \mathbf{u}_2 - \mathbf{w}\}$, then $\{\mathbf{v}_1, \mathbf{v}_2\}$ is orthogonal to each other, therefore linearly independent, and forms an orthogonal basis of V.
The relation between $\{\mathbf{u}_1, \mathbf{u}_2\}$ and $\{\mathbf{v}_1, \mathbf{v}_2\}$ can be written as

$$\begin{aligned}\mathbf{u}_1 &= \mathbf{v}_1 \\ \mathbf{u}_2 &= \mathbf{w} + \mathbf{v}_2 = \left(\frac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right)\mathbf{v}_1 + \mathbf{v}_2\end{aligned}$$

If we create a matrix with $\mathbf{u}_1, \mathbf{u}_2$ as its first and second columns, and another one with $\mathbf{v}_1, \mathbf{v}_2$ as its first and second columns, then the above relations can be written as

$$[\mathbf{u}_1\ \mathbf{u}_2] = [\mathbf{v}_1\ \mathbf{v}_2]\begin{bmatrix}1 & \frac{\mathbf{u}_2\cdot\mathbf{v}_1}{\mathbf{v}_1\cdot\mathbf{v}_1} \\ 0 & 1\end{bmatrix}.$$

The two columns of the matrix $[\mathbf{v}_1\ \mathbf{v}_2]$ are orthogonal to each other, and this property can be encoded in

$$[\mathbf{v}_1\ \mathbf{v}_2]^{\mathrm{T}}[\mathbf{v}_1\ \mathbf{v}_2] = \begin{bmatrix}\mathbf{v}_1 \cdot \mathbf{v}_1 & 0 \\ 0 & \mathbf{v}_2 \cdot \mathbf{v}_2\end{bmatrix}$$

being a diagonal matrix.
If we set $\mathbf{w}_1$ to be the unit vector in the direction of $\mathbf{v}_1 : \mathbf{w}_1 = \mathbf{v}_1/\|\mathbf{v}_1\|$, and $\mathbf{w}_2$ to be the unit vector in the direction of $\mathbf{v}_2 : \mathbf{w}_2 = \mathbf{v}_2/\|\mathbf{v}_2\|$, then we find $\left(\frac{\mathbf{u}_2\cdot\mathbf{v}_1}{\mathbf{v}_1\cdot\mathbf{v}_1}\right)\mathbf{v}_1 = (\mathbf{u}_2\cdot\mathbf{w}_1)\,\mathbf{w}_1$, and the relation between $\{\mathbf{u}_1,\mathbf{u}_2\}$ and $\{\mathbf{v}_1,\mathbf{v}_2\}$ can be turned into a relation between $\{\mathbf{u}_1,\mathbf{u}_2\}$ and $\{\mathbf{w}_1,\mathbf{w}_2\}$ as follows:

$$\begin{aligned}\mathbf{u}_1 &= \|\mathbf{v}_1\|\mathbf{w}_1\\ \mathbf{u}_2 &= (\mathbf{u}_2\cdot\mathbf{w}_1)\,\mathbf{w}_1 + \|\mathbf{v}_2\|\mathbf{w}_2\end{aligned}$$

and in matrix form we have

$$[\mathbf{u}_1\,\mathbf{u}_2] = [\mathbf{w}_1\,\mathbf{w}_2]\begin{bmatrix}\|\mathbf{v}_1\| & \mathbf{u}_2\cdot\mathbf{w}_1\\ 0 & \|\mathbf{v}_2\|\end{bmatrix}. \tag{2.1}$$

The matrix $[\mathbf{w}_1\,\mathbf{w}_2]$ has an even nicer property: its columns are *orthonormal*, and that property is now encoded in

$$[\mathbf{w}_1\,\mathbf{w}_2]^{\mathrm{T}}[\mathbf{w}_1\,\mathbf{w}_2] = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}. \tag{2.2}$$

The discussion above works for $\mathbf{u}_1,\mathbf{u}_2 \in \mathbb{R}^n, n \geq 2$, then $[\mathbf{u}_1\,\mathbf{u}_2], [\mathbf{w}_1\,\mathbf{w}_2]$ are $n\times 2$ matrices. Another way of expressing the relation between $\{\mathbf{u}_1,\mathbf{u}_2\}$ and $\{\mathbf{w}_1\,\mathbf{w}_2\}$ is that the latter is an orthonormal basis in the span of $\{\mathbf{u}_1,\mathbf{u}_2\}$; in terms of matrices, it is an orthonormal basis of the column space of $[\mathbf{u}_1,\mathbf{u}_2]$.
When $n = 2$, $[\mathbf{w}_1\,\mathbf{w}_2]$ is a square matrix, then (2.2) implies that $[\mathbf{w}_1\,\mathbf{w}_2]^{\mathrm{T}}$ is the inverse matrix of $[\mathbf{w}_1\,\mathbf{w}_2]$, thus we also have

$$[\mathbf{w}_1\,\mathbf{w}_2][\mathbf{w}_1\,\mathbf{w}_2]^{\mathrm{T}} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

which implies that the columns of $[\mathbf{w}_1\,\mathbf{w}_2]^{\mathrm{T}}$, namely, the rows of $[\mathbf{w}_1\,\mathbf{w}_2]$ are also orthonormal. A square matrix satisfying (2.2) is called an *orthogonal** matrix. (2.1) is called a QR-factorization of the matrix $[\mathbf{u}_1\,\mathbf{u}_2]$, with Q referring to the orthogonal matrix $[\mathbf{w}_1\,\mathbf{w}_2]$, and R referring to the upper triangular matrix $\begin{bmatrix}\|\mathbf{v}_1\| & \mathbf{u}_2\cdot\mathbf{w}_1\\ 0 & \|\mathbf{v}_2\|\end{bmatrix}$.

*One would expect such a matrix should be called an orthonormal matrix, but it has become custom to call such a matrix an orthogonal matrix.

Example 2.7.1

Apply the Gram-Schmidt process to the set of vectors $\left\{\begin{bmatrix}2\\-1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\}$ to obtain a set of orthogonal basis, then turn it into an orthonormal basis. Finally, record the relations between the given vectors and the set of orthonormal vectors obtained as a QR factorization.

According to the Gram-Schmidt process, we set $\mathbf{v}_1 = \begin{bmatrix}2\\-1\end{bmatrix}$, and

$$\mathbf{v}_2 = \begin{bmatrix}1\\1\end{bmatrix} - \left(\frac{\begin{bmatrix}1\\1\end{bmatrix}\cdot\mathbf{v}_1}{\mathbf{v}_1\cdot\mathbf{v}_1}\right)\mathbf{v}_1 = \begin{bmatrix}1\\1\end{bmatrix} - \frac{1}{5}\begin{bmatrix}2\\-1\end{bmatrix} = \begin{bmatrix}3/5\\6/5\end{bmatrix}.$$

Thus $\left\{\begin{bmatrix}2\\-1\end{bmatrix}, \begin{bmatrix}3/5\\6/5\end{bmatrix}\right\}$ is the set of orthogonal vectors obtained by applying directly the Gram-Schmidt process.

To get a set of orthonormal vectors, we normalize the orthogonal vectors to make each a unit vector:

$$\mathbf{w}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \frac{\mathbf{v}_1}{\sqrt{5}} = \begin{bmatrix}\frac{2}{\sqrt{5}}\\-\frac{1}{\sqrt{5}}\end{bmatrix};$$

$$\mathbf{w}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \frac{\mathbf{v}_2}{\sqrt{(3/5)^2+(6/5)^2}}.$$

But we could have computed $\mathbf{w}_2$ as the unit vector in the direction of $5/3\,\mathbf{v}_2 = \begin{bmatrix}1\\2\end{bmatrix}$, which gives $\begin{bmatrix}\frac{1}{\sqrt{5}}\\\frac{2}{\sqrt{5}}\end{bmatrix}$.

The relations between the given vectors and $\{\mathbf{w}_1, \mathbf{w}_2\}$ are recorded as

$$\begin{bmatrix}2\\-1\end{bmatrix} = \sqrt{5}\mathbf{w}_1 \qquad\qquad = \sqrt{5}\begin{bmatrix}\frac{2}{\sqrt{5}}\\-\frac{1}{\sqrt{5}}\end{bmatrix};$$

$$\begin{bmatrix}1\\1\end{bmatrix} = \frac{1}{5}\mathbf{v}_1 + \mathbf{v}_2 = \frac{1}{\sqrt{5}}\mathbf{w}_1 + \frac{3}{\sqrt{5}}\mathbf{w}_2 \quad = \frac{1}{\sqrt{5}}\begin{bmatrix}\frac{2}{\sqrt{5}}\\-\frac{1}{\sqrt{5}}\end{bmatrix} + \frac{3}{\sqrt{5}}\begin{bmatrix}\frac{1}{\sqrt{5}}\\\frac{2}{\sqrt{5}}\end{bmatrix}.$$

In terms of the matrix with these vectors as columns we get

$$\begin{bmatrix}2 & 1\\-1 & 1\end{bmatrix} = \begin{bmatrix}\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}}\\-\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}}\end{bmatrix}\begin{bmatrix}\sqrt{5} & \frac{1}{\sqrt{5}}\\0 & \frac{3}{\sqrt{5}}\end{bmatrix}.$$

Here

$$Q = \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}$$

is an orthogonal matrix with

$$Q^{\mathrm{T}}Q = \begin{bmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 1\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$R = \begin{bmatrix} \sqrt{5} & \frac{1}{\sqrt{5}} \\ 0 & \frac{3}{\sqrt{5}} \end{bmatrix}$$

is an upper triangular matrix in the QR factorization.

For the case of a basis of three vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, the construction for $\{\mathbf{v}_1, \mathbf{v}_2\}$ is the same:

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{u}_1 \\ \mathbf{v}_2 &= \mathbf{u}_2 - \left(\frac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right) \mathbf{v}_1. \end{aligned}$$

To get $\mathbf{v}_3$, we choose it as $\mathbf{u}_3 - \mathbf{w}$, where $\mathbf{w}$ is the orthogonal projection of $\mathbf{u}_2$ in $\mathrm{Span}\{\mathbf{u}_1, \mathbf{u}_2\}$. The key here is, instead of writing $\mathbf{w}$ as a linear combination of $\mathbf{u}_1, \mathbf{u}_2$, we write it as a linear combination of $\mathbf{v}_1, \mathbf{v}_2$, as $\mathrm{Span}\{\mathbf{u}_1, \mathbf{u}_2\} = \mathrm{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal to each other!
Thus we write $\mathbf{w} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2$, and it is chosen so that $\mathbf{u}_3 - \mathbf{w} \perp \mathbf{v}_1, \mathbf{v}_2$, namely,

$$\begin{aligned} (\mathbf{u}_3 - [c_1\mathbf{v}_1 + c_2\mathbf{v}_2]) \cdot \mathbf{v}_1 = 0, \\ (\mathbf{u}_3 - [c_1\mathbf{v}_1 + c_2\mathbf{v}_2]) \cdot \mathbf{v}_2 = 0. \end{aligned}$$

Using $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, we find

$$\begin{aligned} c_1 &= \frac{\mathbf{u}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}, \\ c_2 &= \frac{\mathbf{u}_3 \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2}. \end{aligned}$$

So we now conclude that

$$\mathbf{v}_3 = \mathbf{u}_3 - \left(\frac{\mathbf{u}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right) \mathbf{v}_1 - \left(\frac{\mathbf{u}_3 \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2}\right) \mathbf{v}_2.$$

We can turn around these relations to express $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ in terms of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$,

similarly as in the case of two vectors case.

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 \\
\mathbf{u}_2 &= \left(\frac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right) \mathbf{v}_1 + \mathbf{v}_2 \\
\mathbf{u}_3 &= \left(\frac{\mathbf{u}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right) \mathbf{v}_1 + \left(\frac{\mathbf{u}_3 \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2}\right) \mathbf{v}_2 + \mathbf{v}_3
\end{aligned}$$

and these relations can be encoded in the matrix relation

$$[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3] = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3] \begin{bmatrix} 1 & \frac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1} & \frac{\mathbf{u}_3 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1} \\ 0 & 1 & \frac{\mathbf{u}_3 \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2} \\ 0 & 0 & 1 \end{bmatrix}.$$

The three columns of $[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]$ are orthogonal to each other and above matrix relation also encodes the recursive algorithm to compute $\mathbf{v}_2, \mathbf{v}_3$ as well as the coefficients of the upper triangular matrix as the second factor above.
If we set $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ to be the unit vectors in the direction of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ respectively, then we can substitute $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ in terms of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$, and note that $[\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]$ is now an orthogonal matrix, and the above matrix relation becomes

$$[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3] = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]R$$

for some upper triangular matrix R. This gives the QR-factorization of the $n \times 3$ matrix $[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3]$.

Remark 2.7.1

The above procedure can be generalized in a straight ward way to any finite number of vectors, where we define

$$\mathbf{v}_j = \mathbf{u}_j - \left[\left(\frac{\mathbf{u}_j \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\right) \mathbf{v}_1 + \cdots + \left(\frac{\mathbf{u}_j \cdot \mathbf{v}_{j-1}}{\mathbf{v}_{j-1} \cdot \mathbf{v}_{j-1}}\right) \mathbf{v}_{j-1}\right].$$

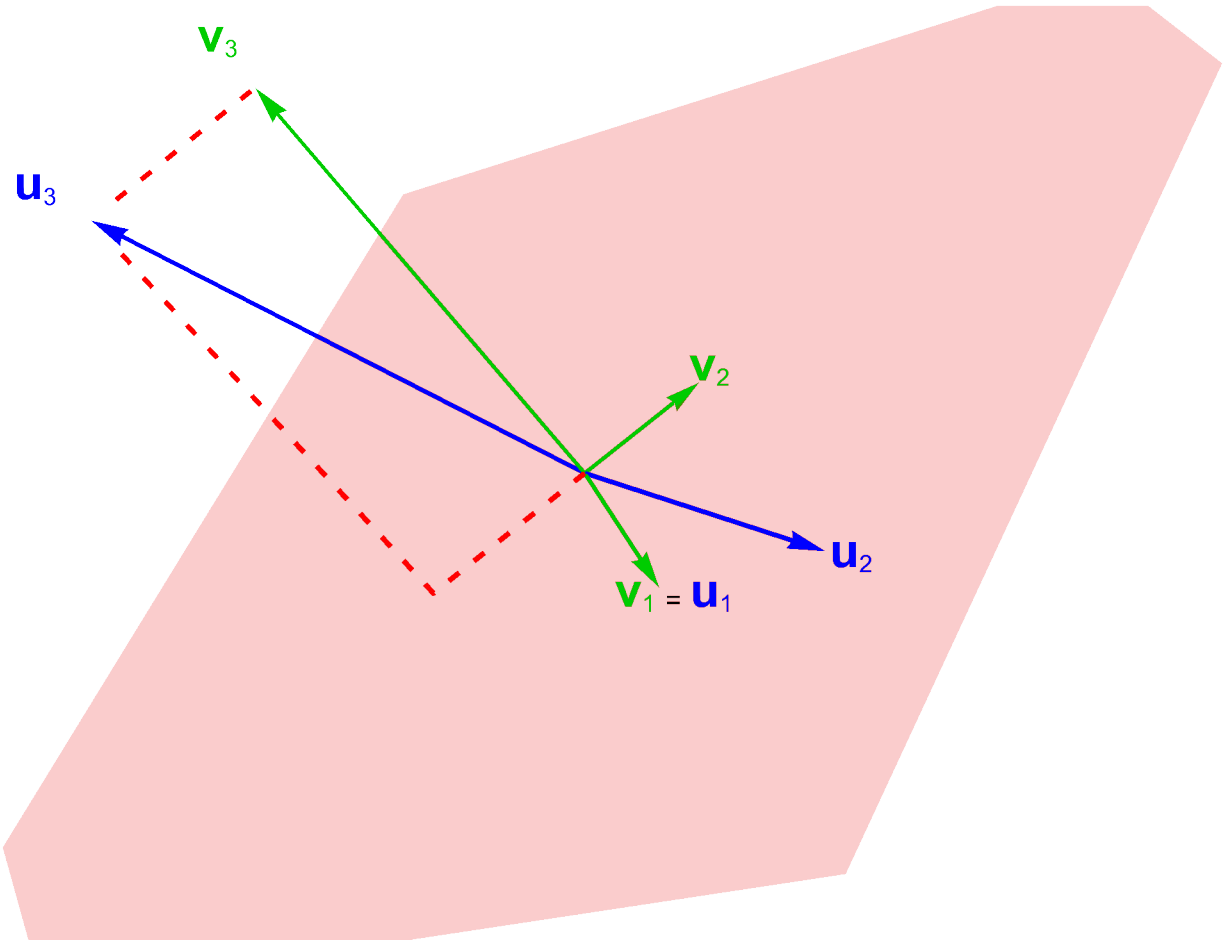
One could also modify the Gram-Schmidt process to transform each newly obtained $\mathbf{v}_j$ directly into a unit vector $\mathbf{w}_j$, and use the $\mathbf{w}_i$'s for $i \leq j$ in the next stage:

$$\mathbf{v}_{j+1} = \mathbf{u}_{j+1} - (\mathbf{u}_{j+1} \cdot \mathbf{w}_1)\, \mathbf{w}_1 - \cdots - (\mathbf{u}_{j+1} \cdot \mathbf{w}_j)\, \mathbf{w}_j.$$

Here we used the relations

$$\left(\frac{\mathbf{u}_{j+1} \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i}\right) \mathbf{v}_i = (\mathbf{u}_{j+1} \cdot \mathbf{w}_i)\, \mathbf{w}_i.$$

Figure 2.1: Illustration of the Gram-Schmidt algorithm as applied to three given vectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ to obtain three orthogonal ones $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_2\}$.



Using this approach in the previous example, we first get $\mathbf{w}_1 = \mathbf{v}_1/\|\mathbf{v}_1\| = \frac{1}{\sqrt{5}}\mathbf{v}_1$, *then*

$$\mathbf{v}_2 = \begin{bmatrix}1\\1\end{bmatrix} - \left(\begin{bmatrix}1\\1\end{bmatrix}\cdot \mathbf{w}_1\right)\mathbf{w}_1 = \begin{bmatrix}1\\1\end{bmatrix} - \frac{1}{\sqrt{5}}\mathbf{w}_1.$$

This gives

$$\begin{bmatrix}1\\1\end{bmatrix} = \frac{1}{\sqrt{5}}\mathbf{w}_1 + \mathbf{v}_2 = \frac{1}{\sqrt{5}}\mathbf{w}_1 + \|\mathbf{v}_2\|\mathbf{w}_2 = \frac{1}{\sqrt{5}}\mathbf{w}_1 + \frac{3}{\sqrt{5}}\mathbf{w}_2.$$

This modification is usually used in the computing implementation. For hand computations, it is easier to work out the orthonormal vectors and use the relation $[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3] = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]R$ *to find* R *by* $R = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]^T[\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3]$, *using* $[\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]^T[\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3] = I_3$.

Remark 2.7.2

If the starting set of vectors $\{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ *is linearly dependent, then some* $\mathbf{u}_j$ *is a linear combination of the vectors* $\{\mathbf{u}_1, \cdots, \mathbf{u}_{j-1}\}$. *This implies that at the stage of computing* $\mathbf{v}_j$, *it becomes* $\mathbf{0}$*! The algorithm as described has to come to a stop here. Some modifications have to be made to account for this possibility. One possibility is to move on to the next (pivot) column vector to find the next non-zero vector orthogonal to all previous ones. If this happens, we would still get an upper triangular matrix* R, *but its diagonal entries may no longer be all positive, and may have some zeroes.*

2.7.2 QR Decomposition

We already saw examples of a matrix decomposed as the product of a matrix whose columns are orthonormal and an upper triangular matrix. This can be done for any matrix after we apply the Gram-Schmidt algorithm to its columns vectors.

Example 2.7.2

Apply the Gram-Schmidt process to the set of vectors $\left\{\begin{bmatrix}2\\-1\\1\end{bmatrix}, \begin{bmatrix}3\\0\\3\end{bmatrix}, \begin{bmatrix}1\\1\\2\end{bmatrix}\right\}$ to obtain a set of orthogonal basis, then turn it into an orthonormal basis. Finally, record the relations between the given vectors and the set of orthonormal

vectors obtained as a QR factorization.

Set $\mathbf{v}_1 = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$ and

$$\mathbf{v}_2 = \begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix} - \left(\frac{\begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix} \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1} \right) \mathbf{v}_1 = \begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix} - \frac{3}{2} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3/2 \\ 3/2 \end{bmatrix}.$$

Then

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} - \left(\frac{\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1} \right) \mathbf{v}_1 - \left(\frac{\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \cdot \mathbf{v}_2}{\mathbf{v}_2 \cdot \mathbf{v}_2} \right) \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 3/2 \\ 3/2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

This indicates that $\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$ is in the span of $\{\mathbf{v}_1, \mathbf{v}_2\}$, which is the same as the span of $\left\{ \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix} \right\}$. The relations between the given columns and $\mathbf{v}_1, \mathbf{v}_2$ are encoded in

$$\begin{bmatrix} 2 & 3 & 1 \\ -1 & 0 & 1 \\ 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ -1 & 3/2 \\ 1 & 3/2 \end{bmatrix} \begin{bmatrix} 1 & 3/2 & 1/2 \\ 0 & 1 & 1 \end{bmatrix}.$$

Now we define

$$\mathbf{w}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\| = \begin{bmatrix} \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}, \quad \mathbf{w}_2 = \mathbf{v}_2 / \|\mathbf{v}_2\| = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix},$$

then $\mathbf{w}_1, \mathbf{w}_2$ is the sought after orthonormal vectors, and the relations between the given vectors and $\mathbf{w}_1, \mathbf{w}_2$ are encoded in the QR factorization

$$\begin{bmatrix} 2 & 3 & 1 \\ -1 & 0 & 1 \\ 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 3\sqrt{6}/2 & \sqrt{6}/2 \\ 0 & \frac{3}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{bmatrix}.$$

Remark 2.7.3

In applying the QR factorization to an $m \times n$ matrix A, according to the algorithm we have discussed, the number of columns in the Q factor is the dimension of the column space of A, namely, the rank of A. When the rank equals m, namely, A has a pivot in each row, then Q is an $m \times m$ orthogonal matrix: $Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I_m$. In such a situation, solving $A\mathbf{x} = \mathbf{b}$ is equivalent to solving $QR\mathbf{x} = \mathbf{b}$. A solution $\mathbf{x}$ to the latter would solve $Q^{\mathrm{T}}QR\mathbf{x} = Q^{\mathrm{T}}\mathbf{b}$, namely, $R\mathbf{x} = Q^{\mathrm{T}}\mathbf{b}$. R is an upper triangular matrix, and has m pivots in such a situation (why?), we can solve this system easily. Conversely, if $\mathbf{x}$ solves $R\mathbf{x} = Q^{\mathrm{T}}\mathbf{b}$, then $QR\mathbf{x} = QQ^{\mathrm{T}}\mathbf{b}$. Since $QQ^{\mathrm{T}} = I_m$, we conclude that $A\mathbf{x} = \mathbf{b}$.
In the above discussion, it is important that the rank of A equals its number of rows. If we try to apply a similar argument to our example above, suppose $\mathbf{x}$ is a solution of

$$\begin{bmatrix} 2 & 3 & 1 \\ -1 & 0 & 1 \\ 1 & 3 & 2 \end{bmatrix} \mathbf{x} = \mathbf{b},$$

then it also solves

$$\begin{bmatrix} \sqrt{6} & 3\sqrt{6}/2 & \sqrt{6}/2 \\ 0 & \frac{3}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathrm{T}} \mathbf{b}$$

The coefficient matrix has a pivot in each of its rows, so has a solution for any $\mathbf{b}$. However, we know the original coefficient matrix has rank 2, so there must be $\mathbf{b}$'s for which the system is inconsistent. What has caused this conflicting conclusion?
The issue is that, even if $\mathbf{x}$ solves this latter system, we can only imply that

$$\begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 3\sqrt{6}/2 & \sqrt{6}/2 \\ 0 & \frac{3}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathrm{T}} \mathbf{b}.$$

But since the Q factor is no longer a square matrix,

$$\begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \neq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$\mathbf{x}$ *turns out to solve*

$$\begin{bmatrix} 2 & 3 & 1 \\ -1 & 0 & 1 \\ 1 & 3 & 2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \mathbf{b}$$

instead of the original system. We will return to this topic in the next section when discussing why the "best fit" solution $\bar{x}$ gives rise to $\bar{b}' := A\bar{x}$ as the orthogonal projection of $\bar{b}$ onto the column space of A.

Remark 2.7.4

The Matlab implementation of the QR factorization is not done exactly as using the Gram-Schmidt algorithm. Its default calling would return a square orthogonal matrix Q, some of whose column vectors are not in the span of the column space of the given matrix. The diagonal entries of its returned R may not be positive or non-negative as described according to the algorithm given. I will post a `Matlab` *live-script file on this.*

2.8 An Optimization-Centric View of Linear Systems

The main focus of this section is how to define and find an "optimal" approximate solution to $A\bar{x} = \bar{b}$ when a solution does not exists and an "optimal" (approximate) solution when there are an infinite number of possible solutions.
One aspect of the optimization problem is to make $\|A\bar{x} - \bar{b}\|$ smallest possible, if we can't make it equal 0. Another aspect is to make $\|\bar{x}\|$ smallest possible when more than one $\bar{x}$ would make $\|A\bar{x} - \bar{b}\|$ smallest possible.
$\min_{\bar{x}} \|A\bar{x} - \bar{b}\| > 0$ when $\bar{b}$ is not in the column space of A. This can happen only when the rank r of A is $< n$ (assuming A is $n \times d$).
If $\bar{x}$ makes $\|A\bar{x} - \bar{b}\|$ smallest possible, then for any vector $\bar{v}$ and scalar t, $f(t) := \|A(\bar{x} + t\bar{v}) - \bar{b}\|^2$ satisfies $f(t) \geq f(0)$. Since $f(t) = \left[A(\bar{x} + t\bar{v}) - \bar{b}\right]^{\mathrm{T}} \left[A(\bar{x} + t\bar{v}) - \bar{b}\right]$

is a quadratic function of t, it is differentiable in t and we must have $f'(0) = 0$. This gives

$$f'(0) = \bar{v}^{\mathrm{T}} A^{\mathrm{T}} \left[A\bar{x} - \bar{b}\right] + \left[A\bar{x} - \bar{b}\right]^{\mathrm{T}} A\bar{v} = 0.$$

Since $\bar{v}^{\mathrm{T}} A^{\mathrm{T}} \left[A\bar{x} - \bar{b}\right] = \left[A\bar{x} - \bar{b}\right]^{\mathrm{T}} A\bar{v}$, it follows that $\bar{v}^{\mathrm{T}} A^{\mathrm{T}} \left[A\bar{x} - \bar{b}\right] = 0$. Since this holds for all $\bar{v}$, we conclude that $A^{\mathrm{T}} \left[A\bar{x} - \bar{b}\right] = 0$. This is our familiar equation $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}}\bar{b}$.

Once we find a solution to $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}}\bar{b}$, we can see that it indeed makes $\|A\bar{x} - \bar{b}\|$ smallest possible, for, taking any vector $\bar{y}$, we can write $\bar{y} = \bar{x} + \bar{v}$ for some $\bar{v}$, and derive

$$\|A\bar{y} - \bar{b}\|^2 = \|A\bar{x} - \bar{b} + A\bar{v}\|^2 = \|A\bar{x} - \bar{b}\|^2 + \|A\bar{v}\|^2 + 2(A\bar{v})^{\mathrm{T}}(A\bar{x} - \bar{b}) \geq \|A\bar{x} - \bar{b}\|^2$$

using $(A\bar{v})^{\mathrm{T}}(A\bar{x} - \bar{b}) = \bar{v}^{\mathrm{T}} A^{\mathrm{T}}(A\bar{x} - \bar{b}) = 0$ and $\|A\bar{v}\|^2 \geq 0$.

We have argued earlier that this system always has a solution. Whether it has a unique solution is determined by whether $\mathrm{Null}(A^{\mathrm{T}} A) = \{\bar{0}\}$. But $\mathrm{Null}(A^{\mathrm{T}} A) = \mathrm{Null}(A)$, so the question boils down to whether $\mathrm{Null}(A) = \{\bar{0}\}$, which is equivalent to whether the columns of A are linearly independent.

When $\mathrm{Null}(A) \neq \{\bar{0}\}$, when once we find one solution $\bar{x}_0$ to $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}}\bar{b}$, all its solutions are given in the form of $\bar{x}_0 + \bar{z}$ for some $\bar{z}$ satisfying $A\bar{z} = \bar{0}$.

In such a situation we would like to find $\bar{z}$ satisfying $A\bar{z} = \bar{0}$ to make $\|\bar{x}_0 + \bar{z}\|$ smallest possible. Namely, to solve

$$\arg\min_{\bar{z}: A\bar{z} = \bar{0}} \|\bar{x}_0 + \bar{z}\|.$$

This can be done by setting $g(t) = \|\bar{x} + t\bar{v}\|^2$, assuming $\bar{x} = \bar{x}_0 + \bar{z}$ is a solution (the existence of such an $\bar{x}$ relies on a property of calculus equivalent to the so called Balzano-Weierstrass property or completeness of the set of real numbers, or can be verified directly after we identify it as a solution to a matrix equation below and show its existence), and $\bar{v}$ is an arbitrary vector satisfying $A\bar{v} = \bar{0}$. Then using the criterion $g(t) \geq g(0)$ for all t we have implies $g'(0) = 0$. This leads to $\bar{x}^{\mathrm{T}}\bar{v} = \bar{x} \cdot \bar{v} = 0$ for all such $\bar{v}$. This is saying that $\bar{x} \perp \mathrm{Null}(A)$. Recall that $\mathrm{Row}(A)^{\perp} = \mathrm{Null}(A)$. But $\mathrm{Row}(A) = \mathrm{Col}(A^{\mathrm{T}})$, so $\mathrm{Col}(A^{\mathrm{T}})^{\perp} = \mathrm{Null}(A)$. This implies by Proposition 2.3.5 that the $\bar{x}$ above must be in $\mathrm{Col}(A^{\mathrm{T}})$, so must be of the form $A^{\mathrm{T}}\bar{y}$ for some $\bar{y} \in \mathcal{R}^n$.

We put together the two criteria $A^{\mathrm{T}} A\bar{x} = A^{\mathrm{T}}\bar{b}$ and $\bar{x} = A^{\mathrm{T}}\bar{y}$ to find that $A^{\mathrm{T}} AA^{\mathrm{T}}\bar{y} = A^{\mathrm{T}}\bar{b}$.

If the rows of A are linearly independent (this necessarily implies that $n \leq d$), then the $n \times n$ matrix AA^{T} is invertible and we find $\bar{y} = (AA^{\mathrm{T}})^{-1}\bar{b}$ to be the unique solution, which leads to $\bar{x} = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}\bar{b}$. The $d \times n$ matrix $A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$ here satisfies $A\left(A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}\right) = I_n$ and is called a right inverse of A.

Suppose that $\bar{x} = A^{\mathrm{T}}\bar{y}$ is such that $AA^{\mathrm{T}}\bar{y} = \bar{b}$. Then for any $\bar{v}$ with $A\bar{v} = \bar{0}$, using the quadratic nature of $\|\bar{x} + \bar{v}\|^2$, we see that

$$\|\bar{x} + \bar{v}\|^2 = \|\bar{x}\|^2 + \|\bar{v}\|^2 + 2\bar{x} \cdot \bar{v} \geq \|\bar{x}\|^2$$

using $\bar{x} \cdot \bar{v} = A^{\mathrm{T}}\bar{y} \cdot \bar{v} = \bar{y}^{\mathrm{T}}A\bar{v} = 0$ and $\|\bar{v}\|^2 \geq 0$. This confirms that this $\bar{x}$ has the minimum norm.

2.8.1 Moore-Penrose Pseudoinverse

When neither the rows nor the columns of A are linearly independent, one modifies the above criterion to minimize $\|A\bar{x} - \bar{b}\|^2 + \lambda\|\bar{x}\|^2$, where $\lambda > 0$ is a parameter.
A similar algorithm shows that a minimizer $\bar{x}$ must satisfy

$$\bar{v}^{\mathrm{T}}\left[A^{\mathrm{T}}\left(A\bar{x} - \bar{b}\right) + \lambda\bar{x}\right] = 0 \text{ for all } \bar{v}.$$

This leads to $\left(A^{\mathrm{T}}A + \lambda I_d\right)\bar{x} = A^{\mathrm{T}}\bar{b}$. Note that for $\lambda > 0$, the matrix $\left(A^{\mathrm{T}}A + \lambda I\right)$ is always invertible. Thus we have found our solution to be $\bar{x} = \left(A^{\mathrm{T}}A + \lambda I\right)^{-1}A^{\mathrm{T}}\bar{b}$.
The *push-through identity* says that $\left(A^{\mathrm{T}}A + \lambda I_d\right)^{-1}A^{\mathrm{T}} = A^{\mathrm{T}}\left(AA^{\mathrm{T}} + \lambda I_n\right)^{-1}$. Thus we can use either to compute $\bar{x}$.
$\lambda > 0$ can be arbitrary. For many applications we would choose a small $\lambda > 0$, or rather, take the limit as $\lambda \to 0+$. It turns out that

$$\lim_{\lambda \to 0+}\left(A^{\mathrm{T}}A + \lambda I_d\right)^{-1}A^{\mathrm{T}} = \lim_{\lambda \to 0+}A^{\mathrm{T}}\left(AA^{\mathrm{T}} + \lambda I_n\right)^{-1}$$

exists. The limiting matrix is called the Moore-Penrose pseudoinverse of A. This is seen by using a QR decomposition of $A = QR$, where Q is an $n \times r$ matrix with orthonormal columns ($r = \mathrm{rank}(A)$) and R is an $r \times d$ upper triangular matrix with linearly independent rows. Then $Q^{\mathrm{T}}Q = I_r$ and $A^{\mathrm{T}}A = R^{\mathrm{T}}R$. The push-through identity implies that

$$\left(A^{\mathrm{T}}A + \lambda I_d\right)^{-1}A^{\mathrm{T}} = \left(R^{\mathrm{T}}R + \lambda I_d\right)^{-1}R^{\mathrm{T}}Q^{\mathrm{T}} = R^{\mathrm{T}}\left(RR^{\mathrm{T}} + \lambda I_r\right)^{-1}Q^{\mathrm{T}}.$$

Since R^{T} has linearly independent columns, RR^{T} invertible. Therefore $\lim_{\lambda \to 0+}\left(RR^{\mathrm{T}} + \lambda I_r\right)^{-1} = \left(RR^{\mathrm{T}}\right)^{-1}$. We now conclude that

$$\lim_{\lambda \to 0+}\left(A^{\mathrm{T}}A + \lambda I_d\right)^{-1}A^{\mathrm{T}} = R^{\mathrm{T}}\left(RR^{\mathrm{T}}\right)^{-1}Q^{\mathrm{T}}.$$

This limit is the Moore-Penrose pseudoinverse of A, and is denoted as A^{+}.
Note that $\bar{x} = A^{+}\bar{b}$ satisfies $A\bar{x} = QRR^{\mathrm{T}}\left(RR^{\mathrm{T}}\right)^{-1}Q^{\mathrm{T}}\bar{b} = QQ^{\mathrm{T}}\bar{b}$. When $r < n$, $QQ^{\mathrm{T}} \neq I_n$. $QQ^{\mathrm{T}}\bar{b}$ turns out to be the orthogonal projection of $\bar{b}$ in the columns space of A.

2.8.2 The Projection Matrix

We review how the orthogonal projection onto a subspace is defined and computed. When an orthonormal basis $\{\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k\}$ of $\mathcal{V}$ is known, then the orthogonal projection of any vector $\mathbf{x}$ is

$$P_{\mathcal{V}}\mathbf{x} = \sum_{i=1}^{k} (\mathbf{x} \cdot \mathbf{v}_i)\, \mathbf{v}_i = [\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k] \begin{bmatrix} \mathbf{x} \cdot \mathbf{v}_1 \\ \vdots \\ \mathbf{x} \cdot \mathbf{v}_k \end{bmatrix} = [\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k]\, [\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k]^{\mathrm{T}}\, \mathbf{x}.$$

This can be written as $P_{\mathcal{V}}\mathbf{x} = QQ^{\mathrm{T}}\mathbf{x}$, if we set $Q = [\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k]$.
When $\{\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k\}$ is a general basis of $\mathcal{V}$, we set A be the matrix with $\{\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k\}$ as its columns: $A = [\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k]$. Then for any vector $\mathbf{u}$, its orthogonal projection $\mathbf{v} = P_{\mathcal{V}}(\mathbf{u})$ onto $\mathcal{V}$ must be a linear combination of $\mathbf{v}_1, \cdots, \mathbf{v}_{k-1}, \mathbf{v}_k$,

$$\mathbf{v} = x_1\mathbf{v}_1 + \cdots + x_{k-1}\mathbf{v}_{k-1} + x_k\mathbf{v}_k = A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix},$$

such that

$$\{\mathbf{u} - \mathbf{v}\} \cdot \mathbf{v}_j = \left\{ \mathbf{u} - A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \right\} \cdot \mathbf{v}_j = 0, \quad \text{for } j = 1, 2, \cdots, k.$$

This can be written in terms of row vectors ${\mathbf{v}_j}^{\mathrm{T}}$ as

$$(\mathbf{v}_j)^{\mathrm{T}} \left\{ \mathbf{u} - A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \right\} = 0, \quad \text{for } j = 1, 2, \cdots, k,$$

which can further be written in a single matrix equation as

$$A^{\mathrm{T}} \left\{ \mathbf{u} - A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \right\} = 0.$$

This linear system simplifies to

$$A^T A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = A^T \mathbf{u}.$$

Solving for $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$ in terms of $\mathbf{u}$, we obtain

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} \mathbf{u}, \text{ provided that } A^{\mathrm{T}} A \text{ is invertible.}$$

We do know that $A^{\mathrm{T}} A$ is invertible when the columns of A are linearly independent. Thus we have

$$P_{\mathcal{V}}(\mathbf{u}) = \mathbf{v} = A \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = A(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} \mathbf{u},$$

namely, $P_{\mathcal{V}}(\mathbf{u})$ is given by the matrix multiplication of $\mathbf{u}$ by the matrix $A(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}$. When we work with the column space $\mathcal{V}$ of a matrix A without knowing wether its columns are linearly independent, we can use the Moore-Penrose pseudoinverse A^{T} of A to identify the projection $\mathbf{v} = AA^{+}\mathbf{u} = QQ^{\mathrm{T}}\mathbf{u}$ of $\mathbf{u}$ in $\mathcal{V}$, where Q is the orthogonal factor in the QR decomposition $A = QR$ of A.
The orthogonal projection $P_{\mathcal{V}}$ onto $\mathcal{V}$ has the following properties.

(i). $P_{\mathcal{V}}\bar{v} = \bar{v}$ for any $\bar{v} \in \mathcal{V}$.

(ii). $P_{\mathcal{V}}\bar{v} = \bar{0}$ for any $\bar{v} \in (\mathcal{V})^{\perp}$.

(iii). $P_{\mathcal{V}} P_{\mathcal{V}} \bar{v} = P_{\mathcal{V}}\bar{v}$ for any $\bar{v}$; equivalently, $P_{\mathcal{V}} P_{\mathcal{V}} = P_{\mathcal{V}}$.

(iv). $P_{\mathcal{V}}$ is a symmetric matrix and $P_{\mathcal{V}}\bar{v} \cdot \bar{w} = \bar{v} \cdot P_{\mathcal{V}}\bar{w} = P_{\mathcal{V}}\bar{v} \cdot P_{\mathcal{V}}\bar{w}$.

These properties characterize an orthogonal projection in the sense that any matrix P satisfying the properties in (iii) and (iv) is the orthogonal projection matrix onto some subspace $\mathcal{V}$: $\mathcal{V} = \mathrm{Null}(I - P)$. This can be seen by noting that (iii) is equivalent to $(I - P)P = P(I - P) = 0$.

2.9 Ill-Conditioned Matrices and Systems

Our theoretical discussions have a clearly defined criterion on when $A\bar{x} = \bar{b}$ is solvable and when A is invertible. In practical computations, round-off errors occur and data may also have errors. This makes it impossible to tell whether a tiny pivot is meant to be 0 but occurs due to computational errors or is meant to be non-zero; when a vector $\bar{b}'$ is used instead of $\bar{b}$ the error between the computed solution $\bar{x}'$ to $A\bar{x}' = \bar{b}'$ and the actual intended solution $\bar{x}$ to $A\bar{x} = \bar{b}$ needs to be estimated, but in certain situations this kind of error becomes unacceptable.
A more appropriate measure of error is the relative error,

$$\frac{\|\bar{x} - \bar{x}'\|}{\|\bar{x}\|} = \frac{\|A^{-1}(\bar{b} - \bar{b}')\|}{\|\bar{x}\|}.$$

Here we assume that A is invertible.
The operator norm of a matrix A is defined as

$$\|A\|_{\text{op}} := \sup_{\bar{x}\neq\bar{0}} \frac{\|A\bar{x}\|}{\|\bar{x}\|}.$$

It gives the maximum magnification factor of the action of $A\bar{x}$. Thus

$$\|A^{-1}(\bar{b} - \bar{b}')\| \leq \|A^{-1}\|_{\text{op}}\|\bar{b} - \bar{b}'\| \text{ and } \|\bar{b}\| = \|A\bar{x}\| \leq \|A\|_{\text{op}}\|\bar{x}\|,$$

from which we obtain

$$\frac{\|\bar{x} - \bar{x}'\|}{\|\bar{x}\|} \leq \|A^{-1}\|_{\text{op}}\|A\|_{\text{op}}\frac{\|\bar{b} - \bar{b}'\|}{\|\bar{b}\|}.$$

It turns out that an appropriate choice of $\bar{b}$ and $\bar{b}'$ can attain the ratio $\|A^{-1}\|_{\text{op}}\|A\|_{\text{op}}$. This number is called the *conditional number* of matrix A. It is often denoted as $\kappa(A)$. When this number is very large, say 10^6 or more, computations in $A\bar{x} = \bar{b}$ could magnify the relative error in $\bar{b}$ by more than 10^6 fold, which is usually considered unacceptable. Such a matrix is called ill-conditioned.
Since

$$\|A^{-1}\|_{\text{op}} := \sup_{\bar{b}\neq\bar{0}} \frac{\|A^{-1}\bar{b}\|}{\|\bar{b}\|} = \sup_{\bar{x}\neq\bar{0}} \frac{\|\bar{x}\|}{\|A\bar{x}\|} = \left(\inf_{\bar{x}\neq\bar{0}} \frac{\|A\bar{x}\|}{\|\bar{x}\|}\right)^{-1},$$

$\|A^{-1}\|_{\text{op}}\|A\|_{\text{op}}$ is the ratio between the maximum magnification and minimum magnification factor of A. Using $\|A\|_{\text{op}} \geq \|A(:, j)\|$ for any j, we see that $\kappa(A)$ is no less than the ratio between the norms of any two columns, therefore, whenever two columns have vastly different sizes, the matrix is ill-conditioned.

We will discuss how to compute or more accurately estimate the condition number later on. For now, we will just alert the student that numerical computations involving ill-conditioned matrices become very unreliable. One can compute the inverse, QR-factorization, `rref`, and rank of

$$\begin{bmatrix} 123456789 & 123456788 \\ 123456790 & 123456789 \end{bmatrix}$$

in `Matlab` to see the effects of this ill-conditioned matrix. I will post a `Matlab` live-script file on this in canvas.

2.10 Inner Products: A Geometric View

The textbook first takes the approach of defining the general inner product (on $\mathcal{R}^n$) using the usual dot product and an invertible matrix S.

Definition 2.10.1: (Inner Products: Restricted Definition)

A mapping from $\bar{x}, \bar{y} \in \mathcal{R}^n$ to $\langle x, y\rangle \in \mathcal{R}$ is an inner product if and only if $\langle \bar{x}, \bar{y}\rangle$ is always equal to the dot product between $S\bar{x}$ and $S\bar{y}$ for some $n \times n$ non-singular matrix S. The inner product $\langle x, y\rangle$ can also be expressed using the Gram matrix $G = S^{\mathrm{T}}S$:

$$\langle \bar{x}, \bar{y}\rangle = (S\bar{x})^{\mathrm{T}}(S\bar{y}) = \bar{x}^{\mathrm{T}}[S^{\mathrm{T}}S]\bar{y} = \bar{x}^{\mathrm{T}}G\bar{y}.$$

The Cauchy-Schwarz inequality continues to hold for this inner product in the form of

$$|\langle \bar{x}, \bar{y}\rangle| \leq \sqrt{\bar{x}^{\mathrm{T}}G\bar{x}}\sqrt{\bar{y}^{\mathrm{T}}G\bar{y}}.$$

Such an inner product induces a norm defined by $\sqrt{\bar{x}^{\mathrm{T}}G\bar{x}}$. The triangle inequality for this norm also holds.

The simplest case is when $S = \mathrm{diag}[\lambda_1, \cdots, \lambda_n]$ is a diagonal matrix with positive diagonal entries λ_j. This amounts to scaling each x_j into $u_j = \lambda_j x_j$ before applying the standard Euclidean dot product on $(u_1, \cdots, u_n)$.

If we need to find $\bar{x}$ to make

$$\langle A\bar{x} - \bar{b}, A\bar{x} - \bar{b}\rangle = (A\bar{x} - \bar{b})^{\mathrm{T}}G(A\bar{x} - \bar{b}) = (\bar{x}^{\mathrm{T}}A^{\mathrm{T}} - \bar{b}^{\mathrm{T}})G(A\bar{x} - \bar{b})$$

as small as possible, then we would find

$$A^{\mathrm{T}}G(A\bar{x} - \bar{b}) = \bar{0},$$

and the best approximation (in this new metric) to $\bar{b}$ would be given by $A(A^{\mathrm{T}}GA)^{-1}A^{\mathrm{T}}\bar{b}$.

A more general and conceptual definition of inner product is the following.

Definition 2.10.2: (Inner-Product: General Definition)

The real value $\langle \bar{u}, \bar{v} \rangle$ is an inner product between $\bar{u}$ and $\bar{v}$, if it satisfies the following axioms for all $\bar{u}$ and $\bar{v}$:

Additivity: $\langle \bar{u}, \bar{v} + \bar{w} \rangle = \langle \bar{u}, \bar{v} \rangle + \langle \bar{u}, \bar{w} \rangle, \langle \bar{v} + \bar{w}, \bar{u} \rangle = \langle \bar{v}, \bar{u} \rangle + \langle \bar{w}, \bar{u} \rangle$.

Multiplicativity: $\langle c\bar{u}, \bar{v} \rangle = c\langle \bar{u}, \bar{v} \rangle, \langle \bar{u}, c\bar{v} \rangle = c\langle \bar{u}, \bar{v} \rangle \; \forall c \in \mathcal{R}$.

Commutativity: $\langle \bar{u}, \bar{v} \rangle = \langle \bar{v}, \bar{u} \rangle$.

Positive definiteness: $\langle \bar{u}, \bar{u} \rangle \geq 0$, with equality only for the zero vector.

The Cauchy-Schwarz inequality continues to hold for this general inner product. It also induces a norm in the form of $\sqrt{\langle \bar{u}, \bar{u} \rangle}$.
Every finite-dimensional inner product $\langle \bar{x}, \bar{y} \rangle$ in $\mathcal{R}^n$ satisfying the above axioms can be shown to be equivalent to $\bar{x}^{\mathrm{T}} S \bar{y}$ for some carefully chosen Gram matrix $G = S^{\mathrm{T}} S$.

2.11 Complex Vector Spaces

In applications we often work with vector spaces over the field of complex numbers (such as when working with Fourier series). The notion of inner product on a vector space over the field of complex numbers has a natural definition, modifying the inner product on a vector space over the field of real numbers in two places.

Definition 2.11.1: (Inner-Product on $\mathcal{C}^n$)

The complex value $\langle \bar{u}, \bar{v} \rangle$ is an inner product between $\bar{u}$ and $\bar{v}$ in $\mathcal{C}^n$, if it satisfies the following axioms for all $\bar{u}$ and $\bar{v}$:

Additivity: $\langle \bar{u}, \bar{v} + \bar{w} \rangle = \langle \bar{u}, \bar{v} \rangle + \langle \bar{u}, \bar{w} \rangle, \langle \bar{v} + \bar{w}, \bar{u} \rangle = \langle \bar{v}, \bar{u} \rangle + \langle \bar{w}, \bar{u} \rangle$.

Multiplicativity[a]: $\langle c\bar{u}, \bar{v} \rangle = c^* \langle \bar{u}, \bar{v} \rangle, \langle \bar{u}, c\bar{v} \rangle = c\langle \bar{u}, \bar{v} \rangle \; \forall c \in \mathcal{C}$.

Conjugate Commutativity: $\langle \bar{u}, \bar{v} \rangle = \langle \bar{v}, \bar{u} \rangle^*$.

Positive definiteness: $\langle \bar{u}, \bar{u} \rangle$ becomes real and nonnegative, and equals 0 only for the zero vector.

In the above the superscript "*" stands for the complex conjugate of a complex number: $(a + bi)^* = a - bi$ for $a, b \in \mathcal{R}$.

[a] This convention is used more often in the applied community; the math community tends to use the convention that $\langle c\bar{u}, \bar{v} \rangle = c\langle \bar{u}, \bar{v} \rangle$, $\langle \bar{u}, c\bar{v} \rangle = c^* \langle \bar{u}, \bar{v} \rangle$.

Remark 2.11.1

If one prefers to work with real-valued inner product, then $\Re\langle\bar{u},\bar{v}\rangle$ satisfies all the property of an inner product on a vector space over the field of real numbers (since $\langle\bar{u},\bar{v}\rangle = \langle\bar{v},\bar{u}\rangle^*$, $\langle\bar{u},\bar{v}\rangle$ and $\langle\bar{v},\bar{u}\rangle$ have equal real parts), but one loses the multiplicativity such as $\Re\langle c\bar{u},\bar{v}\rangle = c^*\Re\langle\bar{u},\bar{v}\rangle$ when c is not real. Since scalar multiplication by a complex scalar is prevalent in such a context and we prefer that the inner product reflects its interaction with this operation, it's easier to work with the complex-valued inner product as defined here. Another implicit property of this inner product is

$$\langle i\bar{u}, i\bar{v}\rangle = \langle\bar{u},\bar{v}\rangle,$$

which implies that multiplication by i preserves the inner product, therefore preserves the length of a vector and angles between vectors. Here the length of a complex vector $\bar{v}$ is still defined as $\sqrt{\langle\bar{v},\bar{v}\rangle}$; the Cauchy-Schwarz and triangle inequalities continue to hold in this context.

A most commonly used inner product on $\mathcal{C}^n$ is defined as $\langle\bar{u},\bar{v}\rangle = \sum_{j=1}^n u_j^* v_j$. We use $\bar{u}^* = [u_1^*, \cdots, u_n^*]^{\mathrm{T}}$ to denote this row vector, which is the transpose of the conjugate of vector $\bar{u}$, then $\langle\bar{u},\bar{v}\rangle = \bar{u}^*\bar{v}$.
In general, we have

Definition 2.11.2: (Conjugate Transpose of Vector and Matrix)

The conjugate transpose $\bar{v}^*$ of a complex vector $\bar{v}$ is obtained by transposing the vector and replacing each entry with its complex conjugate. The conjugate transpose $V^*$of a complex matrix V is obtained by transposing the matrix and replacing each entry with its complex conjugate.

The conjugate transpose satisfies $(AB)^* = B^*A^*$ and $(A^*)^* = A$.
If the columns of a matrix U are orthonormal, then $U^*U = I$. A square matrix with this property is called a *unitary* matrix. It has the property $\langle U\bar{u}, U\bar{v}\rangle = \langle\bar{u},\bar{v}\rangle$.
A more general inner product on $\mathcal{C}^n$ is defined by

$$\langle\bar{u},\bar{v}\rangle = (S\bar{u})^* S\bar{v} = \bar{u}^* S^* S\bar{v} = \bar{u}^* G\bar{v},$$

where S is an invertible $n \times n$ matrix and $G = S^*S$ satisfies $G^* = G$ (a matrix with this property is called a Hermitian matrix). It is convenient to denote this inner product as $\langle\bar{u},\bar{v}\rangle_G$ to signal its relation to G.

Definition 2.11.3: (Orthogonality in $\mathcal{C}^n$)

Two vectors $\bar{u}, \bar{v}$ are said to be orthogonal if $\langle \bar{u}, \bar{v} \rangle = 0$.

The notion of a set of orthonormal vectors continues to make sense. If $\{\bar{v}_j : j = 1, \ldots, l\}$ is a set of orthonormal vectors and $\bar{u} = \sum_{j=1}^{l} c_j \bar{v}_j$, then $\langle \bar{v}_j, \bar{u} \rangle = c_j$ (the order between $\bar{v}_j$ and $\bar{u}$ is important here; $\langle \bar{u}, \bar{v}_j \rangle$ would not work).
The Pythagorean Theorem in $\mathcal{C}^n$ holds under a modified orthogonal relation: If $\bar{u}, \bar{v}$ is such that $\langle \bar{u}, \bar{v} \rangle$ is purely imaginary, then

$$\|\bar{u} + \bar{v}\|^2 = \|\bar{u}\|^2 + \|\bar{v}\|^2.$$

This follows from noting that

$$\begin{aligned} \|\bar{u} + \bar{v}\|^2 &= \langle \bar{u}, \bar{u} \rangle + \langle \bar{u}, \bar{v} \rangle + \langle \bar{v}, \bar{u} \rangle + \langle \bar{v}, \bar{v} \rangle \\ &= \langle \bar{u}, \bar{u} \rangle + 2\Re\langle \bar{u}, \bar{v} \rangle + \langle \bar{v}, \bar{v} \rangle. \end{aligned}$$

When $\langle \bar{u}, \bar{v} \rangle$ is purely imaginary, $\Re\langle \bar{u}, \bar{v} \rangle$ holds, from which the Pythagorean relation follows. When $\Re\langle \bar{u}, \bar{v} \rangle = 0$, the Pythagorean relation indicates that $\bar{u}$ and $\bar{v}$ form a 90° angle. But when we say two vectors $\bar{u}, \bar{v}$ are orthogonal in a complex inner product, we refer to the relation $\langle \bar{u}, \bar{v} \rangle = 0$, not the relation $\Re\langle \bar{u}, \bar{v} \rangle = 0$.
The Gram-Schmidt algorithm can be adapted to this context: the orthogonal component of $\bar{v}$ with respect to $\bar{u}$ is $\bar{v} - \frac{\langle \bar{u}, \bar{v} \rangle}{\langle \bar{u}, \bar{u} \rangle} \bar{u}$, etc.

Example 2.11.1

The vectors $\begin{bmatrix} 1+i \\ 1-i \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ satisfy

$$\begin{aligned} \langle \begin{bmatrix} 1+i \\ 1-i \end{bmatrix}, \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} \rangle &= \begin{bmatrix} 1-i & 1+i \end{bmatrix} \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} = 4, \\ \langle \begin{bmatrix} 1+i \\ 1-i \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rangle &= \begin{bmatrix} 1-i & 1+i \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1-i, \\ \langle \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} \rangle &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} \rangle = 1+i, \\ \langle \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rangle &= 1, \end{aligned}$$

so the Gram-Schmidt algorithm would give

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1-i}{4} \begin{bmatrix} 1+i \\ 1-i \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{i}{2} \end{bmatrix}.$$

After normalizing these two orthogonal vectors, we obtain an orthonormal basis

$$\left\{ \begin{bmatrix} \frac{1+i}{2} \\ \frac{1-i}{2} \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} \end{bmatrix} \right\}.$$

The relations between the given vectors and these orthonormal vectors can be expressed as the following matrix factorization, similar to the QR factorization, except that the Q factor satisfies $Q^*Q = I$:

$$\begin{bmatrix} 1+i & 1 \\ 1-i & 0 \end{bmatrix} = \begin{bmatrix} \frac{1+i}{2} & \frac{1}{\sqrt{2}} \\ \frac{1-i}{2} & \frac{i}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 2 & \frac{1-i}{2} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Observe that if $\{\bar{u}_1, \cdots, \bar{u}_k\}$ is orthonormal and ω_j, $1 \leq j \leq k$, are complex numbers satisfying $|\omega_j| = 1$, then $\{\omega_1 \bar{u}_1, \cdots, \omega_k \bar{u}_k\}$ continues to be orthonormal. Since $\begin{bmatrix} \frac{1+i}{2} \\ \frac{1-i}{2} \end{bmatrix} = \frac{1+i}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} \end{bmatrix}$, and $|\frac{1+i}{\sqrt{2}}| = 1$, we could also use

$$\left\{ \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} \end{bmatrix} \right\}$$

as a set of orthonormal vectors.
The Gram matrix associated with $\begin{bmatrix} 1+i & 1 \\ 1-i & 0 \end{bmatrix}$ is the following Hermitian matrix

$$G = \begin{bmatrix} 1+i & 1 \\ 1-i & 0 \end{bmatrix}^* \begin{bmatrix} 1+i & 1 \\ 1-i & 0 \end{bmatrix} = \begin{bmatrix} 1-i & 1+i \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1+i & 1 \\ 1-i & 0 \end{bmatrix} = \begin{bmatrix} 4 & 1-i \\ 1+i & 1 \end{bmatrix}.$$

If we use this G to define an inner product, then

$$\langle \bar{u}, \bar{u} \rangle_G = \begin{bmatrix} u_1^* & u_2^* \end{bmatrix} \begin{bmatrix} 4 & 1-i \\ 1+i & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 4|u_1|^2 + 2\Re\left[(1-i)u_1^* u_2\right] + |u_2|^2.$$

Example 2.11.2

The equation $z^n = 1$ has n roots with the properties:

(a). $z = e^{2ik\pi/n}$ for some $k = 0, 1, \cdots, n-1$ with $|z| = 1$.

(b). Any root $z \neq 1$ satisfies $1 + z + \cdots + z^{n-1} = 0$.

(c). If z is a root, then so is $z^* = z^{-1}$.

(d). If z_1 and z_2 are two roots, then so is $z_1 z_2$.

Based on these properties we see that for any two roots $z_1 \neq z_2$, the vectors $(1, z_1, \cdots, z_1^{n-1})^{\mathrm{T}}$ and $(1, z_2, \cdots, z_2^{n-1})^{\mathrm{T}}$ are orthogonal:

$$\langle \begin{bmatrix} 1 \\ z_1 \\ \vdots \\ z_1^{n-1} \end{bmatrix}, \begin{bmatrix} 1 \\ z_2 \\ \vdots \\ z_2^{n-1} \end{bmatrix} \rangle = \begin{bmatrix} 1 & z_1^* & \cdots & (z_1^*)^{n-1} \end{bmatrix} \begin{bmatrix} 1 \\ z_2 \\ \vdots \\ z_2^{n-1} \end{bmatrix} = 1+(z_1^* z_2)+\cdots+(z_1^* z_2)^{n-1} = 0,$$

as $z_1^* z_2 \neq 1$.
All of these vectors have lengths equal to $\sqrt{n}$, so we can write down an $n \times n$ unitary matrix B whose jth column is $\begin{bmatrix} 1 & z_{j-1} & \cdots & (z_{j-1})^{n-1} \end{bmatrix}^{\mathrm{T}} / \sqrt{n}$, with $z_j = e^{2ij\pi/n}$. This matrix B turns out to be symmetric.
In the context of discrete Fourier transform, one is supposed to find $\bar{x} \in \mathcal{C}^n$ such that $B\bar{x} = \bar{s}$ for some given $\bar{s} \in \mathcal{C}^n$ representing the values of a time series. Since $B^* B = I$, we find $\bar{x} = B^* \bar{s}$ easily. The interpretation of $B\bar{x} = \bar{s}$ is that $f(z_{j-1}) = s_j$ for $j = 1, 2, \cdots, n$, where $f(z) = (x_1 + x_2 z + \cdots + x_n z^{n-1})/\sqrt{n}$.

Chapter 3

Eigenvectors and Diagonalizable Matrices

3.1 Introduction

Suppose that A is an $n \times n$ matrix. A non-zero vector $\mathbf{x} \in \mathcal{R}^n$ satisfying $A\mathbf{x} = \lambda\mathbf{x}$ is called an eigenvector of the matrix A, and the corresponding λ is called an eigenvalue of A.

If matrix A has a set of n eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ which is linearly independent, namely, the $n \times n$ matrix V with $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ as its columns has rank n, and with $\{\lambda_1, \cdots, \lambda_n\}$ as its corresponding eigennvalues, then for any vector $\mathbf{x} \in \mathcal{R}^n$, we can find its coordinates in $\{\mathbf{v}_1, \cdots, \mathbf{v}_n\}$ as

$$\mathbf{x} = x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n,$$

then

$$A\mathbf{x} = A(x_1\mathbf{v}_1 + \cdots + x_n\mathbf{v}_n) = x_1A\mathbf{v}_1 + \cdots + x_nA\mathbf{v}_n = \lambda_1x_1\mathbf{v}_1 + \cdots + \lambda_nx_n\mathbf{v}_n.$$

This gives a much easier interpretation for $A\mathbf{x}$ than the algebraic definition of the matrix-vector product $A\mathbf{x}$ gives.

The discussions above can be written in matrix form as follows: $AV = VD$, where D is the diagonal matrix with $\{\lambda_1, \cdots, \lambda_n\}$ as entries on its diagonal. Since V is invertible, it follows that $V^{-1}AV = D$, or equivalently, $A = VDV^{-1}$. In such a situation, we say that A is diagonalizable.

When a matrix A is diagonalizable, its action $A\mathbf{x}$ can also be encoded in matrix form:

$$A\mathbf{x} = A[\mathbf{v}_1, \cdots, \mathbf{v}_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = VD \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = V \begin{bmatrix} \lambda_1x_1 \\ \vdots \\ \lambda_nx_n \end{bmatrix}.$$

We investigate in this chapter criteria for a matrix to be diagonalizable.

3.2 Determinants

A geometric motivation for the notion of the determinant of a general $n \times n$ matrix is to use it to represent the (signed) volume of parallelepiped formed by the column vectors of the matrix as its edges.
Let's first review the situation in $\mathbb{R}^2$. Let $[\mathbf{u}, \mathbf{v}]$ denote the parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$,. Analytically, this means the set of vectors $\{s\mathbf{u} + t\mathbf{v} : 0 \leq s, t \leq 1\}$. Let $A[\mathbf{u}, \mathbf{v}]$ denote the area of this parallelogram. Then we have

$$A[\mathbf{u}, c\mathbf{v}] = cA[\mathbf{u}, \mathbf{v}], \quad \text{for all } \mathbf{u}, \mathbf{v}, \tag{a1}$$

at least for $c \geq 0$,

$$A[\mathbf{u}, \mathbf{v} + c\mathbf{u}] = A[\mathbf{u}, \mathbf{v}], \quad \text{for all } \mathbf{u}, \mathbf{v}, \text{ and } c. \tag{a2}$$

(a2) is a reflection of the geometric principle that *parallelograms with the same base and equal heights have equal areas*. Here $[\mathbf{u}, \mathbf{v} + c\mathbf{u}]$ and $[\mathbf{u}, \mathbf{v}]$ share a common base $\mathbf{u}$, and have equal heights with base $\mathbf{u}$. The following is a Desmos graph illustration* of areas of parallelograms with adjacent edges$[\mathbf{u}, \mathbf{v}]$ and $[\mathbf{u}, \mathbf{v} + c\mathbf{u}]$.
In order to extend (a1) to all c, we can either replace c on the right by $|c|$, or replace $A[\mathbf{u}, \mathbf{v}]$ by the *signed* area $\mathrm{sgA}[\mathbf{u}, \mathbf{v}]$. $\mathrm{sgA}[\mathbf{u}, \mathbf{v}]$ satisfies (a1) for all c, and satisfies (a2) as well; furthermore, we also have

$$\mathrm{sgA}[\mathbf{u}, \mathbf{v}+\mathbf{w}] = \mathrm{sgA}[\mathbf{u}, \mathbf{v}]+\mathrm{sgA}[\mathbf{u}, \mathbf{w}] \text{ and } \mathrm{sgA}[\mathbf{u}+\mathbf{v}, \mathbf{w}] = \mathrm{sgA}[\mathbf{u}, \mathbf{w}]+\mathrm{sgA}[\mathbf{v}, \mathbf{w}] \tag{a3}$$

for all $\mathbf{u}, \mathbf{v}, \mathbf{w}$ (Construct corresponding parallelograms to give a geometric proof to the above). We would rather keep the very useful properties (a1) and (a3) and accept that $\mathrm{sgA}[\mathbf{u}, \mathbf{v}]$ can take on negative values, than insisting on working with a nonnegative area function; we can always define, in the end, that $A[\mathbf{u}, \mathbf{v}] = |\mathrm{sgA}[\mathbf{u}, \mathbf{v}]|$.
The textbook gives several definitions of the determinant, the first one of which is based on similar consideration. But the textbook definitions require the validity of several assertions made there (e.g., it assumes that the volume defined by the row vectors and the volume defined by the column vectors of a square matrix are equal, which requires a proof). One also needs to verify that the different definitions are consist. The geometric definition is the most intuitive and provides effective algorithm for computing the determinant of a matrix by performing elementary column/row operations. The only unsatisfying aspect is that it presumes the notion of the volume of a

*https://www.desmos.com/calculator/895gqiflo9

parallelepiped which satisfies (a1–3). The recursive definition and the explicit formula do not require any geometric construction but involve tedious algebraic computations which are not practical even for computer implementation for large matrices. The textbook definitions also imply the validity of the formula along any column in (3.2) and any row in (3.6), both of which require proofs. Logically one should define the determinant using either the recursive definition or the explicit formula, then establishes properties (a1–3) from the definition and uses (a1–3) subsequently in practical computations.

Based on (a1–3) and the normalization requirement that $A[\mathbf{e}_1, \mathbf{e}_2] = 1$, it is easy to derive that sgA$[\mathbf{u}, \mathbf{v}]$ must be given by the usual formula for the determinant of the 2×2 matrix $[\mathbf{u}, \mathbf{v}]$ with $\mathbf{u}, \mathbf{v}$ as its columns (we are abusing notation using the same $[\mathbf{u}, \mathbf{v}]$ to denote both the geometric parallelogram and the matrix). Here is how the argument goes. Suppose $\mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} c \\ d \end{bmatrix}$. Let's first assume that $a \neq 0$, then $\mathbf{u} = a \begin{bmatrix} 1 \\ b/a \end{bmatrix}$, so according to (a1),

$$\det[\mathbf{u}, \mathbf{v}] = a \det \begin{bmatrix} 1 & c \\ b/a & d \end{bmatrix}.$$

Next $\mathbf{v}$ subtracting $c*$ the first column of the matrix on the right hand gives $\begin{bmatrix} 0 \\ d - cb/a \end{bmatrix}$, and according to (a2) and (a1),

$$\det \begin{bmatrix} 1 & c \\ b/a & d \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ b/a & d - cb/a \end{bmatrix} = (d - cb/a) \det \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix}.$$

Finally, the first column of the above matrix subtracting $b/a*$ the second column produces $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, so

$$\det \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 1.$$

Putting these together, we get

$$\det[\mathbf{u}, \mathbf{v}] = a(d - cb/a) = ad - bc.$$

When $a = 0$ and $c \neq 0$, or when $a = c = 0$, we just need to adjust the above arguments and in both cases we arrive at the same conclusion.

The same argument also works for a 3×3 matrix, and can be used to define the determinant of any $n \times n$ matrix, namely,

$$\det[\mathbf{v}_1, \cdots, \mathbf{v}_n] = \text{signed volume}[\mathbf{v}_1, \cdots, \mathbf{v}_n],$$

where signed volume$[\mathbf{v}_1, \cdots, \mathbf{v}_n]$ is the n-dimensional analogue of parallelepiped and obeys similar properties as in (a1–3).
(a2) means that $\det A$ is unchanged when the columns of A undergo the kind of "elementary column operations" as described in (a2). (a1) implies that if A has a column of 0's, then $\det A = 0$. These together also imply that when any column of A can be expressed as a linear combination of some other columns (namely when A is not invertible), then $\det A = 0$, as one can then perform a number of operations as in (a2) to produce a column of 0's. Furthermore, for any diagonal matrix D with its diagonal entries $d_1, \cdots, d_n$,

$$\det D = d_1 \cdots d_n.$$

This follows by applying (a1) multiple times to produce

$$\det D = d_1 \cdots d_n \cdot \det I_n = d_1 \cdots d_n,$$

where we still demand the normalization condition that $\det I_n = 1$.
For any upper triangular matrix R, if it has n pivots, then we can repeatedly apply the column operations in (a2) to reduce R to the diagonal matrix whose entries on the diagonal are the same as those of R, therefore conclude that $\det R =$ product of its diagonal entries.
At this point we don't have a definition for the volume or signed volumes of a general region in $\mathbb{R}^n$, but for any region spanned by n vectors in $\mathbb{R}^n$, the properties (a1–a3) are natural ones for the notion of signed volume.
A consequence of (a1–a3) is that if one swaps two columns (or rows) of matrix A to obtain B, then $\det B = -\det A$. Here is an argument for a 2×2 matrix $A = [\mathbf{u}\mathbf{v}]$ which also works for the general case. We already mentioned that (a1–a3) imply that $\det[\mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}] = 0$. Property (a3) then implies that

$$0 = \det[\mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}] = \det[\mathbf{u}, \mathbf{u}]+\det[\mathbf{u}, \mathbf{v}]+\det[\mathbf{v}, \mathbf{u}]+\det[\mathbf{v}, \mathbf{v}] = \det[\mathbf{u}, \mathbf{v}]+\det[\mathbf{v}, \mathbf{u}],$$

from which it follows that $\det[\mathbf{u}, \mathbf{v}] = -\det[\mathbf{v}, \mathbf{u}]$.
$\det A$ also plays the role of magnifying factor between the signed volumes of a region U and its image $A(U)$ under the linear function defined by A:

$$\text{signed volume of } A(U) = \det A \text{ (signed volume of } U).$$

This interpretation forces $\det(AB) = \det A \det B$, as

$$\begin{aligned}\text{signed volume of } AB(U) &= \det A \text{ (signed volume of } B(U)) \\ &= \det A \det B \text{ (signed volume of } U).\end{aligned}$$

As a consequence of this property,

$$\det(AB) = \det A \det B = \det(BA),$$

even though $AB \neq BA$ may happen. In addition, when A is invertible,

$$1 = \det I_n = \det(AA^{-1}) = \det A \det A^{-1},$$

so $\det(A^{-1}) = 1/\det(A)$.
Another useful fact is that $\det A^T = \det A$.
Based on the interpretation that $\det Q$ represents the signed volume of the parallelepiped with the columns of Q as its edges, $\det Q = \pm 1$ for any orthogonal matrix Q, as any two adjacent edges form a unit square. This can be justified rigorously using $Q^{\mathrm{T}}Q = I$ and

$$1 = \det I = \det(Q^{\mathrm{T}}Q) = \det Q^{\mathrm{T}} \det Q = (\det Q)^2,$$

where we have used the property that $\det(Q^{\mathrm{T}}Q) = \det Q^{\mathrm{T}} \det Q$ and $\det Q^{\mathrm{T}} = \det Q$.
As this point, when A has a QR factorization $A = QR$, the above properties imply that $\det A = \det Q \det R$, where $\det Q = \pm 1$, and $\det R$ is the product of its diagonal entries. The only remaining issue is to determine how to tell whether $\det Q = 1$ or -1.
As a consequence of the above properties, when a symmetric matrix A is diagonalized by an orthogonal matrix Q: $A = QDQ^T$,

$$\det A = \det Q \det D \det Q^T = (\det Q)^2 \det D = \det D,$$

where $\det D$ is the product of its diagonal entries, which is the product of the eigenvalues of A.
A most relevant property for us is

Theorem 3.2.1: (Invertibile matrix in terms of its determinant)

An $n \times n$ matrix A is invertible if and only if $\det A \neq 0$. Equivalently, An $n \times n$ matrix A is singular if and only if $\det A = 0$.

Proof. Suppose that A is invertible. Then there exists some $n \times n$ matrix B such that $AB = I_n$. It then follows that $1 = \det I_n = \det(AB) = \det A \det B$, which shows that $\det A \neq 0$.
Conversely, suppose $\det A \neq 0$. We aim to show that A is invertible. Suppose not, then the columns of A must be linearly dependent, for, a square matrix with linearly independent columns would be invertible. This means that there exist coefficients $c_1, \ldots, c_n$, not all 0, such that

$$c_1 A[:, 1] + \cdots + c_n A[:, n] = \bar{0}.$$

We may assume that $c_1 \neq 0$.Then, if we multiply the first column of A by c_1 to get a new matrix A_1, we would have $\det A_1 = c_1 \det A \neq 0$. But, according to (a2), if we add $c_i A[:, i]$ to column 1 of A_1 for each $i = 2, \cdots, n$, we would get a matrix B such that $\det B = \det A_1$. On the other hand, the first column of B is now $\bar{0}$. This forces $\det B = 0$ by (a1). This shows that, when $\det A \neq 0$, A must be invertible. □

Exercise 1. Applying (a1-3) to calculate

$$\det \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Exercise 2. Applying (a1-3) and/or using the interpretation that the determinant equals the signed volume of the parallelepiped with the columns of the matrix as edges to calculate the determinant of the Jacobian matrix of the map from cylindrical coordinates (r, θ, z) to the rectangular coordinates (x, y, z):

$$\det \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(HINT: The columns are orthogonal to each other.)

Exercise 3. Applying (a1-3) and/or using the interpretation that the determinant equals the signed volume of the parallelepiped with the columns of the matrix as edges to calculate the determinant of the Jacobian matrix of the map from spherical polar coordinates (r, θ, ϕ) to the rectangular coordinates (x, y, z):

$$\det \begin{bmatrix} \sin\phi\cos\theta & -r\sin\phi\sin\theta & r\cos\phi\cos\theta \\ \sin\phi\sin\theta & r\sin\phi\cos\theta & r\cos\phi\sin\theta \\ \cos\phi & 0 & -r\sin\phi \end{bmatrix}.$$

(HINT: The columns are orthogonal to each other.)

3.3 Diagonalizable Transformations and Eigenvectors

Here is again the definition of an eigenvector and an eigenvalue.

Definition 3.3.1: (Eigenvectors and Eigenvalues)

A d-dimensional *non-zero* column vector $\bar{x}$ is said to be an eigenvector of $d \times d$ matrix A, if the following relationship is satisfied for some scalar λ:

$$A\bar{x} = \lambda\bar{x}. \tag{3.11}$$

The scalar λ is referred to as its eigenvalue.

Suppose that $\{\bar{v}_1, \cdots, \bar{v}_d\}$ are eigenvectors of A with $\lambda_1, \cdots, \lambda_d$ be the corresponding eigenvalues. Then

$$A\bar{v}_i = \lambda_i\bar{v}_i, \quad i = 1, \cdots, d. \tag{3.12}$$

In matrix form this becomes

$$A[\bar{v}_1, \cdots, \bar{v}_d] = [\lambda_1\bar{v}_1, \cdots, \lambda_d\bar{v}_d] = [\bar{v}_1, \cdots, \bar{v}_d]\text{diag}[\lambda_1, \cdots, \lambda_d] = V\Delta,$$

with $V = [\bar{v}_1, \cdots, \bar{v}_d]$, $\Delta = \text{diag}[\lambda_1, \cdots, \lambda_d]$. Namely,

$$AV = V\Delta. \tag{3.13}$$

If the eigenvectors $\{\bar{v}_1, \cdots, \bar{v}_d\}$ are *linearly independent*, then V is invertible and the above equation is then equivalent to

$$A = V\Delta V^{-1}. \tag{3.14}$$

In such a case we say that matrix A is diagonalizable. Note that, whenever (3.13) holds, it implies that the columns of V are eigenvectors of A, and that (3.14) is equivalent to A having d linearly independent eigenvectors.

Remark 3.3.1

The textbook has a statement on p. 106 that "choosing V to have unit columns (which results in V^{-1} having unit rows)". This seems a mis-statement or a typo.

Note that the ith eigenvector $\bar{v}_i$ belongs to the null space of $A - \lambda_i I$ because $(A - \lambda_i I)\bar{v}_i = \bar{0}$ for the non-zero $\bar{v}$. In other words, the determinant of $A - \lambda_i I$ must be zero. This polynomial expression that yields the eigenvalue roots is referred to as the characteristic polynomial of A.

Definition 3.3.2: (Characteristic Polynomial)

The characteristic polynomial of a $d \times d$ matrix A is the degree-d polynomial in λ obtained by expanding $\det(A - \lambda I)$.

Observation 3.3.1 The characteristic polynomial $f(\lambda)$ of $d \times d$ matrix A is a polynomial in λ of the following form, where $\lambda_1, \cdots, \lambda_d$ are eigenvalues of A:

$$\det(A - \lambda I) = (\lambda_1 - \lambda) \cdots (\lambda_d - \lambda). \tag{3.15}$$

Therefore, the eigenvalues and eigenvectors of a matrix A can be computed as follows:

1. The eigenvalues of A can be computed by expanding $\det(A-\lambda I)$ as a polynomial expression in λ, setting it to zero, and solving for λ.

2. For each root λ_i of this polynomial, we solve the system of equations $(A-\lambda I)\bar{v} = \bar{0}$ in order to obtain the eigenvectors. We can pick a set of linearly independent eigenvectors with eigenvalue λ_i as a basis of the right null space of $(A - \lambda I)$.

In practice, it's not practical to compute the polynomial $\det(A - \lambda I)$ explicitly for large matrices; even if it is available, the roots of $\det(A - \lambda I) = 0$ can only be found approximately, so $\det(A - \lambda I)$ will be nearly but not exactly 0 for the computed λ_i; then $(A - \lambda_i I)\bar{v} = \bar{0}$ will only have $\bar{0}$ as a solution theoretically, so one has to develop tools to find approximate eigenvectors that approximate the true eigenvectors. These are not easy tasks; we will briefly discuss some aspects later on.

Example 3.3.1

The matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

has its eigenvalues given by $\det(A - \lambda I_3) = (1 - \lambda)^3 = 0$. So $\lambda = 1$ is the only eigevalue (we say that it has algebraic multiplicity 3 due to $(1-\lambda)^3$ in the factorization of the characteristic polynomial).
To find its eigenvectors, we solve

$$(A - I_3)\bar{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \bar{x} = \bar{0}.$$

This system has only one pivot and two free variables, with the solutions given

as

$$\bar{x} = \begin{bmatrix} x \\ 0 \\ z \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + z \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The eigenspace of A corresponding to $\lambda = 1$ is two-dimensional.
One interesting observation is that $(A - I_3)^2 = 0$ and that this matrix is not diagonalizable. For, if A were diagonalizable, then the diagonal entries must be the eigenvalues of A; in this case they must be 1's. So we would have $A = VIV^{-1} = I$, which can't be the case.

Exercise 1. Find the characteristic polynomial, the eigenvalues, and the eigenvectors of the matrix

$$\begin{bmatrix} -1 & -2 \\ 4 & 5 \end{bmatrix}.$$

3.3.1 Complex Eigenvalues

It is possible for the characteristic polynomial of a real-valued matrix to have complex roots. In fact, a rotation matrix with rotation angle not equal 0 or π has non-real eigenvalues.

Example 3.3.2

The eigenvalues of the rotation matrix

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

is given by solving

$$\det \begin{bmatrix} \cos\theta - \lambda & -\sin\theta \\ \sin\theta & \cos\theta - \lambda \end{bmatrix} = (\cos\theta - \lambda)^2 + \sin^2\theta = 0,$$

which gives $\lambda = \cos\theta \pm i\sin\theta$.
If we take $\lambda = \cos\theta + i\sin\theta$, its corresponding eigenvectors solve

$$\begin{bmatrix} -i\sin\theta & -\sin\theta \\ \sin\theta & -i\sin\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

whose solutions are given by $-ix - y = 0$ in the cases $\theta \neq 0, \pi$. Thus they are multiples of $[1, -i]^{\mathrm{T}}$.

If we take $\lambda = \cos\theta - i\sin\theta$, its corresponding eigenvectors solve

$$\begin{bmatrix} i\sin\theta & -\sin\theta \\ \sin\theta & i\sin\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

whose solutions are given by $ix - y = 0$ in the cases $\theta \neq 0, \pi$. Thus they are multiples of $[1, i]^{\mathrm{T}}$.
Using Euler's formla

$$e^{i\theta} = \cos\theta + i\sin\theta, e^{-i\theta} = \cos\theta - i\sin\theta,$$

the above eigenvector-eigenvalue relations can be written as

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix}.$$

Note that the pair of eigenvectors

$$\left\{ \begin{bmatrix} 1 \\ -i \end{bmatrix}, \begin{bmatrix} 1 \\ i \end{bmatrix} \right\}$$

are orthogonal in the sense of complex inner product:

$$\left\langle \begin{bmatrix} 1 \\ -i \end{bmatrix}, \begin{bmatrix} 1 \\ i \end{bmatrix} \right\rangle = \begin{bmatrix} 1 \\ -i \end{bmatrix}^{*} \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} 1 & i \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} = 0.$$

If we normalize each column to obtain

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{i}{\sqrt{2}} \end{bmatrix},$$

then $U^{-1} = U^{*}$, and we have

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} U = U \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix}, \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = U \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix} U^{-1}.$$

In general, if all the entries of A are real and A has non-real eigenvalues, then they must appear in complex conjugate pairs: if $a + bi$ is an eigenvalue of A for $a, b \in \mathcal{R}$, then so is $a - bi$; furthermore, if $\bar{p} + i\bar{q}$ is an eigenvector of A corresponding to $a + bi$, then $\bar{p} - i\bar{q}$is an eigenvector of A corresponding to $a - bi$.

The relation

$$A(\bar{p} + i\bar{q}) = (a + bi)(\bar{p} + i\bar{q})$$

is equivalent to the set of two matrix-vector equations:

$$A\bar{p} = a\bar{p} - b\bar{q}, A\bar{q} = b\bar{p} + a\bar{q}.$$

Remark 3.3.2

On p. 107, the textbook has the statement "the complex vector $u = a + ib$ is always orthogonal to the vector $v = i[a + ib]$". One can give a sense of the orthogonal relation, but not in the sense of Definition 2.11.3 of the textbook. For a non-zero complex number c, we have a sense that ic makes a 90° angle with c, but $\langle c, ic\rangle = c^*(ic) = |c|^2 i \neq 0$ is a purely imaginary number. $u = a+ib$ and $v = i[a + ib]$ have a similar relation:

$$\langle a + ib, i[a + ib]\rangle = i(|a|^2 + |b|^2),$$

which reflects that, geometrically, $a + ib$ and $i[a + ib]$ make a 90° angle.

3.3.2 Left Eigenvectors and Right Eigenvectors

A left eigenvector is a *non-zero row vector* $\bar{y}$, such that $\bar{y}A = \lambda\bar{y}$ for some scalar λ. A necessary and sufficient condition for A to have a left eigenvector with λ as eigenvalue is still $\det(A - \lambda I) = 0$.
Based on the relation $\bar{y}A = \lambda\bar{y}$, it becomes clear that $A^{\mathrm{T}}\bar{y}^{\mathrm{T}} = \lambda\bar{y}^{\mathrm{T}}$, so $\bar{y}^{\mathrm{T}}$ is a right eigenvector of A^{T} when $\bar{y}$ is a left eigenvector of A.
Suppose that A is a $d \times d$ diagonalizable matrix with

$$A = V\Lambda V^{-1}. \tag{3.16}$$

Then the columns of V are right eigenvectors of A. But the above relation is equivalent to

$$V^{-1}A = \Lambda V^{-1},$$

which shows that the rows of V^{-1} are left eigenvectors of A. Furthermore,

$$A^{\mathrm{T}} = (V^{-1})^{\mathrm{T}}\Delta V^{\mathrm{T}},$$

which shows that the columns of $(V^{\mathrm{T}})^{-1} = (V^{-1})^{\mathrm{T}}$, namely, the rows of V^{-1} are right eigenvectors of A^{T}.

3.3.3 Existence and Uniqueness of Diagonalization

The key condition for a $d \times d$ matrix to be diagonalizable is that there exists d linearly independent eigenvectors. So our basic strategy is (i). find all eigenvalues of the given matrix A, including complex ones; (ii). for each eigenvalue λ, find a basis of the null space $\text{Null}(A - \lambda I)$, which is a maximal set of linearly independent eigenvectors of A corresponding to λ; and (iii). examine whether the union of these different basis vectors are still linearly independent and whether there are d such basis vectors in total.

(i) and (ii) have clearly defined algorithms. For (iii), we first have

Lemma 3.3.1

The eigenvectors belonging to distinct eigenvalues are linearly independent.

As a consequence of this, we have

Lemma 3.3.2

When the roots of the characteristic polynomial are distinct, one can find d linearly independent eigenvectors. Therefore, a (possibly complex-valued) diagonalization $A = V\Delta V^{-1}$ of a real-valued matrix A with d distinct roots always exists.

When the characteristic polynomial of a matrix A are not all distinct, namely, when the characteristic polynomial has a factorization $\prod_i (\lambda_i - \lambda)^{r_i}$ with some $r_i > 1$, r_i is called the *algebraic multiplicity* of the eigenvalue λ_i. In such a case, the dimension of $\text{Null}(A - \lambda_i I)$ could be smaller than r_i. When this occurs, since $\sum_i r_i = d$, we would only have fewer than d linearly independent eigenvectors of A to work with. The matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

in Example 3.3.1 is one such example: $\lambda = 1$ has algebraic multiplicity 3 but the eigenspace corresponding to $\lambda = 1$ only has dimension 2.

If the dimension of $\text{Null}(A - \lambda_i I)$ is smaller than r_i for some eigenvalue λ_i, then a diagonalization does not exist. The closest we can get to a diagonalization is the Jordan normal form (see Section 3.3.4). Such a matrix is said to be *defective*. The dimension of $\text{Null}(A - \lambda_i I)$ is called the geometric multiplicity of λ_i.

When a matrix A is diagonalizable, the choice of V and Δ are not unique; but the diagonal entries of Δ must be the eigenvalues of A arranged in some order.

Exercise 2. Show that the matrix

$$A = \begin{bmatrix} -1 & -2 \\ 4 & 5 \end{bmatrix}$$

is diagonalizable and use it to construct the general solution of

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

Exercise 3. Let $A = \begin{bmatrix} -1/2 & 3/2 \\ 0 & 1 \end{bmatrix}$.

(a). Verify that $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ are eigenvectors of A and provide the corresponding eigenvalues.

(b). Prove that A is diagonalizable and write out a diagonalization of A.

(c). Compute A^6 using the above diagonalization of A and determine $\lim_{k\to\infty} A^k$.

3.3.4 Existence and Uniqueness of Triangulization

When the dimension of $\text{Null}(A - \lambda_i I)$ is smaller than r_i for some eigenvalue λ_i, it turns out that the inclusion relations

$$\text{Null}(A - \lambda_i I) \subset \text{Null}(A - \lambda_i I)^2 \subset \cdots \subset \text{Null}(A - \lambda_i I)^l \subset \text{Null}(A - \lambda_i I)^{l+1} \subset \mathcal{R}^d$$

will have a first l at which $\text{Null}(A - \lambda_i I)^l = \text{Null}(A - \lambda_i I)^{l+1}$. In such a case $\dim \text{Null}(A - \lambda_i I)^l$ turns out to match r_i. Furthermore, if we set $s_j = \dim \text{Null}(A - \lambda_i I)^{j+1} - \dim \text{Null}(A - \lambda_i I)^j$ for $j = 1, \cdots, l-1$, then $s_j > 0$. We first choose $\bar{v}_1, \cdots, \bar{v}_{s_{l-1}}$ linearly independent vectors in $\text{Null}(A - \lambda_i I)^l \setminus \text{Null}(A - \lambda_i I)^{l-1}$. Then, for each $m = 1, \cdots, s_{l-1}$, the vectors

$$\{(A - \lambda_i I)^{l-1} \bar{v}_m, \cdots, (A - \lambda_i I)\bar{v}_m, \bar{v}_m\}$$

are linearly independent, and are said to form a *Jordan chain*. Note that the first vector in this chain, $(A - \lambda_i I)^{l-1} \bar{v}_m$, is an eigenvector of A, due to $\bar{v}_m \in \text{Null}(A - \lambda_i I)^l$. The remaining vectors in this chain are called *generalized eigenvectors* of A in the sense that they are not annihilated by $A - \lambda_i I$ but by a power of $A - \lambda_i I$.

Using the relation that $(A-\lambda_i I)(A-\lambda_i I)^n = (A-\lambda_i I)^{n+1}$, and forming a matrix with the above vectors as columns in that order, we get

$$\begin{aligned}&(A-\lambda_i)[(A-\lambda_i I)^{l-1}\bar{v}_m, \cdots, (A-\lambda_i I)\bar{v}_m, \bar{v}_m] \\ =&[(A-\lambda_i I)^{l-1}\bar{v}_m, \cdots, (A-\lambda_i I)\bar{v}_m, \bar{v}_m] \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.\end{aligned}$$

This relation can also be written as

$$\begin{aligned}&A[(A-\lambda_i I)^{l-1}\bar{v}_m, \cdots, (A-\lambda_i I)\bar{v}_m, \bar{v}_m] \\ =&[(A-\lambda_i I)^{l-1}\bar{v}_m, \cdots, (A-\lambda_i I)\bar{v}_m, \bar{v}_m] \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix}.\end{aligned}$$

The upper triangular bi-diagonal matrix

$$\begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix} = \lambda_i I_l + \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} =: \lambda_i I_l + N$$

with λ_i on the diagonal, 1 in the $(j, j+1)$ position, is called a Jordan block. Note that

$$N^l = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}^l = 0.$$

This matrix is called Nilponent. This property allows us to compute

$$\left(\lambda_i I_l + \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \right)^n$$

using binomial expansion easily.
It turns out that if repeats this process successfully to $\text{Null}(A-\lambda_i)^m$ for $m = l-1, \cdots, 1$, keeping $(A-\lambda_i)^{l-m}\bar{v}_p$ obtained at previous stages as part of the linearly

independent vectors in $(A-\lambda_i)^m \setminus (A-\lambda_i)^{m-1}$ and augmenting additional linearly independent ones if $s_{m-1} > s_m$, then the collection of these vectors would have r_i vectors and are still linearly independent. Using these vectors to form an invertible matrix V, we would get $AV = VJ$, for some J consisting of stacking the various Jordan blocks along its diagonal. Such a matrix J is called a Jordan canonical form.

The purpose of diagonalizing a matrix, and when that is not possible, making $A = VJV^{-1}$ for a Jordan canonical form, is to simplify computations involving iterations of powers of A. As $A^k = V\Delta^k V^{-1}$ when $A = V\Delta V^{-1}$ and Δ^k is easily to compute for a diagonal matrix; J^k is also relatively easy to compute for a Jordan form.

In problems where one needs A^k for large k, the most dominant terms in Δ^k or J^k are associated with the eigenvalues with the largest absolute value. In order for a system not to have exponential growth, none of the eigenvalues should have absolute value greater than 1.

Another application is in solving a system of ODEs given by $\mathbf{x}'(t) = A\mathbf{x}(t)$. Making a change of variables $\mathbf{x}(t) = V\mathbf{y}(t)$, then the system reduces to $V\mathbf{y}'(t) = AV\mathbf{y}(t)$, so $\mathbf{y}'(t) = \Delta\mathbf{y}(t)$ when $A = V\Delta V^{-1}$ is diagonalizable. The system in $\mathbf{y}(t)$ is completely decoupled, and can be solved easily. When $A = VJV^{-1}$, we would need to work with $\mathbf{y}'(t) = J\mathbf{y}(t)$, which is still relatively easy to solve.

Let's take the case of J being a $s \times s$ Jordan block

$$\lambda I_l + N = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}.$$

Then we can solve $\mathbf{y}'(t) = J\mathbf{y}(t)$ by first solving the last component using $y_s'(t) = \lambda y_s(t)$ to get $y_s(t) = c_s e^{\lambda t}$, then plugging this into the immediate equation above:

$$y_{s-1}'(t) = \lambda y_{s-1}(t) + y_s(t) = \lambda y_{s-1}(t) + c_s e^{\lambda t}$$

to get $y_{s-1}(t) = (c_{s-1} + c_s t)e^{\lambda t}$. One can continue with this procedure to find the rest of the components.

A more streamlined procedure is to have

$$\mathbf{y}(t) = e^{Jt}\mathbf{y}(0),$$

where

$$e^{Jt} = e^{(\lambda I+N)t} = e^{\lambda t}e^{Nt} = e^{\lambda t}\left(I + Nt + \frac{N^2t^2}{2!} + \cdots + \frac{N^{s-1}t^{s-1}}{(s-1)!}\right)$$
$$= e^{\lambda t}\begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{s-1}}{(s-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{s-2}}{(s-2)!} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

where in the expansion for e^{Nt}, we used $N^k = 0$ for $k \geq s$. We caution that in general $e^{A+B} \neq e^Ae^B$ for matrices A and B due to $AB \neq BA$ in general. But $IN = NI = I$, so $e^{(\lambda I+N)t} = e^{\lambda t}e^{Nt}$.

Remark 3.3.3

For solutions to $\mathbf{x}'(t) = A\mathbf{x}(t)$, the behavior of $\mathbf{x}(t)$ as $t \to \infty$ is dictated by $e^{\lambda_i t}$ for various eigenvalues of A. But $|e^{\lambda_i t}| = e^{\Re(\lambda_i)t}$, so it is the real parts of the eigenvalues of A that determine the grow or decay of solutions of such a system, and often we only need to focus on λ with the largest real part and need not work out the precise Jordan form of a matrix.

3.3.5 Similar Matrix Families Sharing Eigenvalues

Definition 3.3.3

Two matrices A and B are said to be similar when $B = VAV^{-1}$ for some invertible V.

A basic property of similar matrices is that *they have the same eigenvalues (and their corresponding multiplicities)*. Furthermore, if $\bar{v}$ is an eigenvector of A, then $V\bar{v}$ is an eigenvector of B corresponding to the same eigenvalue. This is seen by using

$$A\bar{v} = \lambda\bar{v}, \bar{v} \neq \bar{0}, B(V\bar{v}) = V(A\bar{v}) = \lambda(V\bar{v}), \text{ and } V\bar{v} \neq \bar{0}.$$

Definition 3.3.4: (Trace of a Square Matrix)

The trace of the $d \times d$ matrix $A = [a_{ij}]$ is defined to be $\sum_{i=1}^{d} a_{ii}$, namely, the sum of the diagonal entries of A. It is denoted as $\mathrm{tr}A$.

A basic property regarding trace is that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Using this property, we see that if $B = VAV^{-1}$, then

$$\mathrm{tr}(B) = \mathrm{tr}(V^{-1}VA) = \mathrm{tr}(A).$$

In particular, if A is diagonalizable: $A = V\Delta V^{-1}$, we have $\mathrm{tr}(A) = \mathrm{tr}(\Delta)$ = sum of the eigenvalues of A.

3.3.6 Diagonalizable Matrix Families Sharing Eigenvectors

Definition 3.3.5: (Simultaneous Diagonalizability)

Two diagonalizable matrices A and B are said to be simultaneously diagonalizable, if a $d \times d$ invertible matrix V exists, such that the columns of V are the eigenvectors of both A and B in the sense[a]:

$$A = V\Delta_1 V^{-1}$$
$$B = V\Delta_2 V^{-1}$$

[a]The textbook makes a glaring mistake here, writing V^{T} in place of V^{-1} in the relations below.

The above relations mean that each column of V is a common eigenvector of A and B, with perhaps different eigenvalues. The definition says that A and B share a set of common eigenvectors which form a basis of $\mathcal{R}^d$, but does not say that any eigenvector of A must be an eigenvector of B. For instance, any diagonalizable matrix A and the identity matrix I are simultaneously diagonalizable, but any non-zero vector is an eigenvector of I, and it may not be an eigenvector of A.

Lemma 3.3.3

Diagonalizable matrices are simultaneously diagonalizable if and only if they are commutative.

Proof. Suppose that A and B are simultaneously diagonalizable, therefore satisfy the relations above. Then

$$AB = V\Delta_1 V^{-1}V\Delta_2 V^{-1} = V\Delta_1\Delta_2 V^{-1}, BA = V\Delta_2 V^{-1}V\Delta_1 V^{-1} = V\Delta_2\Delta_1 V^{-1}.$$

Since the diagonal matrices Δ_1 and Δ_2 satisfy $\Delta_1\Delta_2 = \Delta_2\Delta_1$, it follows that $AB = BA$.

Conversely, suppose that $AB = BA$. Let λ be an eigenvalue of A and $\bar{v}$ be an associated eigenvector of A: $A\bar{v} = \lambda\bar{v}$. Then

$$A(B\bar{v}) = B(A\bar{v}) = \lambda(B\bar{v})$$

shows that $B\bar{v} \in \text{Null}(A-\lambda I)$. We say that $\text{Null}(A-\lambda I)$ is invariant under B. It can be shown, under the assumption that B is diagonalizable, that there exists a basis of $\text{Null}(A-\lambda I)$ consisting of eigenvectors of B. These are then shared eigenvectors of A and B. One does that for every eigenvalue of A, and uses the diagonalizability of A to show that the collection of all such shared eigenvectors form a basis of $\mathcal{R}^d$, proving that A and B are simultaneously diagonalizable. □

3.3.7 Symmetric Matrices

One of the fundamental properties of (real) symmetric matrices is that they are always diagonalizable, and have orthonormal eigenvectors. This result is referred to as the spectral theorem:

Theorem 3.3.1: (Spectral Theorem)

Let A be a $d \times d$ symmetric matrix with real entries. Then, A is always diagonalizable with real eigenvalues and has orthonormal, real-valued eigenvectors. In other words, A can be diagonalized in the form $A = V\Delta V^{\mathrm{T}}$ with orthogonal matrix V.

The key ingredients and steps are

(i). An eigenvalue λ of A is characterized by the condition that $(A-\lambda I_n)\mathbf{x} = \mathbf{0}$ has non-zero solutions; i.e., $A-\lambda I_n$ does not have an inverse. An algebraic condition is that the determinant of $A-\lambda I_n$ must be 0. For a symmetric matrix A with real entries, its eigenvalues are all real.

(ii). For each λ such that the determinant of $A-\lambda I_n$ is 0, find the general solutions to $(A-\lambda I_n)\mathbf{x} = \mathbf{0}$. It will be spanned by a set of solution vectors. Apply the Gram-Schmidt Algorithm to produce a set of orthonormal spanning set for $\text{Null}(A-\lambda I_n)$.

(iii). If $\lambda_1 \neq \lambda_2$ are two distinct eigenvalues of A, with $\mathbf{x}_1$ and $\mathbf{x}_2$ being corresponding eigenvectors, then $\mathbf{x}_1 \perp \mathbf{x}_2$.

(iv). If we collect all the eigenvectors constructed above, it forms an orthonormal basis of $\mathbb{R}^n$: $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$, with $A\mathbf{q}_j = \lambda_j \mathbf{q}_j$, $1 \leq j \leq n$. Setting

$Q = [\mathbf{q}_1 \dots \mathbf{q}_n]$, then $Q^TQ = I_n$, $QQ^T = I_n$, and

$$AQ = Q\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} = QD; \qquad \text{thus } Q^{-1}AQ = Q^TAQ = D.$$

In implementing the above algorithm, the first step is to identify all eigenvalues of the matrix A, and for each eigenvalue λ of A to find an orthonormal basis for the space of solutions of $(A - \lambda I)\mathbf{x} = \mathbf{0}$. The second step is usually accomplished by first finding a basis for the space of solutions of $(A - \lambda I)\mathbf{x} = \mathbf{0}$, then applying the Gram-Schmidt Algorithm to obtain an orthonormal basis.

For a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, its eigenvalues are roots of the quadratic polynomial of $\det(A - tI) = t^2 - (a + d)t + ad - bc = 0$. For a 3×3 matrix A, $\det(A - tI)$ is a cubic polynomial in t. Theoretically there is a formula for the roots of a cubic polynomial, but it is not used in practice, as it is algebraically too complicated.

Example 3.3.3

For

$$A = \begin{bmatrix} 41 & -12 \\ -12 & 34 \end{bmatrix},$$

one can verify that

$$A\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 41 & -12 \\ -12 & 34 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 75 \\ 100 \end{bmatrix} = 25\begin{bmatrix} 3 \\ 4 \end{bmatrix},$$

so $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ is an eigenvector corresponding to eigenvalue $\lambda = 25$.

To find the full set of eigenvalues, one solves

$$\det(A - tI) = \det\begin{bmatrix} 41 - t & -12 \\ -12 & 34 - t \end{bmatrix} = t^2 - 75t + 1250 = 0$$

to find $t = 25$ and $t = 50$.

Next we find all solution of

$$(A - 50I)\mathbf{x} = \begin{bmatrix} -9 & -12 \\ -12 & -16 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which are given by solving $-3x - 4y = 0$. Thus all solutions are multiples of $\begin{bmatrix} 4 \\ -3 \end{bmatrix}$.

Note that the eigenvectors $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ -3 \end{bmatrix}$ are orthogonal to each other. To produce an orthogonal matrix we need to make each a unit vector. Thus the two columns of the following matrix are orthonormal eigenvectors of the given matrix A

$$Q = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix},$$

and the eigenvector relations are encoded in

$$AQ = Q \begin{bmatrix} 25 & 0 \\ 0 & 50 \end{bmatrix},$$

or equivalently, $Q^T AQ = \begin{bmatrix} 25 & 0 \\ 0 & 50 \end{bmatrix}$.

Note that in solving for $(A-50I)\mathbf{x} = \mathbf{0}$, we could have said that all solutions are multiples of $\begin{bmatrix} -4 \\ 3 \end{bmatrix}$. This would have resulted in a different orthogonal matrix

$$\begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}.$$

This matrix has determinant equal to 1, and represents a rotation matrix, while the previous choice has its determinant equal to -1 and represents a reflection matrix.

Proofs of (i) and (iii). (i). Let A be a real symmetric matrix and λ be an eigenvalue of A and $\bar{v}$ be an associated eigenvector: $A\bar{v} = \lambda\bar{v}$. At this point both λ and $\bar{v}$ could be non-real valued. However $\bar{v}^*\bar{v} > 0$ is real valued, and $\lambda\bar{v}^*\bar{v} = \bar{v}^*A\bar{v}$. The complex conjugate of the scalar $\bar{v}^*A\bar{v}$ can be computed as $(\bar{v}^*A\bar{v})^* = \bar{v}^*A^*\bar{v}$. Since A is real symmetric matrix, it follows that $A^* = A$ and $(\bar{v}^*A\bar{v})^* = (\bar{v}^*A\bar{v})$, proving that $\bar{v}^*A\bar{v}$ is real valued. Thus λ can be solved from $\lambda\bar{v}^*\bar{v} = \bar{v}^*A\bar{v}$ as the quotient of two real numbers, therefore is real.

(iii). Let $\bar{v}, \bar{w}$ be two eigenvectors of A: $A\bar{v} = \lambda\bar{v}, A\bar{w} = \mu\bar{w}$, with $\lambda \neq \mu$. Then

$$\bar{w}^*A\bar{v} = \lambda\bar{w}^*\bar{v}, \bar{w}^*A^*\bar{v} = \mu\bar{w}^*\bar{v}.$$

Since $A = A^*$, it follows that $\lambda\bar{w}^*\bar{v} = \mu\bar{w}^*\bar{v}$. Using $\lambda \neq \mu$, we conclude that $\bar{w}^*\bar{v} = 0$, namely, $\bar{v} \perp \bar{w}$. □

Remark 3.3.4

The above proof only used $A = A^*$. A matrix with this property is called a Hermitian matrix. (i)–(iii) hold for any Hermitian matrix; (iv) also holds with the modification that the orthonormal vectors $\mathbf{q}_j \in \mathbb{C}^n$ and the resulting Q satisfies $Q^*Q = I_n$. A matrix Q satisfying $Q^*Q = I_n$ is called a unitary matrix. When A is a real matrix and λ is a real-valued eigenvalue of A. If $\mathbf{p} + i\mathbf{q}$ is an associated eigenvector: $A(\mathbf{p} + i\mathbf{q}) = \lambda(\mathbf{p} + i\mathbf{q})$, then one can separate the real and imaginary parts of the two sides to conclude that $A\mathbf{p} = \lambda\mathbf{p}$ and $A\mathbf{q} = \lambda\mathbf{q}$. This means that $\mathbf{p}$ and $\mathbf{q}$ are themselves are eigenvectors, provided that they are not $\bar{0}$. In other words, in such a situation, one can always work with real-valued eigenvectors of A.

Remark 3.3.5

The statements after the definition of A-orthogonality are confusing, and not quite correct. I will clarify them later.

Exercise 4. For the matrix

$$A = \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix},$$

find an orthogonal matrix Q such that $AQ = QD$ for some diagonal matrix. Identify D.

Exercise 5. Find the eigenvalues of the matrices

$$Q_1 = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \\ \frac{4}{5} & -\frac{3}{5} \end{bmatrix} \text{ and } Q_2 = \begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix}.$$

Exercise 6. Given that the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix}$$

are -3 and 3.

(a). Diagonalize the matrix A, *i.e.*, find an invertible matrix $\mathbf{C}$ such that $\mathbf{C^{-1}A\,C = D}$, with $\mathbf{D}$ a diagonal matrix.
(b). For the same matrix $\mathbf{A}$, use your work in part (a) to find an orthogonal matrix $\mathbf{C_1}$ such that $\mathbf{C_1^{-1}A\,C_1 = D}$, with $\mathbf{D}$ a diagonal matrix.
(c). Use your work in part (a) or part (b) to compute $\mathbf{A}^4$.

Exercise 7. For the symmetric matrix $\begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}$ and the Hermitian matrix $\begin{bmatrix} 1 & i \\ -i & -1 \end{bmatrix}$, determine their respective eigenvalues and a basis for each eigenspace; then determine whether one or both are diagonalizable.

3.3.8 Positive Semidefinite Matrices

Definition 3.3.6: (Positive Semidefinite Matrix)

A $d \times d$ symmetric matrix A is positive semidefinite if and only if for any vector $\bar{x} \in \mathcal{R}^d$, the following is true:

$$\bar{x}^{\mathrm{T}} A \bar{x} \geq 0. \tag{3.17}$$

Here is a characterization of a positive semidefinite matrix in terms of its eigenvalues.

Lemma 3.3.4

A symmetric matrix is positive semidefinite if and only if all its eigenvalues are non-negative.

Proof. Suppose that A is positive semidefinite. Let λ be an eigenvalue of A and $\bar{v}$ be an associated eigenvector. We already know that λ and $\bar{v}$ are real-valued because A is real symmetric. Then

$$\lambda(\bar{v}^{\mathrm{T}}\bar{v}) = \bar{v}^{\mathrm{T}}(\lambda\bar{v}) = \bar{v}^{\mathrm{T}} A \bar{v} \geq 0,$$

and using $\bar{v}^{\mathrm{T}}\bar{v} > 0$, we conclude that $\lambda \geq 0$.
Conversely, suppose that all the eigenvalues of the symmetric matrix A are non-negative. Then the spectral theorem gives

$$A = P\mathrm{diag}[\lambda_1, \cdots, \lambda_d]P^{\mathrm{T}}$$

for some orthogonal matrix P with λ_j being the eigenvalues of A, and

$$\bar{x}^{\mathrm{T}} A \bar{x} = \bar{x}^{\mathrm{T}} P \mathrm{diag}[\lambda_1, \cdots, \lambda_d] P^{\mathrm{T}} \bar{x}.$$

If we set $\bar{y} = P^{\mathrm{T}}\bar{x}$, then $\bar{y}^{\mathrm{T}} = \bar{x}^{\mathrm{T}}P$ and

$$\bar{x}^{\mathrm{T}}A\bar{x} = \bar{y}^{\mathrm{T}}\mathrm{diag}[\lambda_1, \cdots, \lambda_d]\bar{y} = \lambda_1 y_1^2 + \cdots + \lambda_d y_d^2.$$

Since we assumed $\lambda_j \geq 0$, we conclude that $\bar{x}^{\mathrm{T}}A\bar{x} \geq 0$, proving that A is positive semidefinite. □

Definition 3.3.7: (Positive Definite Matrix)

A $d \times d$ symmetric matrix A is positive definite if and only if for any *non-zero* vector $\bar{x} \in \mathcal{R}^d$, the following is true:

$$\bar{x}^{\mathrm{T}}A\bar{x} > 0. \tag{3.18}$$

Corollary 3.3.2

If a $d \times d$ matrix A is positive semidefinite, then $A[i,i] \geq 0$ for every $i = 1, \cdots, d$. If A is positive definite, then $A[i,i] > 0$ for every $i = 1, \cdots, d$.

Proof. This follows by using $\bar{x}^{\mathrm{T}}A\bar{x} = \sum_{i,j=1}^{d} A[i,j]x_i x_j \geq 0$ and setting $\bar{x} = \bar{e}_i$ for some $i, 1 \leq i \leq d$, then $\bar{e}_i^{\mathrm{T}}A\bar{e}_i = A[i,i] \geq 0$ for a positive semidefinite A; when A is positive definite, $\bar{e}_i^{\mathrm{T}}A\bar{e}_i = A[i,i] > 0$. □

Note that the conditions $A[i,i] > 0$ for $i = 1, \cdots, d$ are not sufficient to make a symmetric A positive definite. For example

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -2 < 0,$$

even though all entries of A are positive. It turns out that a sufficient condition for a symmetric A to be positive definite is that $\det[A[i,j]]_{1 \leq i,j \leq k} > 0$ for each $k = 1, \cdots, d$. But in practice one does not use this criterion as computing the determinant of all these sub matrices is expensive.

Lemma 3.3.5

A symmetric matrix is positive definite if and only if all its eigenvalues are positive.

Lemma 3.3.6

A $d \times d$ matrix is positive semi-definite if and only if it can be expressed in the form $B^T B$ for some matrix B.
A $d \times d$ matrix is positive definite if and only if it can be expressed in the form $B^T B$ for some invertible matrix B.

Proof. We first show that $B^{\mathrm{T}}B$ is positive semidefinite, and positive definite for an invertible B. This follows from

$$\bar{x}^{\mathrm{T}}B^{\mathrm{T}}B\bar{x} = (B\bar{x})^{\mathrm{T}}(B\bar{x}) \geq 0,$$

for any $\bar{x} \in \mathcal{R}^d$, and when B is invertible and $\bar{x} \neq \bar{0}$, then $B\bar{x} \neq \bar{0}$, which leads to $(B\bar{x})^{\mathrm{T}}(B\bar{x}) > 0$.
Next, suppose that A is positive semidefinite. Then the Spectral Theorem

$$A = P\mathrm{diag}[\lambda_1, \cdots, \lambda_d]P^{\mathrm{T}}$$

for some orthogonal matrix P with λ_j being the eigenvalues of A. Since we know that $\lambda_j \geq 0$, we can define $B = \mathrm{diag}[\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_d}]P^{\mathrm{T}}$. It then follows that $A = B^{\mathrm{T}}B$. When A is positive definite, all $\lambda_j > 0$, then the matrix B is invertible. □

3.3.9 Cholesky Factorization: Symmetric LU Decomposition

The Cholesky Factorization writes a positive definite matrix A as LL^{T} where L is a lower-triangular matrix, and all diagonal entries of L are positive. The entries of L can be computed recursively as follows.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} & \cdots & l_{n1} \\ 0 & l_{22} & l_{32} & \cdots & l_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & l_{nn} \end{bmatrix},$$

and evaluating the first row

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \end{bmatrix} = l_{11} \begin{bmatrix} l_{11} & l_{21} & l_{31} & \cdots & l_{n1} \end{bmatrix}$$

gives

$$a_{11} = l_{11}^2, a_{1j} = l_{11}l_{j1}, j = 2, \cdots, n.$$

Since $a_{ii} > 0$, it follows that

$$l_{11} = \sqrt{a_{11}}, l_{j1} = a_{1j}/\sqrt{a_{11}}.$$

This determines all entries in the first column of L. Now look at the second row

$$\begin{bmatrix} a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \end{bmatrix} = l_{21} \begin{bmatrix} l_{11} & l_{21} & l_{31} & \cdots & l_{n1} \end{bmatrix} + l_{22} \begin{bmatrix} 0 & l_{22} & l_{32} & \cdots & l_{n2} \end{bmatrix}.$$

We conclude that $a_{22} = l_{21}^2 + l_{22}^2$, from which we solve

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{a_{22} - \frac{a_{12}^2}{a_{11}}}.$$

l_{22} has a positive solution when $a_{11}a_{22} - a_{12}^2 > 0$. Once l_{22} is determined, then l_{j2}, for $j > 2$, can be determined through

$$l_{j2} = \left(a_{2j} - l_{21}l_{j1}\right)/l_{22}.$$

This process can be continued as long as each subsequent l_{jj} can be solved as a positive solution. If this process stops at some stage, it would mean that the given matrix is not positive definite.

Once the existence of such an L is guaranteed, then the above algorithm is guaranteed to work. To show the existence of such an L, we use the factorization $A = B^{\mathrm{T}}B$, where B is invertible, and the QR factorization of $B = QR$. Then $B^{\mathrm{T}} = R^{\mathrm{T}}Q^{\mathrm{T}}$, and $A = R^{\mathrm{T}}Q^{\mathrm{T}}QR = R^{\mathrm{T}}R$. The factor R^{T} is lower triangular with positive diagonal entries .

3.4 Machine Learning and Optimization Applications

We already discussed the main applications of diagonalization of a matrix in the context of computing powers of a matrix or in solving a differential equation. This section is mostly focused on how a symmetric real matrix and a positive (semi)definite one could arise from applications.

For any $n \times d$ matrix D, the matrix $S = DD^{\mathrm{T}}$ is positive semidefinite, and is called the similarity matrix of D. It encodes the relations between the rows of D: $S[i, j] = D[i; :] \cdot D[j, ;]$.

Suppose that each column of D has mean 0, then the positive semidefinite matrix $C = D^{\mathrm{T}}D/n$ is called the covariance matrix of D. $C[i, j] = D[:, i] \cdot D[:, j]/n$ represents the covariance of the mean-centered columns $D[:, i]$ and $D[:, j]$ for $i \neq j$, and the variance of the column $D[:, i]$ when $i = j$.

One question discussed is to what extent S or C determine D. D certainly can't be determined from S or C, as for any $d \times d$ orthogonal matrix P, the similarity matrix of DP is $(DP)(DP)^{\mathrm{T}} = DPP^{\mathrm{T}}D^{\mathrm{T}} = DD^{\mathrm{T}}$ the same as that of D; and for any $n \times n$

orthogonal matrix Q, the covariance matrix of QD is $(QD)^{\mathrm{T}}QD = D^{\mathrm{T}}D$ is the same as that of D.

In optimization problems of a differentiable function in multi-dimensions, one often uses its Taylor expansion of order two to approximate the function. The Taylor expansion at some $\bar{x}_0$ takes the form

$$f(\bar{x}_0) + \nabla f(\bar{x}_0) \cdot (\bar{x} - \bar{x}_0) + \frac{1}{2}(\bar{x} - \bar{x}_0)^{\mathrm{T}}[\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)](\bar{x} - \bar{x}_0).$$

We will take $\bar{x}_0$ after a shift of the variable. Such a degree two polynomial attains a minimum or maximum at some $\bar{x}$, if its gradient is $\bar{0}$ at $\bar{x}$ and its Hessian matrix is positive definite (for a minimum) and negative definite (for a maximum). But the gradient of this function is (taking $\bar{x}_0 = \bar{0}$)

$$\nabla f(\bar{x}_0) + [\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)]\bar{x}$$

and its Hessian matrix is $[\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)]$. We need to investigate whether this Hessian matrix is positive definite, negative definite, or neither. Based on our discussions, this is answered by whether its eigenvalues are all positive, all negative, or neither. To answer more detailed questions such as the geometry of the level set

$$\{\bar{x} : \bar{x}^{\mathrm{T}}[\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)]\bar{x} = c\}$$

we use the Spectral Theorem on the Hessian matrix

$$[\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)] = P\mathrm{diag}[\lambda_1, \cdots, \lambda_n]P^{\mathrm{T}}$$

and introduce the new variable $\bar{y} = P^{\mathrm{T}}\bar{x}$, then $\bar{y}^{\mathrm{T}} = \bar{x}^{\mathrm{T}}P$ and

$$\bar{x}^{\mathrm{T}}[\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}_0)]\bar{x} = \bar{y}^{\mathrm{T}}\mathrm{diag}[\lambda_1, \cdots, \lambda_n]\bar{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2.$$

The set $\{\bar{y} : \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 = c\}$ can be described easily. Note that $\bar{x} = P\bar{y}$, so $\bar{y} = \bar{e}_j$ corresponds to $\bar{x} = P[:, j]$.

Exercise 1. For the quadratic function defined by

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 41 & -12 \\ -12 & 34 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

perform the change of variable

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

and find the function in terms of u, v.

Example 3.4.1

Consider the quadratic function $x^2 - xy + y^2$, which can be written as $[x\ y]A[x\ y]^T$, where $A = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$. We apply our algorithm above.

(i). An eigenvalue λ of A makes $A - \lambda I_2 = \begin{bmatrix} 1-\lambda & -\frac{1}{2} \\ -\frac{1}{2} & 1-\lambda \end{bmatrix}$ to have no inverse. This is equivalent to $\begin{bmatrix} 1-\lambda \\ -\frac{1}{2} \end{bmatrix} \parallel \begin{bmatrix} -\frac{1}{2} \\ 1-\lambda \end{bmatrix}$. But this is equivalent to $(1-\lambda)^2 - (-\frac{1}{2})^2 = 0$. This gives $\lambda = \frac{1}{2}$ or $\lambda = \frac{3}{2}$.

(ii). For $\lambda = \frac{1}{2}$, we need to find all solutions to

$$(A - \frac{1}{2}I_2)\mathbf{x} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The solutions $\mathbf{x} = (x\ \ y)^T$ are given by $x - y = 0$. Choosing y as a free variable, we see that the solutions are

$$\begin{bmatrix} y \\ y \end{bmatrix} = y \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We choose a normalized vector, which in this case is $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

Then for $\lambda = \frac{3}{2}$, we solve

$$(A - \frac{3}{2}I_2)\mathbf{x} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The solutions $\mathbf{x} = (x\ \ y)^T$ are given by $x + y = 0$. Choosing y as a free variable, we see that the solutions are

$$\begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

We choose a normalized vector, which in this case is $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$.

(iii). We note that $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ are orthonormal. If we set

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix},$$

then $Q^TQ = QQ^T = I_2$, and $AQ = Q\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}$. If we make the change of variables

$$\begin{bmatrix} x \\ y \end{bmatrix} = Q\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} x' \\ y' \end{bmatrix},$$

then $\begin{bmatrix} x & y \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} Q^T$, and

$$\begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} Q^TAQ \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2.$$

So under this change of variables, $x^2 - xy + y^2 = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2$.
For any $c > 0$, the level curve $x^2 - xy + y^2 = c$ becomes $\frac{1}{2}(x')^2 + \frac{3}{2}(y')^2 = c$ in the x'-y' coordinates. Since the relation between (x, y) and (x', y') coordinates here are of a rotation ($(x', y') = (1, 0)$ gets mapped to $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$, which means that the x'-axis gets mapped to the line $y = x$; and $(x', y') = (0, 1)$ gets mapped to $\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$, which means that the x'-axis gets mapped to the line $y = -x$). This change of variable allows us to see (a). the level curves of $x^2 - xy + y^2$ are ellipses, and (b). The function $x^2 - xy + y^2 = \frac{1}{2}(x')^2 + \frac{3}{2}(y')^2$ is non-negative, and can grow unbounded.

Remark 3.4.1

In step (ii) above for computing the eigenvectors associated with $\lambda = \frac{3}{2}$, if we had chosen x as a free variable, we would have gotten the eigenvectors as $x\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, and with $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ as its normalized eigenvector;

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} x' \\ y' \end{bmatrix}$$

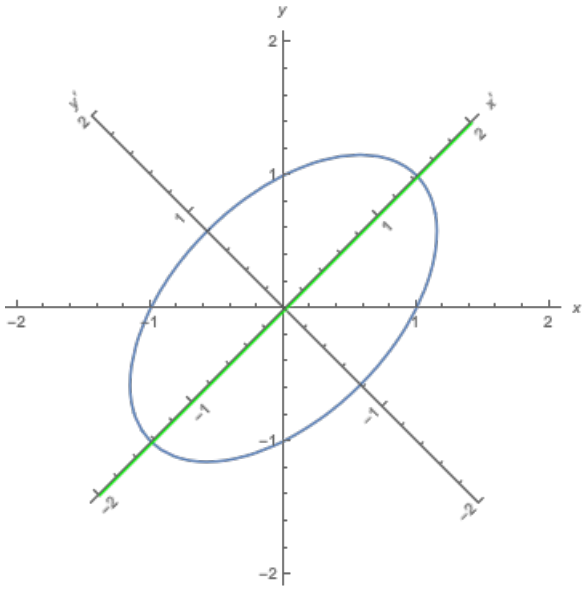


Figure 3.1: Contour plot of $x^2 - xy + y^2 = 1$ in the original x-y axes and in the rotated x'-y' axes.

would still have been an legitimate orthogonal change of variables, but this time, the y'-axis would have been mapped to be along the direction of the vector $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$, *making the x'-y' coordinate system left handed. This change of variables is through a reflection, instead of a rotation.*
Algebraically a rotation is represented by an orthogonal matrix with a positive determinant, while a reflection is represented by an orthogonal matrix with a negative determinant.

Exercise 2.

(i) For $2x^2 + 4xy + 5y^2$, perform a change of variables of the form

$$\begin{bmatrix} x \\ y \end{bmatrix} = Q \begin{bmatrix} u \\ v \end{bmatrix}$$

where Q is some orthogonal matrix, so that $2x^2 + 4xy + 5y^2$ transforms into $\lambda_1 u^2 + \lambda_2 v^2$. Identify Q, λ_1 and λ_2.

(ii) Given that $9, -36, 81$ are eigenvalues of the matrix

$$A = \begin{bmatrix} 21 & -42 & 36 \\ -42 & 36 & -6 \\ 36 & -6 & -3 \end{bmatrix}.$$

Find their corresponding eigenvectors. Then construct an orthonormal matrix Q such that $Q^T AQ$ is a diagonal matrix.

(iii) Use the result of the previous problem to perform a change of variables of the form

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = Q \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

to transform the quadratic function $21x^2 + 36y^2 - 3z^2 - 84xy - 12yz + 72xz$ into a quadratic function in u, v, w. Then describe the geometry of the level set $21x^2 + 36y^2 - 3z^2 - 84xy - 12yz + 72xz = 81$.

3.4.5 Eigenvectors in Norm-Constrained Quadratic Programming

When A is a real symmetric matrix, its eigenvalues and eigenvectors are related to the optimization problem:

$$\text{Optimize} \bar{x}^{\mathrm{T}} A \bar{x}, \text{ subject to } \|\bar{x}\|^2 = 1.$$

Using the orthogonal diagonalization $A = P\Lambda P^{\mathrm{T}}$ of P, with P orthogonal and $\Lambda = \text{diag}[\lambda_1, \cdots, \lambda_d]$, if we set $\bar{y} = P^{\mathrm{T}} \bar{x}$, then

$$\bar{x}^{\mathrm{T}} A \bar{x} = \bar{y}^{\mathrm{T}} \Lambda \bar{y} = \sum_{j=1}^{d} \lambda_j y_j^2,$$

and the constraint $\|\bar{x}\|^2 = 1$ is preserved: $\|\bar{y}\|^2 = 1$, as P and P^{T} preserve the norm. From this it is clear that

$$\max\{\bar{x}^{\mathrm{T}} A \bar{x} : \|\bar{x}\|^2 = 1\} = \max\{\lambda_j\}, \quad \min\{\bar{x}^{\mathrm{T}} A \bar{x} : \|\bar{x}\|^2 = 1\} = \min\{\lambda_j\}.$$

Furthermore, suppose that $\lambda_1 = \max\{\lambda_j\}$, then

$$\max\{\sum_{j=1}^{d} \lambda_j y_j^2 : \|\bar{y}\|^2 = 1\}$$

is attained when $\bar{y} = \bar{e}_1$, which corresponds to $\bar{x} = P\bar{e}_1$ equals the eigenvector of A from the first column of P. Similarly for the $\max\{\lambda_j\}$.
In fact the intermediate eigenvalues also have a characterization in terms of some optimization problem. If we order the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots$, then

$$\lambda_k = \max_{V_k} \min\{\bar{x}^{\mathrm{T}} A \bar{x} : \|\bar{x}\|^2 = 1, \bar{x} \in V_k, \text{ a } k\text{-dimensional subspace}\},$$

and

$$\lambda_1 + \cdots + \lambda_k = \max\{\mathrm{tr}\left(V^{\mathrm{T}} A V\right) : V \text{ is } n \times d \text{ with orthonormal columns}\}.$$

Note that if the columns of V are $\{\bar{v}_1, \cdots, \bar{v}_k\}$, then $\mathrm{tr}\left(V^{\mathrm{T}} A V\right) = \sum_{j=1}^{k} \bar{v}_j^{\mathrm{T}} A \bar{v}_j$.

3.5 Numerical Algorithms for Finding Eigenvectors

3.5.2 The Power Method for Finding Dominant Eigenvectors

The power method finds the eigenvector with the largest absolute eigenvalue of a matrix, which is also referred to as its *dominant eigenvector* or *principal eigenvector*. Consider a d × d matrix A, which is diagonalizable with real eigenvalues and with a basis of $\mathcal{R}^d$ consisting of eigenvectors $\bar{v}_1, \cdots, \bar{v}_d$ of A. We order these eigenvectors so that their corresponding eigenvalues are from the largest to the smallest in their absolute values: $|\lambda_1| \geq |\lambda_2| \geq \cdots$. Let's suppose that the largest one $|\lambda_1|$ has a one dimensional eigenspace spanned by $\bar{v}_1$. Start with a vector $\bar{x}$, which can be written as

$$\bar{x} = \sum_{j=1}^{d} \alpha_j \bar{v}_j$$

for some coefficients $\alpha_1, \cdots, \alpha_d$. If we pick $\bar{x}$ randomly, then $\alpha_1 \neq 0$. The key feature of the power method is that

$$A^k \bar{x} = \sum_{j=1}^{d} \alpha_j \lambda_j^k \bar{v}_j,$$

and the term $\alpha_1 \lambda_1^k \bar{v}_1$ becomes the dominant one in the sense that

$$\sum_{j=1}^{d} \alpha_j \lambda_j^k \bar{v}_j = \lambda_1^k \left(\alpha_1 \bar{v}_1 + \sum_{j=2}^{d} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \bar{v}_j \right).$$

Since $|\lambda_j|/|\lambda_1| < 1$ for $j \geq 2$, the terms in the summation quickly converges to 0. Namely,

$$A^k \bar{x} / \lambda_1^k \to \alpha_1 \bar{v}_1,$$

which is an eigenvector of A corresponding to λ_1.
Since we don't know λ_1 in advance, we can modify the last step above by examining the limit of

$$\lim_{k\to\infty} A^k\bar{x}/\|A^k\bar{x}\| = \bar{w}.$$

We see that $\bar{w}$ is a unit vector parallel to $\bar{v}_1$. We can then compute $A\bar{w}$ and compare it with $\bar{w}$ to find λ_1.
Note that in the implementation of this algorithm, one does not compute the power A^k of the matrix, instead, one computes $A^k\bar{x}$ as $A(A(\cdots(A\bar{x})))$. One may not use the normalization $A^k\bar{x}/\|A^k\bar{x}\|$, but use some other normalization, say, dividing by the largest magnitude of the components to make that component equal ± 1.
The idea can be modified to find other eigenvalues and eigenvectors of a diagonalizable matrix. Note that the ratio $|\lambda_2|/|\lambda_1| < 1$ determines how fast this algorithm converges. One idea of modification is to *shift* the matrix A by some ρI so as to make the modified ratio closer to 0. This is often done in combination with the *inverse power method*. Say, one has some approximate value ρ for λ_1, then $|\lambda_j - \rho|$ is much smaller than $|\lambda_j - \rho|$ for $j \neq 1$, which means that $|(\lambda_j - \rho)^{-1}|$ is much bigger than $|(\lambda_j - \rho)^{-1}|$ for $j \neq 1$! $(\lambda_j - \rho)^{-1}$ are the eigenvalues of $(A - \rho I)^{-1}$. So one applies the power method to this matrix $(A - \rho I)^{-1}$, namely, starting with some $\bar{x}_0$, one constructs $\bar{x}_{k+1}$ from $(A - \rho I)\bar{y}_{k+1} = \bar{x}_k$ and $\bar{x}_{k+1} = \bar{y}_{k+1}/\sigma_{k+1}$, where σ_{k+1} is a normalizing factor.

Chapter 4

Optimization Basics: A Machine Learning View

4.1 Introduction

Due to time constraint, we will not follow the development of the textbook carefully, but will only sketch a few main ideas. Since we have already covered SVD of a matrix, during which we also encountered several optimization problems, we will use those as examples in our discussion.

4.2 The Basics of Optimization

An optimization problem has an *objective function* that is defined in terms of a set of variables, referred to as *optimization variables*. The goal of the optimization problem is to compute the values of the variables at which the objective function is either maximized or minimized. In the setting of machine learning, objective function is often referred to as a *loss function*.

The objective function is often assumed to have continuous derivatives, and sometimes continuous second derivatives. Assume the objective function is $J(\bar{w})$, where $\bar{w} = (w_1, \cdots, w_d)$ lies in a domain U in $\mathcal{R}^d$. Here are some basic objects that will play a role in characterizing/finding the values of $\bar{w}$ at which J is either maximized or minimized:

Definition 4.2.1: Gradient and Hessian

Suppose that J has continuous derivatives in U and $\bar{w} \in U$, then the gradient

of J at $\bar{w}$ is the vector

$$\nabla J(\bar{w}) = (\frac{\partial J}{\partial w_1}(\bar{w}), \cdots, \frac{\partial J}{\partial w_d}(\bar{w}))^{\mathrm{T}} \in \mathcal{R}^d.$$

Suppose that J has continuous second derivatives in U and $\bar{w} \in U$, then the Hessian matrix of J at $\bar{w}$ is the $d \times d$ matrix

$$\mathrm{Hess}(J)(\bar{w}) := \left[\frac{\partial^2 J}{\partial w_i \partial w_j}(\bar{w})\right].$$

J is said to attain its maximum in U at $\bar{w}$ if $J(\bar{w}) \geq J(\bar{v})$ for any $\bar{v} \in U$.
Similarly, J is said to attain its minimum in U at $\bar{w}$ if $J(\bar{w}) \leq J(\bar{v})$ for any $\bar{v} \in U$.
J is said to attain a local maximum at $\bar{w} \in U$ if there exists a ball $B \subset U$ containing $\bar{w}$ such that J to attain its maximum in B at $\bar{w}$.
J is said to attain a local mnimum at $\bar{w} \in U$ if there exists a ball $B \subset U$ containing $\bar{w}$ such that J to attain its minimum in B at $\bar{w}$.

Here is the basic characterization of the values of $\bar{w}$ at which J is either maximized or minimized.

Theorem 4.2.1: Criteria for a Maximizer or a Minimizer

Suppose that J attains either the maximum or the minimum of J in U at $\bar{w} \in U$, then the gradient of J at $\bar{w}$ must be the zero vector $\bar{0}$ if J is differentiable at $\bar{w}$.
Suppose further that J has continuous second derivatives in U, then the Hessian matrix of J at $\bar{w}$ must be positive semidefinite at $\bar{w}$ if J attains the minimum at $\bar{w}$, and the Hessian matrix of J at $\bar{w}$ must be negative semidefinite at $\bar{w}$ if J attains the maximum at $\bar{w}$.
Conversely, suppose that J has continuous second derivatives in U, and at some $\bar{w} \in U$, the gradient $\nabla J(\bar{w}) = \bar{0}$ and its Hessian matrix $\mathrm{Hess}(J)(\bar{w})$ is positive definite, then $\bar{w}$ is a local maximum of J.
Suppose that J has continuous second derivatives in U, and at some $\bar{w} \in U$, the gradient $\nabla J(\bar{w}) = \bar{0}$ and its Hessian matrix $\mathrm{Hess}(J)(\bar{w})$ is negative definite, then $\bar{w}$ is a local minimum of J.

The possibility of J attaining a maximum or minimum at a point $\bar{w}$ where J is not differentiable needs to be watched out, as in the case of $J(x) = |x|$ on $-1 \leq x \leq 1$.

Proof for the second part uses the Taylor expansion of second order of J at $\bar{w}$: for a

sufficiently small ball B containing $\bar{w}$ and $\bar{v} \in B$,

$$J(\bar{v}) = J(\bar{w}) + \nabla J(\bar{w})^{\mathrm{T}}(\bar{v} - \bar{w}) + \frac{1}{2}(\bar{v} - \bar{w})^{\mathrm{T}}\mathrm{Hess}(J)(\bar{u})(\bar{v} - \bar{w}), \tag{4.1}$$

where $\bar{u}$ is some vector in B depending on $\bar{v}$. The continuity assumption on the second derivatives of J imply that for any $\epsilon > 0$, there exists sufficiently small ball B containing $\bar{w}$, $\|\mathrm{Hess}(J)(\bar{w}) - \mathrm{Hess}(J)(\bar{u})\|_{\mathcal{F}} < \epsilon$. As a consequence, if $\mathrm{Hess}(J)(\bar{w})$ is positive definite and $\epsilon > 0$ is sufficiently small, then $\mathrm{Hess}(J)(\bar{u})$ is also positive definite for all $\bar{u} \in B$. The Taylor approximation (4.1) also has an analogue for $\nabla J(\bar{v})$:

$$\|\nabla J(\bar{v}) - \nabla J(\bar{w}) - \mathrm{Hess}(J)(\bar{w})\,(\bar{v} - \bar{w})\,\| \leq\leq \epsilon\|\bar{v} - \bar{w}\|.$$

In other words, the quadratic function

$$Q(\bar{v}) := J(\bar{w}) + \nabla J(\bar{w})^{\mathrm{T}}(\bar{v} - \bar{w}) + \frac{1}{2}(\bar{v} - \bar{w})^{\mathrm{T}}\mathrm{Hess}(J)(\bar{w})(\bar{v} - \bar{w})$$

of $\bar{v}$ provides an approximation for $J(\bar{v})$ for $\bar{v} \in B$, with an error which is no greater than $\frac{\epsilon}{2}\|\bar{v} - \bar{w}\|^2$, and the gradient of $Q(\bar{v})$ also provides an approximation for the gradient of $J(\bar{v})$ with an error which is no greater than $\epsilon\|\bar{v} - \bar{w}\|$, using

$$\nabla Q(\bar{v}) = \nabla J(\bar{w}) + \mathrm{Hess}(J)(\bar{w})(\bar{v} - \bar{w}), \quad \mathrm{Hess}(Q)(\bar{v}) = \mathrm{Hess}(J)(\bar{w}).$$

If we only assume that J has continuous derivatives, then we also have a Taylor expansion of order 1:

$$J(\bar{v}) = J(\bar{w}) + \nabla J(\bar{u})^{\mathrm{T}}(\bar{v} - \bar{w}), \tag{4.2}$$

where $\bar{u}$ is some vector in B depending on $\bar{v}$.

One basic method to study a function of several variables is to study its restriction to any one-dimensional line: set $f(t) = J(\bar{w} + t\bar{z})$ for any fixed $\bar{z}$ (as a direction vector). Then

$$f'(t) = \nabla J(\bar{w} + t\bar{z})^{\mathrm{T}}\bar{z}, \quad f''(t) = \bar{z}^{\mathrm{T}}\mathrm{Hess}(J)(\bar{w} + t\bar{z})\bar{z}.$$

In particular,

$$f'(0) = \nabla J(\bar{w})^{\mathrm{T}}\bar{z}, \quad f''(0) = \bar{z}^{\mathrm{T}}\mathrm{Hess}(J)(\bar{w})\bar{z}.$$

The formulae above imply that if J attains a (local) maximum or minimum at $\bar{w}$, then

$$\nabla J(\bar{w})^{\mathrm{T}}\bar{z} = 0, \quad \bar{z}^{\mathrm{T}}\mathrm{Hess}(J)(\bar{w})\bar{z} \geq 0, \text{ for any } \bar{z} \text{ in the case of } J(\bar{w}) \text{ being a minimum.}$$

As a consequence, when $J(\bar{w})$ is a minimum, $\nabla J(\bar{w}) = \bar{0}$ and $\mathrm{Hess}(J)(\bar{w})$ is positive semidefinite; when $J(\bar{w})$ is a maximum, $\nabla J(\bar{w}) = \bar{0}$ and $\mathrm{Hess}(J)(\bar{w})$ is negative semidefinite.

Furthermore,

$$\min\{f'(0) : \|\bar{z}\| = 1\} = -\|\nabla J(\bar{w})\|,$$

which is attained when $\bar{z} = -\nabla J(\bar{w})/\|\nabla J(\bar{w})\|$ (if $\|\nabla J(\bar{w})\| \neq 0$). $-\nabla J(\bar{w})$ is referred to as the direction of greatest descent of J at $\bar{w}$.

Remark 4.2.1: Quadratic Objective Functions

We have already encountered several quadratic objective functions. In the least squares problem, we need to find the vector(s) $\bar{w}$ that minimizes the following quadratic function

$$J(\bar{w}) = \|A\bar{w} - \bar{b}\|^2 = (A\bar{w} - \bar{b})^T(A\bar{w} - \bar{b}) = \bar{w}^T A^T A\bar{w} - 2\bar{b}^T A\bar{w} + \bar{b}^T\bar{b}$$

using $\bar{w}^T A^T\bar{b} = \bar{b}^T A\bar{w}$. For this quadratic function we see that

$$\nabla J(\bar{w}) = 2A^T A\bar{w} - 2A^T\bar{b}, \quad Hess(J)(\bar{w}) = 2A^T A.$$

So its Hessian matrix is a constant symmetric positive semidefinite matrix $2A^T A$.
If A itself is real and symmetric, we can consider the quadratic function

$$Q(\bar{w}) = \frac{1}{2}\bar{w}^T A\bar{w} - \bar{b}^T\bar{w}.$$

Note that

$$\nabla Q(\bar{w}) = A\bar{w} - \bar{b}, \quad Hess(Q)(\bar{w}) = A.$$

So any $\bar{w}$ satisfying $\nabla Q(\bar{w}) = \bar{0}$ solves $A\bar{w} = \bar{b}$. In particular, if Q has a maximum or minimum at some $\bar{w}$, then $A\bar{w} = \bar{b}$.
These relations between the maximizer/minimizer of a quadratic function and a linear algebra problem provides an option for solving the linear algebra problem by algorithms which computes solutions to these optimization problems.
The best rank-k approximation problem for an $n \times d$ matrix A appears to be minimizing a quadratic function

$$\|A - UV^T\|_{\mathcal{F}}^2,$$

in the entries of the $n \times k$ matrix U and the $d \times k$ matrix V, as one only sees some quadratic terms in U and V in these entries after expanding out the above squared norm, but some terms involve the product of quadratic terms in entries U with quadratic terms in entries V, which make this function of quartic function of the unknowns.

We are concerned with the following questions.

(I). What conditions on a function J and its domain U would guarantee that J attains a minimum either at some interior point $\bar{w} \in U$ (preferably a unique one), or at some boundary point $\bar{w} \in \partial U$?

(II). Are there efficient algorithms to compute the points where J attains its minimum?

4.3 Convex Functions and Convex Sets

It turns out that the class of *convex functions* gives good answers to question (I). The notion of a convex set in $\mathbb{R}^n$ is more general than that of a convex function.

Definition 4.3.1

A set C in $\mathbb{R}^n$ is called convex, if for any $\mathbf{a}, \mathbf{b} \in C$ and any $t \in \mathbb{R}, 0 \leq t \leq 1$, we have $(1-t)\mathbf{a} + t\mathbf{b} \in C$.

Geometrically, the set $\{(1-t)\mathbf{a} + t\mathbf{b} : 0 \leq t \leq 1\}$ is the line segment in $\mathbb{R}^n$ with $\mathbf{a}, \mathbf{b}$ as its ends. A convex set needs not have any interior point.

Definition 4.3.2

A real-valued function f defined on a convex set C is called convex, if for any $\mathbf{a}, \mathbf{b} \in C$ and any $t \in \mathbb{R}, 0 \leq t \leq 1$, we have

$$f((1-t)\mathbf{a} + t\mathbf{b}) \leq (1-t)f(\mathbf{a}) + tf(\mathbf{b}).$$

f is called strictly convex if we have the strict inequality

$$f((1-t)\mathbf{a} + t\mathbf{b}) < (1-t)f(\mathbf{a}) + tf(\mathbf{b}) \text{ for any } 0 < t < 1.$$

f is called concave if $-f$ is convex. Equivalently, the defining inequality above is reversed for a concave function.

Geometrically, if we construct a line in $\mathbb{R}^n \times \mathbb{R} \supset C \times \mathbb{R}$ through the points $(\mathbf{a}, f(\mathbf{a})), (\mathbf{b}, f(\mathbf{b}))$, then it has parametric equation

$$\mathbf{x} = ((1-t)\mathbf{a} + t\mathbf{b}, (1-t)f(\mathbf{a}) + tf(\mathbf{b})),$$

so $(1-t)f(\mathbf{a}) + tf(\mathbf{b})$ is the "height of the line above the point" $(1-t)\mathbf{a} + t\mathbf{b}$. When f is convex, $f((1-t)\mathbf{a} + t\mathbf{b})$ stays below the height at $(1-t)\mathbf{a} + t\mathbf{b}$ of the above line segment for $0 \leq t \leq 1$. Here* is a Desmos page illustrating this geometric property.

*www.desmos.com/calculator/mbdd9ymzug

f is a convex function iff the set $\{(\mathbf{x}, y) : \mathbf{x} \in C, y \geq f(\mathbf{x})\}$ in $\mathbb{R}^n \times \mathbb{R}$, called the *epigraph* of f, is convex. Another characterization of a convex function is that for every real number c, the *sub level set* of f defined by $\{\mathbf{x} : f(\mathbf{x}) \leq c\}$ is a convex set. Because of this relation, properties of convex functions can often be studied as properties of convex functions. The notion of *supporting hyperplane* of a convex set and that of the graph of a convex function plays an important role in studying properties of a convex set and a convex function.
Here are two useful Lemmas.

Lemma 4.3.1

Suppose that C_α for $\alpha \in \Lambda$ is a family of convex sets in $\mathbb{R}^d$ and that $\cap_{\alpha\in\Lambda}$ is non-empty, then it is also a convex set.
Suppose that f_α for $\alpha \in \Lambda$ is a family of convex functions defined on a convex set $C \subset \mathbb{R}^d$ and that $\sup_\alpha f_\alpha(\bar{x})$ is not ∞ for $\bar{x} \in C$, then it is a convex function on C.

Lemma 4.3.2

Suppose that f and g are convex functions on C, ϕ is an increasing and convex function defined on $\mathbb{R}$, and $\alpha : K \mapsto C$ is an affine map, namely,

$$\alpha((1-t)\mathbf{a} + t\mathbf{b}) = (1-t)\alpha(\mathbf{a}) + t\alpha(\mathbf{b})$$

for any $\mathbf{a}, \mathbf{b} \in K$ and $0 < t < 1$, where K is also a convex set. Then

(a). $f + g$ is convex on C.

(b). $\phi \circ f$ is convex on C.

(c). $f \circ \alpha$ is convex on K.

(d). Suppose that f has continuous second derivatives, then its Hessian matrix $Hess(f)(\mathbf{a})$ is positive semidefinite. Conversely, if f has continuous second derivatives and its Hessian matrix $Hess(f)(\mathbf{a})$ is positive semidefinite (definite) for every $\mathbf{a} \in C$, then f is convex (strictly convex) in C.

(e). if f is strictly convex on C, then it can attain its minimum in C at most at one point.

Proof. We only give a proof for (b) and (d). For any $\mathbf{a}, \mathbf{b} \in C$ and any $t \in \mathbb{R}, 0 \leq t \leq 1$, we have

$$f((1-t)\mathbf{a} + t\mathbf{b}) \leq (1-t)f(\mathbf{a}) + tf(\mathbf{b}).$$

Since ϕ is assumed to be increasing, it follows that

$$\phi \circ f((1-t)\mathbf{a} + t\mathbf{b}) \leq \phi\left((1-t)f(\mathbf{a}) + tf(\mathbf{b})\right).$$

But ϕ is also assumed to be convex, it follows that

$$\phi\left((1-t)f(\mathbf{a}) + tf(\mathbf{b})\right) \leq (1-t)\phi \circ f(\mathbf{a}) + t\phi \circ f(\mathbf{b}).$$

Combining these two inequalities shows that $\phi \circ f$ is convex.
For (d), if f has continuous second derivatives, then the function of t, $h(t) := f((1-t)\mathbf{a} + t\mathbf{b})$ is a convex function of $t \in [0,1]$ with two derivatives, so $f''(t) \geq 0$ for all $t \in [0,1]$. But

$$f''(t) = (\mathbf{b}-\mathbf{a})^{\mathrm{T}}\mathrm{Hess}(f)((1-t)\mathbf{a} + t\mathbf{b})(\mathbf{b}-\mathbf{a}).$$

At any $\mathbf{a} \in U$, there exists some $\delta > 0$ such that for any unit vector $\mathbf{u}$, $\mathbf{a} + \delta\mathbf{u} \in U$. Then we choose $\mathbf{b} = \mathbf{a} + \delta\mathbf{u}$ and use $f''(0) \geq 0$ to imply that $\delta^2\mathbf{u}^{\mathrm{T}}\mathrm{Hess}(f)(\mathbf{a})\mathbf{u} \geq 0$. This shows that $\mathrm{Hess}(f)(\mathbf{a})$ is positive semidefinite.
Conversely, suppose that f has continuous second derivatives and $\mathrm{Hess}(f)(\mathbf{a})$ is positive semidefinite for any $\mathbf{a} \in U$. Then the function of one variable t, $h(t) := f((1-t)\mathbf{a} + t\mathbf{b})$ is a convex function of $t \in [0,1]$ due to $f''(t) \geq 0$. But this shows that

$$f((1-t)\mathbf{a} + t\mathbf{b}) = h(t) \leq (1-t)h(0) + th(1) = (1-t)f(\mathbf{a}) + tf(\mathbf{b}).$$

□

When each $f_\alpha(\bar{x})$ is an affine function, $\sup_\alpha f_\alpha(\bar{x})$ is called the *envelope* of this family of affine functions. It turns out that every convex function is the envelope of its (affine) support functions.

Example 4.3.1

If $C \subset \mathbb{R}^n$ and A is a $n \times d$ matrix and $\bar{b} \in \mathbb{R}^n$, then $\alpha(\bar{x}) = A\bar{x} + \bar{b}$ defines an affine map from $\mathbb{R}^d$ to $\mathbb{R}^n$; furthermore, $K := \{\bar{x} \in \mathbb{R}^d : \alpha(\bar{x}) \in K\}$ is a convex set of $\mathbb{R}^d$.
Any norm on $\mathbb{R}^d$ is a convex function of $\bar{x} \in \mathbb{R}^d$. This follows from the triangle inequality and degree-1 homogeneity of a norm:

$$\|(1-t)\mathbf{a} + t\mathbf{b}\| \leq \|(1-t)\mathbf{a}\| + \|t\mathbf{b}\| = (1-t)\|\mathbf{a}\| + t\|\mathbf{b}\|$$

for $0 < t < 1$. For any $p > 1$, the l^p-norm $\|\bar{x}\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ is strictly convex. The l^1-norm is not strictly convex, even in 1-dimension, as demonstrated by $|x|$.

As a consequence, $\|A\bar{x} - \bar{b}\|_p$ for $p \geq 1$ and the quadratic function $\|A\bar{x} - \bar{b}\|_2^2$ are convex functions of $\bar{x}$.
If A itself is real and symmetric, then the quadratic function $Q(\bar{w}) = \frac{1}{2}\bar{w}^{\mathrm{T}} A\bar{w} - \bar{b}^{\mathrm{T}}\bar{w}$ is convex if and only if the matrix A is positive semidefinite. This is because $\mathrm{Hess}(Q)(\bar{w}) = A$ for any $\bar{w}$.

Note that a convex function, even a strictly convex one, may not attain its minimum on its domain or boundary, as in the case of $x^2 + by$ on $\mathbb{R}^2$ or $\ln x^{-1}$ on $\mathbb{R}^+$. But A is positive definite, then the quadratic function $Q(\bar{w}) = \frac{1}{2}\bar{w}^{\mathrm{T}} A\bar{w} - \bar{b}^{\mathrm{T}}\bar{w}$ attains its minimum on $\mathbb{R}^d$ at a unique point $\bar{x}$, which is the unique solution to $A\bar{x} = \bar{b}$.

Remark 4.3.1

Note that a convex set in $\mathcal{R}^d$ needs not have interior points, as are the cases of a straight-line or a plane in $\mathcal{R}^3$. Many optimization problems involve finding the minimum of an objective function whose domain is constrained by one or more equalities or inequalities of linear expressions of $\bar{x} \in \mathcal{R}^d$ for the form $g_i(\bar{x}) := a_{i1}x_1 + \cdots + a_{id}x_d \leq (\text{or} =) b_i$. Each $\{\bar{x} \in \mathcal{R}^d : a_{i1}x_1 + \cdots + a_{id}x_d \leq (\text{or} =) b_i\}$ is convex in $\mathcal{R}^d$, so the constraints define the intersection of a collection of convex sets, which is a convex set, and we need to solve an optimization problem for a function on such a convex set.
In simple cases, the constraining convex set can be described (parametrized) by a number of unconstrained variables. For example, the constraint $\{(x, y) : ax + by = c\}$, where $b \neq 0$, can be described as $\{(x, y) : y = -\frac{a}{b}x + \frac{c}{b}\}$, while the constraint $\{(x, y) : x + 2y = 3, x \geq 0, y \geq 0\}$ can be described as $\{(x, y) : 0 \leq y \leq \frac{3}{2}, x = 3 - 2y\}$. One can substitute the solved variable in terms of the unconstrained variable into the objective function to reduce the problem to an optimization problem without constraints.

Exercise 1. Show that for $\bar{x}, \bar{y} \in \mathcal{R}^d$, the function $(c - \bar{x}^{\mathrm{T}}\bar{y})^2$ is convex as a function of either $\bar{x}$ or $\bar{y}$, but not convex as a joint function of $(\bar{x}, \bar{y})$.

Exercise 2. Show that for any $c > 0$ and any $x \in \mathcal{R}$, the function $c|y| + \frac{1}{2}(y - x)^2$ of y on $\mathcal{R}$ is strictly convex, and attains its minimum at

$$y_* = S_c(x) := \begin{cases} x - c & \text{if } x > c; \\ 0 & \text{if } c \geq x \geq -c; \\ x + c & \text{if } x < -c. \end{cases}$$

$S_c(x)$ is called the shrinkage function in statistics and this method of finding a minimum of a regularized $|y|$ is called soft thresholding. HINT: Apply the criteria that a function attains its minimum either at any interior point where its derivatives vanish or are not defined, or at a boundary point.

Exercise 3. Assume that $a, b \neq 0$. Consider the constrained minimization problem

$$\min\{|x| + |y| : ax + by = c\}$$

Show that the minimum is attained at either $(c/a, 0)$ or $(0, c/b)$. HINT: You may apply the same strategy as suggested in the previous problem.

Exercise 4. Assume that $a, b, c \neq 0$. Consider the constrained minimization problem

$$\min\{|x| + |y| + |z| : ax + by + cz = d\}$$

Show that the minimum is attained at either $(d/a, 0, 0)$ or $(0, d/b, 0)$, or $(0, 0, d/c)$.

Exercise 5. Find $\min\{|x| + |y| + |z| : x + 2y + 3z = 6, 3x + 2y + z = 12\}$.

Exercise 6. Find $\min\{x + y + z : x + 2y + 3z = 6, x, y, z \geq 0\}$ and $\max\{x + y + z : x + 2y + 3z = 6, x, y, z \geq 0\}$.

Exercise 7. Compare the gradients of the functions

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \text{ and } g(x, y) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Exercise 8. Let $\bar{X}_i, i = 1, \cdots, m$, be a collection of data vectors in $\mathcal{R}^d$. Find $\bar{W}$ which minimizes $\sum_{i=1}^m \|\bar{X}_i - \bar{W}\|^2$.

Exercise 9. Let C be an $n \times d$ matrix, A be an $n \times m$ matrix, and B be a $k \times d$ matrix. One seeks an $m \times k$ matrix X to minimize $J(X) := \|C - AXB\|_{\mathcal{F}}^2$. Compute the gradient of J with respect to X and set it to be $\bar{0}$ to identify the conditions for X to be a minimizer and find solutions to this equation. Compare them with the solutions given by using SVD method. HINT: $X = A^+CB^+$ is a solution.

4.4 Gradient Descent Algorithms

4.4.1 Line Search and Direct Gradient Descent

Our discussion earlier shows that the restriction of J to the line through $\bar{w}$ in the direction of $\bar{p}$, $f(t) = J(\bar{w} + t\bar{p})$ satisfies

$$f'(t) = \nabla J(\bar{w} + t\bar{p})^{\mathrm{T}}\bar{p}, \quad f'(0) = \nabla J(\bar{w})^{\mathrm{T}}\bar{p}.$$

As a consequence

$$\min\{f'(0) : \|\bar{p}\| = 1\} = -\|\nabla J(\bar{w})\|,$$

which is attained when $\bar{p} = -\nabla J(\bar{w})/\|\nabla J(\bar{w})\|$ (if $\|\nabla J(\bar{w})\| \neq 0$), the direction of greatest descent. We will not force the normalization $\|\bar{p}\| = 1$ and will take $\bar{p} = \bar{r} := -\nabla J(\bar{w})$. For the quadratic function $J = \frac{1}{2}\bar{w}^{\mathrm{T}}A\bar{w} - \bar{b}^{\mathrm{T}}\bar{w}$, since $\nabla J(\bar{w}+t\bar{p}) = A(\bar{w} + t\bar{p}) - \bar{b}$, $\bar{r} := \bar{b} - A\bar{w}$, called the residue of the proposed approximate solution $\bar{w}$.

To compute the location of a minimum of J, we start at some trial point $\bar{w}_0$ and search in the direction of some $\bar{p}_0$. One algorithm, called the *line search*, finds $t > 0$ such that $f(t) = J(\bar{w}_0 + t\bar{p}_0)$ reaches its minimum (only one for a convex J) at some $t_0 > 0$, which is characterized by $f'(t_0) = \nabla J(\bar{w}_0 + t_0\bar{p}_0)^{\mathrm{T}}\bar{p}_0 = 0$. Then one sets $\bar{w}_1 = \bar{w}_0 + t_0\bar{p}_0$ and restarts the search from $\bar{w}_1$ in the direction of some new $\bar{p}_1$.

In the standard gradient descent, one takes $\bar{p}_0 = \bar{r}_0 = -\nabla J(\bar{w}_0)$ and $\bar{p}_1 = \bar{r}_1 = -\nabla J(\bar{w}_1)$. Note that this new direction is orthogonal to the earlier direction $\bar{p}_0$.

When J is the quadratic function $\frac{1}{2}\bar{w}^{\mathrm{T}}A\bar{w} - \bar{b}^{\mathrm{T}}\bar{w}$ (A is real and symmetric), the condition $f'(t_0) = \nabla J(\bar{w}_0 + t_0\bar{p}_0)^{\mathrm{T}}\bar{p}_0 = 0$ can be solved as

$$t_0 = -\frac{\bar{p}_0^{\mathrm{T}}\nabla J(\bar{w}_0)}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0} = \frac{\bar{p}_0^{\mathrm{T}}(\bar{b} - A\bar{w}_0)}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0} = \frac{\bar{p}_0^{\mathrm{T}}\bar{r}_0}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0} \text{ using } \bar{r}_0 = \bar{b} - A\bar{w}_0.$$

Thus the line search algorithm gives

$$\bar{w}_1 = \bar{w}_0 + \left(\frac{\bar{p}_0^{\mathrm{T}}\bar{r}_0}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0}\right)\bar{p}_0 = \bar{w}_0 + \left(\frac{\bar{p}_0^{\mathrm{T}}\bar{r}_0}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0}\right)(\bar{b} - A\bar{w}_0), \qquad \text{when we choose } \bar{p}_0 = \bar{r}_0 = \bar{b} - A\bar{w}_0.$$

Note that we have used A being positive definite here to make sure that $\bar{p}_0^{\mathrm{T}}A\bar{p}_0 > 0$ when $\bar{z}_0 \neq \bar{0}$, which also assures that $t_0 > 0$. $s_0 := \frac{\bar{p}_0^{\mathrm{T}}\bar{r}_0}{\bar{p}_0^{\mathrm{T}}A\bar{p}_0}$ is called the *step size*, or *learning rate* in the context of machine learning. It is chosen so that the restriction of J along the ray $\bar{w}_0 + s\bar{p}_0$ reaches its minimum. After one-step iteration of the line search, we find the updated residue becomes

$$\bar{r}_1 = \bar{b} - A\bar{w}_1 = \bar{r}_0 - s_0 A\bar{p}_0.$$

This process can be iterated.
It turns out that the gradient descent as directly applied may not converge at a fast rate. One way to speed up the convergence is to *pre-condition* the equations to speed up the convergence. This is referred to as *Feature Preprocessing* in the context of machine learning.

Example 4.4.1

For $J(x,y) = \frac{1}{2}(x^2 + by^2)$, $0 < b \leq 1$, $\nabla J(x,y) = (x, by)^{\mathrm{T}}$ and the coefficient matrix is $A = \begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix}$. So the gradient descent algorithm gives

$$\bar{r}_k = -\nabla J(x_k, y_k) = -\begin{bmatrix} x_k \\ by_k \end{bmatrix}, s_k = \frac{x_k^2 + b^2 y_k^2}{x_k^2 + b^3 y_k^2},$$

and

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \frac{x_k^2 + b^2 y_k^2}{x_k^2 + b^3 y_k^2} \begin{bmatrix} x_k \\ by_k \end{bmatrix} = \begin{bmatrix} x_k \frac{(b-1)b^2 y_k^2}{x_k^2 + b^3 y_k^2} \\ y_k \frac{(1-b)b^2 x_k^2}{x_k^2 + b^3 y_k^2} \end{bmatrix}.$$

Note that the relation $x_k^2 = b^2 y_k^2$ is preserved under iteration, so if we start with $x_0 = b, y_0 = 1$, we find

$$x_k = b\left(\frac{b-1}{b+1}\right)^k, \quad y_k = \left(\frac{1-b}{b+1}\right)^k,$$

which tends to 0 at the rate of $\left(\frac{1-b}{b+1}\right)^k$. This is good when $b < 1$ is close to 1, but is not very good when $b < 1$ is close to 0. For example, when $b = 0.1$, $\frac{1-b}{b+1} = 9/11$, whose 10th power reaches 0.1344. Note also that the algorithm causes a zigzag of the search paths when $b < 1$, changing direction perpendicularly at each step.

Remark 4.4.1

Note that in the example above, $J(x,y) = \frac{1}{2}(x^2 + by^2)$ is the sum of a function of x and a function of y, the roles of x and y are completely "decoupled" here, referred as "Additive Separability" in the textbook, and one can find the minimum of the one-variable function $\frac{1}{2}x^2$ in one-step, and the minimum of the one-variable function of $\frac{b}{2}y^2$ also in one-step. This illustrates that, although $-\nabla J(x,y)$ is the direction of greatest descent at (x,y), line-searching along such directions may not be the most efficient way to find the minimizer.

If one can make a transformation of variables to reduce a given function as the sum of functions of separate variables, one can always find the minimizer of the given function by finding the minimizer of each summand function of the separate variable.
In principle this is always possible for a quadratic function given in terms of $\frac{1}{2}\bar{x}^T A\bar{x} - \bar{b}^T\bar{x}$, *where one seeks* $\bar{x} = P\bar{y}$ *so that*

$$J(\bar{x}) = \frac{1}{2}\bar{x}^T A\bar{x} - \bar{b}^T\bar{x} = \frac{1}{2}\bar{y}^T P^T AP\bar{y} - \bar{b}^T P\bar{y}$$

becomes the sum of functions of the separate variables y_i. *This happens when* $P^T AP$ *becomes a diagonal matrix. This is equivalent to the columns of* P *being* A*-orthogonal in the sense that* $\bar{p}_i^T A\bar{p}_j = 0$ *for* $i \neq j$, *where* $\bar{p}_i = P(:, i)$.
In fact, this idea can be used even when there may not be enough known A*-orthogonal vectors to make an invertible* P.
Suppose that there are k A*-orthogonal non-zero vectors* $\{\bar{p}_1, \cdots, \bar{p}_k\}$. *Since* A *is positive definite,* $(\bar{x}, \bar{y})_A := \bar{x}^T A\bar{y}$ *defines an inner product, and all properties of a general inner product are valid for this one. In particular, the set of non-zero orthogonal vectors in this inner product,* $\{\bar{p}_1, \cdots, \bar{p}_k\}$, *must be linearly independent. When we look for a minimum of* $J(\bar{x})$ *restricted to the* k*-dimensional space Span*$\{\bar{p}_1, \cdots, \bar{p}_k\}$, *we can write such a vector as* $\bar{x} = y_1\bar{p}_1 + \cdots + y_k\bar{p}_k$, *then*

$$\bar{x}^T A\bar{x} = (y_1\bar{p}_1 + \cdots + y_k\bar{p}_k)^T A(y_1\bar{p}_1 + \cdots + y_k\bar{p}_k) = \sum_{i=1}^{k} \mu_i y_i^2$$

becomes decoupled in terms of y_i, *where we have used* $\bar{p}_i^T A\bar{p}_j = \mu_i\delta_{ij}$. J *as a function of* $(y_1, \cdots, y_k)$ *also becomes decoupled and one can find its minimum easily.*
For a symmetric matrix A, *we know that we can do this by some orthogonal matrix* P. *But that requires computing the eigenvectors of* A, *which is more expensive than solving* $A\bar{x} = \bar{b}$.
One can find a set of A*-orthogonal non-zero vectors by applying a generalized Gram-Schmidt orthogonalization process to a set of linearly independent vectors, where the orthogonality is in the sense of* A*-orthogonality. It turns out that the conjugate gradient method, to be introduced next, provides an efficient algorithm to accomplish this.*

Example 4.4.2

Suppose that A is positive definite and $\mathbf{u}_1, \mathbf{u}_2$ is a pair of linearly independent vectors. We would like to take $\mathbf{v}_1 = \mathbf{u}_1$ and construct $\mathbf{v}_2 = \mathbf{u}_2 - a\mathbf{v}_1$ so that $\mathbf{v}_2$ is A-orthogonal to $\mathbf{v}_1$. This amounts to

$$(\mathbf{u}_2 - a\mathbf{u}_1) \cdot A\mathbf{v}_1 = 0.$$

It follows from this that $a = \frac{\mathbf{u}_2 \cdot A\mathbf{v}_1}{\mathbf{v}_1 \cdot A\mathbf{v}_1}$.
If $\mathbf{u}_3$ is given, one can construct $\mathbf{v}_3 = \mathbf{u}_3 - a_1\mathbf{v}_1 - a_2\mathbf{v}_2$ such that

$$(\mathbf{u}_3 - a_1\mathbf{v}_1 - a_2\mathbf{v}_2) \cdot A\mathbf{v}_i = (\mathbf{u}_3 - a_1\mathbf{v}_1 - a_2\mathbf{v}_2)^{\mathrm{T}} A\mathbf{v}_i = 0, i = 1, 2.$$

Using $\mathbf{v}_i A\mathbf{v}_j = 0$ for $i \neq j$, one finds that

$$a_1 = \frac{\mathbf{u}_3 \cdot A\mathbf{v}_1}{\mathbf{v}_1 \cdot A\mathbf{v}_1}, \quad a_2 = \frac{\mathbf{u}_3 \cdot A\mathbf{v}_2}{\mathbf{v}_2 \cdot A\mathbf{v}_2}.$$

Exercise 1. Write out pseudo-codes for implementing the gradient descent method on the function $x^2 - xy + y^2 + x + y$. Then find (x_i, y_i) for $i = 1, 2, 3$ with $(x_0, y_0) = (0, 0)$.

Exercise 2. For

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix},$$

apply the modified Gram-Schmidt process from $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ to construct a set of A-orthogonal vectors.

4.4.2 Conjugate Gradient Method

The conjugate gradient method modifies the choice of the subsequent search direction from $-\nabla J(\bar{w}_{k+1})$ to $-\nabla J(\bar{w}_{k+1}) + \beta\bar{p}_k$, where $\bar{p}_k$ is the search direction of the previous step, and β is a parameter. For the case of J being the quadratic function $J = \frac{1}{2}\bar{w}^{\mathrm{T}} A\bar{w} - \bar{b}^{\mathrm{T}}\bar{w}$, this turns into

$$\bar{w}_{k+1} = \bar{w}_k + s_k\bar{p}_k, \tag{4.3}$$
$$\bar{r}_{k+1} = \bar{r}_k - s_k A\bar{p}_k, \tag{4.4}$$
$$\bar{p}_{k+1} = \bar{r}_{k+1} + \beta_k\bar{p}_k, \tag{4.5}$$

where $\beta_k = \bar{r}_{k+1}^{\mathrm{T}}\bar{r}_{k+1}/\bar{r}_k^{\mathrm{T}}\bar{r}_k$ measures the improvement of the residue at this step, and s_k is modified into $\frac{\bar{p}_k^{\mathrm{T}}\bar{r}_k}{\bar{p}_k^{\mathrm{T}}A\bar{p}_k}$, still dictated by

$$-\nabla J(\bar{w}_k + s_k\bar{p}_k)\cdot\bar{p}_k = \bar{p}_k^{\mathrm{T}}\bar{r}_{k+1} = 0. \tag{4.6}$$

The choice of β_k guarantees that

$$\{\bar{p}_0, \bar{p}_1, \cdots, \bar{p}_{k+1}\} \text{ is a set of } A\text{-orthogonal vectors;} \tag{4.7}$$
$$\mathrm{Span}\{\bar{p}_0, \bar{p}_1, \cdots, \bar{p}_{k+1}\} = \mathrm{Span}\{\bar{r}_0, \bar{r}_1, \bar{r}_{k+1}\}; \tag{4.8}$$
$$\{\bar{r}_0, \bar{r}_1, \cdots, \bar{r}_{k+1}\} \text{ is a set of Euclidean orthogonal vectors.} \tag{4.9}$$

(4.7), (4.8), and (4.9) can be proved using induction on k; one can further prove that, for $l < k$,

$$\bar{p}_l^{\mathrm{T}}\bar{r}_{k+1} = 0, \tag{4.10}$$

as $\bar{p}_l$ is a linear combination of $\{\bar{r}_0, \cdots, \bar{r}_l\}$ by (4.8).
Technically, the requirement $(\bar{p}_{k+1}, \bar{p}_k)_A = 0$ turns into

$$(\bar{r}_{k+1} + \beta_k\bar{p}_k)^{\mathrm{T}}A\bar{p}_k = s_k^{-1}(\bar{r}_{k+1} + \beta_k\bar{p}_k)^{\mathrm{T}}(\bar{r}_k - \bar{r}_{k+1}) = 0, \tag{4.11}$$

where we have used (4.4). Using (4.5) and (4.6), we find

$$\bar{p}_k^{\mathrm{T}}\bar{r}_k = (\bar{r}_k + \beta_{k-1}\bar{p}_{k-1})^{\mathrm{T}}\bar{r}_k = \bar{r}_k^{\mathrm{T}}\bar{r}_k,$$

and

$$\bar{r}_{k+1}^{\mathrm{T}}\bar{r}_k = \bar{r}_{k+1}^{\mathrm{T}}(\bar{p}_k - \beta_{k-1}\bar{p}_{k-1}) = 0.$$

Using these relations in (4.11), we find $\beta_k = \bar{r}_{k+1}^{\mathrm{T}}\bar{r}_{k+1}/\bar{r}_k^{\mathrm{T}}\bar{r}_k$.
Our discussion here assumes that $\bar{r}_j \neq \bar{0}$ for $j = 0, 1, \cdots, k$. If $\bar{r}_{k+1} = \bar{0}$, it would mean that $\bar{b} - A\bar{w}_{k+1} = \bar{r}_{k+1} = \bar{0}$ and that $\bar{w}_{k+1}$ is a solution to our problem. Suppose that A is $n \times n$, then either (4.9) or (4.10) implies that $\bar{r}_{k+1}$ is orthogonal to the $(k+1)$ vectors $\bar{r}_0, \cdots, \bar{r}_k$, but a non-zero vector in $\mathcal{R}^n$ can't be orthogonal to more than $(n-1)$ non-zero vectors, so $\bar{r}_{k+1} = \bar{0}$ before k reaches n, which means that this algorithm should be able to find the minimum within n steps. But for large matrices, one would like to find a reasonably good approximate solution well before n steps.
This method is called momentum-based learning, as the term $\beta_k\bar{p}_k$ carries some momentum (memory in other contexts) from the previous step.
Let $\bar{x}_*$ denotes the exact solution $A\bar{x}_* = \bar{b}$, then $r_k = A(\bar{x}_* - \bar{x}_k)$ is one measure of how close $\bar{x}_k$ is to $\bar{x}_*$. Another measure is

$$||\bar{x}_* - \bar{x}_k||_A := \sqrt{(\bar{x}_* - \bar{x}_k, \bar{x}_* - \bar{x}_k)_A} = \sqrt{(\bar{x}_* - \bar{x}_k)^{\mathrm{T}}A(\bar{x}_* - \bar{x}_k)}.$$

It can be proved that

$$||\bar{x}_* - \bar{x}_k||_A \leq 2||\bar{x}_* - \bar{x}_0||_A \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\right)^k,$$

where $\kappa_2(A)$ is the condition number of A in the Euclidean norm. For $A = \begin{bmatrix} 1 & 0 \\ 0 & b \end{bmatrix}$, $0 < b < 1$, $\kappa_2(A) = b^{-1}$. The convergence rate is an improvement over the direct gradient descent method.

Exercise 3. Write out pseudo-codes for implementing the conjugate gradient method on the function $x^2 - xy + y^2 + x + y$. Then find (x_i, y_i) for $i = 1, 2, 3$ with $(x_0, y_0) = (0, 0)$.

4.4.3 Newton's Method

Newton's method to find the zero of a function $F(\bar{w})$ near $\bar{w}_0$ is to approximate $F(\bar{w})$ by $F(\bar{w}_0) + F'(\bar{w}_0)(\bar{w} - \bar{w}_0)$ and set it to be $\bar{0}$, which gives

$$\bar{w}_1 = \bar{w}_0 - (F'(\bar{w}_0))^{-1} F(\bar{w}_0)$$

as the next step of approximation.
In the context of locating a minimizer of $J(\bar{w})$, we apply Newton's method to $F(\bar{w}) = \nabla J(\bar{w})$. Then $F'(\bar{w}) = \text{Hess}(J)(\bar{w})$ becomes the Hessian matrix of J at $\bar{w}$, and the algorithm becomes

$$\bar{w}_1 = \bar{w}_0 - (\text{Hess}(J)(\bar{w}_0))^{-1} \nabla J(\bar{w}_0).$$

For the case of the quadratic function $J = \frac{1}{2}\bar{w}^{\mathrm{T}} A \bar{w} - \bar{b}^{\mathrm{T}} \bar{w}$, $\text{Hess}(J)(\bar{w}) = A$ is the constant matrix A and $\nabla J(\bar{w}) = A\bar{w} - \bar{b}$, so the algorithm becomes

$$\bar{w}_1 = \bar{w}_0 - A^{-1}(A\bar{w}_0 - \bar{b}).$$

But this algorithm involves computing A^{-1} multiplied to a vector, which is equivalent to computing solutions to $A\bar{x} = \bar{b}$, so Newton's algorithm in this context does not give an alternative method for solving such a linear system.

4.5 Comments on Stochastic Gradient Descent, Tikhonov Regularization, Optimization Models for Binary Targets, and the Support Vector Machine

Stochastic Gradient Descent Instead of computing the full gradient at each step, one samples a subset of the directions then carry out the gradient descent in the

sampled directions. This is referred to as *mini-batch* stochastic gradient descent. The textbook also discusses *Coordinate Descent*, which performs the gradient descent in a *single* variable at a time (and sequentially updating the variables).

Tikhonov Regularization The gradient descent method applied to a quadratic function works only when the coefficient matrix of the quadratic term is positive definite. When this condition is not known, one adds the regularization term $\lambda\|\bar{W}\|^2$ to the objective function to find a minimizer of

$$\frac{1}{2}\|D\bar{W} - \bar{y}\|^2 + \frac{\lambda}{2}\|\bar{W}\|^2.$$

The coefficient matrix of the quadratic term becomes $D^{\mathrm{T}}D + \lambda I$, which is positive definite for $\lambda > 0$. This approach is used in defining the Moore-Penrose pseudoinverse of a matrix. Such a regularization term is often used in the subsequent discussions in the textbook.

Optimization Models for Binary Targets In least-squares classification on binary targets, the target vector $\bar{y} = [y_1, \cdots, y_n]^{\mathrm{T}}$ only takes values -1 or 1, while the d-dimensional row vectors in the data matrix D, $\bar{X}_1, \cdots, \bar{X}_n$, take real values. One can pretend that the binary targets are real-valued and applies the least-square regression. Using $y_i^2 = 1$, the loss function

$$\frac{1}{2}\sum_i |\bar{W} \cdot \bar{X}_i - y_i|^2 = \frac{1}{2}\sum_i |\bar{W} \cdot (y_i\bar{X}_i) - 1|^2,$$

is equivalent to one with $y_i\bar{X}_i$ as the ith row of a modified data matrix $\tilde{D}$ and $[1, \cdots, 1]^{\mathrm{T}}$ as the modified target vector.
One drawback of this approach is that it gives weight to the magnitude to $|\bar{W} \cdot \bar{X}_i - y_i|^2$, but the classification problem should not substantially differentiate the terms $|\bar{W} \cdot \bar{X}_i - y_i|^2$ when $\bar{W} \cdot \bar{X}_i$ has the same sign as y_i (e.g., $\bar{W} \cdot \bar{X}_i = 2$ and $\bar{W} \cdot \bar{X}_i = 20$ should give the same classification).

The support vector machine (SVM) treats well-separated points in the loss function in a more careful way by not penalizing them at all. Note that a point is correctly classified by the least-squares classification model when $\bar{W} \cdot (y_i\bar{X}_i) > 0$. We consider the point well separated when $\bar{W} \cdot (y_i\bar{X}_i) > 1$, and can set the loss function to be 0 when this happens. We define the L^2-SVM loss function as

$$J(W) = \sum_{i=1}^{n} \Big(\max\{0, 1 - \bar{W} \cdot (y_i\bar{X}_i)\}\Big)^2 + \frac{\lambda}{2}\|\bar{W}\|^2.$$

Here, $\lambda > 0$ and the term $\frac{\lambda}{2}\|\bar{W}\|^2$ provides some regularization for the first term in the sense that it makes the loss function strictly convex—Note that each $\bar{W} \mapsto \max\{0, 1 - \bar{W} \cdot (y_i \bar{X}_i)\}$, and its square, is convex, but not strictly convex.
The L^1-version of the above loss function is

$$J(W) = \sum_{i=1}^{n} \max\{0, 1 - \bar{W} \cdot (y_i \bar{X}_i)\} + \frac{\lambda}{2}\|\bar{W}\|^2$$

and is called the Hinge-loss SVM function.

Exercise 1. Determine the gradient of the L^2-SVM loss function, and respectively, of the Hinge-loss SVM function.

Chapter 6

CONSTRAINED OPTIMIZATION

6.1 Introduction

In many optimization problems the variables defining the loss functions are subject to some constraints, either in the form of some equality constraints or some inequality constraints, or a combination of both. The domain of definition of the loss function in such cases are called feasible region. Examples include

$$\min \|\bar{x}\| \text{ subject to } A\bar{x} = \bar{b},$$

for some given matrix A and vector $\bar{b}$, or

$$\min \|\bar{x}\| \text{ subject to } A\bar{x} = \bar{b}, x_i \geq 0 \text{ for some designated or all } i.$$

The constraints in the examples here are given in terms of linear expressions in $\bar{x}$, but one may also encounter constraints given in terms of nonlinear expressions in $\bar{x}$. When the constraints are given in terms of a number of linear expressions in $\bar{x}$, they can be formulated in the form of $\mathbf{a}_i^{\mathrm{T}}\bar{x} \leq (=)b_i$ for a collection of i's. The feasible regions in such situations are still convex.
Commonly used methods to solve such optimization problems include the method of (augmented) Lagrange multipliers and dual optimization problem.

6.2 The method of Lagrange multipliers for optimization with equality constraints

When the loss function $f(\bar{x})$ and the constraints given in the form of $g_i(\bar{x}) = 0$ for $i \in I$ have differentiable $f(\bar{x})$ and $g_i(\bar{x})$, any minimizer $\bar{x}$ subject to the given

constraints satisfy

$$\nabla f(\bar{x}) \in \text{Span}\{\nabla g_i(\bar{x}), i \in I\},$$

provided that $\{\nabla g_i(\bar{x}), i \in I\}$ is linearly independent. This translates into

$$\nabla f(\bar{x}) = \sum_{i \in I} \lambda_i \nabla g_i(\bar{x})$$

for some Lagrange multipliers λ_i. Analytically we should find solutions to the system

$$\begin{cases} \nabla f(\bar{x}) = \sum_{i \in I} \lambda_i \nabla g_i(\bar{x}), \\ g_i(\bar{x}) = 0, i \in I. \end{cases} \tag{6.1}$$

as candidates for a minimizer. If $\bar{x} \in \mathcal{R}^d$, then this system contains $d + |I|$ equations for the unknowns $\bar{x}, \lambda_1, \cdots, \lambda_{|I|}$.
Formally one can introduce a function of $(\bar{x}, \lambda_1, \cdots, \lambda_{|I|})$, called the Lagrangian,

$$L(\bar{x}, \lambda_1, \cdots, \lambda_{|I|}) = f(\bar{x}) - \sum_{i \in I} \lambda_i g_i(\bar{x}).$$

Then

$$\begin{cases} \dfrac{\partial L}{\partial x_j} = \dfrac{\partial f(\bar{x})}{\partial x_j} - \sum_{i \in I} \lambda_i \dfrac{\partial g_i(\bar{x})}{\partial x_j}, \\ \dfrac{\partial L}{\partial \lambda_i} = -g_i(\bar{x}), i \in I. \end{cases}$$

So (6.1) is equivalent to $\frac{\partial L}{\partial x_j} = 0$ for $j = 1, \cdots, d$ and $\frac{\partial L}{\partial \lambda_i} = 0$ for $i \in I$.
When the constraints are linear equations in the form of $A\bar{x} = \bar{b}$ and $f(\bar{x})$ is convex for $\bar{x} \in \mathcal{R}^d$, $L(\bar{x}, \Lambda) = f(\bar{x}) - \Lambda^{\mathrm{T}}(A\bar{x} - \bar{b})$ is convex in $(\bar{x}, \Lambda)$. Techniques for convex optimization such as gradient descents can be applied. It turns out that such a function in general does not have a finite minimum, as for each fixed $\bar{x}$, $L(\bar{x}, \Lambda)$ is a linear function of Λ so is unbounded unless $A\bar{x} = \bar{b}$. However, if $f(\bar{x})$ is strictly convex in $\bar{x}$ and grows faster than linear as $\bar{x} \to \infty$, then for each fixed λ, $L(\bar{x}, \Lambda)$ as a function of $\bar{x}$ has a unique minimizer $\bar{x} = S(\Lambda)$ and the resulting function $L(S(\Lambda), \Lambda)$ becomes a concave function of Λ*, so one seeks a maximum of this function. So the solution scheme can be described as

$$\begin{cases} \bar{x} = S(\Lambda) = \arg\min_{\bar{w}} L(\bar{w}, \Lambda), \\ \Lambda = \arg\max_{\Lambda} L(S(\Lambda), \Lambda), \end{cases}$$

*for each $\bar{x}$, $L(\bar{x}, \Lambda)$ is a linear function of Λ so is concave in Λ, therefore, $L(S(\Lambda), \Lambda)$ as the minimum of a family of concave functions of Λ is concave in Λ.

namely, as a solution of the max-mini scheme:

$$\max_{\Lambda} \min_{\bar{w}} L(\bar{w}, \Lambda).$$

The gradient search algorithm should be adjusted so that it would search for the next step of $\lambda_i(k+1)$ in the gradient ascent direction:

$$\begin{cases} x_j(k+1) = x_j(k) - s_k \dfrac{\partial L(\bar{x}(k), \Lambda(k))}{\partial x_j} & = x_j(k) - s_k \left(\dfrac{\partial f(\bar{x}(k))}{\partial x_j} - \displaystyle\sum_{i \in I} \lambda_i(k) \dfrac{\partial g_i(\bar{x}(k))}{\partial x_j} \right) \\ \lambda_i(k+1) = \lambda_i(k) + s_k \dfrac{\partial L(\bar{x}(k), \Lambda(k))}{\partial \lambda_i} & = \lambda_i(k) - s_k g_i(\bar{x}(k)) \end{cases}$$

for a step-size $s_k > 0$ determined by the line search criterion or $\bar{x}(k+1)$ is constructed directly as

$$\bar{x}(k+1) = \arg\min_{\bar{w}} L(\bar{w}, \Lambda(k)).$$

In the latter case, we would update $\Lambda(k+1)$ by

$$\lambda_i(k+1) = \lambda_i(k) + \alpha_k \frac{\partial L(\bar{x}(k+1), \Lambda(k))}{\partial \lambda_i}$$

for some step-size $\alpha_k > 0$.

Example 6.2.1

When $f(\bar{x})$ is a quadratic function $\frac{1}{2}\bar{x}^{\mathrm{T}} B \bar{x} - \bar{c}^{\mathrm{T}} \bar{x}$ with a positive definite B and the constraints are linear equations in the form of $A\bar{x} = \bar{b}$, the Lagrangian becomes

$$L(\bar{x}, \Lambda) = \frac{1}{2}\bar{x}^{\mathrm{T}} B \bar{x} - \bar{c}^{\mathrm{T}} \bar{x} - \Lambda^{\mathrm{T}}(A\bar{x} - \bar{b}).$$

$A\bar{x} = \bar{b}$ is equivalent to $g_i(\bar{x}) = A(i,:)\bar{x} - b_i = 0$, and $\nabla g_i(\bar{x}) = A(i,:)^{\mathrm{T}}$, so (6.1) turns into

$$\begin{cases} B\bar{x} - \bar{c} - A^{\mathrm{T}}\Lambda = \bar{0}, \\ A\bar{x} - \bar{b} = \bar{0}. \end{cases}$$

In this case $S(\Lambda) = B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c})$ and

$$\begin{aligned}
&L(S(\Lambda), \Lambda) \\
=&\frac{1}{2}[B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c})]^{\mathrm{T}}(A^{\mathrm{T}}\Lambda + \bar{c}) - \bar{c}^{\mathrm{T}}B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) - \Lambda^{\mathrm{T}}[AB^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) - \bar{b}] \\
=&\frac{1}{2}(A^{\mathrm{T}}\Lambda + \bar{c})^{\mathrm{T}}B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) - \bar{c}^{\mathrm{T}}B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) - (A^{\mathrm{T}}\Lambda)^{\mathrm{T}}B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) + \Lambda^{\mathrm{T}}\bar{b} \\
=&-\frac{1}{2}(A^{\mathrm{T}}\Lambda + \bar{c})^{\mathrm{T}}B^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) + \Lambda^{\mathrm{T}}\bar{b}
\end{aligned}$$

which is concave in Λ, as B^{-1} is also positive definite.
The gradient search algorithm becomes

$$\begin{cases}
\bar{x}(k+1) = \bar{x}(k) - s_k \left(B\bar{x}(k) - \bar{c} - A^{\mathrm{T}}\Lambda(k)\right), \\
\Lambda(k+1) = \Lambda(k) - s_k \left(A\bar{x}(k) - \bar{b}\right).
\end{cases}$$

In the case here, the problem can be reduced to a linear algebra problem of solving for Λ from

$$A\bar{x} - \bar{b} = AS(\Lambda) - \bar{b} = AB^{-1}(A^{\mathrm{T}}\Lambda + \bar{c}) - \bar{b} = \bar{0}.$$

Typically A is a wide matrix with full row rank, then A^{T} is a tall matrix with $\mathrm{Null}(A^{\mathrm{T}}) = \{\bar{0}\}$, and $AB^{-1}A^{\mathrm{T}}$ is invertible, so Λ can be solved from the above system, which can then be substitute into $\bar{x} = S(\Lambda)$ to obtain $\bar{x}$.

Exercise 1. Prove that $AB^{-1}A^{\mathrm{T}}$ in the context of the above example is invertible.

Exercise 2. Find $\min \|\bar{x}\|_2$ subject to $\mathbf{a} \cdot \bar{x} = 1$, where $\mathbf{a} \neq \bar{0} \in \mathcal{R}^d$ is given.

Exercise 3. Find $\min \|(x, y, z)\|_2$ subject to $x + y + z = 5$ and $x + y - 2z = 2$.

6.3 Dual Variational Problem

In the case of linear constraints in the form of $A\bar{x} = \bar{b}$ and a convex loss function $f(\bar{x})$, the constrained minimization problem is equivalent to the maximization problem of a dual function defined as follows.

Definition 6.3.1: Legendre transform

Suppose that $f(\bar{x})$ is convex on $\mathcal{R}^d$, then its Legendre transform $f^*(\bar{y})$ is a function of $\bar{y} \in \mathcal{R}^d$ defined by

$$f^*(\bar{y}) = \sup_{\bar{x}} \{\bar{x}^{\mathrm{T}}\bar{y} - f(\bar{x})\}.$$

This $f^*(\bar{y})$, as the supremum of a family of affine functions of $\bar{y}$, is a convex function of $\bar{y} \in \mathcal{R}^{\mathrm{T}}$. By the definition,

$$f^*(\bar{y}) \geq \bar{x}^{\mathrm{T}}\bar{y} - f(\bar{x}) \text{ for all } \bar{x} \in \mathcal{R}^d, \text{ with equality iff } \bar{y} = \nabla f(\bar{x}).$$

Example 6.3.1

If $f(x) = \frac{1}{p}|x|^p$ for $x \in \mathcal{R}$ and some $p > 1$, then $xy - \frac{1}{p}|x|^p$ attains its maximum at x^* at which the derivative of this function of x, $y - |x|^{p-2}x$, equals 0. Thus $x^* = \text{sign}(y)|y|^{\frac{1}{p-1}}$ and

$$f^*(y) = |y|^{\frac{p}{p-1}} - \frac{1}{p}|y|^{\frac{p}{p-1}} = \frac{p-1}{p}|y|^{\frac{p}{p-1}}.$$

In the set-up of finding a critical point of the Lagrangian $L(\bar{x}, \Lambda)$, the unique minimizer $\bar{x} = S(\Lambda)$of $L(\bar{x}, \Lambda) = f(\bar{x}) - \Lambda^{\mathrm{T}}(A\bar{x} - \bar{b})$ as a function of $\bar{x}$ is used to define the function $L(S(\Lambda), \Lambda)$, but $f(S(\Lambda)) - \Lambda^{\mathrm{T}}AS(\Lambda) = -f^*(A^{\mathrm{T}}\Lambda)$, so $L(\bar{x}, \Lambda) = -f^*(A^{\mathrm{T}}\Lambda) + \Lambda^{\mathrm{T}}\bar{b}$ and we can potentially search for a maximizer of this function of Λ directly.

Remark 6.3.1

One advantage of the dual variational formulation is that when $f(\bar{x})$ is additively separable and the constraints are linear, the minimization problem $\min_{\bar{x}} L(\bar{x}, \Lambda)$ is often easy to solve.
The simplest example is $f(\bar{x}) = x_1^2 + \cdots + x_d^2$ subject to $a_1x_1 + \cdots + a_dx_d = c$. Then

$$\begin{aligned} L(\bar{x}, \lambda) =& x_1^2 + \cdots + x_d^2 - \lambda(a_1x_1 + \cdots + a_dx_d - d) \\ =& (x_1^2 - \lambda a_1x_1) + \cdots + (x_d^2 - \lambda a_dx_d) + \lambda d, \end{aligned}$$

where, for a fixed λ, each of $(x_i^2 - \lambda a_ix_i)$ has an easily determined minimizer x_i.

6.4 Rayleigh Quotients and Optimization with a Quadratic Constraint

An optimization problem of a quadratic function with a quadratic constraint arises in connection with the Rayleigh quotient

$$\frac{\bar{x}^{\mathrm{T}} A \bar{x}}{\bar{x}^{\mathrm{T}} \bar{x}}$$

of a real symmetric $d \times d$ matrix A for $\bar{x} \neq \bar{0} \in \mathcal{R}^d$.
It turns out that the smallest eigenvalue λ_1 of A and the largest eigenvalue λ_d of A are characterized by

$$\lambda_1 = \min_{\bar{x} \neq \bar{0} \in \mathcal{R}^d} \frac{\bar{x}^{\mathrm{T}} A \bar{x}}{\bar{x}^{\mathrm{T}} \bar{x}}, \quad \lambda_d = \max_{\bar{x} \neq \bar{0} \in \mathcal{R}^d} \frac{\bar{x}^{\mathrm{T}} A \bar{x}}{\bar{x}^{\mathrm{T}} \bar{x}}.$$

Because both the numerator and denominator of the Rayleigh quotient are homogeneous degree 2 polynomials in $\bar{x}$, the minimization and maximization problems are equivalent to

$$\min\{\bar{x}^{\mathrm{T}} A \bar{x} : \bar{x}^{\mathrm{T}} \bar{x} = 1\}, \quad \text{and respectively,} \max\{\bar{x}^{\mathrm{T}} A \bar{x} : \bar{x}^{\mathrm{T}} \bar{x} = 1\}.$$

We can thus consider the constrained optimization problem for the function $f(\bar{x}) := \bar{x}^{\mathrm{T}} A \bar{x}$ subject to $g(\bar{x}) := \bar{x}^{\mathrm{T}} \bar{x} - 1 = 0$.
We can consider the Lagrangian

$$L(\bar{x}, \lambda) := \bar{x}^{\mathrm{T}} A \bar{x} - \lambda(\bar{x}^{\mathrm{T}} \bar{x} - 1).$$

But both $\bar{x}^{\mathrm{T}} A \bar{x}$ and $\bar{x}^{\mathrm{T}} \bar{x} - 1$ are now quadratic in $\bar{x}$, so $L(\bar{x}, \lambda)$ may not have a finite minimum or maximum for a given λ. But it still makes sense to consider critical points of $L(\bar{x}, \lambda)$:

$$\frac{\partial L(\bar{x}, \lambda)}{\partial x_j} = 2A(j, :)\bar{x} - 2\lambda x_j,$$

so if $\bar{x}$ is a critical point of $L(\bar{x}, \lambda)$ in the $\bar{x}$ variable, then

$$A\bar{x} = \lambda \bar{x}.$$

We are of course only interested in a solution such that $\bar{x} \neq \bar{0}$. Our knowledge about the eigenvalues of A implies that only when λ equals one of the eigenvalues of A, will this constrained optimization problem have a non-zero critical point.

Remark 6.4.1

It turns out that the intermediate eigenvalues of A can be characterized as follows:

$$\lambda_{j+1} = \max_{S_j} \left[\min_{\bar{x} \perp S_j, \neq \bar{0}} \frac{\bar{x}^T A \bar{x}}{\bar{x}^T \bar{x}} \right],$$

where S_j runs over all j-dimensional subspaces of $\mathcal{R}^d$, and

$$\lambda_{n-j} = \min_{S_j} \left[\max_{\bar{x} \perp S_j, \neq \bar{0}} \frac{\bar{x}^T A \bar{x}}{\bar{x}^T \bar{x}} \right].$$

Using these characterizations, to prove $\lambda_{j+1} \geq b$, it suffices to find one j-dimensional subspace S_j such that $\min_{\bar{x} \perp S_j, \neq \bar{0}} \frac{\bar{x}^T A \bar{x}}{\bar{x}^T \bar{x}} \geq b$, while to prove $\lambda_j \leq a$, it suffices to find one $(n-j)$-dimensional subspace S_{n-j} such that $\max_{\bar{x} \perp S_{n-j}, \neq \bar{0}} \frac{\bar{x}^T A \bar{x}}{\bar{x}^T \bar{x}} \leq a$.

6.5 Some Optimization Problems involving L^1-norms

There has been strong interests in several optimization problems involving the L^1-norm $\|\bar{x}\|_1 := \sum_{i=1}^d |x_i|$ of $\bar{x} \in \mathcal{R}^d$. $\|\bar{x}\|_1$ is still a convex function of $\bar{x}$ but fails to be differentiable whenever some $x_i = 0$.

Here are a few commonly encountered problems.

Least Absolute Deviations For a given $n \times d$ matrix A and $n \times 1$ vector $\bar{b}$, one tries to find $\bar{x}$ which minimizes $\|A\bar{x} - \bar{b}\|_1$. (Typically $n > d$ here.)

Basis Pursuit For a given $n \times d$ matrix A and $n \times 1$ vector $\bar{b}$, one tries to find $\min \|\bar{x}\|_1$ subject to the constraint $A\bar{x} = \bar{b}$. (Typically $n < d$ here.)

Lasso For a given $n \times d$ matrix A, an $n \times 1$ vector $\bar{b}$, and some $\lambda > 0$, one tries to find $\bar{x}$ which minimizes $\frac{1}{2}\|A\bar{x} - \bar{b}\|_2^2 + \lambda \|\bar{x}\|_1$. (Typically $n < d$ here.)

A commonly used approach for solving the Least Absolute Deviations problem is to introduce an auxiliary variable $\bar{z} = A\bar{x} - b$ to turn the problem into a constrained optimization problem:

$$\min\{\|\bar{z}\|_1 : A\bar{x} - \bar{z} = \bar{b}\}.$$

One introduces the regularized Lagrangian

$$L_\rho(\bar{x}, \bar{z}, \bar{y}) := \|\bar{z}\|_1 + \bar{y}^{\mathrm{T}}(A\bar{x} - \bar{z} - \bar{b}) + \frac{\rho}{2}\|A\bar{x} - \bar{z} - \bar{b}\|_2^2$$

for some $\rho > 0$. One introduces the scaled dual variable $\bar{u} = \rho^{-1}\bar{y}$, then

$$\bar{y}^{\mathrm{T}}(A\bar{x} - \bar{z} - \bar{b}) + \frac{\rho}{2}\|A\bar{x} - \bar{z} - \bar{b}\|_2^2 = \frac{\rho}{2}\|A\bar{x} - \bar{z} - \bar{b} + \bar{u}\|_2^2 - \frac{\rho}{2}\|\bar{u}\|_2^2,$$

and one often works with the scaled Lagrangian

$$\hat{L}_\rho(\bar{x}, \bar{z}, \bar{u}) := \|\bar{z}\|_1 + \frac{\rho}{2}\|A\bar{x} - \bar{z} - \bar{b} + \bar{u}\|_2^2 - \frac{\rho}{2}\|\bar{u}\|_2^2.$$

$\hat{L}_\rho(\bar{x}, \bar{z}, \bar{u})$ is differentiable in $\bar{x}$ and $\bar{u}$ so at a critical point $(\bar{x}^*, \bar{z}^*, \bar{u}^*)$, we must have

$$\nabla_{\bar{x}}\hat{L}_\rho(\bar{x}^*, \bar{z}^*, \bar{u}^*) = \rho A^{\mathrm{T}}\left(A\bar{x}^* - \bar{z}^* - \bar{b} + \bar{u}^*\right) = \bar{0}, \tag{6.2}$$

$$\nabla_{\bar{u}}\hat{L}_\rho(\bar{x}^*, \bar{z}^*, \bar{u}^*) = \rho\left(A\bar{x}^* - \bar{z}^* - \bar{b}\right) = \bar{0}. \tag{6.3}$$

The critical condition in relation to the z variable is expressed in terms of the sub gradient $\partial\|\bar{z}\|_1$ of $\|\bar{z}\|_1$ due to its loss of differentiability when some $z_i = 0$:

$$\bar{0} \in \partial\|\bar{z}^*\|_1 - \rho\left(A\bar{x}^* - \bar{b} - \bar{z}^* + \bar{u}^*\right). \tag{6.4}$$

In combination with (6.3), (6.4) is equivalent to

$$\bar{0} \in \partial\|\bar{z}^*\|_1 - \rho\bar{u}^*. \tag{6.5}$$

One commonly used algorithm is the Alternating Direction Method of Multipliers (ADMM), which holds $\bar{z}_k$ and $\bar{u}_k$ first at each step, updates $\bar{x}$ by finding $\bar{x}_{k+1}$ that minimizes $\hat{L}_\rho(\bar{x}, \bar{z}_k, \bar{u}_k)$, then updates $\bar{z}$ by find $\bar{z}$ that minimizes $\hat{L}_\rho(\bar{x}_{k+1}, \bar{z}, \bar{u}_k)$, and finally updates $\bar{u}$ by finding $\bar{u}_{k+1}$ by requiring that $(\bar{x}_{k+1}, \bar{z}_{k+1}, \bar{u}_{k+1})$ satisfies (6.5). In the case here, $\bar{x}_{k+1}$ solves the least square problem

$$\min_{\bar{x}} \|A\bar{x} - \bar{z}_k - \bar{b} + \bar{u}_k\|_2^2,$$

so is given by solving $A^{\mathrm{T}}(A\bar{x} - \bar{z}_k - \bar{b} + \bar{u}_k) = \bar{0}$:

$$\bar{x}_{k+1} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}(\bar{z}_k + \bar{b} - \bar{u}_k).$$

Denote $\bar{r}_{k+1} = A\bar{x}_{k+1} - \bar{b}$, then

$$\hat{L}_\rho(\bar{x}_{k+1}, \bar{z}, \bar{u}_k) = \{\|\bar{z}\|_1 + \frac{\rho}{2}\|\bar{r}_{k+1} + \bar{u}_k - \bar{z}\|_2^2\},$$

which is the sum of the single variable functions $|z_i| + \frac{\rho}{2}|(\bar{r}_{k+1} + \bar{u}_k)_i - z_i|^2$, whose minimizer is given by the soft threshold solution $S_{\rho^{-1}}((\bar{r}_{k+1} + \bar{u}_k)_i)$. We write $\bar{z}_{k+1} = S_{\rho^{-1}}(\bar{r}_{k+1} + \bar{u}_k)$ and understand this as component-wise. In terms of the sub gradient criterion, we have

$$\bar{0} \in \partial\|\bar{z}_{k+1}\|_1 - \rho\left(\bar{r}_{k+1} + \bar{u}_k - \bar{z}_{k+1}\right). \tag{6.6}$$

Finally, the update of $\bar{u}_k$ is dictated by (6.5) and (6.6) and given by

$$\bar{u}_{k+1} = \bar{u}_k + \bar{r}_{k+1} - \bar{z}_{k+1}.$$

The basis pursuit problem can be approached using the augmented Lagrangian

$$\|\bar{x}\|_1 + \bar{y}^{\mathrm{T}}(A\bar{x} - \bar{b}) + \frac{\rho}{2}\|A\bar{x} - \bar{b}\|_2^2,$$

but the $\bar{x}$-minimization problem here does not have a direct solution as in the soft threshold problem. There is a variant of the ADMM approach, but we will not discuss it here.

One expects to find a sparse vector as a minimizer for the basis pursuit problem, as suggested by the case of one linear constraint in $\mathcal{R}^2$. In general, here is an argument to see this. The vectors $\bar{x}$ satisfying the constraint $A\bar{x} = \bar{b}$ has the structure of $\bar{x}_0 + \bar{u}$, where $\bar{u} \in \mathrm{Null}(A)$. For the typical case of this problem $\dim(\mathrm{Null}(A)) = k$ is large. We will argue that $\min \|\bar{x}_0 + \bar{u}\|_1$ for $\bar{u} \in \mathrm{Null}(A)$ is attained when k-coordinates of $\bar{x}_0 + \bar{u}$ are zero, namely, it is a sparse vector.

$\|\bar{x}\|_1$ fails to be differentiable when some $x_i = 0$. Each such set is a coordinate hyperplane, and the d such hyperplanes partition $\mathcal{R}^d$ into 2^d parts, in each of which the gradient of $\|\bar{x}\|_1$ is a constant non-zero vector. If this vector is not orthogonal to $\mathrm{Null}(A)$, then the gradient of the function $\|\bar{x}_0 + \bar{u}\|_1$ is not zero anywhere in this part, so a minimizer of $\|\bar{x}_0 + \bar{u}\|_1$ can't occur at a point where none of its coordinate is zero. If this vector is orthogonal to $\mathrm{Null}(A)$, then the gradient of the function $\|\bar{x}_0 + \bar{u}\|_1$ is zero in this part of the constraint, therefore, $\|\bar{x}_0 + \bar{u}\|_1$ takes a constant value in this part and one can always find a representative one or more of the coordinates are zero.

Suppose that one finds a minimizer whose first coordinate, say, is zero, then its remaining coordinates form a vector in $\mathcal{R}^{d-1}$ and is a minimizer of $\|\bar{x}'\|_1$ for $\bar{x}' \in \mathcal{R}^{d-1}$ with the affine constraints $A\bar{x} = \bar{b}, \bar{x} = (0; \bar{x}')$. The null space of the relevant matrix here is at least $(k-1)$ dimensional. By an induction argument, one sees that there must be a minimizer $\bar{x}'$ with at least $(k-1)$ zero coordinates.

6.6 Some Optimization Problems with Inequality Constraints

Some optimization problems involve constraints in the form of linear inequalities $A\bar{w} \leq \bar{b}$ for some matrix A and vector $\bar{b}$. When such constraints take on the simple form of $l_i \leq w_i \leq u_i$ for one or more indices i, they are called box constraints, and the optimization problem is one formulated on a closed box; the usual algorithms for finding an optimal value of a function can be adapted easily to this situation.

For general constraints of the form $A\bar{w} \leq \bar{b}$, one can introduce the Lagrangian relaxation

$$L(\bar{\alpha}) = \min_{\bar{w}} \{f(\bar{w}) + \bar{\alpha}^{\mathrm{T}}(A\bar{w} - \bar{b})\},$$

where one requires the Lagrangian multipliers $\alpha_i \geq 0$ for $i = 1, \cdots, m$, with $\bar{\alpha} = [\alpha_1, \cdots, \alpha_m]^{\mathrm{T}}$. This set-up penalizes $\bar{w}$ when $A\bar{w} \leq \bar{b}$ is not satisfied. One then finds $\bar{\alpha}$ which maximizes $L(\bar{\alpha})$ subject to the box constraints $\alpha_i \geq 0$ for $i = 1, \cdots, m$.
We illustrate this approach by turning the Hinge-loss SVM optimization

$$\min_{\bar{W}} J(W) = \min_{\bar{W}} \left\{ \sum_{i=1}^{n} \max\{0, 1 - \bar{W} \cdot (y_i \bar{X}_i)\} + \frac{\lambda}{2} \|\bar{W}\|^2 \right\}$$

into an optimization problem with inequality constraints: introduce slack variables ξ_i subject to

$$\xi_i \geq 1 - \bar{W} \cdot (y_i \bar{X}_i), \tag{6.7}$$
$$\xi_i \geq 0 \tag{6.8}$$

and the modified Lagrangian

$$J_{md}(\bar{W}, \xi) := \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2} \|\bar{W}\|^2.$$

Our original problem is equivalent to

$$\min_{\bar{W}, \xi} J_{md}(\bar{W}, \xi) \quad \text{subject to (6.7) and (6.8).}$$

The relaxed Lagrangian is then

$$J_{rl} := \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2} \|\bar{W}\|^2 + \sum_{i=1}^{n} \alpha_i (1 - \bar{W} \cdot (y_i \bar{X}_i) - \xi_i) - \sum_{i=1}^{n} \gamma_i \xi_i$$

subject to $\alpha_i, \gamma_i \geq 0$.
For any fixed such α_i, γ_i, we solve

$$\min_{\bar{W}, \xi} J_{rl},$$

from which we obtain

$$\frac{\partial J_{rl}}{\partial \bar{W}} = \lambda \bar{W} - \sum_{i=1}^{n} \alpha_i y_i \bar{X}_i^{\mathrm{T}} = 0, \tag{6.9}$$
$$\frac{\partial J_{rl}}{\partial \xi_i} = 1 - \alpha_i - \gamma_i = 0. \tag{6.10}$$

Substituting $\bar{W} = \lambda^{-1} \sum_{i=1}^{n} \alpha_i y_i \bar{X}_i^{\mathrm{T}}$ and $\gamma_i = 1 - \alpha_i$ into J_{rl}, and noting that the ξ_i has disappeared due to (6.10), we obtain the dual function

$$L_D(\bar{\alpha}) = -\frac{1}{2\lambda} \| \sum_{i=1}^{n} \alpha_i y_i \bar{X}_i^{\mathrm{T}} \|_2^2 + \sum_{i=1}^{n} \alpha_i.$$

The constraints $\gamma_i \geq 0$ also imply that $\alpha_i \leq 1$. We now need to find a maximizer of $L_D(\bar{\alpha})$ subject to $0 \leq \alpha_i \leq 1$. Once we have found such an $\bar{\alpha}$, we can find the minimizer to our original problem from $\bar{W} = \lambda^{-1} \sum_{i=1}^{n} \alpha_i y_i \bar{X}_i^{\mathrm{T}}$.

Simplex Algorithm in Linear Programming The problem seeks to find a maximizer to the objective function $\bar{c}^{\mathrm{T}}\bar{x}$ subject to $A\bar{x} \leq \bar{b}$ and $x_i \geq 0$. Here $\bar{c} = (c_1, \cdots, c_n)^{\mathrm{T}}$ and A is $p \times n$. The feasible set $\{\mathbf{x} \in \mathcal{R}^n : A\bar{x} \leq \bar{b}, x_i \geq 0\}$ is a convex polytope. Geometric argument suggests that a maximizer lies on a vertex (or corner) point of this polytope. The simplex algorithm is often used to find a maximizer by following a certain algorithm to search among the vertices of the polytope.
First, how do we identify a vertex of a convex polytope? Just as a face of a polytope is identified by setting only one of the given linear inequality constraints into equality, a vertex of a polytope is identified by setting n of the $(n+p)$ linear inequality constraints into equalities, so that solving these n linear equalities would give a unique solution as vertex (in the generic case). How does one determine edges through a given vertex and then identifies the neighboring vertex along this edge? Geometrically, one removes one of the n linear equations, resulting in the remaining $(n-1)$ linear equations in the n variables to have one free variable; adjusting that free variable would keep these $(n-1)$ linear equations satisfied, but would change the value of the left hand side of the removed equation from being zero to some non-zero value and the value of the left hand side of the remaining p equations. One stops if one of these values, which were initially non-zero, reaches zero. This reaches the neighboring vertex.
Which edge to follow from one vertex? The rate of change of the objective function $\bar{c}^{\mathrm{T}}\bar{x}$ along any edge is a constant. One chooses an edge along which this rate of change is positive if there is one such edge. If this rate of change along any edge is non-positive, then the objective function has reached its maximum at this vertex.
The algorithm is implemented by introducing p slack variables $s_i = b_i - A(i,:)\bar{x}$, so the constraints $A\bar{x} \leq \bar{b}$ turns into $s_i \geq 0$. The constraints have been transformed into p linear equations in the $(n+p)$ variables $\bar{x}, s_i$ subject to $x_j, s_i \geq 0$. Let's rename the variables as $\bar{x} \in \mathcal{R}^{n+p}$ and rename the matrix so that the constraints take the form of

$$A\bar{x} = \bar{b}, x_i \geq 0, i = 1, \cdots, n+p.$$

A still has p rows, and we may assume that $\mathrm{rank}(A) = p$. We also take $\bar{c}$ to be a vector in $\mathcal{R}^{n+p}$. The main purpose of this set up is not to distinguish the original variables and the slack variables in the algorithm.
Then any vertex corresponds to setting n of the $(n+p)$ variables x_i equal to 0. Allowing one of these n variables to increase from 0 to positive values would give an edge from that vertex. In the simplex algorithm, the column corresponding to this chosen variable is called a pivot column, and this variable is called the entering variable. The edge would reach another vertex when one of the remaining p variables,

currently positive, reaches 0 as a result of this chosen variable increasing from 0. That variable is called the leaving variable.

Chapter 7

SINGULAR VALUE DECOMPOSITION

7.1 Introduction

We first sketch the definition of *singular value decomposition* and its geometric interpretation, then remark on its applications in *low-rank matrix approximation* and *data compression*. The details of the development will be filled-in in subsequent sections. Given an $n \times d$ matrix B, singular value decomposition decomposes it as follows:

$$B = Q\Sigma P^{\mathrm{T}}.$$

Here, Q is an $n \times n$ matrix with orthonormal columns, Σ is an $n \times d$ rectangular diagonal matrix with nonnegative entries at (i, i) positions arranged in non-increasing order as i increases, and P is a $d \times d$ matrix with orthonormal columns. The columns of Q and the columns of P are referred to as *left singular vectors* and *right singular vectors*, respectively. The positive entries of Σ are referred to as *singular values* of B:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix},$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, $\sigma_{r+1} = 0$ (if $r < \min\{n, d\}$), where $1 \leq r \leq \min\{n, d\}$ is the same as the rank of B.
Note that, once a singular value decomposition of B is given: $B = Q\Sigma P^{\mathrm{T}}$, we get

$$B^{\mathrm{T}} = P\Sigma^{\mathrm{T}}Q^{\mathrm{T}},$$

which is a singular value decomposition of B^{T}; furthermore, they share the same singular values, and any right singular vector of B becomes a left eigenvector of B^{T}, and any left singular vector of B becomes a right eigenvector of B^{T}.

Singular value decomposition does not require B to be a square matrix and can be considered as a generalization of the diagonalization of a real symmetric matrix by orthogonal matrices.

Geometric interpretation of singular value decomposition Since P in the singular value decomposition is invertible, the equation $B = Q\Sigma P^{\mathrm{T}}$ is equivalent to $BP = Q\Sigma$. But when $d \leq n$, the latter is equivalent to $BP[:, j] = \sigma_j Q[;, j]$ for $j = 1, \cdots, d$, and has the interpretation that **(a)**. the columns of P, $\{P[:, j], j = 1, \cdots, d\}$, form an orthonormal basis of $\mathcal{R}^d$, **(b)**. $\{BP[:, j], j = 1, \cdots, d\}$ are orthogonal to each other. Namely, the linear map $B : \mathcal{R}^d \mapsto \mathcal{R}^n$ has an orthonormal basis $\{\bar{v}_j, j = 1, \cdots, d\}$ such that $\{B\bar{v}_j, j = 1, \cdots, d\}$ continues to be orthogonal. When $n < d$, the equations $BP[:, j] = \sigma_j Q[;, j]$ for $j = r + 1, \cdots, d$, are interpreted as $BP[:, j] = 0Q[;, j]$. In fact, these relations can be summarized as $BP[:, j] = \sigma_j Q[;, j]$ for $j = 1, \cdots, r$, and $BP[:, j] = \bar{0}$ for $j > r$, which imply that the columns $r + 1$ through d of P form an orthonormal basis of the null space of B, and columns 1 through r of Q form an orthonormal basis of the column space of B!
What about columns $r+1$ through n of Q? It turns out that they form an orthonormal basis of the null space of B^{T}.
Let $\bar{v}_j = P[:, j]$ for $j = 1, \cdots, d$, and $\bar{w}_k = Q[:, k]$ for $k = 1, \cdots, n$. Then any $\bar{x} \in \mathcal{R}^d$ can be written as $\bar{x} = \sum_{j=1}^{d} c_j \bar{v}_j$ for some c_j, $j = 1, \cdots, d$, and $B\bar{x} = \sum_{j=1}^{d} c_j \sigma_j \bar{w}_j$ (interprete $\sigma_j = 0$ when $j > r$). Since $\{\bar{v}_j, j = 1, \cdots, d\}$ is orthonormal in $\mathcal{R}^d$ and $\{\bar{w}_k, k = 1, \cdots, n\}$ is orthonormal in $\mathcal{R}^n$, we find

$$\|\bar{x}\|^2 = \sum_{j=1}^{d} c_j^2, \quad \|B\bar{x}\|^2 = \sum_{j=1}^{r} c_j^2 \sigma_j^2.$$

Thus

$$\frac{\|B\bar{x}\|}{\|\bar{x}\|} = \left(\frac{\sum_{j=1}^{r} c_j^2 \sigma_j^2}{\sum_{j=1}^{d} c_j^2} \right)^{1/2}$$

reaches its maximum possible value of σ_1 when $\bar{x}$ is a multiple of $\bar{v}_1$ and its minimum possible value of σ_r when $r = d \leq n$, and of 0 when $r < d$. Namely, σ_1 is the maximum possible magnifying factor of B and σ_r when $r = d \leq n$ is the minimum possible magnifying factor of B.

Algebraic interpretation of singular value decomposition The equation $B = Q\Sigma P^{\mathrm{T}}$ implies

$$B = \sum_{j=1}^{r} \sigma_j Q[:, j] P[:, j]^{\mathrm{T}}$$

Note that each $Q[:, j]P[:, j]^{\mathrm{T}}$ is a rank-1 $n \times d$ matrix. It follows that

$$B^{\mathrm{T}} = \sum_{j=1}^{r} \sigma_j P[:, j] Q[:, j]^{\mathrm{T}},$$

from which we see that

$$B^{\mathrm{T}}Q[:,k] = \begin{cases} \sigma_k P[:,k] & \text{for } k \le r, \\ \bar{0} & \text{for } k > r. \end{cases}$$

It follows that

$$B^{\mathrm{T}}BP[:,j] = \begin{cases} \sigma_j^2 P[:,j] & \text{for } j \le r, \\ \bar{0} & \text{for } k > r, \end{cases} \quad BB^{\mathrm{T}}Q[:,k] = \begin{cases} \sigma_k^2 Q[:,k] & \text{for } k \le r, \\ \bar{0} & \text{for } k > r. \end{cases}$$

Namely, the columns of P are orthonormal eigenvectors of $B^{\mathrm{T}}B$, and columns of Q are orthonormal eigenvectors of BB^{T}, furthermore, the singular values σ_j, $j = 1, \cdots, r$, are the non-zero eigenvalues of $B^{\mathrm{T}}B$ and BB^{T}.

Questions to be addressed: (a). What are the significance of the singular values σ_j's? **(b).** Are there any relations between the singular values of a square matrix and its eigenvalues? What are the differences of their roles in applications? **(c).** How to compute the singular values (and eigenvalues) numerically?

7.2 SVD: A Linear Algebra Perspective

7.2.1 Singular Value Decomposition of a Square Matrix

First we note that if A is a real symmetric matrix with orthogonal diagonalization $A = P\Lambda P^{\mathrm{T}}$, then the columns of P are orthonormal, and $AP[:,j] = \lambda_j P[:,j]$ is orthogonal to $AP[:,k] = \lambda_k P[:,k]$ for $j \neq k$. To get a singular value decomposition for A, we only need to make $AP[:,j]$ a non-negative multiple of a unit vector. This is done by arranging the eigenvalues such that $|\lambda_j|$ is non-increasing in j and setting

$$Q = P\mathrm{diag}[\mathrm{sign}(\lambda_1), \cdots, \mathrm{sign}(\lambda_n)], \quad \Sigma = \mathrm{diag}[|\lambda_1|, \cdots, |\lambda_n|],$$

where we set $\mathrm{sign}(\lambda) = 1$ when $\lambda = 0$.
For any $n \times d$ matrix B, $B^{\mathrm{T}}B$ and BB^{T} are positive semidefinite and symmetric (but of sizes $d \times d$ and $n \times n$ respectively). The following lemma is stated for a square matrix, but it also holds for a rectangular matrix.

Lemma 7.2.1

Let B be a square, $m \times m$ matrix. Then, the following results are true:

1. *If $\bar{p}$ is a unit eigenvector of $B^{\mathrm{T}}B$ with non-zero eigenvalue λ, then $B\bar{p}$*

is an eigenvector of BB^T with the same eigenvalue λ. Furthermore, the norm of $B\bar{p}$ is $\sqrt{\lambda}$ (recall that $\lambda \geq 0$).

2. *If $\bar{q}$ is a unit eigenvector of BB^T with non-zero eigenvalue λ, then $B^T\bar{q}$ is an eigenvector of B^TB with the same eigenvalue λ. Furthermore, the norm of $B^T\bar{q}$ is $\sqrt{\lambda}$.*

Corollary 7.2.1: (Eigenvector Pairing)

Let B be a square, $m \times m$ matrix. Then, the matrices $B^{\mathrm{T}}B$ and BB^{T} have the same set of m eigenvalues $\lambda_1, \cdots, \lambda_m$. Let the m orthonormal eigenvectors of the symmetric matrix $B^{\mathrm{T}}B$ be denoted by $\bar{p}_1, \cdots, \bar{p}_m$ with eigenvalues $\lambda_1, \cdots, \lambda_m$. Then, it is possible to find m orthonormal eigenvectors $\bar{q}_1, \cdots, \bar{q}_m$ of BB^{T}, such that the following holds:

$$\bar{q}_i\sqrt{\lambda_i} = B\bar{p}_i$$

Theorem 7.2.2: (Existence of SVD)

Let the columns of the $m \times m$ matrix P contain the m orthonormal eigenvectors of the $m \times m$ matrix $B^{\mathrm{T}}B$, and let Σ be an $m \times m$ diagonal matrix with diagonal entries containing the square-root of the corresponding eigenvalues. By convention, the columns of P and Σ are ordered, so that the eigenvalues values are in non-increasing order. Then, it is possible to find an $m \times m$ orthogonal matrix Q containing the orthonormal eigenvectors of BB^{T} , such that the following holds:

$$B = Q\Sigma P^{\mathrm{T}}.$$

Lemma 7.2.2: (Polar Decomposition)

Any square matrix can be expressed in the form US, where U is an orthogonal matrix, and S is a symmetric positive semidefinite matrix.

7.2.2 Square SVD to Rectangular SVD via Padding

The argument of the previous subsection can be adapted to rectangular matrices easily, so there is no advantage of reducing the argument to the square matrix case by paddling a rectangular one into a square one.

7.2.3 Several Definitions of Rectangular Singular Value Decomposition

The textbook gives the following definition, but it should be treated as a theorem.

Definition 7.2.1: (Singular Value Decomposition)

Consider an $n \times d$ matrix D with real-valued entries. Such a matrix can always be factorized into three matrices as follows:

$$D = Q\Sigma P^{\mathrm{T}}.$$

Here, Q is an $n \times n$ matrix with orthonormal columns containing the left singular vectors, Σ is an $n \times d$ rectangular "diagonal" matrix with diagonal entries containing the singular values in non-increasing order, and P is a $d \times d$ matrix with orthonormal columns containing the right singular vectors.

The textbook gives the following two definitions, but they should be treated as a corollary of the Theorem from the previous subsection and the respective Q's and P's are only part of the original Q's and P's in the *full* singular value decomposition.

Definition 7.2.2: (Economy Singular Value Decomposition)

Consider an $n \times d$ matrix D with real-valued entries. Such a matrix can always be factorized into three matrices as follows:

$$D = Q\Sigma P^{\mathrm{T}}.$$

Here, $p = \min\{n, d\}$, Q is an $n \times p$ matrix with orthonormal columns containing the left singular vectors, Σ is an $p \times p$ diagonal matrix with diagonal entries containing the singular values in non-increasing order, and P is a $d \times p$ matrix with orthonormal columns containing the right singular vectors.

Definition 7.2.3: (Compact Singular Value Decomposition)

Consider an $n \times d$ matrix D with real-valued entries. Such a matrix can always be factorized into three matrices as follows:

$$D = Q\Sigma P^{\mathrm{T}}.$$

Here, r is the rank of D, Q is an $n \times r$ matrix with orthonormal columns containing the left singular vectors, Σ is an $p \times p$ diagonal matrix with diagonal entries consisting the singular values in non-increasing order, and P is a $d \times r$

matrix with orthonormal columns containing the right singular vectors.

Example 7.2.1

For the matrix

```
B = 3x6
     1     2     3     3     5     6
     6     5     4     3     2     1
     7     7     7     6     7     7
```

its singular value decomposition gives

```
[U1,S1,V1]=svd(B)
```

```
U1 = 3x3
   -0.3971   -0.7134   -0.5774
   -0.4193    0.7006   -0.5774
   -0.8164   -0.0128    0.5774

S1 = 3x6
   20.5328         0         0         0         0         0
         0    5.8656         0         0         0         0
         0         0    0.0000         0         0         0

V1 = 6x6
   -0.4202    0.5798    0.6728   -0.0583    0.1279    0.1217
   -0.4191    0.3387   -0.6786   -0.1268    0.1689    0.4523
   -0.4180    0.0976   -0.1589   -0.3393   -0.5874   -0.5747
   -0.3578   -0.0196   -0.0800    0.9122   -0.1209   -0.1359
   -0.4159   -0.3845   -0.0018   -0.1380    0.7059   -0.4023
   -0.4148   -0.6256    0.2350   -0.1195   -0.3117    0.5196
```

The `"econ" option`

```
[Uc,Sc,Vc]=svd(B,"econ")
```

gives the same U but truncates S and V to the 3×3 matrices

```
Sc = 3x3
   20.5328         0         0
         0    5.8656         0
         0         0    0.0000

Vc = 6x3
   -0.4202    0.5798    0.6728
   -0.4191    0.3387   -0.6786
   -0.4180    0.0976   -0.1589
   -0.3578   -0.0196   -0.0800
   -0.4159   -0.3845   -0.0018
   -0.4148   -0.6256    0.2350
```

The rank of B is 2, so the first two columns of $U1$ form an orthonormal basis of the column space of B, and the transpose of the first two columns of $V1$ form an orthonormal basis of the row space of B. One gets the compact singular value decomposition of B using those relevant rows and columns

```
Uc(:,1:2)*Sc(1:2,1:2)*Vc(:,1:2)'
```

which gives back the B:

```
ans = 3x6
    1.0000    2.0000    3.0000    3.0000    5.0000    6.0000
    6.0000    5.0000    4.0000    3.0000    2.0000    1.0000
    7.0000    7.0000    7.0000    6.0000    7.0000    7.0000
```

If one decides to get a rank-1 approximation using the largest singular value, one gets

```
Uc(:,1)*Sc(1,1)*Vc(:,1)'
```

```
ans = 3x6
    3.4262    3.4174    3.4086    2.9179    3.3910    3.3822
    3.6173    3.6080    3.5987    3.0806    3.5801    3.5708
    7.0435    7.0254    7.0073    5.9985    6.9712    6.9531


This is an example of a truncated singular value decomposition to
be introduced next.
```

7.2.4 Truncated Singular Value Decomposition

Recall that the singular value decomposition $B = Q\Sigma P^{\mathrm{T}}$ implies

$$B = \sum_{j=1}^{r} \sigma_j Q[:, j]P[:, j]^{\mathrm{T}},$$

where r is the rank of B. If we take $k \leq r$, and define

$$D_k = \sum_{j=1}^{k} \sigma_j Q[:, j]P[:, j]^{\mathrm{T}} \tag{7.1}$$

then D_k has rank k (why?) and provides an approximation to D in the sense that $D - D_k = \sum_{j=k+1}^{r} \sigma_j Q[:, j]P[:, j]^{\mathrm{T}}$, where σ_j, for $j = k+1, \cdots, r$, are $\leq \sigma_r \leq \cdots \leq \sigma_1$. We can rewrite this D_k in the form of

$$Q_k \Sigma_k P_k^{\mathrm{T}}, \tag{7.2}$$

where Q_k, Σ_k, P_k consist of the first k columns (and the first k rows in the case of Σ_k), respectively, of Q, Σ, P. D_k is called a *truncated singular value decomposition* of D.

A truncated singular value decomposition of D is often used as a form of compression of the data matrix D. Another interpretation is to interpret each row of U_k in

$$U_k = DP_k = Q_k \Sigma_k \tag{7.3}$$

as the coordinates in the orthonormal basis $\{P[:, 1], \cdots, P[:, k]\}$ of the orthogonal projection of the corresponding row vector of D into $\mathrm{Span}\{P[:, 1], \cdots, P[:, k]\}$, as

$$U_k[j, :] = D[j, :]P_k = (D[j, :]P[:, 1], \cdots, D[j, :]P[:, k]),$$

so $P_k U_k[j, :]^{\mathrm{T}}$ represents the orthogonal projection of $D[j, :]$ into $\mathrm{Span}\{P[:, 1], \cdots, P[:, k]\}$. The scatter plot of the row vectors of U_k in $\mathcal{R}_k$ lies in $\mathrm{Span}\{P[:, 1], \cdots, P[:, k]\}$, with each vector representing the orthogonal projection of the corresponding row vector of D into $\mathrm{Span}\{P[:, 1], \cdots, P[:, k]\}$. Figure 7.2 gives an illustration of this.

Relating Truncation Loss to Singular Values

Here we want to measure the loss in the approximation $D - D_k$ using the Frobenius norm of a matrix.

Definition 7.2.4: (Frobenius Inner Product and Orthogonality)

he Frobenius inner product $\langle A, B\rangle_{\mathcal{F}}$ between $n \times d$ matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ is equal to $\sum_i \sum_j a_{ij} b_{ij}$, which is equal to the trace of $A^{\mathrm{T}}B$:

$$\langle A, B\rangle_{\mathcal{F}} = \langle B, A\rangle_{\mathcal{F}} = \mathrm{tr}(A^{\mathrm{T}}B) = \mathrm{tr}(AB^{\mathrm{T}}).$$

Two matrices are Frobenius orthogonal if their Frobenius inner product is 0. The Frobenius norm of A is defined as $\|A\|_{\mathcal{F}} = \sqrt{\langle A, A\rangle_{\mathcal{F}}}$.

Of course, one has to verify that the above definitions indeed satisfy the conditions for an inner product and norm respectively.
If P is a $d \times d$ orthogonal matrix, then

$$\langle AP, BP\rangle_{\mathcal{F}} = \mathrm{tr}(APP^{\mathrm{T}}B) = \mathrm{tr}(AB) = \langle A, B\rangle_{\mathcal{F}},$$

which means that right-multiplication by a $d \times d$ orthogonal matrix preserves the inner product. Likewise, left-multiplication by a $n \times n$ orthogonal matrix preserves the inner product. As a consequence, if Q is a $n \times n$ orthogonal matrix, then $\|QAP\|_{\mathcal{F}} = \|A\|_{\mathcal{F}}$. In particular, if B has a singular value decomposition $B = Q\Sigma P^{\mathrm{T}}$, then

$$\|B\|_{\mathcal{F}} = \|\Sigma\|_{\mathcal{F}} = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}.$$

Lemma 7.2.3

Let A and B be two $n \times d$ matrices that are Frobenius orthogonal. Then, the squared Frobenius norm of $(A + B)$ can be expressed in terms of the Frobenius norms of A and B as follows:

$$\|A + B\|_{\mathcal{F}}^2 = \|A\|_{\mathcal{F}}^2 + \|B\|_{\mathcal{F}}^2.$$

Corollary 7.2.3

Let $A_1, \cdots, A_k$ be any set of k matrices of the same size that are all Frobenius orthogonal to one another. Then, the squared Frobenius norm of the sum of these matrices can be expressed in terms of the Frobenius norms of the

individual matrices as follows:

$$\|\sum_{j=1}^{k} A_j\|_{\mathcal{F}}^2 = \sum_{j=1}^{k} \|A_j\|_{\mathcal{F}}^2$$

Corollary 7.2.4

Let $A_1, \cdots, A_k$ be any set of k matrices of the same size that are all Frobenius orthogonal to one another. Then, the squared Frobenius norm of a linear combination of these matrices can be expressed in terms of the Frobenius norms of the individual matrices as follows:

$$\|\sum_{j=1}^{k} \sigma_j A_j\|_{\mathcal{F}}^2 = \sum_{j=1}^{k} \sigma_j^2 \|A_j\|_{\mathcal{F}}^2$$

Lemma 7.2.4

Let $\bar{q}_i$ and $\bar{q}_j$ be orthogonal to one another, and also let $\bar{p}_i$ and $\bar{p}_j$ be orthogonal. Then, the rank-1 matrices $D_i = \bar{q}_i \bar{p}_i^T$ and $D_j = \bar{q}_j \bar{p}_j^T$ are Frobenius orthogonal.

As a consequence

$$\|D - D_k\|_{\mathcal{F}}^2 = \sum_{j=k+1}^{r} \sigma_j^2 \tag{7.4}$$

using

$$\|Q[:, j]P[:, j]^{\mathrm{T}}\|_{\mathcal{F}}^2 = 1.$$

It turns out that if we take any rank-k $n \times d$ matrix W,

$$\|D - W\|_{\mathcal{F}}^2 \geq \|D - D_k\|_{\mathcal{F}}^2.$$

Namely, the truncated D_k gives the best approximation of D in terms of the Frobenius norm among rank-k matrices.

7.2.7 Two-Way Versus Three-Way Decompositions

The singular value decomposition $Q\Sigma P^{\mathrm{T}}$ of D can be treated as a two-way decomposition UV^{T}, with $U = Q\Sigma$ having orthogonal columns and $V = P$ having orthonormal columns. The same can be done with the rank-k approximation $D_k = Q_k \Sigma_k P_k^{\mathrm{T}} = U_k P_k^{\mathrm{T}}$, with

$$U_k = Q_k \Sigma_k = D_k P_k = D P_k$$

representing the coordinates of the orthogonal projection of rows of D into the first k columns of P.

Exercise 1. For the matrix

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

(i). Compute the eigenvalues and a basis of their respective eigenspace for $A^{\mathrm{T}}A$ and AA^{T} and verify that they share their non-zero eigenvalues.

(ii). Find a singular value decomposition for A and identify its singular values, right singular vectors and left singular eigenvectors.

(iii). Identify a set of orthonormal basis for the row space of A and, respectively, the columns space of A. Also identify a basis for the null space of A and of $\mathrm{Null}(A^{\mathrm{T}})$ (if any).

(iv). Find an economic singular value decomposition for A.

(v). Write A as the sum of two rank-1 matrices using the singular value decomposition for A.

(vi). Find the rank-1 truncated singular value decomposition of A.

(vii). Use the singular value decomposition of A to find the pseudo inverse A^{+} of A.

(viii). Find the shortest solution to $A\bar{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

(ix). Find a least-square solution to $A^{\mathrm{T}}\bar{x} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$ and use the solution to construct a best-fitting line to scatter plots of the row vectors of $B = \begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 0 & 0 \end{bmatrix}$.

(x). Find a singular value decomposition of (column) mean-centered matrix from B and identify its right singular vectors. How does the rank-1 approximation of B through this PCA decomposition relate to the best-fitting line of the previous part?

Exercise 2. Construct a 3×4 matrix A of rank-1 which satisfies

$$A \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}.$$

Identify its singular value(s), right and left singular vector(s).

Exercise 3. If the $m \times n$ matrix A has orthogonal columns with lengths $\sigma_1, \cdots, \sigma_n$, identify its singular value(s), right and left singular vector(s).

Exercise 4. If the 2×2 symmetric matrix A has eigenvalues 3 and -2, with respective unit eigenvectors $\bar{u}_1, \bar{u}_2$, identify its singular value(s), right and left singular vector(s).

Exercise 5. Determine the pseudo inverse of A when $A = s\bar{u}\bar{v}^{\mathrm{T}}$ for $s \neq 0$ in $\mathbb{R}$ and unit vectors $\bar{u} \in \mathbb{R}^l$, $\bar{v} \in \mathbb{R}^m$. Use this to examine the relation between $(AB)^+$ and B^+A^+ when A and B are rank-1 matrices.

Exercise 6. Show that both AA^+ and A^+A are projection matrices. Identify the fundamental subspaces associated with these matrices.

Exercise 7. Show that $(A^{\mathrm{T}}A)A^+ = A^{\mathrm{T}}$.

Exercise 8. Suppose that A is an $n \times n$ invertible matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. Show that any eigenvalue λ of A satisfies $\sigma_1 \geq |\lambda| \geq \sigma_n$. Construct an example where strict inequalities hold for all eigenvalues of A.

7.3 SVD: An Optimization Perspective

The focus of this section is the following:

> Truncated SVD provides the best possible rank-k approximation of a matrix in terms of squared error.

If D is $n \times d$, $k \leq \min\{n, d\}$, U is a rank-k $n \times k$ matrix and V is a rank-k $d \times k$ matrix, then we treat UV^{T} as a a rank-k approximation of D. (Why does UV^{T} have rank k?) The squared error is given by

$$\|D - UV^{\mathrm{T}}\|_{\mathcal{F}}^2.$$

Suppose that D has a singular value decomposition $D = Q\Sigma P^{\mathrm{T}}$. Then its truncated rank-k SVD is $D_k = \sum_{i=1}^{k} \sigma_i Q(:,i)P(:,i)^{\mathrm{T}}$, and

$$\|D - D_k\|_{\mathcal{F}}^2 = \sum_{i=k+1}^{r} \sigma_i^2.$$

Since the Frobenius norm is invariant under multiplication by an orthogonal matrix, we have

$$\|D - UV^{\mathrm{T}}\|_{\mathcal{F}}^2 = \|\Sigma - Q^{\mathrm{T}}UV^{\mathrm{T}}P\|_{\mathcal{F}}^2,$$

where $\hat{U} = Q^{\mathrm{T}}U$ continues to be a rank-k $n \times k$ matrix and $\hat{V} = P^{\mathrm{T}}V$ a rank-k $d \times k$ matrix. The only non-zero entries of Σ are its r singular values, if we denote the matrix $Q^{\mathrm{T}}UV^{\mathrm{T}}P$ by S, then

$$\begin{aligned}\|\Sigma - Q^{\mathrm{T}}UV^{\mathrm{T}}P\|_{\mathcal{F}}^2 &= \sum_{i=j=1}^{r} (\sigma_i - S(i,i))^2 + \sum_{i=j=r+1}^{\min\{n,d\}} S(i,i)^2 + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{d} S(i,j)^2 \\ &\geq \sum_{i=j=1}^{r} (\sigma_i - S(i,i))^2.\end{aligned}$$

Among S such that $S(i,j) = 0$ for $i \neq j$ and with rank equal to k, the minimum of $\sum_{i=j=1}^{r}(\sigma_i - S(i,i))^2$ is $\sum_{i=k+1}^{r}\sigma_i^2$, which is attained when $S(i,i) = \sigma_i$ for $1 \leq i \leq k$ and $S_{i,i} = 0$ for $k+1 \leq i \leq \min\{n,d\}$. We get such an S when we take $U = Q\mathrm{diag}[\sqrt{\sigma_1}, \cdots, \sqrt{\sigma_k}]$ and $V = P\mathrm{diag}[\sqrt{\sigma_1}, \cdots, \sqrt{\sigma_k}]$.

7.3.1 A Maximization Formulation with Basis Orthogonality

When V is a $d \times k$ matrix with orthonormal columns, each row of DV is the coordinate of the orthogonal projection of that row into the column space of V, so $\|DV\|_{\mathcal{F}}^2$ reflects the energy of this projection—$\|D\|_{\mathcal{F}}^2 - \|DV\|_{\mathcal{F}}^2$ is the energy of the residual. To minimize this residual energy is equivalent to maximize $\|DV\|_{\mathcal{F}}^2$. So we formulate the optimization problem

$$\text{Maximize}_V \|DV\|_{\mathcal{F}}^2, \text{ subject to } V^{\mathrm{T}}V = I_k.$$

The problem can be solved in a similar way using $D = Q\Sigma P^{\mathrm{T}}$ and

$$\|DV\|_{\mathcal{F}}^2 = \|\Sigma\hat{V}\|_{\mathcal{F}}^2,$$

where $\hat{V} = P^{\mathrm{T}}V$ continues to satisfy $\hat{V}^{\mathrm{T}}\hat{V} = I_k$. Since $\|\cdot\|_{\mathcal{F}}$ in invariant under multiplication by an orthogonal matrix, we can right-multiply $\hat{V}$ by a sequence of

Householder reflections successively to make $\hat{V}$ lower triangular—this process also maintains the orthonormal property of the columns. Using

$$\|\Sigma\hat{V}\|_{\mathcal{F}}^2 = \sum_{i=1}^{r} \sigma_i^2 \sum_{j=1}^{\min\{i,k\}} \hat{V}(i,j)^2 = \sum_{j=1}^{k}\sum_{i=j}^{r} \sigma_i^2 \hat{V}(i,j)^2$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$, $\sum_i V(i,j)^2 = 1$, we see that

$$\sum_{j=1}^{k}\sum_{i=j}^{r} \sigma_i^2 \hat{V}(i,j)^2 \leq \sum_{j=1}^{k} \sigma_j^2,$$

and equality can be attained if $V(:,j) = \bar{e}_j$ for $j = 1, \cdots, k$. Thus the maximum of $\|DV\|_{\mathcal{F}}^2$, subject to $V^{\mathrm{T}}V = I_k$, is $\sum_{j=1}^{k} \sigma_j^2$.

7.3.2 A Minimization Formulation with Residuals

Here is a reformulation of the approximation problems as an unconstrained optimization problem:

$$\text{Minimiz}_{U,V}\|D - UV^{\mathrm{T}}\|_{\mathcal{F}}^2, U \text{ is } n \times k, \text{ and } V \text{ is } d \times k.$$

The hidden constrain is that the rank of U and V, as well as UV^{T} is at most k. One reaches the same conclusion in a similar fashion.

7.3.4 Principal Component Analysis

Given an $n \times d$ data matrix D, let $\bar{\mu}$ be the $1 \times d$ matrix representing the mean of columns of D, then

$$M = D - \bar{1}\bar{\mu}$$

is the mean-centered data set. In Principal Component Analysis, one does an SVD on M to identify its principal right singular vectors and singular values: $M = \hat{Q}\Sigma V^{\mathrm{T}}$, which is equivalent to getting an orthogonal diagonalization of the covariance matrix $C = M^{\mathrm{T}}M/n = V\Delta V^{\mathrm{T}}$ with $\Delta = \Sigma^2/n$.
Note that $n\bar{\mu} = \bar{1}D$, so

$$M^{\mathrm{T}}M = (D^{\mathrm{T}} - \bar{\mu}^{\mathrm{T}}\bar{1}^{\mathrm{T}})(D - \bar{1}\bar{\mu}) = D^{\mathrm{T}}D - n\bar{\mu}^{\mathrm{T}}\bar{\mu}.$$

This shows that the singular values of M are no greater than those of D, and that the residual of rank-k approximation of M is less than or equal to the residual of rank-k approximation of D.

Once we obtain a rank-k approximation UV^{T} of M, we define $\bar{1}\bar{\mu} + UV^{\mathrm{T}}$ as a corresponding approximation of D, and denote it as D_{pca}.

Exercise 1. Let $B = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 0 & 1 \end{bmatrix}$. (a). Find the residual of the optimal rank-1 approximation of B. (b) Find the residual of the optimal rank-1 approximation of the (column) mean-centered matrix from B.

Exercise 2. Let $C = \begin{bmatrix} 1 & 1 & 2 \\ -1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$. Find the maximum and minimum values of $\|CV\|_{\mathcal{F}}$, where V is a 3×2 matrix with orthonormal columns.

7.4 Applications of Singular Value Decomposition

The textbooks gives a survey of various applications of singular value decomposition. We have already seen the ideas of *dimensionality reduction* in using truncated singular value decomposition and how the four fundamental subspaces of a matrix relate to its singular value decomposition. We will take some time to discuss its application in computing the pseudo inverse of a matrix.

7.4.4 Moore-Penrose Pseudoinverse

The Moore-Penrose pseudoinverse of a matrix was defined through a limiting process and we found a formula for computing it using the QR factorization of the matrix. It can also be computed efficiently using the singular value decomposition of a matrix. Consider the compact SVD of an $n \times d$ matrix D of rank $k \leq \min\{n, d\}$: $D = Q\Sigma P^{\mathrm{T}}$. Here, Q is an $n \times k$ matrix with orthonormal columns, Σ is a $k \times k$ diagonal matrix with only positive diagonal entries, and P is a $d \times k$ matrix orthonormal columns. Then the pseudoinverse D^+ of D is given by

$$\begin{aligned} D^+ = &\lim_{\lambda \to 0+} (D^{\mathrm{T}}D + \lambda I_d)^{-1} D^{\mathrm{T}} = \lim_{\lambda \to 0+} P(\Sigma^2 + \lambda I_k)^{-1} \Sigma Q^{\mathrm{T}} \\ = &P\Sigma^{-1}Q^{\mathrm{T}} \end{aligned}$$

Condition Number of a Matrix in terms of its Singular Values

Since the condition number $\kappa(D)$ of a matrix D is defined as the ratio of the largest value of $\|D\bar{x}\|/\|\bar{x}\|$ to the smallest value of $\|D\bar{x}\|/\|\bar{x}\|$, when $\mathrm{Null}(D) = \{\bar{0}\}$, we find

$$\kappa(D) = \sigma_1(D)/\sigma_k(D), k = \mathrm{rank}(D) = \ \#\text{ of columns of } D.$$

7.4.5 Solving Linear Equations and Linear Regression

If we use the full singular value decomposition $D = Q\Sigma P^{\mathrm{T}}$, then solving $D\bar{x} = \bar{b}$ is equivalent to solving

$$\Sigma\bar{y} = Q^{\mathrm{T}}\bar{b}, \quad \bar{x} = P\bar{y}.$$

When Σ has a row of 0's, then the first system may not have a solution unless the corresponding row of $Q^{\mathrm{T}}\bar{b}$ is also 0; when Σ has a column of 0's, then the first system won't have a unique solution, if it has any solutions.
The best fitting solution in terms of $\bar{y}$ uses the first k of the equations for $\bar{y}$, where k is the rank of Σ (and of D), and sets $y_j = 0$ for $j > k$. The residual of this approximate solution comes from $Q(:, j)^{\mathrm{T}}\bar{b}$ for $j > k$, which are the coordinates of the orthogonal component of $\bar{b}$ with respect to the column space of $D = \mathrm{Span}\{Q(:, 1), \cdots, Q(:, k)\}$. This process is equivalent to using the compact singular value decomposition $Q_c\Sigma_c P_c^{\mathrm{T}}$ of D to get $\bar{y}_c = \Sigma_c^{-1}Q_c^{\mathrm{T}}\bar{b} \in \mathcal{R}^k$ and $\bar{x} = P_c\bar{y}_c$ ($P_c\bar{y}_c = P\bar{y}$, if we amend 0's to the end of $\bar{y}_c$ to make it as a vector $\bar{y} \in \mathcal{R}^d$). The end result is the same as given by the Moore-Penrose pseudo inverse: $\bar{x} = P_c\Sigma_c^{-1}Q_c^{\mathrm{T}}\bar{b} = D^{+}\bar{b}$.

7.5 Numerical Algorithms for SVD

Our construction for the SVD of a matrix D using the eigenvalues and eigenvectors of $D^{\mathrm{T}}D$ (or DD^{T}) provides a way to find the SVD of D, provided we have a good algorithm to compute the eigenvalues and eigenvectors of a symmetric matrix.
Since D and $V^{\mathrm{T}}DU$ share the same singular values for any orthogonal matrices U and V (of the appropriate sizes), if we can find orthogonal matrices U and V to reduce D to a simpler $D_1 = V^{\mathrm{T}}DU$ whose SVD can be computed efficiently $D_1 = Q_1\Sigma P_1^{\mathrm{T}}$, then $D = VQ_1\Sigma P_1^{\mathrm{T}}U^{\mathrm{T}}$ gives an SVD of D.
It turns out that one can apply appropriate Householder reflections successively to obtain a D_1 of the bi-diagonal form

$$\begin{bmatrix} s_1 & * & 0 & \cdots & 0 \\ 0 & s_2 & * & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

This uses the elementary fact that for any non-zero vector $\bar{v}$ one can construct a canonical Householder reflection H_j such that $H_j\bar{v} = \|\bar{v}\|\bar{e}_j$. We first apply this algorithm to the first column of D to produce an $n \times n$ Householder reflection H_1 such that

$$D_1 = H_1 D = \begin{bmatrix} s_1 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Then we focus the $(n-1) \times (d-1)$ matrix by removing the first row and column of D_1 and find an $(n-1) \times (n-1)$ Householder reflection H_2 which has the same effect on the first column of this sub matrix. If we define

$$G_2 = \begin{bmatrix} 1 & \bar{0} \\ \bar{0} & H_2 \end{bmatrix},$$

then it is an orthogonal $n \times n$ matrix and

$$G_2 D_1 = \begin{bmatrix} s_1 & * & * & \cdots & * \\ 0 & s_2 & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

After at most $\min\{n, d\}$ steps, we have reduced D to an upper triangular matrix U. Then we construct a $(d-1) \times (d-1)$ Householder reflection R_1 such that $[U(1,2), \cdots, U(1,d)]R_1 = t_1[1, 0, \cdots]$. Now we define

$$S_1 = \begin{bmatrix} 1 & \bar{0} \\ \bar{0} & R_1 \end{bmatrix}.$$

It satisfies

$$US_1 = \begin{bmatrix} s_1 & t_1 & 0 & \cdots & 0 \\ 0 & \hat{s}_2 & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

This process can be continued, which eventually reduces the matrix to an upper bi-diagonal matrix. There are efficient algorithms to compute the singular values of bi-diagonal matrices.