Applied Linear Theory

Adam Boucher

Contents

Part 1

Theory

CHAPTER 1

Linear Spaces

In mathematics we use the word 'space' in a wide array of different contexts. In most of these contexts the abstraction of a mathematical space is a useful tool for framing mathematical problems. A space typically describes a set of objects along with an associated collection of operations or transformations. The axioms, operations, and rules of a mathematical space are intrinsic properties of the space itself. These rules form the defining features of the mathematical space, and characterize the operations and activities one can perform within that space. Essentially these axioms determine the rules of the game. In abstract study the axioms and properties are defined at the outset and taken for 'granted' in the work that follows. This approach is helpful in two ways. First from the mathematical perspective the axioms provide a foundation for more sophisticated theorems and conjectures to be developed. Second from the perspective of applications, once the axioms are assumed or verified for a system of interest then one inherits all the theorems and results which have been worked on in the abstract case. Thus the theorems and results are independent of how the mathematical space is instantiated in reality. By studying the properties of mathematical spaces in a systematic and logical way, one can deduce important and interesting results which capture general structural properties valid in every case, rather than trying to use the particulars of an application to reason inductively about other applications which appear to have similar structure. In this sense the study of abstract mathematical spaces is purely deductive.
By maintaining an awareness and sensitivity for the structural (specifically algebraic) properties of the measurements of physical systems, anyone with a knowledge of mathematical theory can immediately transfer that knowledge to **any** physical models exhibiting the same structural properties. This intrinsic versatility of mathematical knowledge is unique to the subject, and is one of the most powerful attractions to mathematics.

It may be helpful to think of a mathematical space as an abstract game, where the elements of the set are the game pieces and the operations and transformations embody the 'rules' of the game. When mathematicians study the properties of different spaces, they try to find useful and interesting properties and connections which can be deduced purely from the structure of the game, without any dependence on the particular pieces at hand. For example, any legal chess move, would be legal in **any** game of chess, regardless of the size, color or material of the chess pieces. Similarly, any theorem we can prove about general linear spaces will be true for **any** linear space whether the underlying elements of the space are n-tuples of numbers, continuous functions, or some other collection of objects

with a linear structure.

From a conceptual standpoint, linear spaces are spaces where the objects may be arbitrarily reconfigured in any desired combination, and in any amount. We will also use the word 'linear' to describe transformations of the objects within a linear space. Again conceptually speaking, a linear transformation is one whose action or operation on any collection of objects is the same as its action on the parts of that collection in any chosen subdivision. More concretely linear spaces are structured around two essential operations. The first operation is an addition operation which shares all of the algebraic properties of arithmetic addition. When we deal with the elements of a linear space, we must have a sensible way of adding those elements together. Regardless of what the elements actually are, it is essential that given any two elements we can uniquely compute the sum of those two elements in a consistent and repeatable manner. Furthermore the addition operation must have the associativity and commutativity of regular arithmetic addition, so we are justified in asserting:

$$(\mathbf{x}+\mathbf{y})+\mathbf{z}=\mathbf{x}+(\mathbf{y}+\mathbf{z})$$
$$\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$$

any time $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ belong to the same linear space.

Incidently, one usually refers to the elements of a linear space as *vectors* as a shorthand description. Since the word *vector* is used as a general description for linear elements, it can be used to describe several different kinds of object, even in the same mathematical statement.

The second operation which must be well defined for a linear space is that of scalar multiplication. For the most common applications we will allow vectors to be multiplied by real numbers, but in certain applications it is perfectly reasonable to multiply vectors by complex numbers, or by numbers drawn from arbitrary mathematical fields. When using the real or complex numbers as the underlying scalar field, scalar multiplication behaves like regular multiplication. In general scalar multiplication must be associative and commutative with vectors and must distribute over vector addition. Thus whenever working within a linear space we are allowed to assert:

$$\alpha\beta\mathbf{x}=\mathbf{x}\beta\alpha,$$
$$\alpha\left(\mathbf{x}+\mathbf{y}\right)=\alpha\mathbf{x}+\alpha\mathbf{y},$$

whenever α and β are scalars, and $\mathbf{x}$ and $\mathbf{y}$ are vectors.

The operations of addition and scalar multiplication set the ground rules for linear spaces and they provide all the underlying structure necessary to describe a wide variety of important mathematical models. The definitions above are given as an informal description of the essential properties of a linear space, however students familiar with abstract

algebra may benefit from a more precise definition.

Definition: *A* **linear space** $\mathcal{V}$ *over a field* K*, consists of a set of objects called vectors which form an abelian group under the vector addition operation, and for which multiplication by elements of the field* K *is an associative, and distributive operation with a unit, denoted by 1.*

When studying linear space from a mathematical perspective each of these properties is assumed to hold by hypothesis, and all of the algebraic properties are taken as axioms.

Remark: Several different conventions are used to distinguish vectors and scalars in modern notation. We will use lowercase boldface type (e.g. $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, or $\boldsymbol{\phi}$, $\boldsymbol{\psi}$) to denote vectors and normal case type to denote scalars (e.g. a, b, c or x, y, z).

1. Construction and Representation in Linear Spaces

We begin our study of linear spaces by defining the core concepts of linear algebra. Initially these concepts might seem abstract and arbitrary, but we will see that they appear repeatedly in the study of linear systems. As we develop these concepts you will see definitions, theorems and lemmas which are moderately rigorous. These components are important to mathematicians for they give a framework to the theory that we develop. As we work through these mathematical results we will discuss the framework itself and try to give you a coherent understanding of the need for proof and justification and some confidence that our results are at least internally consistent and correct. My level of rigor and detail varies considerably from result to result, but feel confident that each theorem can be proven with more precision and technical detail if required.

A **subspace** of a linear vector space is a subset of the vectors of another linear space which is self-contained under the addition and scalar multiplication operations induced by the parent space. Very often one needs to focus one's attention on restricted classes of elements, rather than all possible vectors of a particular kind. The concept of subspace allows us to talk about restricted collections of vectors where all of the axioms of linear algebra still hold. If we were to restrict ourselves only to subsets of elements of a vector space, we might lose the linear structure and might not be able to make sense of arbitrary addition and scalar multiplication operations. Since subspaces are themselves, linear spaces they preserve the linear structure of the parent space.
Using the concepts of linear combinations and span we can easily generate and represent subspaces for given linear vector spaces.

We begin with the concept of a linear combination. If $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ are vectors drawn from some linear space, then any vector, $\mathbf{v}$ which can be represented by a scaled sum of these vectors is a linear combination of said vectors.

Definition: *The vector,* $\mathbf{v}$ *is said to be a* **linear combination** *of the vectors* $\{\boldsymbol{\phi}_i\}_{i=1}^n$ *if and only if there exist scalars* α_i*,* $i = 1..n$ *such that:*

$$\mathbf{v} = \alpha_1\boldsymbol{\phi}_1 + \alpha_2\boldsymbol{\phi}_2 + \ldots \alpha_n\boldsymbol{\phi}_n$$

We can use the linear combination concept to construct linear spaces. We use the term **span** to define the set of all possible linear combinations of a given set of vectors.

Definition: *Let $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \ldots \boldsymbol{\phi}_n\}$ be a set of vectors with a well defined addition operation, and a well defined scalar multiplication operation (with elements from some field, K). We define the* **span** *of $\{\boldsymbol{\phi}_i\}_{i=1}^n$ as the set of all possible linear combinations. Algebraically we have:*

$$Sp(\{\boldsymbol{\phi}_i\}) \equiv \left\{ \mathbf{v} \middle| \exists \{\alpha_i\} \left(\mathbf{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i \right) \right\}.$$

Spanning Lemma: *For any set of vectors $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \ldots \boldsymbol{\phi}_n\}$, the set $Sp(\{\boldsymbol{\phi}_i\})$ forms a linear vector space over the field K under the induced addition and scalar multiplication operations.*

Proof:

When trying to prove that a particular set of vectors forms a vector space, the proof typically hinges on three important factors. One must establish that the set is closed under vector addition and scalar multiplication, one must establish that the zero vector is an element of the set, and one must ensure that the set contains appropriate inverse elements. Usually the underlying vector addition and scalar multiplication operation are familiar, and one assumes these operations possess the required algebraic properties without additional proof.

To demonstrate that $Sp(\{\boldsymbol{\phi}_i\})$ is closed under vector addition we select two arbitrary elements from the set, $\mathbf{u}$ and $\mathbf{v}$, we then need to establish that their sum is also an element of $Sp(\{\boldsymbol{\phi}_i\})$. Since $\mathbf{u}$ and $\mathbf{v}$ are both equal to linear combinations of the $\boldsymbol{\phi}_i$ we write them in terms of these representations and take their sum.

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i, \qquad \mathbf{v} = \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}_i$$

using these representations and the algebraic properties of the addition and scalar multiplication operations we find that:

$$\mathbf{u} + \mathbf{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i + \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}_i,$$

$$\mathbf{u} + \mathbf{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i + \beta_i \boldsymbol{\phi}_i,$$

$$\mathbf{u} + \mathbf{v} = \sum_{i=1}^{n} (\alpha_i + \beta_i) \boldsymbol{\phi}_i.$$

Because K is a field the sums $\alpha_i + \beta_i$ must each be elements of K. We conclude that $\mathbf{u}+\mathbf{v}$ may be written as a linear combination of the $\boldsymbol{\phi}_i$. Consequently,

$$\forall \mathbf{u}, \mathbf{v}\Big(\mathbf{u}, \mathbf{v} \in Sp(\{\boldsymbol{\phi}_i\}) \rightarrow \mathbf{u}+\mathbf{v} \in Sp(\{\boldsymbol{\phi}_i\})\Big).$$

which is exactly closure under addition. Closure under scalar multiplication is left as an exercise.

To demonstrate that $Sp(\{\boldsymbol{\phi}_i\})$ possesses the zero vector we note that

$$\emptyset = 0\boldsymbol{\phi}_1 + 0\boldsymbol{\phi}_2 + \cdots + 0\boldsymbol{\phi}_n,$$

so $\emptyset \in Sp(\{\boldsymbol{\phi}_i\})$.

To demonstrate the span is closed under inverses, we take an arbitrary $\mathbf{u}$, and construct the inverse element from the same set. Let $\mathbf{u} \in Sp(\{\boldsymbol{\phi}_i\})$ be given. Since $\mathbf{u}$ is a linear combination of the $\boldsymbol{\phi}_i$ we can write:

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i$$

we now construct the additive inverse for $\mathbf{u}$ by taking the linear combination:

$$-\mathbf{u} = \sum_{i=1}^{n} -\alpha_i \boldsymbol{\phi}_i$$

where the $-\alpha_i$ are the appropriate additive inverses drawn from the field K. One easily verifies that:

$$\mathbf{u} + (-\mathbf{u}) = \emptyset,$$

as desired.
Now that all the conditions for a linear space are verified, we conclude that for any collection of vectors, $\{\boldsymbol{\phi}_i\}_{i=1}^n$ the set $Sp(\{\boldsymbol{\phi}_i\})$ forms a linear vector space over K.

Q.E.D

Spanning Corollary: *If every element* $\mathbf{v}$ *contained in a linear vector space* $\mathcal{V}$ *over a field* K *may be written as a linear combination of some set of vectors* $\{\boldsymbol{\phi}_i\}$ *also contained in* $\mathcal{V}$*, then*

$$\mathcal{V} = Sp(\{\boldsymbol{\phi}_i\}).$$

Proof:

This proof is left as an exercise.

Remark: When a set of vectors $\{\boldsymbol{\phi}_i\}$ has the property that $Sp(\{\boldsymbol{\phi}_i\}) = \mathcal{V}$ we use the word span as a verb and say that the $\boldsymbol{\phi}_i$ **span** the vector space $\mathcal{V}$.

The spanning lemma and its corollary are important for several reasons, first if we have a vector addition operation and compatible scalar field in mind, we can use the spanning lemma to construct linear vector spaces from scratch. Secondly, in light of the spanning corollary we can use a spanning set to obtain **representations** of all of the elements within a linear space. We will find that many choices of representation are possible and some representations are more 'efficient' than others. As we pass from purely abstract study to more specialized applications you may find it helpful to think of representations as possible maps or coordinate systems for linear spaces. You may also think of a representation of a vector as a particular list of ingredients. Picking a particular representation for the elements in a linear vector space has important implications in every application.

In order to define what we mean by efficiency in the context of vector representation we introduce the concept of **linear dependence**. This concept gives us a formal and mathematical handle on redundancy in vector representations.

Definition: *A set of vectors* $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n\}$ *is said to be* **linearly dependent** *if and only if it is possible to create a non-trivial linear combination of the* $\boldsymbol{\phi}_i$ *which sums to the zero vector. Formally we have:*

The set $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n\}$ is linearly dependent if and only if:

$$\exists \{\alpha_i\}_{i=1}^n \text{ and } \exists j \in [1, 2, \ldots, n] \text{ such that } \alpha_j \neq 0, \text{ and } \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i = \boldsymbol{\emptyset}.$$

This definition captures redundancy because we can use the formula in the definition to show that at least one of the vectors in the set $\{\boldsymbol{\phi}_i\}$ can be represented (i.e. written as a linear combination) of the other vectors. Explicitly, since we have taken α_j to be non-zero, we can manipulate the linear dependence relation as follows:

$$\begin{aligned} \alpha_1 \boldsymbol{\phi}_1 + \alpha_2 \boldsymbol{\phi}_2 \cdots + \alpha_j \boldsymbol{\phi}_j + \ldots \alpha_n \boldsymbol{\phi}_n &= \boldsymbol{\emptyset}, \\ \alpha_1 \boldsymbol{\phi}_1 + \alpha_2 \boldsymbol{\phi}_2 \cdots + \alpha_n \boldsymbol{\phi}_n &= -\alpha_j \boldsymbol{\phi}_j, \\ -\frac{\alpha_1}{\alpha_j} \boldsymbol{\phi}_1 - \frac{\alpha_2}{\alpha_j} \boldsymbol{\phi}_2 \cdots - \frac{\alpha_n}{\alpha_j} \boldsymbol{\phi}_n &= \boldsymbol{\phi}_j \end{aligned}$$

In light of this dependence relation we can now systematically remove any reference to $\boldsymbol{\phi}_j$ and construct new representations of every vector in $Sp(\{\boldsymbol{\phi}_i\})$ without using $\boldsymbol{\phi}_j$. In terms of linear vector spaces, such a dependence relation implies that the linear space generated by taking the span of $\{\boldsymbol{\phi}_i\}_{i=1}^n$ is the same as the linear space generated by the smaller set $\{\boldsymbol{\phi}_i\}_{i=1}^n \backslash \{\boldsymbol{\phi}) j\}$.

When a set of vectors is **not** linearly dependent, we say that set is **linearly independent**. Linear independence is also an important property in linear algebra, and it can be given a formal definition, which is just the logical negation of the definition of linear dependence.

Definition: *A set of vectors* $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n\}$*, is said to be* **linearly independent** *if*

and only if it is impossible to find a non-trivial linear combination of the $\boldsymbol{\phi}_i$ which sums to the zero vector. Formally we can write:
The set $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n\}$ is linearly independent if and only if:

$$\forall \{\alpha_i\}_{i=1}^{n} \in K \left(\sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i = \boldsymbol{\emptyset} \rightarrow \forall j \in [1, 2, \ldots n], \quad \alpha_j = 0 \right).$$

The formal definitions of linear dependence and linear independence use rather complex logically quantified statements which can be tricky to use properly unless you have some competence with logic and mathematical proof. On a conceptual level, you should think of these definitions as providing guarantees for what can or cannot happen with particular sets of vectors.

Suppose that we state the set of vectors $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ is a linearly **dependent** set. This statement **means** several different things, first it means that there is a way to choose an 'interesting' set of scalars α, β, γ which make the equation:

$$\alpha\mathbf{x} + \beta\mathbf{y} + \gamma\mathbf{z} = \boldsymbol{\emptyset}$$

true. By 'interesting' here we mean that is is possible to choose these numbers so that α, β and γ are not all equal to zero. Secondly, as a consequence of the existence of those numbers, it is also possible to write at least one of these vectors as a linear combination of the other two. As stated earlier this can be interpreted as a kind of redundancy in the collection of the vectors, where we can think of at leat one vector as being a composition derived from a linear combination of the other vectors. This vector will not contribute anything new to the span, and can be removed and the remaining collection of vectors will retain the ability to represent all of the elements of exactly the same space.

Alternately, suppose that we state that the set of vectors $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ is a linearly **independent** set. This statement means several different things. First we know that if we select any scalars: α, β and γ which are not all zero, then the linear combination:

$$\alpha\mathbf{u} + \beta\mathbf{v} + \gamma\mathbf{w} \neq \boldsymbol{\emptyset}.$$

As a consequence of that result, we can conclude that it is **impossible** to create any of these vectors $\mathbf{u}, \mathbf{v}$, or $\mathbf{w}$ out of linear combinations of the other two. This means that in a linearly independent set, each vector contributes something which is essentially new to the collection, something which cannot be captured exactly by linear combinations of the other elements.

Before proceeding to the next section we prove a short technical lemma which turns out to provide an important arithmetic connection between the concepts of span and linear independence.

Finite Span-Independence Lemma: *Suppose the set* $\{\mathbf{x}_i\}_{i=1}^n$ *spans a linear vector space,* $\mathcal{X}$*, and the set* $\{y_j\}_{j=1}^m$ *is a linearly independent set in* $\mathcal{X}$*.*

then:

$$m \leq n$$

Proof:

The idea of the proof is rather simple, but requires a full mastery of all of the definitions used thus far. We proceed by construction and use the representations provided by the $\mathbf{x}_i$ to create a new spanning set out of the $\mathbf{y}_j$, we then use this new spanning set to show that $m > n$ implies a dependence relation among the $\{\mathbf{y}_j\}$.

We first note that since the $\mathbf{y}_j$'s are linearly independent none of the $\mathbf{y}_j = \emptyset$.
Next we write $\mathbf{y}_1$ as a linear combination of the $\mathbf{x}_i$.

$$\mathbf{y}_1 = \sum_{i=1}^{n} \alpha_{1,i}\mathbf{x}_i.$$

Since $\mathbf{y}_1 \neq \emptyset$ at least one of the $\alpha_{1,i} \neq 0$. Call this element $\alpha_{1,p}$. We can then write:

$$\mathbf{x}_p = \frac{1}{\alpha_{1,p}}\mathbf{y}_1 - \sum_{i=1,\neq k}^{n} \frac{\alpha_{1,i}}{\alpha_{1,p}}\mathbf{x}_i$$

Remark: In this proof, I have glossed over the fact that at each step there must be at least one non-zero element $\mathbf{x}_q$ present in the representative linear combination. I leave it as an exercise to prove that the set of $\mathbf{y}_j$ must be linearly dependent if no such $\mathbf{x}_q$ exists.

Using this formula we can systematically replace the appearance of $\mathbf{x}_p$ by using $\mathbf{y}_1$ and the other $\mathbf{x}_i$. This new set will still be a spanning set for $\mathcal{X}$, and consequently we can repeat this process with $\mathbf{y}_2$. Similarly, manipulating the representation of $\mathbf{y}_2$ we can use $\mathbf{y}_2$ to replace another element, $\mathbf{x}_q$. Again, obtaining a new spanning set which involves only $\mathbf{y}_1, \mathbf{y}_2$ and the remaining $\mathbf{x}_i$. At each step we can use the new spanning set to replace another element of the $\mathbf{x}_i$, and after n steps we will have constructed a new spanning set which contains only the $\{\mathbf{y}_j\}_{j=1}^n$, if there are any $\mathbf{y}_j$, then they can be represented as a linear combination of the previous n elements, thus they must be linearly dependent, which is a contradiction.
Q.E.D.

Exercises:

(1) Let $\mathcal{V}$ be a linear space over a field K, and let $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots \boldsymbol{\phi}_n\} \subset \mathcal{V}$. Show that the set $Sp(\{\boldsymbol{\phi}_i\}_{i=1}^n)$ is closed under scalar multiplication.

(2) Prove the spanning corollary. (Remark: Because the algebraic properties of vector addition and scalar mulitplication are taken as part of the axioms of linear algebra, this proof is essentially a proof of set equality.)

(3) Let a linear space $\mathcal{X}$ be spanned by the set of vectors: $\mathcal{S} \equiv \{\mathbf{y}_1, \ldots \mathbf{y}_{n-1}, \mathbf{x}_1, \ldots \mathbf{x}_m\}$. Let the representation of $\mathbf{y}_n$ in terms of the elements of $\mathcal{S}$ be given by:

$$\mathbf{y}_n = \sum_{i=1}^{n-1} \alpha_i \mathbf{y}_i + \sum_{j=1}^{m} \beta_j \mathbf{x}_j$$

Prove that if the set of vectors: $\{\mathbf{y}_1, \ldots \mathbf{y}_{n-1}, \} \cup \{\mathbf{y}_n\}$ is linearly independent, then at least one of the $\beta_j \neq 0$.

2. Efficient Representation in Linear Spaces

In this section we introduce the concepts of **basis** and **dimension** which capture the ideas of an efficient representation, and a useful size and complexity measure for linear spaces.

Definition: *A set of vectors* $\{\boldsymbol{\phi}_i\}_{i=1}^n$ *is called a* **basis** *for a linear vector space,* $\mathcal{V}$ *if this set of vectors spans* $\mathcal{V}$*, and the set* $\{\boldsymbol{\phi}_i\}_{i=1}^n$ *is linearly independent.*

There are many possible bases for any particular vector space, but each basis shares several important characteristics.

Unique Representation Theorem: *Let* $\{\boldsymbol{\phi}_i\}_{i=1}^n$ *be a basis for a linear space V, then the representation of each* $\mathbf{v} \in \mathcal{V}$ *in terms of the* $\boldsymbol{\phi}_i$ *basis is* **unique***.*

Proof:

Uniqueness in mathematics is typically demonstrated by showing that any two objects exhibiting the properties of the unique object must be equal. In our case we let $\mathbf{v} \in \mathcal{V}$ be given, arbitrary and fixed and we suppose that $\mathbf{v}$ possesses two different representations in the $\boldsymbol{\phi}_i$ basis. Thus:

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i, \text{ and } \mathbf{v} = \sum_{i=1}^{n} \beta_i \boldsymbol{\phi}_i$$

Taking the difference of these two representations we arrive at the equation:

$$\sum_{i=1}^{n} (\alpha_i - \beta_i) \boldsymbol{\phi}_i = 0$$

Since the $\boldsymbol{\phi}_i$ form a basis, they must be linearly independent and consequently this equation only holds if $\alpha_i - \beta_i = 0$ for each i. We conclude that the two representations of $\mathbf{v}$ must have been identical.

Q.E.D.

The previous theorem helps us to understand why certain representations might be 'better' than others. Suppose that we consider a vector space $\mathcal{V}$ which is spanned by two different collections of vectors, one collection is assumed to be a basis, $\{\mathbf{b}_i\}_{i=1}^N$, while the other collection is assumed to be a spanning set which is not a basis, $\{\boldsymbol{\phi}_n\}_{n=1}^M$. By the definitions of basis and spanning set every $v \in \mathcal{V}$ can be expressed uniquely in terms of the basis, and can be represented by at least one linear combination of the spanning set. Given that we are able to write v in terms of either representation:

$$\mathbf{v} = \sum_{i=1}^{N} \beta_i \mathbf{b}_i \qquad \text{or } \mathbf{v} = \sum_{n=1}^{M} \alpha_n \boldsymbol{\phi}_n,$$

which representation is 'better?' The answer here comes down to clarity of communication. If you are trying to perform calculations and communicate your results with others, then the basis representation is the **only** choice, (this furthermore specifies the addition operation uniquely as the component-wise addition operation inherited from the scalar field). By selecting a basis for your space, you define a unique representation for each and every vector. The uniqueness ensures that you and anyone else using the same basis can agree on whether two vectors are the same or different just by inspection. If you use a spanning set with linearly dependent elements to create your representations, then you must always check all equivalent representations whenever you wish to compare two vectors. Furthermore the operation of addition would not be well defined with respect to the representations, there would be an infinite collection of possible addition rules which would agree with the correct addition of the underlying vectors.

Basis Theorem: *Let $\mathcal{V}$ be a vector space spanned by a finite set of vectors. Any two bases for $\mathcal{V}$ have the same number of vectors.*

Proof:

Suppose that $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_j\}_{i=1}^m$, are both bases for $\mathcal{V}$. Since each set is both a spanning set and linearly independent we can apply the Finite-Span independence lemma twice to obtain:

$$n \leq m, \text{ and } m \leq n.$$

we conclude that $n = m$ as desired.

Q.E.D.

Since the number of vectors contained in a basis for a fixed vector space is independent of the particular choice of basis, this number may give us some insight into the intrinsic character of the vector space itself.

Definition: *We define the number of vectors contained in a basis of a vector space the* **dimension** *of that vector space. Numerically the dimension is given by:*

$$\dim(\mathcal{V}) \equiv \min_{\{\mathcal{S} \subset \mathcal{V} \;\mid\; Sp(\mathcal{S}) = \mathcal{V}\}} |\mathcal{S}| \,.$$

(The above formal definition avoids any reference to linear independence by defining the dimension in terms of the cardinality of the smallest size spanning sets for $\mathcal{V}$.)

Any vector space with a finite basis is called a **finite dimensional vector space**.

The concept of dimension turns out to be an extremely powerful way of capturing, analyzing and predicting the behavior of objects in linear spaces. Many results involving

the structure of domains, ranges and images of mappings and functions between linear spaces can be handled with a simple arithmetic on the dimensions of the spaces involved. A simple result using such dimensional arithmetic is given below.

Subspace Theorem: *(1) All subspaces of finite dimensional vector spaces are finite dimensional. (2) Given a finite dimensional vector space $\mathcal{X}$ and a subspace $\mathcal{Y}$, there is a complementary subspace $\mathcal{Z}$ so that every $\mathbf{x} \in \mathcal{X}$ may be written uniquely as:*

$$\mathbf{x} = \mathbf{y} + \mathbf{z}$$

In addition we have the result that:

$$\dim(\mathcal{X}) = \dim(\mathcal{Y}) + \dim(\mathcal{Z})$$

Proof:
Let $\mathcal{X}$ be a finite dimensional vector space, and let $\mathcal{Y}$ be a subspace of $\mathcal{X}$.

We will prove (1) incidentally as we work through the proof of (2). $\mathcal{X}$ is a finite dimensional vector space, so it must possess a basis, $\{\boldsymbol{\phi}_i\}_{i=1}^n$.

Since the elements of $\mathcal{Y}$ are contained in $\mathcal{X}$, each element can be represented by the finite basis. We construct a basis for $\mathcal{Y}$ as follows: we begin by selecting some non-zero element $\mathbf{y}_1 \in \mathcal{Y}$ and express $\mathbf{y}_1$ as a linear combination of the $\boldsymbol{\phi}_i$ basis.

$$\mathbf{y}_1 = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i.$$

We now seek to replace one of the $\boldsymbol{\phi}_i$ with $\mathbf{y}_1$. Since $\mathbf{y}_i \neq \emptyset$ there exists at least one $\alpha_i \neq 0$, call this coefficient α_p. Then note that:

$$\boldsymbol{\phi}_p = \frac{1}{\alpha_{1,p}} \mathbf{y}_1 - \sum_{i=1,\neq k}^{n} \frac{\alpha_{1,i}}{\alpha_{1,p}} \boldsymbol{\phi}_i$$

Given this formula we can use the set $\{\mathbf{y}\} \cup \{\boldsymbol{\phi}_{i\neq p}\}$ as a new basis for $\mathcal{X}$. Next if possible find another vector $\mathbf{y}_2 \in \mathcal{Y}$ which lies in $Sp(\{\boldsymbol{\phi}_{i\neq p})$, use this vector to replace another element $\boldsymbol{\phi}_q$ in the remaining set of original basis vectors. Consequently we can use the set $\{\mathbf{y}_1, \mathbf{y}_2\} \cup \{\boldsymbol{\phi}_{i\neq,p,q}\}$ as a basis for $\mathcal{X}$. Continue this process until it is impossible to find any vectors in $\mathcal{Y}$ which belong to the span of the remaining original basis vectors.

This process is guaranteed to terminate after no more than n steps since the original space was finite dimensional. Suppose that the process terminates after j steps.

Finally, we claim that the set $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j\}$ forms a basis for $\mathcal{Y}$. To establish this claim let $\mathbf{u} \in \mathcal{Y}$. Since $\mathcal{Y}$ is a subspace $\mathbf{u}$ possesses a unique representation in terms of the mixed basis $\{\mathbf{y}_1, \ldots, \mathbf{y}_j, \boldsymbol{\phi}_{j+1}, \ldots, \boldsymbol{\phi}_n\}$ (The $\boldsymbol{\phi}$ may need to be renumbered.)

$$\mathbf{u} = \sum_{i=1}^{j} \beta_i \mathbf{y}_i + \sum_{i=j+1}^{n} \beta_i \boldsymbol{\phi}_i$$

Since $\mathcal{Y}$ is a subspace it is closed under vector addition, and since each $\mathbf{y}_i \in Y$ we have:

$$\mathbf{u} - \sum_{i=1}^{j} \beta_i \mathbf{y}_i = \sum_{i=j+1}^{n} \beta_i \boldsymbol{\phi}_i \in \mathcal{Y}$$

Since it was impossible to exhibit any non-zero elements of $\mathcal{Y}$ belonging to the span of the remaining $\boldsymbol{\phi}_i$, we conclude that $\beta_{j+1}, \ldots \beta_n$ must all be zero. Since $\mathbf{u} \in \mathcal{Y}$ was arbitrary we find $Sp(\{\mathbf{y}_i\}_{i=1}^{j}) = \mathcal{Y}$. Furthermore, since the mixed basis is linearly independent, the subset of $\mathbf{y}_i$'s must also be linearly independent. We conclude that $\mathcal{Y}$ is finite dimensional and has dimension $j \leq n$. This proves (1).

Next we define $\mathcal{Z} = Sp(\{\boldsymbol{\phi}_i\}_{i=j+1}^{n})$ we find that $\mathcal{Z}$ is also finite dimensional with $\dim(\mathcal{Z}) = n - j$. Each element $\mathbf{x} \in \mathcal{X}$ can be written uniquely in terms of the mixed basis as:

$$\mathbf{x} = \sum_{i=1}^{j} \alpha_i \mathbf{y}_i + \sum_{i=j+1}^{n} \alpha_i \boldsymbol{\phi}_i$$

Since each of these representations refers to unique elements of $\mathcal{Y}$ and $\mathcal{Z}$ respectively (2) follows directly.

Q.E.D.

In addition to having computational properties the concept of dimension can be used to show that for a given dimension and scalar field there is essentially only one vector space structure up to isomorphism. This amazing result tells us that any theorems or results of linear algebra which hold for a particular dimension or scalar field will hold for **all** linear spaces of that dimension over that scalar field regardless of the application or nature of the underlying vectors.

Isomorphism Theorem: *Any two finite dimensional vector spaces with the same dimension, and defined over the same scalar field are isomorphic.*

Proof: We prove the theorem by showing that we can construct an isomorphism between an arbitrary vector space over a field K with dimension n, and the vector space K^n. Since this isomorphism can be created from any n-dimensional vector space over K, we can create an isomorphism between any two n-dimensional vector spaces over K by composing the two instances of the isomorphism with K^n.

First we define K^n as the collection of all $n-tuples$ of elements from the field K, we define the vector addition operation for this collection objects as the regular field addition operation applied component-wise to the n-tuples. We leave it as an exercise to show that this set of objects is actually a finite dimensional vector space over K with dimension, n.

Next Let $\mathcal{V}$ be a finite dimensional vector space over K with $\dim(\mathcal{V}) = n$.
We let: $\{\boldsymbol{\phi}_i\}_{i=1}^n$ be a basis for $\mathcal{V}$. We define a mapping $\Phi : \mathcal{V} \to K^n$ which we will show is an isomorphism. For each $\mathbf{u} \in \mathcal{V}$ we define $\Phi(\mathbf{u}) \in K^n$ by using the representation of $\mathbf{u}$ in the $\boldsymbol{\phi}_i$ basis.

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i$$
$$\Phi(\mathbf{u}) =< \alpha_1, \alpha_2, \ldots, \alpha_n >$$

That is, we take the n-tuple of coordinates of $\mathbf{u}$ in its $\boldsymbol{\phi}$ representation as its image under the mapping Φ. Since the $\boldsymbol{\phi}_i$ form a basis of $\mathcal{V}$ this mapping is well defined, and it is straightforward to demonstrate that Φ is a bijection. To demonstrate that Φ is an isomorphism we compute:

$$\begin{aligned}
&\Phi(\mathbf{u}) = \boldsymbol{\alpha}, \quad \Phi(\mathbf{v}) = \boldsymbol{\beta} \\
&\Phi(\mathbf{u}+\mathbf{v}) =< \alpha_1 + \beta_1, \alpha_2 + \beta_2, \ldots, \alpha_n + \beta_n > \\
&\Phi(\mathbf{u}+\mathbf{v}) =< \alpha_1, \alpha_2, \ldots, \alpha_n > +_K < \beta_1, \beta_2, \ldots, \beta_n > \\
&\Phi(\mathbf{u}+\mathbf{v}) = \boldsymbol{\alpha} +_K \boldsymbol{\beta} \\
&\Phi(\mathbf{u}+\mathbf{v}) = \Phi(\mathbf{u}) +_K \Phi(\mathbf{v})
\end{aligned}$$

Similarly we can show that:

$$\Phi(k\mathbf{u}) = k\Phi(\mathbf{u})$$

we conclude that the linear structure is preserved under this mapping, and that Φ is indeed an isomorphism between $\mathcal{V}$ and K^n. As outlined earlier, this implies that any two n-dimensional vector spaces over K must be isomorphic to one another.

Q.E.D.

Exercises:

(1) Prove the reduction lemma.
Reduction Lemma: *Any finite set $\{\boldsymbol{\phi}_i\}_{i=1}^m$, which spans a linear vector space $\mathcal{V}$ may be reduced to a basis for $\mathcal{V}$.*

(2) Prove the completion lemma:
Completion Lemma: *Any linearly independent set of vectors in a finite dimensional vector space $\mathcal{V}$ may be augmented by additional vectors to form a basis of $\mathcal{V}$.*

(3) Prove that n-tuples of real numbers under component-wise addition form a linear vector space over the field of real numbers. (This is an important special case of the n-component vector space over an arbitrary field K, if you are familiar with the field axioms, proving the general case is structurally identical.)

3. Changing Representations in Linear Spaces

In the previous sections we have defined all of the foundational concepts necessary for working with linear spaces. We have outlined the fundamental properties of linear spaces; we have outlined how to characterize the elements of finite dimensional linear spaces in terms of spanning sets and bases, and we have used the concept of dimension to show that once one specifies both a scalar field and a dimension, that there is only a single structure for a finite dimensional linear space.

Before looking at the applications of finite dimensional linear spaces, we consider one final theoretical construction which will help to motivate our choice of notation and help to explain the conceptual underpinnings of the operation of matrix multiplication, which can seem very confusing and arbitrary when presented in a purely procedural context.

We consider the problem of changing bases in finite dimensional linear spaces. In light of the isomorphism theorem, it is sufficient to only consider the linear space K^n, since any other n dimensional linear space over K is isomorphic to K^n anyway. We will further simplify our lives by restricting ourselves to using the real numbers as our scalar field.

Let us begin by defining $\mathbb{R}^n$ as the following collection of objects:

$$\mathbb{R}^n = \{\langle x_1, x_2, \ldots, x_n\rangle, \Big| \quad x_i \in \mathbb{R}, \quad i = 1, 2, \ldots, n]\}$$

We define the addition operation for these objects as component-wise addition of real numbers, and we define scalar multiplication by a real number, by multiplying all components by the same real scalar. At this point, you should already have proven that this forms a linear space over the field $\mathbb{R}$, so we make the additional claim that this is an finite dimensional linear space over $\mathbb{R}$ with dimension n without explicit proof.

Next we wish to define a shorthand notation for these objects which permits an easy translation from one representation into another representation. To begin we consider the canonical basis given by:

$$\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n\}$$

We define $\mathbf{e}_i$ as an array of n real numbers, with unity in the ith component and zeros in all other components. We then write these arrays vertically, which allows us to look at the components in parallel and helps to save space during hand computation.

Thus if we consider $\mathbb{R}^3$, our 'canonical basis' has three basis vectors given by and notated in the following way:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

This basis gives a very natural representation of the elements of $\mathbb{R}^3$ because the representation of each vector in this basis is exactly the same as the result of vector addition applied to the basis vectors:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix},$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3$$

One cannot underestimate the importance of the calculation above because this provides the cleanest translation from component notation, to the abstract notation of linear combinations of vectors which was introduced and used extensively in the previous sections.

In the abstract linear combination notation, the coefficients x_i are called the **coordinates** of the vector $\mathbf{x}$ with respect to the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. Suppose now that we are given two different bases of $\mathbb{R}^3$ and we wish to translate a representation given in one basis, into a new representation in terms of the other basis. If we wish such a transformation to be unique and well defined, then we must use the canonical basis as an intermediary to connect the two abstract bases to the underlying object. (Without a canonical basis to enforce a unique connection, we could define an infinite collection of possible isomorphisms between these two bases.)

Thus, let us suppose that we have a basis which is different than the canonical basis given by the following vectors: $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3\}$.
Suppose further that we know the representation of each of the $\boldsymbol{\phi}$ basis vectors in terms of the canonical basis. Thus,

$$\boldsymbol{\phi}_j = \phi_{1j}\mathbf{e}_1 + \phi_{2,j}\mathbf{e}_2 + \phi_{3,j}\mathbf{e}_3,$$

with each ϕ_{ij} a known real number.

Now suppose that we are given a representation of a vector $\mathbf{x}$ in terms of the $\boldsymbol{\phi}$ basis, and we wish to change our representation to the canonical basis. To change to the canonical basis we can use the known representations of the $\boldsymbol{\phi}$ vectors in the canonical basis.

$$\mathbf{x} = x_1\boldsymbol{\phi}_1 + x_2\boldsymbol{\phi}_2 + x_3\boldsymbol{\phi}_3$$

$$\mathbf{x} = \sum_{i=1}^{3} x_i\phi_{1i}\mathbf{e}_1 + x_i\phi_{2,i}\mathbf{e}_2 + x_i\phi_{3,i}\mathbf{e}_3$$

Next, since we can translate the canonical basis into it's components we also have:

$$\mathbf{x} = \begin{pmatrix} \sum x_i\phi_{1i} \\ \sum x_i\phi_{2i} \\ \sum x_i\phi_{3i} \end{pmatrix}$$

This equation provides the motivation for the algebraic definition of matrix multiplication.

Note that we have exactly nine quantities which fully specify the $\boldsymbol{\phi}$ basis in terms of the canonical basis. If we organize these quantities into a two dimensional array they become:

$$\begin{pmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \boldsymbol{\phi}_3 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix}$$

And if we write the coordinates of $\mathbf{x}$ in the $\boldsymbol{\phi}$ basis using the same component notation as we adopted for the canonical basis, then we can use the calculation above to motivate the definition of matrix-vector multiplication.

$$\begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \equiv \begin{pmatrix} \sum x_i\phi_{1i} \\ \sum x_i\phi_{2i} \\ \sum x_i\phi_{3i} \end{pmatrix}$$

(**Note:** I have used square brackets around the coordinate components of $\mathbf{x}$ in the $\boldsymbol{\phi}$ basis, to help distinguish between the different representations. To clarify, the meaning of the notation see below:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \equiv x_1\boldsymbol{\phi}_1 + x_2\boldsymbol{\phi}_2 + x_3\boldsymbol{\phi}_3,$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \equiv x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3.$$

As we will see in the chapters that follow, the ability to change representations forms the cornerstone of many techniques and applications of the theory of linear spaces. One typically recognizes linear structure in a given problem, then uses the linear structure to change representations to simplify the problem, then one solves the simplified problem and

reverts to the original representation. We can pose a wide variety of different problems involving representations and changing bases in linear spaces, but we delay these questions until we have studied linear mappings.

Exercises:

(1) Prove that the set $\{\mathbf{e}_1, \mathbf{e}_2, \dots \mathbf{e}_n\}$ is a basis for $\mathbb{R}^n$. (To prove this you must establish two different results: first that the basis vectors span $\mathbb{R}^n$, second the basis vectors are linearly independent.

(2) Prove that $\mathbb{R}^n$ is an n-dimensional linear space. (Remark: If you have proven exercise (1), this result is a trivial corollary, however you can opt to prove this result without proving exercise (1))

(3) Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be the canonical basis for $\mathbb{R}^2$ and define:

$$\boldsymbol{\phi}_1 = \mathbf{e}_1 + \mathbf{e}_2, \quad \boldsymbol{\phi}_2 = \mathbf{e}_1 - \mathbf{e}_2,$$

Using square brackets to denote the components of the $\boldsymbol{\phi}$ representation and parentheses to denote the canonical representation convert the following representations in the $\boldsymbol{\phi}$ basis to the canonical basis.

$$\begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix}, \quad \begin{bmatrix}1\\1\end{bmatrix}, \quad \begin{bmatrix}-2\\2\end{bmatrix}$$

CHAPTER 2

Transformations of Linear Spaces

Mappings between linear spaces are defined as functions which correlate the elements of one linear space, called the **domain** to another linear space called the **co-domain**. Mathematicians often use a shorthand notation for functions which carries over to mappings between linear spaces. This notation describes the mathematical environment surrounding the mapping without actually describing the mapping itself. As such, engineers and physicists may neglect such notation as it plays an organizational role and doesn't play an active computational role in problem solving. For theoretical results and conceptualizing problems however, the notation can be very helpful.

We write:

$$T : \mathcal{X} \to \mathcal{Y}$$

to describe a mapping T which maps or correlates elements from the domain space $\mathcal{X}$ and the co-domain $\mathcal{Y}$. Since we are studying linear spaces, we are particularly interested in linear mappings.

Remark: We have used standard function notation, and assume that the reader is familiar with the use of $T(\mathbf{u})$ to denote the image of $\mathbf{u}$ under the map T, but we remark explicitly that $T(\mathbf{u})$ and $T(\mathbf{v})$ are actually elements of the co-domain, $\mathcal{Y}$, so the vector addition operations on the left and right hand sides of the additivity equation may be different.

1. Fundamentals of Linear Mappings

Definition: *A linear mapping is one which preserves the structure of linear spaces. We say that $T : \mathcal{X} \to \mathcal{Y}$ is a* **linear mapping**. *If T possesses both of the following properties:*

(1) *Additivity: For each $\mathbf{u}$ and $\mathbf{v}$ in the domain, $\mathcal{X}$ we have*

$$T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v}).$$

(2) *Homgeneity: For each $\mathbf{u}$ in the domain, $\mathcal{X}$, and each $k \in K$*

$$T(k\mathbf{u}) = kT(\mathbf{u})$$

Our first result involving linear mappings is a powerful result which shows that the notation we developed in section 1.3 actually gives us a consistent notation which allows us to describe the action of any linear transformation between finite dimensional linear spaces. We prove the result for mappings from $\mathbb{R}^n$ into $\mathbb{R}^m$, but the result applies to all linear mappings between finite dimensional linear spaces over the same field, K.

Matrix Representation Theorem: *Any linear mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ may be represented as an $m \times n$ matrix. Conversely every $m \times n$ matrix is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$.*

Proof:

We proceed by construction. Let $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots \boldsymbol{\phi}_n\}$ be a basis for $\mathbb{R}^n$, and let $\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots \boldsymbol{\psi}_n\}$ be a basis for $\mathbb{R}^m$.

Let T be a well-defined linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Because T is well-defined it maps each $\boldsymbol{\phi}_i$ to a unique image element $T(\boldsymbol{\phi}_i) \in \mathbb{R}^m$. Each of these image elements possesses a representation in terms of the $\boldsymbol{\psi}$ basis. Thus we can write

$$T(\boldsymbol{\phi}_j) = a_{1j}\boldsymbol{\psi}_1 + a_{2j}\boldsymbol{\psi}_2 + \ldots a_{mj}\boldsymbol{\psi}_m$$

Since the $\boldsymbol{\phi}$ form a basis of $\mathcal{R}^n$ we can write any vector $\mathbf{x} \in \mathbb{R}^n$ uniquely in terms of its $\boldsymbol{\phi}$ coordinates. Using this representation and the linear properties of T we find:

$$\begin{aligned} T(\mathbf{x}) &= T(x_1\boldsymbol{\phi}_1 + x_2\boldsymbol{\phi}_2 \cdots + x_n\boldsymbol{\phi}_n) \\ T(\mathbf{x}) &= x_1T(\boldsymbol{\phi}_1) + x_2T(\boldsymbol{\phi}_2) \cdots + x_nT(\boldsymbol{\phi}_n) \end{aligned}$$

Now using the representations of the $T(\boldsymbol{\phi}_i)$ in the $\boldsymbol{\psi}$ basis for $\mathbb{R}^m$ and grouping terms together we obtain:

$$T(\mathbf{x}) = \sum_{j=1}^{n} x_j a_{1j}\boldsymbol{\psi}_1 + \sum_{j=1}^{n} x_j a_{2j}\boldsymbol{\psi}_2 \cdots + \sum_{j=1}^{n} x_j a_{mj}\boldsymbol{\psi}_m$$

Now by taking the view that either, the $\boldsymbol{\psi}$ basis is the canonical basis for $\mathbb{R}^m$, or adopting the shorthand component notation we find that:

$$T(\mathbf{x}) = \begin{pmatrix} \sum_{j=1}^{n} x_j a_{1j} \\ \sum_{j=1}^{n} x_j a_{2j} \\ \vdots \\ \sum_{j=1}^{n} x_j a_{mj} \end{pmatrix}$$

Extracting each x_j from this formula and using the definition of matrix-vector multiplication introduced in section 1.3 we obtain:

$$T(\mathbf{x}) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Because this representation is independent of the choice of $\mathbf{x}$, we can use this notation to define T.

The proof that multiplication by a $n \times m$ matrix is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a consequence of the algebraic properties of our definition of matrix-vector multiplication. We leave the explicit proof of this fact as an exercise.

Q.E.D.

This proof gives us both a notation and a method of performing computations with linear mappings where both the domain and co-domain are finite dimensional vector spaces. Next we turn to the fundamental theory of linear mappings.

The algebraic properties of linear mappings are exactly the properties necessary to preserve the linear structure of the underlying space. Because internal linear structure is preserved, we can expect linear mappings to map subspaces from the domain into subspaces of the range.

Subspace Mapping Lemma:
Let $\mathcal{X}$ and $\mathcal{Z}$ be finite dimensional vector spaces, and let $\mathcal{Y}$ be a subspace of $\mathcal{X}$. If $T : \mathcal{X} \to \mathcal{Z}$ is a linear mapping, then $T(\mathcal{Y})$ is a linear subspace of $\mathcal{Z}$. In addition if, $\mathcal{W}$ is a linear subspace of $\mathcal{Z}$, then the inverse image of $\mathcal{W}$ forms a subspace of the domain $\mathcal{X}$

Proof:
Because T is assumed to be linear, and $\mathcal{X}$ and $\mathcal{Z}$ are both assumed to be linear spaces. We need to show that the linear properties of T force the image of $\mathcal{Y}$ under T to have all of the necessary properties to be a subspace of $\mathcal{Z}$.

To establish closure under the addition operation, consider $\mathbf{u}, \mathbf{v} \in \mathcal{Y}$. Since $\mathcal{Y}$ is a subspace of $\mathcal{X}$, we also have $\mathbf{u} + \mathbf{v} \in \mathcal{Y}$. Thus we find that for each $T(\mathbf{u})$ and $T(\mathbf{v})$ in $T(\mathcal{Y})$ we also have:

$$T(\mathbf{u}) + T(\mathbf{v}) = T(\mathbf{u} + \mathbf{v}) \in T(\mathcal{Y}) \subset \mathcal{Z}$$

Establishing closure under scalar multiplication is similar.

We leave the proof of the inverse image as an exercise.

Q.E.D

Range Corollary: *The image of the whole domain forms a subspace of the co-domain. We call this space the* **range** *of the mapping T. We denote this space with $\mathcal{R}(T)$*

Proof:

The whole domain is a subspace of itself, so the subspace mapping theorem applies.

Q.E.D.

Nullspace Corollary: *The collection of all vectors in the domain which are mapped into the zero vector by a linear map, T form a vector space. This vector space is called the* **nullspace** *and is denoted $\mathcal{N}(T)$.*

Proof:

Because the zero vector is trivially is always a subspace of the codomain, the inverse image of this subspace must be a subspace of the domain by the subspace mapping theorem.

Q.E.D.

Next we turn to one of the most important and applicable result about linear mappings, this result allows us to use the dimension of the of the domain as a bookkeeping quantity which allows us to organize the behavior of linear mappings in a systematic and important way.

Dimension Theorem: *Let $T : \mathcal{X} \to \mathcal{Z}$ be a linear mapping with $\mathcal{X}$ finite dimensional. Then:*

$$\dim(\mathcal{X}) = \dim(\mathcal{R}(T)) + \dim(\mathcal{N}(T))$$

Proof:

By the subspace mapping theorem $\mathcal{N}(T)$ is a linear subspace of $\mathcal{X}$. Since $\mathcal{X}$ is a finite dimensional vector space, $\mathcal{N}(T)$ must be finite dimensional and by the subspace theorem from the previous chapter, every subspace of a finite dimensional vector space has a complementary subspace.

Set $\mathcal{Y} = \mathcal{X} \backslash \mathcal{N}(T)$. By the subspace mapping theorem, this subspace maps into a subspace of $\mathcal{Z}$, and since $T(\mathcal{Y})$ contains everything not mapped to the zero vector this subspace must be $\mathcal{R}(T)$. Using the dimensional arithmetic from the subspace theorem we obtain:

$$\dim(\mathcal{X}) = \dim(\mathcal{Y}) + \dim(\mathcal{N}(T))$$

By establishing $\dim(\mathcal{Y}) = \dim(\mathcal{R}(T))$ we will be done. We claim that $T : \mathcal{Y} \to \mathcal{R}(T)$ generates an isomorphism. We leave it as an exercise to prove that $T : \mathcal{Y} \to \mathcal{R}(T)$ is a bijection, and the isomorphism property holds due to the linearity of T.

Q.E.D.

The importance and power of the preceeding theorem for setting up boundaries and structure for linear transformations cannot be understated. Suppose for example, that one is trying to solve a system of three linear equations in three unknown variables. Such a system can always be represented in terms of a matrix vector equation, and the linear transformation defined by the matrix can be in exactly one of four categories. The matrix could have a three dimensional range and a trivial nullspace, a two dimensional range and a one dimensional null space; a one dimensional range and a two dimensional null space, or a trivial range and a three dimensional null space. These four different classes cover all of the possibilities and each case has different ramifications when one is trying to solve equations involving that linear mapping.

The dimension theorem has several important corollaries regarding the characterization and structure of the nullspace which are useful independently of the main theorem.

Null-space corollaries: *(1) Let $T : \mathcal{X} \to \mathcal{Z}$ be a linear mapping and let $dim(\mathcal{X}) > dim(\mathcal{Z})$, then $dim(\mathcal{N}(T)) > 0$ and there are* **non-zero** *vectors satisfying:*

$$T(\mathbf{x}) = 0.$$

(2) Let $T : \mathcal{X} \to \mathcal{Z}$ be a linear mapping and let $dim(\mathcal{X}) = dim(\mathcal{Z})$, if $T(\mathbf{x}) = 0$ only for the zero vector, then T is an isomorphism between $\mathcal{X}$ and $\mathcal{Z}$.

Proof:
(1) is a direct consequence of the dimension theorem. Since $dim(\mathcal{R}(T)) \leq \dim(\mathcal{Z})$, we conclude that $\dim(\mathcal{N}(T)) > 0$ in order for the dimensional arithmetic to work out.

(2) is proven by noticing that since only $T(\mathbf{\emptyset}) = 0$, the nullspace must have zero dimension. Then using the dimension theorem we conclude that $\mathcal{R}(T) = \mathcal{Z}$.
Q.E.D.

From a practical perspective these corollaries are important because they allow us to derive information about the global behavior of the linear mapping T. In linear algebra, every theorem and corollary eventually hinges on the solution of some system of linear equations, and the results which are posed in terms of hypotheses which can be checked by direct computation are the easiest to implement in practice. If we translate the preceeding two corollaries into matrix, vector notation we find that the hypotheses of (1) may be checked simply by observation, and the hypotheses of (2) may be checked by observation and by **solving** a linear system.

When T is a linear mapping, one can evaluate the truth of: $\dim(\mathcal{X}) = \dim(\mathcal{Z})$ (along with the corresponding inequalities) simply by observing the 'shape' of any matrix representation for T. If T is square with the same number of rows and columns, then the

dimension of the domain and the co-domain are equal. If T is a 'fat' matrix with more columns than rows then the dimension of the domain is greater than the dimension of the co-domain. If T is 'skinny' with more rows than columns, then the domain has a smaller dimension than the co-domain.

In order to determine whether $T(\mathbf{x}) = 0$ has any non-zero solutions we must proceed to solve the simultaneous system of linear equations defined by this matrix vector equation. (**Remark:** For a detailed account of the technique of Gaussian Elimination turn to the Practica Book.)

Exercises:

(1) Prove that the inverse image of any subspace of the co-domain must form a subspace of the domain.

(2) Prove that the nullspace is a subspace of the domain directly without using the subspace mapping theorem.

(3) Prove that any linear mapping defines a bijection between the complement of the null space and the range.

2. Duality, Adjoint and Transpose

In addition to considering transformations which transform vectors into other vectors, we also give special consideration to transformations which transform vectors into scalars.

Definition: *If $\mathcal{X}$ is a linear space over a field K, The* **dual space** *is the collection of all linear, scalar valued functions mapping $\mathcal{X}$ to K.*

The idea of the dual space can be daunting at first, but it allows for a very concise and elegant presentation of some of the more advanced and technical aspects of the subject. Furthermore, the notion is indispensable for obtaining a correct understanding of many aspects of the theory of infinite dimensional linear spaces. Before we proceed to prove any results regarding the dual space we must first introduce and clarify the required notation.

Let $\mathcal{X}$ be a finite dimensional linear space and let $\mathbf{t} : \mathcal{X} \to K$ be a linear mapping. Following standard function notation it is natural to use $\mathbf{t}(\mathbf{x})$ to denote the image of $\mathbf{x}$ under the mapping $\mathbf{t}$. In order to capture the subtle interaction between the dual space, and the original space we also define:

$$(\mathbf{t}, \mathbf{x}) \equiv \mathbf{t}(\mathbf{x}).$$

Using the parenthetic notation emphasizes that elements from the original space and the dual space may actually be considered on equal footing, as two independent arguments for an operation which may be termed a **scalar product.**

Our first result regarding duality will be to prove that any finite dimensional vector space is actually isomorphic to it's dual space, which helps to justify the scalar product notation introduced above.

Duality Theorem: *Let $\mathcal{X}$ be a finite dimensional vector space over K. The dual space, $\mathcal{X}'$ is also a finite dimensional linear space over K, and is isomorphic to $\mathcal{X}$.*

Remark: We note that in infinite dimensional spaces the dual space is generally **not** isomorphic to the original space.

Proof:

Let $\mathbf{t} : \mathcal{X} \to K$ be a linear mapping.

Since $\mathcal{X}$ is finite dimensional, it must have a basis, $\{\boldsymbol{\phi}_i\}$. Using this basis we can represent an arbitrary vector as:

$$\mathbf{x} = x_1\boldsymbol{\phi}_1 + x_2\boldsymbol{\phi}_2 + \cdots + x_n\boldsymbol{\phi}_n.$$

Since $\mathbf{t}$ is linear, for all $\mathbf{x} \in \mathcal{X}$ we can write:

$$\mathbf{t}(\mathbf{x}) = \sum_{i=1}^{n} x_i \mathbf{t}(\boldsymbol{\phi}_i)$$

Thus the action of each linear function $\mathbf{t} \in \mathcal{X}'$ can be fully defined by the n-scalars which act on the chosen basis vectors of $\mathcal{X}$. Furthermore, we can easily show that linear mappings are closed under addition and scalar multiplication by virtue of the algebraic properties of the field, K.

To show that $\mathcal{X}'$ is isomorphic to $\mathcal{X}$ we define the transformation $\Phi_{\boldsymbol{\phi}} : \mathcal{X}' \to \mathcal{X}$ by:

$$\Phi(\mathbf{t}) = \sum_{i=1}^{n} \mathbf{t}(\boldsymbol{\phi}_i)\boldsymbol{\phi}_i$$

We leave it as an exercise to prove that this mapping defines an isomorphism between $\mathcal{X}'$ and $\mathcal{X}$. Q.E.D.

A useful way of thinking of the dual space is by thinking of the elements of the dual space as the set of all possible 'measurements' of the vectors in the underlying space. Depending upon the underlying context different measurements may be more or less meaningful. Suppose for example that a farmer uses a vector with four components to keep track of the amounts of different kinds of grain stored in a barn. Suppose further that the farmer is keeping track of barley, corn, millet and wheat. If the farmer wishes to know how much wheat he has, he simply selects the element of the dual space which will provide the correct measurement. If he wants to know how much feed he has for his animals he selects a more

complex measurement (if, for example, his cows eat both barley and corn and they can survive equally well on either feed, then the farmer might use the measure $\mathbf{t}(\mathbf{x}) = 1\mathbf{b} + 1\mathbf{c}$. If the cows require twice the amount of barley vs. corn to survive, then the measurement might be $\mathbf{t}(\mathbf{x}) = \frac{1}{2}\mathbf{b} + 1\mathbf{c}$.) This perspective on the dual space is especially helpful for correctly understanding the theory of distributions, when one studies infinite dimensional function spaces.

When one is concerned with finite dimensional vector spaces the dual space itself is generally of secondary interest, but the scalar product notation gives us a valuable way of manipulating linear mappings in ways that uncover important technical information.

Definition: *Let $T : \mathcal{X} \to \mathcal{Y}$ be a linear transformation. The* **adjoint** *of T denoted by T'. The adjoint is defined by the following equation:*

$$(\mathbf{y}, T(\mathbf{x}))_{\mathcal{Y}} = (T'(\mathbf{y}), \mathbf{x})_{\mathcal{X}}$$

When working over the field of real numbers the adjoint T' is often called the **transpose**. We leave it as an exercise to verify that the adjoint is obtained by switching the rows and the columns of the matrix representation of T.

Remark: The subscripts on the scalar product notation are included to help distinguish the scalar products in $\mathcal{X}$ and $\mathcal{Y}$ which may be different operations.

Remark: We have introduced two meanings for the superscript $'$. When applied to a linear space, the prime refers to the dual space, and when applied to a linear transformation the prime refers to the adjoint (transpose) operator. If you pay careful attention to context, this overloaded notation should not cause any problems.

Exercises:

(1) Show that the dual space of any finite dimensional linear space is itself a linear space.
(2) Let $\{\boldsymbol{\phi}_i\}_{i=1}^n$ be a basis for the linear space $\mathcal{X}$, and let $\Phi : \mathcal{X}' \to \mathcal{X}$ be defined by:

$$\Phi(\mathbf{t}) = \sum_{i=1}^{n} \mathbf{t}(\boldsymbol{\phi}_i)\boldsymbol{\phi}_i$$

verify that Φ defines an isomorphism between $\mathcal{X}'$ and $\mathcal{X}$. (Hint: Select a basis for $\mathcal{X}'$ which makes it easy to compute the dimension of $\mathcal{X}'$, then use the isomorphism theorem)
(3) Use the definition of the scalar product and the definition of matrix multiplication to show by direct computation that if the elements of T are a_{ij}, that the elements of T' are a_{ji}.

3. Forward and Backward Problems for Linear Mappings

In this section we outline the fundamental computational problems of linear algebra, these problems reappear in every application and mastery of both their computational and theoretical foundations is necessary before one can profess any level of competence over the subject.

Forward Problem: *Given any linear transformation $T : \mathcal{X} \to \mathcal{Y}$ and a particular input vector $\mathbf{x} \in \mathcal{X}$ find a representation of the output vector.*

By specifying a basis for both the domain and codomain, we can construct a matrix representation for the linear transformation T, and by using our definition for matrix, vector multiplication we can readily compute the output vector.

Backward Problem: *Given any linear transformation $T : \mathcal{X} \to \mathcal{Y}$ and a particular* **output** *vector $\mathbf{y} \in \mathcal{Y}$ find all vectors, $\mathbf{x} \in \mathcal{X}$ satisfying:*

$$T\mathbf{x} = \mathbf{y}.$$

While the forward problem is mathematically natural, in practice the backward problem is usually the relevant one. Conceptually, the forward problem is solved simply by seeing what happens and by knowing the transformation we can uncover the correct output, but the backward problem is solved by finding inputs which correspond to the desired output.

4. Existence and Uniqueness Properties of Linear Mappings

In 1902 Jacques Hadamard provided a definition for a well-posed mathematical problem. Hadamard's criteria for a well posed problem has three distinct branches:

- The question of existence.
- The question of uniqueness.
- The question of stability.

By themselves Hadamard's criteria are not necessary or sufficient to deduce whether a mathematical problem is interesting, useful, or important, but they provide a valuable theoretical framework from which can help to guide a mathematical approach almost any applied problem. At the most practical level one wants to find answers to mathematical problems, but without any theoretical structure one could search in vain for a problem which has no solution, furthermore whenever one approaches problems connected with the physical world both approximation errors and measurement errors are the rule rather than the exception. Understanding how errors propagate and transform through mathematical computation is very important and should always be a concern when deciding how much trust to place in your mathematical computations. We will study these questions in detail for both the forward problem and backward problem.

The first criterion of 'well-posedness' gives rise to the question of existence. The existence question simply asks, 'Is it possible to select a vector which makes the matrix-vector equation true?' You should note that the question of existence is one of possibility. One can answer this question in the affirmative by finding a solution, but in many cases it is possible to show that a solution exists without actually finding it. (Such results are important when one tries to solve a problem using computational aids. Many algorithms will happily search forever looking for non-existentent solutions, but if we're waiting for the

results, then it helps to **know** that the algorithm actually has a solution to find.)

The second criterion of 'well-posedness' gives rise to the question of uniqueness. This is a follow-up question to the existence question and is also important for applications. If we know that a problem has a unique solution, then if both you and I try to find the solution, we know our solutions must agree. If a problem has multiple solutions, then it is possible that different techniques may find different solutions.

The final criterion for 'well-posedness' gives rise to the question of stability. This question is of paramount importance when one wants to approximate the solution to a problem, and is studied deeply in the field of numerical analysis. We will delay the answer to this question until chapter 4 when we have discuss the idea of analysis in normed linear spaces.

When studying the forward problem, the questions of existence and uniqueness have very simple answers. Whenever matrix-vector multiplication is well defined (i.e. the matrix representation of T is the correct size matrix), the forward problem always possesses a unique solution.

When studying the backward problem, the questions of existence and uniqueness have more subtle, nuanced answers.

For clarity, we pose these questions explicitly:

ExQ: *What conditions on the matrix, T, and the right hand side, $\mathbf{y}$, will* **guarantee** *the existence of a vector $\mathbf{x}$ satisfying the backward problem:*

$$T\mathbf{x} = \mathbf{y}?$$

UniQ: *What conditions on the matrix, T, and the right hand side, $\mathbf{y}$, will* **guarantee** *that a solution (when it exists) to the backward problem is unique?*

We have now set out on a rather ambitious program, that of completely understanding the solvability and 'well-posedness' of the backward problem for linear operators acting on finite dimensional vector spaces. Using the concepts we have already defined we can answer the uniqueness question quite easily.

Backward Uniqueness Theorem: *Let $T : \mathcal{X} \to \mathcal{Y}$ be a linear mapping, and assume the linear system:*

$$T\mathbf{x} = \mathbf{y}$$

has a solution, $\mathbf{x}$.

Then, $\mathbf{x}$ *is unique if and only if* $\mathbf{z} = \emptyset$ *is the* **only** *solution to:*

$$T\mathbf{z} = \emptyset.$$

Proof:
We proceed by interpreting the matrix-vector equation in terms of the abstract linear combination notation, and using our knowledge of representations in linear spaces.

Using the definition of matrix multiplication we can partially multiply the product $T\mathbf{x}$ to obtain:

$$T\mathbf{x} = \begin{pmatrix} \vdots & \vdots & \dots & \vdots \\ T_{i1} & T_{i2} & \dots & T_{im} \\ \vdots & \vdots & \dots & \vdots \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$T\mathbf{x} = x_1 \begin{pmatrix} \vdots \\ T_{i1} \\ \vdots \end{pmatrix} + x_2 \begin{pmatrix} \vdots \\ T_{i2} \\ \vdots \end{pmatrix} \cdots + x_m \begin{pmatrix} \vdots \\ T_{im} \\ \vdots \end{pmatrix}$$

$$T\mathbf{x} = x_1\mathbf{T}_1 + x_2\mathbf{T}_2 \cdots + x_m\mathbf{T}_m$$

Since $\mathbf{x}$ is a solution to the linear system $T\mathbf{x} = \mathbf{y}$ we have:

$$x_1\mathbf{T}_1 + x_2\mathbf{T}_2 \cdots + x_m\mathbf{T}_m = \mathbf{y}$$

This implies that $\mathbf{y} \in Sp(\{\mathbf{T}_j\})$. This representation is unique if the columns, $\mathbf{T}_j$ are linearly independent. The condition that only the zero vector solves $T\mathbf{z} = \emptyset$ is equivalent to the columns of T being linearly independent.

Q.E.D.

This answers the uniqueness question and again provides a connection between the abstract formulation of linear spaces and a concrete answer to a theoretical question. In order to answer the existence question we must introduce one additional concept.

Definition: *Two vectors are called* **orthogonal** *if the scalar product between those two vectors is zero.*
Using the concept of a dual space and the interpretation of elements of the dual space as 'measurements' two vectors which are orthogonal are vectors which are completely insensitive to each other's measurement. Geometrically, orthogonal vectors are mutually perpendicular (in two and three dimensions this notion is intuitive, but the result generalizes to any finite dimensional vector space if the angle is measured with respect to the plane spanned by the two vectors under consideration.)

Orthogonal Complement Lemma: *Let $\mathcal{X}$ be a finite dimensional linear space, and let $\{\mathbf{w}_i\}$ be a given set of vectors. The set of all vectors orthogonal to this collection of vectors forms a subspace of $\mathcal{X}$.*

Proof:
Define $W = \{\mathbf{w}_i\}_{i=1}^n \subset \mathcal{X}$. We define the **orthogonal complement** to W as:

$$W^c = \left\{ \mathbf{x} \in \mathcal{X} \middle| \ \ (\mathbf{x}, \mathbf{w}_i) = 0, \ \ \forall i \in [1, \dots n \right\}$$

Writing the defining constraints for W^c we find:

$$\begin{pmatrix} \dots & \mathbf{w}_1 & \dots \\ \dots & \mathbf{w}_2 & \dots \\ \dots & \mathbf{w}_n & \dots \end{pmatrix} \mathbf{x} = \mathbf{\emptyset}$$

thus the orthogonal complement to the set W is the null space of a particular linear transformation, and by the nullspace corollary these must form a linear vector space.

Q.E.D.

Next we present the exact answer to the existence question for finite dimensional linear transformations. This theorem is quite powerful, and when posed in the language below possesses a natural extension to infinite dimensional vector spaces. For practical purposes its important to note that the proof of this theorem is not connected with any practical applications of the result itself, so understanding the proof is a luxury rather than a necessity.

Backward Existence Theorem: (Finite Dimensional Fredholm Alternative)
The equation $T\mathbf{x}$=$\mathbf{y}$ has a solution if and only if the right-hand side, $\mathbf{y}$ is orthogonal to some basis for the null space of the adjoint operator T'.

$$(\mathbf{y}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathfrak{N}(T')$$

Proof: To prove this theorem, we begin by assuming that $T\mathbf{x} = \mathbf{y}$ has a solution, we assume $\mathbf{v} \in \mathcal{N}(T')$ then we compute:

$$\begin{aligned} T\mathbf{x} &= \mathbf{y} \\ (\mathbf{v}, T\mathbf{x}) &= (\mathbf{v}, \mathbf{y}), \\ (T'\mathbf{v}, \mathbf{x}) &= (\mathbf{v}, \mathbf{y}), \\ (0, \mathbf{x}) &= (\mathbf{v}, \mathbf{y}), \\ 0 &= (\mathbf{v}, \mathbf{y}). \end{aligned}$$

This proves that $\mathbf{y}$ is orthogonal to the null-space of T', whenever $T\mathbf{x} = \mathbf{y}$ has a solution.

To prove that orthogonality actually implies the existence of a solution we decompose $\mathbf{y}$ into two parts:

$$\mathbf{y} = \mathbf{y}_r + \mathbf{y}_o.$$

Since the range of T is a subspace of the codomain, and since the codomain is finite dimensional we can do this uniquely using the subspace theorem.

Here, $\mathbf{y}_r$ lies in the range of T, and $\mathbf{y}_o$ is the part of $\mathbf{y}$ orthogonal to the range. By showing $\mathbf{y}_o = 0$, we will have shown that $\mathbf{y} = \mathbf{y}_r$ which implies that $\mathbf{y}$ is in the range of T, and $T\mathbf{x} = \mathbf{y}$ necessarily has a solution.
Since $\mathbf{y}_o$ is orthogonal to the range of A:

$$(\mathbf{y}_o, T\mathbf{x}) = 0$$

Utilizing the definition of the adjoint we find:

$$(T'\mathbf{y}_o, \mathbf{x}) = 0 \quad \forall \mathbf{x}$$

This forces $T'\mathbf{y}_o = 0$. This implies that $\mathbf{y}_o$ lies in the null space of T', and that $(\mathbf{y}, \mathbf{y}_o) = 0$. Using the properties of the scalar product notation

$$0 = (\mathbf{y}, \mathbf{y}_o) = (\mathbf{y}_r + \mathbf{y}_o, \mathbf{y}_o) = (\mathbf{y}_r, \mathbf{y}_o) + (\mathbf{y}_o, \mathbf{y}_o)$$

Since $\mathbf{y}_r$ and $\mathbf{y}_o$ are an orthogonal decomposition we are left with:

$$0 = (\mathbf{y}_o, \mathbf{y}_o)$$

the only vector which is insensitive to its own measurement is the zero vector. Thus $\mathbf{y} = \mathbf{y}_r$ and $\mathbf{y}$ must live entirely in the range of T and thus $T\mathbf{x} = \mathbf{y}$ has a solution.

Q.E.D.

Remark: As stated in the preamble to the backward existence theorem, the proof of this theorem is not part of the implementation of the result. The existence theorem allows us to characterize necessary conditions for the existence of solutions to linear systems. In the case of large linear systems, checking these conditions can be much faster than proceeding with a brute force computational approach, only to find that the problem has no exact solution.

Exercises:

(1) Prove the backward uniqueness theorem using the following scheme.

   (a) Assume that $\mathbf{x}$ is a solution to $T\mathbf{x} = \mathbf{y}$ and $\mathbf{z}$ is a non-zero solution to $T\mathbf{z} = \mathbf{\emptyset}$. Show that $\mathbf{x} + \mathbf{z}$ is a second solution to the backward problem.
   (b) Assume that $\mathbf{x}_1$ and $\mathbf{x}_2$ are two **different** solutions to $T\mathbf{x} = \mathbf{y}$. Show that by combining these solutions appropriately one can construct a non-zero solution to $T\mathbf{z} = \mathbf{\emptyset}$.

(2) Consider the following linear system:

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

(a) Express this linear system in terms of two simultaneous linear equations and use your intuition to determine the relationship between y_1 and y_2 which **must** hold for this system to have a solution.

(b) Determine the algebraic form of the condition required by the backward existence theorem (i.e. $\mathbf{y}$ must be orthogonal to the null-space of the adjoint) and show that it is equivalent to the intuitive condition from (a).

(3) Consider the following inverse problem:
Let $\mathcal{X}$ and $\mathcal{Y}$ be finite dimensional linear spaces. Given a collection of input and output pairs from these spaces $\{[\mathbf{x}_i, \mathbf{y}_i]\}$, find a linear mapping $T : \mathcal{X} \to \mathcal{Y}$ which satisfies:

$$T\mathbf{x}_i = \mathbf{y}_i, \text{ for each } i.$$

Show that by organizing the entries of the matrix, T into a vector, one can rewrite this inversion problem in terms of a backward problem where the entries of the matrix are the unknowns. (Note: There are many ways to do this.)

CHAPTER 3

Introduction to Spectral Theory

The solvability theory outlined in the previous chapter gives a complete characterization of the existence and uniqueness of solutions to linear systems, however this theory doesn't lend any insight into the actual transformations which are possible under linear mappings.
The current chapter outlines the fundamentals of spectral theory for linear mappings from $\mathbb{R}^n$ into itself. Spectral theory for linear mappings is concerned with selecting a basis for the domain and range where the effects of the linear transformation are particularly simple and transparent. Using such a representation can lead to immense computational and conceptual simplification, and can greatly increase the tractability of complex linear problems. Before we begin with the study of spectral theory proper we consider a simple explicit example to illustrate the ideas. Suppose we consider a linear transformation in three dimensions whose effect on the canonical basis is given by the following matrix representation:

$$T_{\mathbf{e}} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ -2 & 0 & 2 \end{pmatrix}$$

This looks like a very complicated transformation, but if we change bases from the canonical basis to a different basis in both the domain and range, the representation of this mapping can become:

$$T_{\boldsymbol{\phi}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

If we wish to solve linear systems involving the transformation T in its canonical representation, we must use some kind of algebraic technique to reduce the full matrix to a more tractable form (either gaussian elimination, matrix inversion, or another matrix decomposition technique). However, if we change bases to the more convenient representation involving the $\boldsymbol{\phi}$ basis, solving the linear system is trivial and requires no algebraic work.

As we proceed we will see that uncovering the correct transformation and basis to simplify the representation of T requires a significant amount of work, however in many situations the advantages outweigh the drawbacks, and there are certain situations where we cannot proceed without first making this transformation.

In addition to the computational advantages provided by a diagonal, or almost diagonal matrix the simplified representation gives us a way to decompose and understand the different facets of the linear transformation.

1. The Eigenvalue Problem

The central mathematical problem of spectral theory is called the eigenvalue problem. Given a linear transformation T, the eigenvalue problem is posed as follows:

Eigenvalue Problem: *Find all scalars λ and all vectors $\mathbf{x}$ which satisfy the following matrix vector equation.*

$$T\mathbf{x} = \lambda\mathbf{x}$$

When they exist the scalars, λ, are called the eigenvalues of T, and the corresponding $\mathbf{x}$'s are called the eigenvectors of T. When one studies infinite dimensional spaces where the underlying elements are functions rather than vectors, the $\mathbf{x}$ are unimaginatively called eigenfunctions.

From the perspective of analysis the eigenvalue problem is important because it isolates those vectors for which the linear transformation has a particularly simple effect. If one can construct a basis purely out of the eigenvalues of a given matrix, and use that basis to represent both the domain and co domain, then the action of the matrix can be purely characterized by the scaling effects on each of the basis vectors. We will prove this powerful result below, but first we consider how one might go about solving the eigenvalue problem.

Approaching the eigenvalue equation as if it were a single algebraic equation its natural to try to group all of the instances of the unknown, $\mathbf{x}$ together on one side of the equation.

$$T\mathbf{x} - \lambda\mathbf{x} = \emptyset.$$

Noting that this equation involves a vector difference, we cannot factor $\mathbf{x}$ out without first inserting a placeholder which allows us to make sense of the difference between the **matrix**, A and the **scalar**, λ. When working with matrix-vector equations the identity matrix, I, plays the same role as multiplication by 1. By noting that:

$$\lambda I\mathbf{x} = \lambda\mathbf{x},$$

for any choice of scalar and any choice of vector. We substitute into the grouped eigenvalue equation and factor out the vector $\mathbf{x}$ to obtain:

$$(T - \lambda I)\,\mathbf{x} = 0,$$

We interpret this equation as saying that $\mathbf{x}$ lies in the null space of the matrix $T - \lambda I$. The null-space of a square matrix only contains non-zero vectors when the matrix is singular and the columns are linearly dependent. By computing the determinant of the matrix $T - \lambda I$ we obtain a polynomial equation which must be satisfied by the eigenvalues.

$$\det(T - \lambda I) = 0$$

For an $n \times n$ matrix, T this will be an n-th degree polynomial equation with up to n roots over the real numbers. To complete the solution of the eigenvalue problem, one takes each eigenvalue in turn and solves the linear system:

$$(T - \lambda_i I)\boldsymbol{\phi}_i = \boldsymbol{\emptyset}$$

to determine the eigenvectors. When solving this problem, one is interested in finding representative, **non-zero** eigenvectors for each eigenvalue, the zero solution is not useful for problem solving purposes.

For finite dimensional vector spaces the solutions to the eigenvalue problem are completely specified by the polynomial equation, $\det(T - \lambda I) = 0$. The roots of this equation, sometimes called the characteristic equation, completely characterize the possible matrix representations of the linear operator.

2. Diagonalization using the Eigenbasis

When a $n \times n$ matrix possesses a full set of n linearly independent eigenvectors, these eigenvectors can be used in combination to create a basis where the matrix will have a diagonal representation. This important result is outlined in the following theorem.

Diagonalization Theorem
If there are n linearly independent eigenvectors each of which solves the eigenvalue problem for an $n \times n$ matrix, A, then the following conditions hold:

(1) *The eigenvectors form a basis for $\mathbb{R}^n$.*
(2) *The matrix A is a diagonal matrix, when represented in terms of the eigenbasis.*
(3) *The entries along the diagonal are the corresponding eigenvalues of the matrix, A.*

Proof:

Since $\mathbb{R}^n$ is n-dimensional, any set of n linearly independent vectors may form a basis. This proves 1.

For the second statement we consider the problem:

$$A\mathbf{x} = \mathbf{f}$$

We let Φ be a matrix whose columns are the eigenvectors of the matrix A. To find the coordinates of $\mathbf{f}$ in the Φ basis we left multiply both sides of the equation by Φ^{-1}.

$$\Phi^{-1}A\mathbf{x} = \Phi^{-1}\mathbf{f}.$$

To find the coordinates of $\mathbf{x}$ in the Φ basis, we must insert the product $\Phi\Phi^{-1}$ between A and $\mathbf{x}$. This yields the new representation of the original linear system in the Φ basis.

$$\Phi^{-1}A\Phi\Phi^{-1}\mathbf{x} = \Phi^{-1}\mathbf{f}.$$

We define: $\mathbf{y} = \Phi^1\mathbf{x}$, and $\mathbf{g} = \Phi^{-1}\mathbf{f}$. these definitions yield:

$$\Phi^{-1}A\Phi\mathbf{y} = \mathbf{g}$$

We now look explicitly at the matrix combination $A\Phi$. (In the computations below I have specialized to a 3×3 system, but the computations are analogous for larger matricies). By definition this is:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} \left[\begin{pmatrix} . \\ \phi_1 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_2 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_3 \\ . \end{pmatrix} \right]$$

Since each of the ϕ_i is a solution to the eigenvalue problem, the result of this matrix multiplication will be:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} \left[\begin{pmatrix} . \\ \phi_1 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_2 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_3 \\ . \end{pmatrix} \right] = \left[\begin{pmatrix} . \\ \lambda_1\phi_1 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \lambda_2\phi_2 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \lambda_3\phi_3 \\ . \end{pmatrix} \right].$$

We can rewrite the righthand side of this equation as:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} \left[\begin{pmatrix} . \\ \phi_1 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_2 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_3 \\ . \end{pmatrix} \right] = \left[\begin{pmatrix} . \\ \phi_1 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_2 \\ . \end{pmatrix} \quad \begin{pmatrix} . \\ \phi_3 \\ . \end{pmatrix} \right] \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Defining the diagonal matrix of eigenvalues as Λ, we return to compact notation to find:

$$A\Phi = \Phi\Lambda$$
$$\Phi^{-1}A\Phi = \Phi^{-1}\Phi\Lambda$$
$$\Phi^{-1}A\Phi = \Lambda$$

Thus the new representation of the matrix A in the Φ basis is given by a diagonal matrix with the eigenvalues along the diagonal. This completes the proof of parts 2 and 3.

Q.E.D.

When a matrix does not possess a full set of eigenvectors, either because the characteristic polynomial possesses complex roots, or because the matrix has repeated real roots, without a corresponding increase in eigenvectors it is not possible to fully diagonalize the matrix. By using generalized eigenvectors, one can still select a basis which has a simplified representation in the new basis. One typically selects a augmented eigenbasis in the following way. First one chooses all linearly independent eigenvectors, then one augments these eigenvectors by selecting only enough vectors to create a basis for the domain. Simple representations may be obtained if one selects additional eigenvectors for any repeated eigenvalues by solving the linear system:

$$(A - \lambda_i I)\mathbf{g} = \boldsymbol{\phi}_i$$

Where λ_i is the repeated eigenvalue and $\boldsymbol{\phi}_i$ the corresponding eigenvector.

2.1. Spectral Theory for Symmetric Matrices. The characteristic polynomial of an arbitrary matrix may have any number of roots, however when we impose symmetry on the matrix we obtain some very powerful results regarding the type and configuration of both the eigenvalues and the eigenvectors. In addition symmetric matrices appear often in physical applications, so these results have a practical use as well as theoretical interest.

Spectra for Symmetric Matrices:
If a matrix, A is symmetric (and self-adjoint), then the following statements are all true:

(1) *the scalar product, $(A\mathbf{x}, \mathbf{x})$ is real-valued for all vectors $\mathbf{x}$.*
(2) *All of the eigenvalues are real-valued.*
(3) *Eigenvectors of distinct eigenvalues must be mutually orthogonal.*
(4) *The eigenvectors form an orthonormal basis for $\mathbb{R}^n$.*
(5) *The matrix A can be diagonalized.*

The proofs of (1) and (2) are quite simple and depend upon the definition of the linearity properties for complex inner products.

To prove (3) we consider two *distinct* eigenpairs for the matrix, A:

$$\lambda, \mathbf{x} \quad \text{and} \quad \mu, \mathbf{y}$$

Follow the chain of computation from left to right, using the properties of eigenvectors and the symmetry of A along the way:

$$\lambda(\mathbf{x}, \mathbf{y}) = (\lambda\mathbf{x}, \mathbf{y}) = (A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}) = (\mathbf{x}, \mu\mathbf{y}) = \mu(\mathbf{x}, \mathbf{y})$$

Since we have maintained equality in each step of the calculation, the quantity on the far left and the far right are equal. Taking the difference of these two quantities yields:

$$(\lambda - \mu)(\mathbf{x} \cdot \mathbf{y}) = 0$$

And thus if $\lambda \neq \mu$ we find that $\mathbf{x}$, and $\mathbf{y}$ must be orthogonal.

The proof of (4) is somewhat lengthy and depends upon dimensional calculations to determine that any repeated eigenvalues have a sufficient number of eigenvectors. (Since the eigenvalues are all real, and distinct eigenvectors are orthogonal, we must only verify that repeated eigenvalues have enough linearly independent eigenvectors to complete the basis.) Part (5) simply follows for our previous work relating eigenvectors to diagonalization.
Q.E.D.

The spectral properties for symmetric matrices tell us that any symmetric matrix possesses a natural coordinate system where the matrix is diagonal and only acts by stretching or compressing in the eigen-directions. We will appeal to this result later in our study of linear regression, and when we use singular valued decompositions to analyze the action of rectangular matrices.

CHAPTER 4

Analysis on Linear Spaces

In our discussion of linear spaces thus far we have defined the structural characteristics of the spaces themselves, and given a characterization of transformations between linear spaces which preserve these characteristics. In this chapter we introduce the tools necessary to analyze the elements of linear spaces using externally imposed measures. In this chapter we will restrict our attention to the field of real numbers.

Definition: *The* **norm** *defines the magnitude or length of a vector. The norm operation takes elements from a linear space and maps them to real numbers. A norm must satisfy all of the following algebraic properties:*

(1) Homogeneity:

$$\|\alpha\mathbf{x}\| = |\alpha| \, \|\mathbf{x}\| , \quad \forall\alpha \in \mathbb{R}$$

(2) Strong Positivity:

$$\|\mathbf{x}\| \geq 0, \quad \|\mathbf{x}\| = 0 \rightarrow \mathbf{x} = \emptyset.$$

(3) Triangle inequality:

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| .$$

1. Inner Products and Induced Norms

We first recall the scalar product notation defined in chapter 2 section 2. The scalar product, also called an inner product, often provides the foundation for analysis in linear spaces. We define the scalar product of two vectors in the same finite dimensional linear space as the sum of the component-wise products of each vector.

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i$$

In finite dimensional vector spaces the scalar product is well defined, and since the dual space is identified with the underlying linear space itself the scalar product induces a natural internal measure of length on the vectors within the space. In infinite dimensional spaces the scalar product may or may not be defined.
When the scalar product is well defined we define the norm induced by the inner product as:

$$\|\mathbf{x}\| \equiv (\mathbf{x}, \mathbf{x})^{\frac{1}{2}},$$

In finite dimensional vector spaces this choice of norm is called the **euclidean** norm and is given by:

$$\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}.$$

In two and three spatial dimensions the euclidean norm corresponds to euclidean length (hence the name), but generalizes easily to any finite number of dimensions.

The norm is an important measurement for analysis in linear spaces because once the norm is selected, the norms of any two vectors can be computed and compared to give an idea of their relative size. The adoption of a norm turns a linear space which only has an internal algebraic structure, into a metric space where one can conduct analysis and look at properties such as continuity and convergence. It is only through the imposition of a metric structure that we can apply the ideas of calculus to the elements of linear spaces.

2. Errors Under Linear Transformations

In order to apply the ideas of analysis to linear transformations, we first consider the propagation of error in the forward and backward problems of linear algebra. We first consider the forward problem.

Suppose the vector $\mathbf{x}^*$ is composed of two parts, a true part and and a erroneous part. We will notate these parts as:

$$\mathbf{x}^* = \mathbf{x} + \boldsymbol{\epsilon}$$

Using the additivity of linear mappings we know:

$$T\mathbf{x}^* = T\mathbf{x} + T\boldsymbol{\epsilon}.$$

Thus, if we knew exactly what the errors were we could use the linear structure to isolate and remove them. If, on the other hand, we did not know the error exactly, but only had a bound on the norm of the error, then we could not remove the error, but could only estimate the change in the norm of the error. In order to be able to study error propagation through linear mappings, we introduce the concept of a matrix norm which quantifies the maximum length of any output vector for all possible input vectors of unit length.

$$\|T\| = \max_{\|\mathbf{x}\|=1} \|T\mathbf{x}\|$$

Alternately, this norm can be defined as:

$$\|T\| = \max_{\mathbf{x}} \frac{\|T\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Remark: If A is not a square matrix, the norm operations in the numerator and denominator may not be the same.

The matrix norm allows us to quantify how errors propagate through linear mappings. If we consider the following forward problem, with T and $\mathbf{x}^*$ given then we have:

$$\begin{aligned} T\mathbf{x}^* &= \mathbf{y}^* \\ T\mathbf{x} + T\boldsymbol{\epsilon} &= \mathbf{y}_\mathbf{x} + \mathbf{y}_\epsilon \end{aligned}$$

Using the matrix norm we have:

$$\begin{aligned} \|\mathbf{y}_\epsilon\| &= \|T\boldsymbol{\epsilon}\| \\ \|\mathbf{y}_\epsilon\| &\leq \|T\| \, \|\boldsymbol{\epsilon}\| \end{aligned}$$

Thus the new error is proportional to the original error, with the matrix norm providing a bound on the constant of proportionality. This gives us an absolute bound on the error norm for the forward problem, which may be important in certain cases, however if we wish to use $\mathbf{y}^*$ as an estimate of the desired output $\mathbf{y}_\mathbf{x}$, then we are more interested in the relative error: that is how much of the value of $\mathbf{y}^*$ is due to error, and how much is due to the real value?

Calculating the relative error is more subtle than calculating the absolute error, since not only are we concerned with the possibility that the error may grow in size, we must also be concerned with the possibility that the true output may shrink in size.

The relative error in the output values is given by:

$$E_{out} = \frac{\|\mathbf{y}_\epsilon\|}{\|\mathbf{y}^*\|}$$

Suppose that the relative error of the inputs is known:

$$E_{in} = \frac{\|\boldsymbol{\epsilon}\|}{\|\mathbf{x}^*\|}$$

Using the linear equations defining $\mathbf{y}_\epsilon$, we have:

$$\|\mathbf{y}_\epsilon\| \leq \|T\| \, \|\boldsymbol{\epsilon}\|$$

In order to bound the relative error of the output, we must find a lower bound for the norm of $\mathbf{y}^*$. We can achieve this by assuming that T is invertible and using the matrix norm:

$$\begin{aligned} T^{-1}\mathbf{y}^* &= \mathbf{x}^* \\ \left\|T^{-1}\right\| \|\mathbf{y}^*\| &\geq \|\mathbf{x}^*\| \\ \|\mathbf{y}^*\| &\geq \frac{1}{\|T^{-1}\|} \|\mathbf{x}^*\| \end{aligned}$$

Combining these two bounds we obtain:

$$\frac{\|\mathbf{y}_\epsilon\|}{\|\mathbf{y}^*\|} \leq \|T\| \left\|T^{-1}\right\| \frac{\|\boldsymbol{\epsilon}\|}{\|\mathbf{x}^*\|}$$
$$E_{out} \leq \kappa(T) E_{in}$$

where $\kappa(T) \equiv \|T\| \left\|T^{-1}\right\|$ is called the **condition number** of T. This number is important in numerical analysis and gives a general measure of how well behaved the matrix will be under numerical manipulation. This result is important because it gives one a bound on the size of relative error which can propagate through a single linear transformation. Note that a singular matrix has a condition number of ∞. This is due to the possibility of having te real value in the null space and the error in the range. No matter what the original relative error, the resulting relative error is one, meaning the result is 100% error.

A numerical example may help clarify the issue. Suppose that you obtained measurement data with 1% relative error and you performed a linear operation with a matrix having a condition number of 10, then 10% of the norm of the resulting data **could** be due to error. This bound is an absolute worst case scenario, but it is possible. A great deal of statistical theory is concerned with describing and predicting output when the errors have assumed distributional characteristics.

In the case of symmetric matrices, the matrix norm and the condition number take on particularly simple forms.

Theorem (Maximum Principle for symmetric matricies)
Define $q(\mathbf{x}) = [A\mathbf{x}, \mathbf{x}]$*, for some symmetric matrix* A*. Then:*

(1) *Let* Λ *be the set of eigenvalues for* A*, then:* $\max_{i\in\Lambda} \lambda_i = \max_{||\mathbf{x}||\leq 1} q(\mathbf{x}) = q(\mathbf{x}_i)$*, where* λ_i *is the largest eigenvalue, and* $\mathbf{x}_i$ *is the corresponding eigenvector.*

(2) *Furthermore, if we consider* $\max_{||\mathbf{x}||\leq 1,(*)} q(\mathbf{x})$*, and we define* $(*)$ *by the following orthogonality criterion:*

$$(*) \quad [\mathbf{x}, \mathbf{x}_j] = 0 \quad \forall 1 \leq j \leq k-1,$$

Then we find:

$$\max_{||\mathbf{x}||\leq 1,(*)} q(\mathbf{x}) = q(\mathbf{x}_k)$$

Where $\mathbf{x}_k$ *is the* k*-th eigenvector.*

Corollary: *The matrix norm for a symmetric matrix induced by the euclidean norm is simply the absolute value of the largest eigenvalue, while the condition number is the absolute value of the maximum eigenvalue divided by the minimum eigenvalue.*

Corollary: *If we consider a sequence of matrices* A_k *with* $det(A_k) \to 0$*, then* $cond.(A_k) \to \infty$*. The condition number of a singular matrix is taken to be* ∞*.*

This maximum principle is powerful because it allows us to bound the action of our operator independent of the argument.

3. Approximation in Normed Linear Spaces

Once a norm has been adopted, one can use the norm to make objective measurements of errors. Such a measurement becomes very important in situations where experimental data precludes the existence of an exact solution and an approximation must be made. When approximation is necessary, one naturally wonders how to make the approximation in such a way that errors are minimized.

3.1. Vector Projection. As a first foray into approximation within finite dimensional linear spaces we consider the following problem:

Given a vector $\mathbf{v} \in \mathbb{R}^n$, what is the best (closest) approximation to $\mathbf{v}$ we can make using a single basis vector $\mathbf{u} \in \mathbb{R}^n$?

We can pose this question in terms of linear spaces. Which vector in the span of the vector $\mathbf{u}$ lies nearest to the vector $\mathbf{v}$?

Before we attempt to solve this problem we must first formalize what we mean by 'nearest.' In the work that follows the distance between two vectors will be defined by the norm of the difference of both vectors so:

$$d(\mathbf{u}, \mathbf{v}) \equiv \|\mathbf{u} - \mathbf{v}\|$$

Once we have made this definition, we can unambiguously interpret the idea of 'nearness' in terms of the value of this distance function. The nearest vector in the span of $\mathbf{u}$ will be the vector which minimizes the value of the distance function. Since we will typically work with the euclidean norm, this notion agrees with our normal geometric notions of distance in two and three dimensional space.

Once we have specified $\mathbf{u}$, $\mathbf{v}$ and the distance function, we can approach this problem from multiple perspectives. We will first solve the problem in terms of elementary single variable calculus, then we will re-interpret our solution in terms of concepts from linear algebra.

To proceed we assume that $\mathbf{v}$, and $\mathbf{u}$ are given and fixed. We also assume that the distance function uses the regular euclidean norm. Then to proceed, we only need to parameterize our space of possible approximations and find the one which minimizes the distance.

Since we are only considering vectors in the span of $\mathbf{u}$ we can parameterize our space with a single scalar variable, say t.

$$Sp(\{\mathbf{u}\}) = \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; \mathbf{x} = t\mathbf{u} \right\}$$

Now the approximation problem can be posed as an optimization problem. Select the value of t which minimizes the distance function:

$$f(t) = d(t\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{n} (tu_i - v_i)^2}$$

Recalling our work with single variable calculus, we know that this function will only take on a local minimum value when the first derivative is identically zero. So finding the best approximation is reduced to finding the value of t for which the derivative is equal to zero. Since the square root function is monotonic, we can even work with the easier squared distance function without losing any information. So now we must solve the following problem:

$$\frac{d}{dt} d^2(t\mathbf{u}, \mathbf{v}) = 0$$
$$\frac{d}{dt} \sum_{i=1}^{n} (tu_i - v_i)^2 = 0$$
$$\sum_{i=1}^{n} 2(tu_i - v_i)u_i = 0$$

Solving the above equation for t we find:

$$t = \frac{\sum_i v_i u_i}{\sum_i (u_i)^2}$$

Re-interpreting this formula in terms of the operations of linear algebra we have:

$$t = \frac{(\mathbf{u}, \mathbf{v})}{(\mathbf{u}, \mathbf{u})}$$
$$t = \frac{(\mathbf{u}, \mathbf{v})}{\|\mathbf{u}\|^2}$$

Thus if we wish to select the best approximation of $\mathbf{v}$ in the span of the single vector $\mathbf{u}$ we should select:

$$t\mathbf{u} = \frac{(\mathbf{u}, \mathbf{v})}{\|\mathbf{u}\|^2} \mathbf{u}$$

In linear algebra this quantity is called the **projection** of $\mathbf{v}$ onto $\mathbf{u}$, and this quantity has a geometric interpretation. To understand this interpretation we typically re-write the formula in the following way:

$$Proj_{\mathbf{u}}(\mathbf{v}) \equiv \frac{(\mathbf{v} \cdot \mathbf{u})}{\|\mathbf{u}\|} \frac{\mathbf{u}}{\|\mathbf{u}\|}$$

Writing the formula in this way we can disentangle the different quantities and obtain a clearer conceptual understanding of the formula. The result here is a vector. This vector

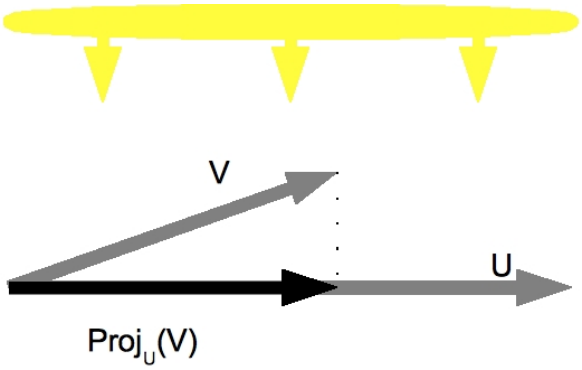

FIGURE 1. The vector projection operator

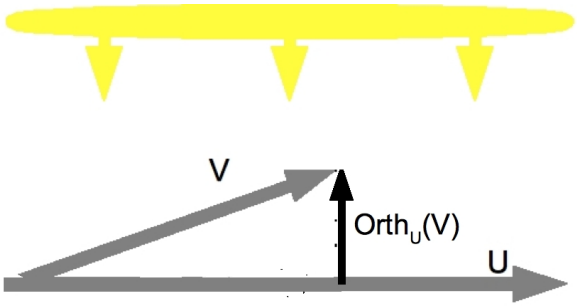

FIGURE 2. The orthogonal component

is composed of two parts, the first part is the magnitude which is defined by the scalar coefficient:

$$\|Proj_{\mathbf{u}}(\mathbf{v})\| = \left(\mathbf{v}, \frac{\mathbf{u}}{\|\mathbf{u}\|}\right),$$

the second part of this vector is the direction defined by the unit vector $\frac{\mathbf{u}}{\|\mathbf{u}\|}$.

The projection of $\mathbf{v}$ in the $\mathbf{u}$ direction lies in the same direction as $\mathbf{u}$, and has the same length as the shadow of the vector $\mathbf{v}$ created by a light source directed perpendicular to $\mathbf{u}$ and pointing directly above $\mathbf{v}$ onto $\mathbf{u}$. (See Figure 1) In addition to the projection operator, one may also want to capture the remainder of a vector after a projection is made. This quantity can be obtained by computing the orthogonal complement)

$$Orth_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - Proj_{\mathbf{u}}(\mathbf{v})$$

If one wants to quantify the size of the error made when using a projection, one simply computes the length of the orthogonal complement. (This the is part of the vector $\mathbf{v}$ which is 'left-out' by the projection.

$$Err(Proj_{\mathbf{u}}(\mathbf{v})) = d(\mathbf{v}, Proj_{\mathbf{u}}(\mathbf{v})) = \|\mathbf{v} - Proj_{\mathbf{u}}(\mathbf{v})\| = \|Orth_{\mathbf{u}}(\mathbf{v})\|$$

3.2. Projections into Subspaces. If one looks at the more general question of approximating a vector using elements from a subspace rather than a single vector, we can approach the idea using ideas from multi-variable calculus, however the problem is more elegantly stated using concepts from linear algebra. We consider the two dimensional approximation problem as an explicit motivating example.

Given a vector $\mathbf{v} \in \mathbb{R}^n$ what is the best approximation to $\mathbf{v}$ contained in the span of two vectors $\mathbf{u}_1$ and $\mathbf{u}_2$?

Approaching this question using calculus, we first adopt the naive approach used in the single vector approximation problem.

We begin by naively parameterizing the admissible approximations with the given spanning set:

$$Sp(\{\mathbf{u}_1, \mathbf{u}_2\}) \equiv \{s\mathbf{u}_1 + t\mathbf{u}_2\}$$

The approximation problem may now be posed as a multi-variable optimization problem over the space of admissible approximations. Choose real numbers s and t to minimize the squared distance between $\mathbf{v}$ and the projection.

$$Proj_{Sp(\mathbf{u}_1,\mathbf{u_2})}(\mathbf{v}) = \min_{s,t\in\mathbb{R}} f(s,t) = \min_{s,t\in\mathbb{R}} \|s\mathbf{u}_1 + t\mathbf{u}_2 - \mathbf{v}\|^2$$

Since this is a multi-variate problem we require that **both** the partial derivatives $\frac{\partial f}{\partial s}$ and $\frac{\partial f}{\partial t}$ be equal to zero simultaneously. These conditions yield the following system of equations to solve for s and t.

$$\frac{\partial f}{\partial s} = \sum_{i=1}^{n} 2(su_{i,1} + tu_{i,2} - v_i)u_{i,1} = 0$$

$$\frac{\partial f}{\partial t} = \sum_{i=1}^{n} 2(su_{i,1} + tu_{i,2} - v_i)u_{i,2} = 0$$

Using linear algebra operations and re-interpreting this as a linear system of two equations and our two unknowns s and t we have:

$$\begin{pmatrix} (\mathbf{u}_1, \mathbf{u}_1) & (\mathbf{u}_1, \mathbf{u}_2) \\ (\mathbf{u}_2, \mathbf{u}_1) & (\mathbf{u}_2, \mathbf{u}_2) \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} (\mathbf{v}, \mathbf{u}_1) \\ (\mathbf{v}, \mathbf{u}_2) \end{pmatrix}$$

By solving this system we obtain the best approximation to $\mathbf{v}$ contained in the span of the given vectors $\mathbf{u}_1$ and $\mathbf{u}_2$.

Looking at this system more critically, we can deduce some very powerful ideas. Consider what happens if $\mathbf{u}_1$ is orthogonal to $\mathbf{u}_2$. In this case the matrix becomes diagonal, and the best approximation is simply the sum of the projections of $\mathbf{v}$ onto $\mathbf{u}_1$ and $\mathbf{u}_2$ respectively. Conversely, if the set $\mathbf{u}_1$ and $\mathbf{u}_2$ are linearly dependent, then the matrix will be singular, and the problem will not have a unique solution (This system will have a solution. The existence condition will automatically be satisfied; proving this is left as an exercise.)

These two observations together raise a question of some importance, what happens if we select a different parameterization of the space of solutions? In linear algebra terms: What happens if we use a different basis for the $Sp(\{\mathbf{u}_1, \mathbf{u}_2\})$?

The answer to this question is of critical importance in applied mathematics. Suppose that rather than use a naive parameterization of $Sp(\{\mathbf{u}_1, \mathbf{u}_2\})$, we **select** a basis of mutually orthogonal vectors: $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$. Then we have the slightly different optimization problem:

$$Proj_{Sp(\mathbf{u}_1,\mathbf{u_2})}(\mathbf{v}) = \min_{s,t\in\mathbb{R}} f(s,t) = \min_{s^*,t^*\in\mathbb{R}} \|s^*\boldsymbol{\phi}_1 + t^*\boldsymbol{\phi}_2 - \mathbf{v}\|^2$$

The basic set-up is the same, however the orthogonality of ϕ_1 and ϕ_2 forces the resulting linear system for the unknown coefficients (s,t) to have diagonal structure.

$$\begin{pmatrix} (\boldsymbol{\phi}_1, \boldsymbol{\phi}_1) & 0 \\ 0 & (\boldsymbol{\phi}_2, \boldsymbol{\phi}_2) \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} (\mathbf{v}, \phi_1) \\ (\mathbf{v}, \phi_2) \end{pmatrix}$$

Furthermore, if we choose $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ to have unit length, then we actually obtain the identity matrix.

We summarize these ideas with the following definitions and theorem:

Definition: Let $\{\boldsymbol{\phi}_i\}$ be a basis for a finite dimensional vector space V. $\{\boldsymbol{\phi}_i\}$ is called an **orthogonal** basis if every pair of basis vectors are orthogonal.

$$(\phi_i, \phi_j) = 0, \text{ if } i \neq j.$$

Definition: Let $\{\phi_i\}$ be a basis for a finite dimensional vector space V. $\{\phi_i\}$ is called an **orthonormal** basis if every pair of basis vectors are orthogonal and each vector has unit norm.

$$(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = \delta_{i,j}, \text{ where } \delta_{i,j} \equiv \begin{cases} 1, & i = j \\ 0 & i \neq j \end{cases}$$

Projection Theorem:
Suppose $\mathbf{v}$ *is a vector in an* n*-dimensional normed vector space,* V*. Let* $U \subset V$ *be an* m *dimensional subspace. Let* $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \ldots \boldsymbol{\phi}_m\}$ *be an orthonormal basis for* U*. Then the*

projection of $\mathbf{v}$ *into* U *is given by:*

$$Proj_U(\mathbf{v}) = (\boldsymbol{\phi}_1, \mathbf{v})\boldsymbol{\phi}_1 + (\boldsymbol{\phi}_2, \mathbf{v})\boldsymbol{\phi}_2 + \cdots + (\boldsymbol{\phi}_m, \mathbf{v})\boldsymbol{\phi}_m$$

The projection of $\mathbf{v}$ *into* U *is the unique vector satisfying:*

$$Proj_U(\mathbf{v}) = \min_{\mathbf{u}\in U} \|\mathbf{u} - \mathbf{v}\|$$

(i.e. The projection is the closest approximation to $\mathbf{v}$ *inside the subspace* U*.)*

Proof:

Parameterizing the projection problem in terms of the $\boldsymbol{\phi}_i$ basis we must select scalars α_i to satisfy the following minimization problem:

$$\min_{\alpha_i\in\mathbb{R}} \left\| \sum_i \alpha_i \boldsymbol{\phi}_i - \mathbf{v} \right\|^2$$

Using elementary multivariable calculus we know that the minimum will satisfy the simultaneous system of equations $\frac{\partial f}{\partial \alpha_i} = 0$ for each $i \in \{1, 2, \ldots m\}$.

Organizing the resulting linear system in terms of the unknown α_i we obtain the following $m \times m$ linear system:

$$\begin{bmatrix} (\boldsymbol{\phi}_1, \boldsymbol{\phi}_1) & (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) & \ldots & (\boldsymbol{\phi}_1, \boldsymbol{\phi}_m) \\ (\boldsymbol{\phi}_2, \boldsymbol{\phi}_1) & (\boldsymbol{\phi}_2, \boldsymbol{\phi}_2) & \ldots & (\boldsymbol{\phi}_2, \boldsymbol{\phi}_m) \\ \vdots & \vdots & \vdots & \vdots \\ (\boldsymbol{\phi}_m, \boldsymbol{\phi}_1) & (\boldsymbol{\phi}_m, \boldsymbol{\phi}_2) & \ldots & (\boldsymbol{\phi}_m, \boldsymbol{\phi}_m) \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} (\mathbf{v}, \boldsymbol{\phi}_1) \\ (\mathbf{v}, \boldsymbol{\phi}_2) \\ \vdots \\ (\mathbf{v}, \boldsymbol{\phi}_m) \end{pmatrix}$$

Using the mutual orthogonality condition $(\phi_i, \phi_j) = 0$ for $i \neq j$ this system reduces to:

$$\begin{bmatrix} (\boldsymbol{\phi}_1, \boldsymbol{\phi}_1) & 0 & \ldots & 0 \\ 0 & (\boldsymbol{\phi}_2, \boldsymbol{\phi}_2) & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & (\boldsymbol{\phi}_m, \boldsymbol{\phi}_m) \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} (\mathbf{v}, \boldsymbol{\phi}_1) \\ (\mathbf{v}, \boldsymbol{\phi}_2) \\ \vdots \\ (\mathbf{v}, \boldsymbol{\phi}_m) \end{pmatrix}$$

Using the normality (i.e. $\|\boldsymbol{\phi}_i\| = 1$) this further reduces to:

$$\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} (\mathbf{v}, \boldsymbol{\phi}_1) \\ (\mathbf{v}, \boldsymbol{\phi}_2) \\ \vdots \\ (\mathbf{v}, \boldsymbol{\phi}_m) \end{pmatrix}$$

This proves the desired result.

Q.E.D.

Lemma: *Every finite dimensional vector space possesses an orthonormal basis.*
Proof:
The lemma follows directly by construction when one uses the Gram-Schmidt Orthogonalization process outlined in the practice section.

The ideas contained in the projection theorem tell us that when we wish to find find approximations in normed linear space, the most computationally efficient manner to proceed it to use orthogonal basis vectors. By selecting an orthogonal basis we decouple the projections from each basis direction and we are free to sum the projections together. If we select an awkward parameterization, then the algebra linking the different parts of the approximation gets much more complicated.

3.3. Overdetermined Systems. The classical problem we solve by using projection and approximation is the overdetermined system. An over-determined system has more equations than unknowns, thus we try to approximate a right-hand side which might live in a high dimensional linear space, but we are restricted to only using the output of a low dimensional input. This type of situation occurs regularly in experimental science as we may be able to repeat experiments many times, but we seldom want to work with a model which is complex enough to capture the experimental data exactly. This is possible, but since experimental measurements contain at least some measurement error, capturing the experimental data exactly is generally discouraged. It is often considered much more valuable to approximately fit experimental data with a curve or model we think accurately captures the dynamics of the underlying situation. If we fit a theoretically justified model to experimental data, we will generally have a much better understanding of the domain of validity of that model and we can more accurately assess the quality of predictions from the model.

The most commonly solved problem involving an overdetermined system is the 'best fit' line. We will use this as an illustrative example.

We consider a given data set of n ordered pairs (x_i, y_i), and we ask the question: 'Which line best 'fits' this data set?' Interpreting this problem with linear algebra as an overarching conceptual framework gives us valuable insights into the answer to this question.

Schematically we assume our model has an affine form:

$$mx + b = y$$

This model takes x as an input, returns y as an output and has two numerical parameters m, and b. To find the best fit line for a given data set we must choose values for the parameters m and b which give the 'best predictions' for the output values. In terms of linear spaces, this model has a two dimensional parameter space, while the desired output lies in an n dimensional linear space (without any prior knowledge, the output y_i could be anything.) By taking each of the given data points and substituting them into this model

we obtain an overdetermined set of equations for the model parameters m and b.

$$\begin{aligned} mx_1 + b &= y_1 \\ mx_2 + b &= y_2 \\ mx_3 + b &= y_3 \\ &\vdots \\ mx_n + b &= y_n \end{aligned}$$

In terms of matrix-vector notation we find:

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

Thus the input data and the choice of model determine the contents of the matrix, the observed output values form the right hand side, and the model parameters form the unknowns we wish to find.
From our theoretical knowledge of solvability theory we can immediately throw out any hope of solving this problem exactly. With that option off the table we turn to approximation. Here we wish to find the 'best' approximation of the given right hand side $\mathbf{y}$ in terms of the two dimensional range of our model. Just like before we can can use calculus to pose this problem in terms of a minimization problem and take partial derivatives with respect to each of the model parameters. Setting each of those derivatives equal to zero simultaneously gives us a system of equations to solve for the optimal model parameters. In this case, however, we can use our knowledge of solvability theory and the idea of projection to short-cut through the multi-variate analysis and obtain the correct system of equations.

We know that a linear system possesses a solution if and only if the right hand side is orthogonal to the null space of the adjoint operator. If we can project $\mathbf{y}$ directly onto the range by removing any part of $\mathbf{y}$ which is not orthogonal to the requisite null space, then we can deal with the resulting linear system with the foreknowledge that that system possesses a solution. Furthermore the projection is the best approximation by construction anyway, so we can bypass the calculus completely in practice.

In order to project $\mathbf{y}$ onto the correct vector space we must remove any part of $\mathbf{y}$ which lies in the null space of the adjoint operator. The easiest way to remove these parts is to multiply by the adjoint operator itself. Anything within the null space is mapped to zero, while anything orthogonal to the null-space is preserved.

Let us test the implications of this idea. By left multiplying by the transpose of our model matrix in the 'best-fit line' case we obtain the following:

$$A\begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$A^T A\begin{pmatrix} m \\ b \end{pmatrix} = A^T \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = A^T \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Performing the matrix product $A^T A$ and the matrix-vector product $A^T \mathbf{y}$ for this particular case we obtain the following:

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

This is extremely promising since we now have a 2×2 linear system to solve for the two model parameters m and b. Solving this system yields **exactly** the same solution obtained through calculus.

This idea works for any linear system. Whenever the linear system:

$$A\mathbf{x} = \mathbf{y}$$

does not possess an exact solution, the best approximation to the solution may be obtained by solving the normal equations given by:

$$A^T A\mathbf{x}^* = A^T \mathbf{y}$$

The exact solution to this problem is best in the sense that the quantity:

$$\|A\mathbf{x}^* - \mathbf{y}\|$$

is minimized. For applications this idea works to find the best approximation whenever all of the model parameters appear linearly in the model equation. Thus, given a data set (x_i, y_i) we can just as easily ask the question what parabola best fits the data. And we can answer this question by solving the normal equations for the given data set with the model equation:

$$\alpha x^2 + \beta x + \gamma = y$$

Explicitly, our overdetermined system is given by:

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_i^2 & x_i^2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n^2 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

We find the best fit quadratic by solving the resulting set of normal equations.

Part 2

Practice

While the theory section of this book is concerned with the mathematical theory surrounding linear spaces, this section is concerned purely with the procedures and techniques which are used in practice. Rather than being presented in a theorem-proof style, the procedures are arranged in self-contained sections which permit the student to study each procedure in isolation, without any reference to the other procedures.

CHAPTER 5

Hand Computation

1. Matrix Multiplication

Matrix multiplication is a natural generalization of the process of matrix vector multiplication briefly defined in Theory, chapter 1, section 3. Matrix multiplication is used to compute compositions of linear mappings, and to compute the output of the forward problem of linear algebra.

Before performing matrix multiplication between given matricies, one must make sure that the operation is actually defined. In order for matrix multiplication to be well defined the number of columns in the matrix on the left must be equal to the number of rows of the matrix on the right. The resultant of any matrix multiplication operation will possess the same number of rows as the left matrix and the same number of columns as the right matrix.

Schematically:

$$A_{n\times m}B_{m\times p} = C_{n\times p}$$

We consider the following explicit example with A a 3×2 matrix and B a 2×3 matrix. If we perform the matrix multiplication AB we obtain a 3×3 matrix.

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}$$

And if we perform the matrix multiplication BA, we obtain a 2×2 matrix.

$$\begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

We remark at this point that the operation of matrix multiplication is **not** commutative, and as the example above suggests, changing the order of two matricies can change the size of the resulting matrix. Matrix multiplication is associative, which one can prove by direct computation on any sized matricies of interest.

The process of matrix multiplication is most easily broken down into steps based upon the rows and columns of the two matrices being multiplied. We will first illustrate matrix multiplication for a set of arbitrary 2×2 matrices, then we will proceed to perform an

explicit example with a 3×3 and a 3×2 matrix.

Abstract Example:
Suppose that we need to perform the following matrix multiplication:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

We already know based upon the earlier discussion that the resultant will have 2 rows and 2 columns.

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} ? & ? \\ ? & ? \end{pmatrix}$$

in order to find the entry in the first row and first column of the resultant we take the first column of the right matrix, and multiply it component-wise with the first row of the left matrix. We sum the results of these computations and place the result in the first row and first column of the resultant.

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & ? \\ ? & ? \end{pmatrix}$$

To fill in the other entries in the resultant we repeat the same process, but choose different rows of the left matrix and columns of the right matrix to obtain the different entries in the resultant. Explicitly the result is given by:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Numerical example:
Suppose that we need to perform the following matrix multiplication:

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 0 \end{pmatrix}$$

We know that the resultant matrix should have 3 rows and 2 columns.

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} ? & ? \\ ? & ? \\ ? & ? \end{pmatrix}$$

To find each of the corresponding values in the resultant we multiply one row from the left matrix and one column from the right matrix component-wise and sum the results. Explicitly to obtain the entry in the first row and first column, we multiply the first row from the left matrix and the first column from the right matrix.

$$0(1) + 1(2) + 2(3) = 8$$

After summing the results we have the entry for the first row and first column.

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 8 & ? \\ ? & ? \\ ? & ? \end{pmatrix}$$

To find the entry in the second row and second column, we multiply the second row from the left matrix with the second column of the right matrix.

$$2(-1) + 1(1) + 0(0) = -1$$

Now we can fill in the second row, second column of the resultant:

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 8 & ? \\ ? & -1 \\ ? & ? \end{pmatrix}$$

Repeating this process for every row and column combination gives us all of the entries in the resultant.

$$\begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 8 & 1 \\ 4 & -1 \\ 2 & 1 \end{pmatrix}$$

When performing matrix multiplication by hand it helps to compute the resultant in an organized fashion, either row by row or column by column. Computing the entries randomly can easily lead to bookkeeping errors.

The key to remembering the process correctly is to remember the schematic relationship:

$$A_{n\times m}B_{m\times p} = C_{n\times p}.$$

The resultant has the same number of rows as the left matrix, and the same number of columns as the right matrix. The individual entries of the resultant are obtained by taking the scalar product of a single row from the left and a single column from the right. Thus, if you can remember this relationship, then it will be very hard to mix up rows and columns.

2. Determinants

The determinant is a real valued function defined for all square matricies. The value of the determinant has a geometric interpretation, but at the elementary level the determinant yields immediate and unambiguous information about linear dependence, invertability, and solvability of certain linear systems. When using the determinant to infer things about a matrix or linear system, the most salient piece of information is whether the determinant is zero or non-zero. Below is a short list of facts which may be deduced from the value of the determinant.

Let A be an $n \times n$ matrix.

- If $\det(A) \neq 0$ then
  - A is called a non-singular matrix.
  - the column vectors that comprise A are linearly independent.
  - the row vectors that comprise A are linearly independent.
  - any linear system of the form $A\mathbf{x} = \mathbf{b}$ always possesses a unique solution.
- If $\det(A) = 0$ then
  - A is called a singular matrix.
  - the column vectors that comprise A are linearly dependent.
  - the row vectors that comprise A are linearly dependent.
  - any linear system of the form $A\mathbf{x} = \mathbf{b}$ either possesses no solution or infinitely many solutions.

In addition to these deductions, the determinant may be used to answer questions posed more theoretically. If one wishes to check whether a set of vectors is linearly dependent, rather than trying to solve the implied linear system, one can collect the vectors into a matrix and compute the determinant of any square sub-matrix. Any collection of vectors covered by a sub-matrix with a non-zero determinant is linearly independent.

To compute a determinant by hand, many procedures are possible we choose to explore the determinant by minors procedure, since this procedure is fast, flexible and allows one to take advantage of any zeros in the matrix to simplify the computation. For 2×2 and 3×3 matrices one can memorize a closed form formula for the determinant, however this method is prone to error, and doesn't scale. Determinant by minors provides a recursively defined procedure for computing the determinant of any square matrix.

We illustrate the process by induction. We often denote the determinant by drawing straight, vertical lines on both sides of the matrix, and define the value of the determinant of a 2×2 matrix by the following formula:

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

Now to define the determinant of a 3×3 matrix we select any row or column within the matrix and define the determinant in terms of a sum of terms depending upon the minors of that row or column. For illustrative purposes we will compute the determinant of a 3×3 matrix in two different ways, and demonstrate that those two computations yield the same result.

To expand about the first row in the matrix we create the following framework:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \equiv \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix} - b \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix} + c \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix}$$

Note that we have one term for each component in the first row, and the signs of these components alternate sign. Each term is multiplied by a two by two matrix whose elements

are given by the minor of the expanding element. To obtain the minor for a, eliminate the row and column containing a, and the resulting elements define the minor for a. Explicitly:

$$\begin{vmatrix} \cancel{\mathbf{a}} & \cancel{\mathbf{b}} & \cancel{\mathbf{c}} \\ \cancel{\mathbf{d}} & \mathbf{e} & \mathbf{f} \\ \cancel{\mathbf{g}} & \mathbf{h} & \mathbf{i} \end{vmatrix} \to \begin{vmatrix} e & f \\ h & i \end{vmatrix}$$

For $-b$ we cross out the first row and second column to obtain the correct minor:

$$\begin{vmatrix} \cancel{\mathbf{a}} & \cancel{\mathbf{b}} & \cancel{\mathbf{c}} \\ \mathbf{d} & \cancel{\mathbf{e}} & \mathbf{f} \\ \mathbf{g} & \cancel{\mathbf{h}} & \mathbf{i} \end{vmatrix} \to \begin{vmatrix} d & f \\ g & i \end{vmatrix}$$

For c we cross out the first row and third column to obtain the correct minor:

$$\begin{vmatrix} \cancel{\mathbf{a}} & \cancel{\mathbf{b}} & \cancel{\mathbf{c}} \\ \mathbf{d} & \mathbf{e} & \cancel{\mathbf{f}} \\ \mathbf{g} & \mathbf{h} & \cancel{\mathbf{i}} \end{vmatrix} \to \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

We combine all of these minors into the determinant and compute:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei - afh - bdi + bfg + cdh - ceg$$

If we wish to compute the determinant by selecting another row or column we are free to do so. For example, we might wish to find the determinant of a 3×3 matrix by expanding about the second column. Here we would create the following framework:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \equiv \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = -b \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix} + e \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix} - h \begin{vmatrix} \# & \# \\ \# & \# \end{vmatrix}$$

We fill the minors by crossing out the rows and columns of each of the new expanding elements to find:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \equiv \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = -b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + e \begin{vmatrix} a & c \\ g & i \end{vmatrix} - h \begin{vmatrix} a & c \\ d & f \end{vmatrix}$$

Multiplying out these determinants carefully, you should find that the resulting formula contains exactly the same terms as the previous formula.

Remark: To determine the correct signs for each element in your expansion you can use the formula $\text{sign}(a_{ij}) = (-1)^{i+j}$ or you can use the following visualization for the

pattern:

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}$$

(This pattern continues its alternation for larger matrices.)

Remark: The chief advantage of determinant by minors is the flexibility you have in constructing the formula. When faced with a matrix containing one or more zeros, select the row or column containing the most zeros to create your expansion, then only compute the minors for the non-zero expansion elements.

3. Gaussian Elimination

Solving a linear system is equivalent to solving a system of simultaneous linear equations. You can always take a linear system, convert it to algebraic notation and perform valid algebraic manipulations to isolate each of the variables in turn. For systems involving two or three variables this transformation is reasonable, but for problems involving more variables using matrix notation is faster, more efficient, and allows one to mechanize the solution procedure.

In our exploration of the technique of Gaussian Elimination, we present the algorithm in several stages, and illustrate each stage using explicit numerical examples.

We term the first stage of the algorithm 'Naive Gaussian Elimination.' This algorithm allows us to systematically convert an arbitrary linear system into a triangular linear system, which lends itself to simple algebraic solution. Naive Gaussian elimination is presented as a deterministic algorithm, and illustrates the fundamental technique of systematically removing variables from the whole system of equations.

The second stage of the algorithm is termed 'Full Gaussian Elimination.' Here we introduce the technique of pivoting which allows us to get around any instances of zero division which can arise when using the naive algorithm.

The third stage of the algorithm is termed back substitution. Here we convert the triangular system of equations obtained with either naive or full gaussian elimination into a system of algebraic equations, which can be easily solved to find the values of the components of the unknown vector, $\mathbf{x}$.

We remark at the outset, that it may feel unnatural at first to use an algorithm to solve mathematical problems. We are typically taught to look for patterns or structure in problems and to use that structure to help us find the solution quickly. In the case of large linear systems, unless one is aware of some type of existing structure before hand (i.e. matrix structure such as banded or block diagonal structure), searching for structure is inefficient, and one should simply charge ahead to find solutions. The purpose of learning gaussian

elimination algorithmically is two-fold. First, in practice, one does not solve large linear systems by hand, so detailed knowledge of the actual solution procedure is only employed when programming, and one must treat the process algorithmically in order to program it effectively. Second, one must learn to inure ones-self from intellectual engagement in routine mathematical calculations. As you pass from the study of elementary algebra and calculus, the purpose of mathematical notation changes. In elementary mathematics we use notation to record and display every step in a calculation. The notation helps us visualize different operations and to learn the rules which govern their manipulation. In more advanced study we assume that the rules of manipulation are known, and mathematical notation becomes a compressed shorthand. Many of the computational details are set aside in order to increase the speed and power of the calculations which can be performed. Solving a linear system is no harder than a bunch of elementary arithmetic calculations. To become competent, one must be able to carefully work through lengthy multi-step calculations without making errors. The algorithm allows us to focus carefully on each arithmetic step without worrying about the purpose and direction of the calculations, the algorithm will take us to the solutions if they exist.

3.1. Naive Gaussian Elimination. The purpose of naive gaussian elimination is to perform a sequence of operations on a linear system which systematically zeros the entries of a matrix, while leaving the components of the unknown vector unchanged. Schematically, under ideal circumstances naive gaussian elimination takes us from a matrix with arbitrary entries, to a matrix with zeros in every entry below the main diagonal. (The main diagonal consists of those entries which lie in the same row and column.)

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \leftrightarrow \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

Before we begin the process of gaussian elimination we make a remark about notation. To speed computations and minimize the amount of writing we need to perform, we only need to track the entries of the matrix and the entries of the vector on the right hand side. Since the vector of unknowns, $\mathbf{x}$, is unchanged by any of the manipulations we will perform we can use the order of the columns in the matrix to keep track of the variables. For working notation we use an augmented matrix, shown on the left, and for clarity we also include the corresponding system of algebraic equations on the right.

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & y_1 \\ a_{21} & a_{22} & a_{23} & y_2 \\ a_{31} & a_{32} & a_{33} & y_3 \end{array}\right], \qquad \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= y_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= y_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= y_3 \end{aligned}$$

Abstract Example:

We begin by performing naive gaussian elimination on an abstract example involving a 3×3 matrix. Since everything is abstract, we introduce new variables at logical stopping points during the process, this will be unnecessary when solving numerical examples. As you follow this example you should try to absorb the different steps of the procedure, do

not worry about the arithmetic particulars until you consider numerical examples. When solving a system by hand use the augmented matrix notation on the left, the algebraic notation included on the right is only for reference.

(1) Step 1: Normalize the First Column. Divide every entry in the first row by the diagonal entry a_{11}.

$$\tfrac{1}{a_{11}}R1 \to R1$$

$$\left[\begin{array}{ccc|c} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{y_1}{a_{11}} \\ a_{21} & a_{22} & a_{23} & y_2 \\ a_{31} & a_{32} & a_{33} & y_3 \end{array}\right], \qquad \begin{array}{c} x_1 + \frac{a_{12}}{a_{11}}x_2 + \frac{a_{13}}{a_{11}}x_3 = \frac{y_1}{a_{11}} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = y_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = y_3 \end{array}$$

(2) Step 2: Zero the first Column. Starting with the second row, subtract the first row multiplied by the entry in the first column from the second row, and use the result to replace the second row.

$$R2 - a_{21}R1 \to R2$$

$$\left[\begin{array}{ccc|c} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{y_1}{a_{11}} \\ 0 & a_{22} - \frac{a_{21}a_{12}}{a_{11}} & a_{23} - \frac{a_{21}a_{13}}{a_{11}} & y_2 - \frac{a_{21}y_1}{a_{11}} \\ a_{31} & a_{32} & a_{33} & y_3 \end{array}\right], \qquad \begin{array}{rcl} x_1 + \frac{a_{12}}{a_{11}}x_2 + \frac{a_{13}}{a_{11}}x_3 & = & \frac{y_1}{a_{11}} \\ \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right)x_2 + \left(a_{23} - \frac{a_{21}a_{13}}{a_{11}}\right)x_3 & = & y_2 - \frac{a_{21}y_1}{a_{11}} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 & = & y_3 \end{array}$$

Repeat this process with the third and any remaining rows.

$$R3 - a_{31}R1 \to R3$$

$$\left[\begin{array}{ccc|c} 1 & \frac{a_{12}}{a_{11}} & \frac{a_{13}}{a_{11}} & \frac{y_1}{a_{11}} \\ 0 & a_{22} - \frac{a_{21}a_{12}}{a_{11}} & a_{23} - \frac{a_{21}a_{13}}{a_{11}} & y_2 - \frac{a_{21}y_1}{a_{11}} \\ 0 & a_{32} - \frac{a_{31}a_{12}}{a_{11}} & a_{33} - \frac{a_{31}a_{13}}{a_{11}} & y_3 - \frac{a_{31}y_1}{a_{11}} \end{array}\right], \qquad \begin{array}{rcl} x_1 + \frac{a_{12}}{a_{11}}x_2 + \frac{a_{13}}{a_{11}}x_3 & = & \frac{y_1}{a_{11}} \\ \left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right)x_2 + \left(a_{23} - \frac{a_{21}a_{13}}{a_{11}}\right)x_3 & = & y_2 - \frac{a_{21}y_1}{a_{11}} \\ \left(a_{32} - \frac{a_{31}a_{12}}{a_{11}}\right)x_2 + \left(a_{33} - \frac{a_{31}a_{13}}{a_{11}}\right)x_3 & = & y_3 - \frac{a_{31}y_1}{a_{11}} \end{array}$$

This step has systematically removed x_1 from all but the first row (or the first equation in algebraic notation). We now repeat this process with x_2. To abbreviate our notation in this example we introduce new variables.

$$\left[\begin{array}{ccc|c} 1 & b_{12} & b_{13} & z_1 \\ 0 & b_{22} & b_{23} & z_2 \\ 0 & b_{32} & b_{33} & z_3 \end{array}\right], \qquad \begin{array}{c} x_1 + b_{12}x_2 + b_{13}x_3 = z_1 \\ b_{22}x_2 + b_{23}x_3 = z_2 \\ b_{32}x_2 + b_{33}x_3 = z_3 \end{array}$$

(3) Step 3: Normalize the second column. Divide every entry in the second row by the diagonal entry b_{22}.

$$\tfrac{1}{b_{22}}R2 \to R2$$

$$\left[\begin{array}{ccc|c} 1 & b_{12} & b_{13} & z_1 \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \frac{z_2}{b_{22}} \\ 0 & b_{32} & b_{33} & z_3 \end{array}\right], \qquad \begin{array}{rcl} x_1 + b_{12}x_2 + b_{13}x_3 & = & z_1 \\ x_2 + \frac{b_{23}}{b_{22}}x_3 & = & \frac{z_2}{b_{22}} \\ b_{32}x_2 + b_{33}x_3 & = & z_3 \end{array}$$

(4) Step 4: Zero the second column. Starting with the **third** row, subtract the **second** row multiplied by the entry in the **second** column from the **third** row, and use the result to replace the third row.

$$R3 - b_{32}R2 \to R3$$

$$\left[\begin{array}{ccc|c} 1 & b_{12} & b_{13} & z_1 \\ 0 & 1 & \frac{b_{23}}{b_{22}} & \frac{z_2}{b_{22}} \\ 0 & 0 & b_{33} - \frac{b_{32}b_{23}}{b_{22}} & z_3 - \frac{b_{32}z_2}{b_{22}} \end{array}\right], \qquad \begin{array}{rcl} x_1 + b_{12}x_2 + b_{13}x_3 & = & z_1 \\ x_2 + \frac{b_{23}}{b_{22}}x_3 & = & \frac{z_2}{b_{22}} \\ \left(b_{33} - \frac{b_{32}b_{23}}{b_{22}}\right)x_3 & = & z_3 - \frac{b_{32}z_2}{b_{22}} \end{array}$$

If there were any more rows we would continue this process until the second column was all zeros below the diagonal. For this small example we are done with the second column. Note that we have systematically removed x_2 from everything except the first and second equations.

To abbreviate out notation we again introduce new variables:

$$\left[\begin{array}{ccc|c} 1 & b_{12} & b_{13} & z_1 \\ 0 & 1 & c_{23} & w_2 \\ 0 & 0 & c_{33} & w_3 \end{array}\right], \qquad \begin{array}{c} x_1 + b_{12}x_2 + b_{13}x_3 = z_1 \\ x_2 + c_{23}x_3 = w_2 \\ c_{33}x_3 = w_3 \end{array}$$

(5) Step 5+: Normalize and zero any remaining columns following the same process as above.
For a 3×3 matrix, naive gaussian elimination is now complete, we have taken an arbitrary matrix and replaced it with a matrix which has zeros in every entry below the main diagonal. To find solutions to this system by hand, convert the resulting augmented matrix to algebraic notation and solve the resulting equations by systematic back-substitution. (The last equation contains only x_3, so solve the equation to find the value of x_3. Then proceed to the second to last equation to find x_2 etc.)

Numerical Example:
In this example we perform naive gaussian elimination on a particular 3×3 matrix, we again outline the steps of the algorithm and we include the shorthand notation for the row operations performed at each step.

$$\begin{bmatrix} 2 & 4 & 6 \\ 2 & 9 & 16 \\ 1 & 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ 18 \\ 6 \end{bmatrix}$$

Writing this linear system as an augmented matrix we obtain:

$$\left[\begin{array}{ccc|c} 2 & 4 & 6 & 8 \\ 2 & 9 & 16 & 18 \\ 1 & 3 & 3 & 6 \end{array}\right]$$

(1) Normalizing the first column:

$$\frac{1}{2}R1 \to R1, \quad \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 2 & 9 & 16 & 18 \\ 1 & 3 & 3 & 6 \end{array}\right].$$

(2) Zero the first column: (These are shown as two separate steps for clarity.)

$$R2 - 2R1 \to R2, \quad \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 5 & 10 & 10 \\ 1 & 3 & 3 & 6 \end{array}\right],$$

$$R3 - 1R1 \to R3, \quad \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 5 & 10 & 10 \\ 0 & 1 & 0 & 2 \end{array}\right].$$

(3) Normalize the second column:

$$\frac{1}{5}R2 \to R2, \quad \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 2 \\ 0 & 1 & 0 & 2 \end{array}\right],$$

(4) Zero the second column:

$$R3 - 1R2 \to R3, \quad \left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & -2 & 0 \end{array}\right].$$

At this point we have completed naive Gaussian elimination and we have obtained an upper triangular system equivalent to our original system. In algebraic notation we have:

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 4 \\ x_2 + 2x_3 &= 1 \\ -2x_3 &= 0 \end{aligned}$$

Back-substituting to find each x_i yields the unique solution set: $x_1 = 0$, $x_2 = 2$, $x_3 = 0$.

3.2. Full Gaussian Elimination. The naive gaussian elimination procedure does not work on all linear systems. Even those which have a unique solution. In each normalization step we divide by the current diagonal entry in the matrix, and if this diagonal element is zero, naive gaussian elimination fails. To get around this potential pitfall we introduce the 'pivoting' operation, which consists of switching the current row with one of the later rows in the matrix. Algebraically speaking, pivoting by row switching is a non-operation. The underlying equations before and after the operation are identical, only the order of presentation changes.

To illustrate the idea we present an abstract example on a 4×4 linear system, partway through the gaussian elimination procedure, and we follow with a very simple numerical 3×3 example where a pivot is required.

Abstract Example:

Consider the following augmented matrix, partway through gaussian elimination:

$$\left[\begin{array}{cccc|c} 1 & a_{12} & a_{13} & a_{14} & y_1 \\ 0 & 0 & a_{23} & a_{24} & y_2 \\ 0 & a_{32} & a_{33} & a_{34} & y_3 \\ 0 & a_{42} & a_{43} & a_{44} & y_4 \end{array}\right]$$

Normally, during naive gaussian elimination the next step would be to normalize the second column dividing by the entry in a_{22}. Since the entry is zero in this case, we must try to switch rows to find a suitable element for the diagonal. Here we may have a choice. If the entry a_{23} is non-zero we can switch the second and third rows. **Or** if the entry a_{24} is non-zero we can switch the second and fourth rows. If both a_{23} and a_{24} select the larger one, and switch rows to place that on the diagonal. If both a_{23} and a_{24} are zero, then the matrix is singular and special care must be taken to find all solutions.

To notate a row switch between the second and third rows you can use the following shorthand notation:

$$R2 \leftrightarrow R3, \quad \left[\begin{array}{cccc|c} 1 & a_{12} & a_{13} & a_{14} & y_1 \\ 0 & a_{32} & a_{33} & a_{34} & y_3 \\ 0 & 0 & a_{23} & a_{24} & y_2 \\ 0 & a_{42} & a_{43} & a_{44} & y_4 \end{array}\right].$$

Or to notate a row switch between the second and fourth rows:

$$R2 \leftrightarrow R4, \quad \left[\begin{array}{cccc|c} 1 & a_{12} & a_{13} & a_{14} & y_1 \\ 0 & a_{42} & a_{43} & a_{44} & y_4 \\ 0 & a_{32} & a_{33} & a_{34} & y_3 \\ 0 & 0 & a_{23} & a_{24} & y_2 \end{array}\right].$$

After either pivoting operation, simply continue the naive gaussian elimination algorithm.

Numerical Example:

Here we use the full gaussian elimination procedure on a non-singular matrix where a single pivot operation is required. We continue to outline the steps of naive gaussian elimination, and record the short hand augmented matrix notation.

$$\left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 2 & 2 & 5 & 2 \\ 0 & 1 & 0 & 3 \end{array}\right]$$

(1) The First column is already normalized, so we move on to step (2).
(2) Zero the first column:

$$R2 - 2R1 \to R2, \qquad \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 3 \end{array}\right]$$

After this single row operation, the whole first column is already zeroed, so we try to proceed to step (3), however the entry a_{22} is zero, so we must pivot. Since we only pivot with later rows, the only choice is to switch the second and third rows.
PIVOT

$$R2 \leftrightarrow R3, \qquad \left[\begin{array}{ccc|c} 1 & 1 & 2 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

(3) Normalize the second column. (This is already done.)
(4) Zero the second column. (This is already done.)

This particular example is especially easy, as the pivot operation is the last operation necessary to finish gaussian elimination. In practice one may need to perform more computations after the pivot operation, and in some cases may need to pivot multiple times.

3.3. Gaussian Elimination for Singular Matrices. In the examples considered thus far, the gaussian elimination procedure has always produced a unique solution set. Careful study of Theory chapter 2 section 4, leads one to wonder what happens in situations with either no solution or multiple solutions. To help explore these ideas we will work through a series of representative numerical examples. These are not exhaustive, but they should give the reader a good feel for potential pitfalls and complications which can arise during the elimination procedure. It's important to remark that in practice, one does not know if there will be one solution, many solutions or no solution to a given problem until one starts to analyze it, so the following examples are a little misleading in that the situation is outlined beforehand.

No Solution

The following linear systems have no exact solution, we begin with a simple 2×2 example, and follow with a more complex 3×3 example.

$$\left[\begin{array}{cc|c} 1 & -1 & 1 \\ 1 & -1 & 0 \end{array}\right]$$

Translating this 2×2 into algebraic notation it should become clear that these two equations are inconsistent, and no exact solution is possible. Here we illustrate what happens in gaussian elimination when a linear system has no solution.

Proceeding with naive gaussian elimination we zero the first column to obtain:

$$R2 - R1 \to R2, \quad \left[\begin{array}{cc|c} 1 & -1 & 1 \\ 0 & 0 & -1 \end{array}\right]$$

Here we immediately run into to trouble. Translating the second row into algebraic notation we have:

$$\begin{aligned} 0x_1 + 0x_2 &= -1 \\ 0 &= -1 \end{aligned}$$

No choice of x_1 and x_2 can possibly make this statement true, so this system is inconsistent and has no exact solution.

Next we consider the following more complex linear system:

$$\left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 1 & 1 & 3 & 0 \\ 2 & -2 & 6 & 0 \end{array}\right]$$

Here even when translated into algebraic notation it may not be at all clear or obvious whether this system has a solution or not. We proceed with gaussian elimination and outline how non-existence presents itself.

Zeroing the first column we obtain:

$$\begin{aligned} R2 - R1 \to R2, \quad & \left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 0 & 2 & 0 & -1 \\ 2 & -2 & 6 & 0 \end{array}\right], \\ R3 - 2R1 \to R3, \quad & \left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & -2 \end{array}\right] \end{aligned}$$

Translating the last row into algebraic notation we obtain a patently false statement:

$$0 = -2.$$

Anytime algebraically valid row operations bring us to a false statement, we can immediately conclude the system is inconsistent and possesses no exact solution.

Remark: Its interesting to note that the story doesn't necessarily end here. There are many situations such as fitting mathematical models to experimental data, where inconsistent systems are the rule rather than the exception, so one spends considerable effort finding approximate solutions that 'best fit' the inconsistent system.

Many Solutions

In addition to the possibility of no exact solution, linear systems may suffer from an infinite set of solutions. Again, in simple cases one may be able to deduce the presence of multiple solutions just by observation. But in larger linear systems the presence of multiple solutions may not be obvious.

When performing gaussian elimination the appearance of multiple solutions is typically characterized by the appearance of one or more rows of zeros in the augmented matrix. The row of zeros indicates that several of the rows in the augmented matrix contain redundant information, so satisfying several of the equations, implies that other equations are satisfied automatically.

We proceed to illustrate the idea by using slightly modified versions of the previous two examples.

$$\left[\begin{array}{cc|c} 1 & -1 & 1 \\ 1 & -1 & 1 \end{array}\right]$$

Translating this 2×2 into algebraic notation it should become clear that these two equations are identical.

Proceeding with naive gaussian elimination and zeroing the first column we obtain a row of zeros:

$$R2 - R1 \to R2, \quad \left[\begin{array}{cc|c} 1 & -1 & 1 \\ 0 & 0 & 0 \end{array}\right]$$

The row of zeros by itself does not pose any problem since that equation is automatically true, however when trying to solve for the values of x_1 and x_2 we find that we don't have enough equations to uniquely specify their values.

$$\begin{aligned} x_1 - x_2 &= 1 \\ 0 &= 0 \end{aligned}$$

We need to specify additional information to completely determine x_1 and x_2. Each piece of information we need to specify is termed a 'degree of freedom.' From a bookkeeping standpoint a linear system with m variables, and n non-zero equations after gaussian elimination has $m - n$ degrees of freedom.

In this particular example we have one degree of freedom. Since we have one equation involving both x_1 and x_2, we could specify either variable and the other variable would have a determinate value. Suppose we select $x_2 = k$, then our equation forces:

$$x_1 = 1 + k$$

Looking at this resulting solution in vector notation we have the general formula:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1+k \\ k \end{pmatrix}$$

Connecting the concept of degrees of freedom to our theoretical view yields an important characterization of the resulting set of solutions. Splitting the solution into parts that are independent of k and parts dependent on k we have:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} k \\ k \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + k \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The part of the solution which is independent of k gets mapped completely into the range, while the part of the solution dependent upon k lies completely in the null space of the linear operator.

Next we consider the following more complex linear system:

$$\left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 1 & 1 & 3 & 0 \\ 2 & -2 & 6 & 2 \end{array}\right]$$

Here even when translated into algebraic notation it may not be at all clear or obvious whether this system has a solution or not. We proceed with gaussian elimination and outline how non-existence presents itself.

Zeroing the first column we obtain:

$$\begin{aligned} R2 - R1 \to R2, \quad & \left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 0 & 2 & 0 & -1 \\ 2 & -2 & 6 & 0 \end{array}\right], \\ R3 - 2R1 \to R3, \quad & \left[\begin{array}{ccc|c} 1 & -1 & 3 & 1 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right] \end{aligned}$$

Translating the last row into algebraic notation we obtain a trivially true statement, which results in a degree of freedom in our solution.

$$0 = 0.$$

Here we have three variables and two non-zero equations. Thus we have one degree of freedom. Looking at the remaining algebraic equations carefully, we can see that the value of x_2 is determined exactly, but the values of x_1 and x_3 are not completely specified.

$$\begin{aligned} x_1 - x_2 + 3x_3 &= 1 \\ 2x_2 &= -1 \end{aligned}$$

Like the previous example we can supplement the equations with an additional piece of information to specify the degree of freedom. We select $x_3 = k$, and solve the resulting system to obtain the general solution:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} - 3k \\ -\frac{1}{2} \\ k \end{pmatrix}$$

This general solution has a two part structure, one part fully specified part mapping directly into the range, and a second part determined by the chosen degree of freedom which lies completely in the null space of the matrix.

As a final numerical example we consider a case with multiple degrees of freedom.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ -1 & -1 & -1 & -1 \end{array}\right]$$

Performing gaussian elimination on this matrix results in two rows of zeros.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Here we have three variables and one non-zero equation, thus we have two degrees of freedom. Since the non-zero equation involves all three variables, we are free to choose any consistent information we like in order to fully specify the solution. If the non-zero equation involved only a subset of the variables, we might be restricted when specifying free variables.

Selecting $x_1 = k_1$, and $x_2 = k_2$, the general solution is given by:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \\ 1 - k_1 - k_2 \end{pmatrix}$$

This general solution is structured in terms of a fully specified range part and a null-space part which is specified by the choice of additional information specifying the degrees of

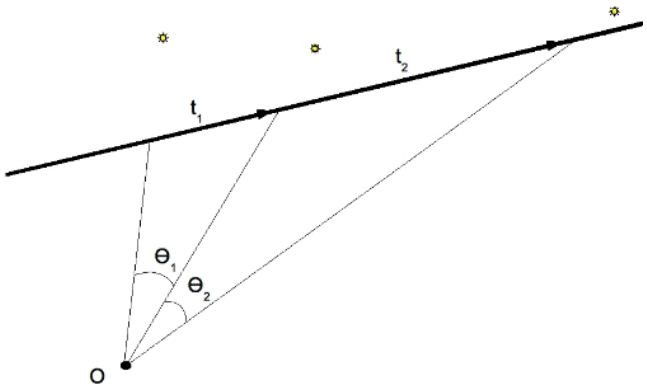

FIGURE 1. Comet Observables

freedom. In cases with multiple degrees of freedom, the null-space is multi-dimensional and can be represented by many different bases, each choice for the degrees of freedom implies a certain basis for the null-space and a particular representation of the solution. If we were to make a different selection for the degrees of freedom, we would obtain a different representation for the general solution.

3.4. Newton's Problem of a Comet's Path. (Adapted and Expanded from P'olya).
This problem was originally posed by Newton in his book on high school algebra. We will use it as an exercise on formulating and constructing linear systems. The problem stated succinctly is: *Determine a function which defines the trajectory of a comet traveling in a straight line at constant speed, using three observations.*

While this completely specifes the problem, a great deal of the complexity come from understanding and unraveling what Newton took to be "observations" of the comet. We suggest the reader read the explanatory paragraph that follows, then attempt to formulate and solve the problem themselves before considering the solution provided.

Terrestrial observations of celestial phenomena are typically made by using an array of background stars which appear fixed relative to local phenomena within the solar system. If one were to collect observations on a passing comet, one would measure the times at which the comet was coincident with three distinct background stars, and measure the difference in observed angle between those fixed stars. We imagine the problem schematically (See Fig 1.).

Given this information there is considerable work that needs to be done to encode this information into a linear system which can be solved to reveal the path of the comet. Based upon your knowledge of linear functions from high school, you may think that three

observations is too much information to determine a line, however note that we are only given knowledge of times and angles rather than the exact locations of the comet, therefore we will need to select a coordinate system and connect the given information to the chosen coordinate system.

Given that the comet is in motion relative to the observer, the observer's location is a reasonable choice for the origin of the coordinate system. Once we select this coordinate system, we impose the assumption that the comet path is a straight line traversed at constant speed to link the observations to the unknown parameters defining the comet's path.

$$C(t) = vt + p$$

Here v is the speed of the comet and p is the initial position of the comet.

4. Matrix Inversion

In this section we concern ourselves with the hand computation of the inverse matrix of a given square matrix A.

The inverse matrix effectively translates the canonical representation of an arbitrary output vector to the correct linear combination of input vectors which will produce that particular output. This allows one to translate a backward problem of linear algebra into a new forward problem, which allows the use of matrix multiplication to compute the answer.

Schematically we can begin by imagining an arbitrary right-hand side as being represented by a linear combination of the canonical basis vectors so:

$$\begin{aligned} A\mathbf{x} &= \mathbf{y} \\ A\mathbf{x} &= y_1\mathbf{e}_1 + y_2\mathbf{e}_2 + \cdots + y_n\mathbf{e}_n \end{aligned}$$

If we can find a solution for each of the basis vectors i.e. find a vector $\mathbf{s}_i$ satisfying $A\mathbf{s}_i = \mathbf{e}_i$ for each i from 1 to n, then we can use the linear properties of A to recover a representation formula for the general solution to this linear system in terms of these individual solutions.

$$\begin{aligned} A\mathbf{x} &= y_1\mathbf{e}_1 + y_2\mathbf{e}_2 + \cdots + y_n\mathbf{e}_n \\ A\mathbf{x} &= y_1A\mathbf{s}_1 + y_2A\mathbf{s}_2 + \cdots + y_nA\mathbf{s}_n \\ A\mathbf{x} &= Ay_1\mathbf{s}_1 + Ay_2\mathbf{s}_2 + \cdots + Ay_n\mathbf{s}_n \\ A\mathbf{x} &= A\left(y_1\mathbf{s}_1 + y_2\mathbf{s}_2 + \cdots + y_n\mathbf{s}_n\right) \\ \mathbf{x} &= y_1\mathbf{s}_1 + y_2\mathbf{s}_2 + \cdots + y_n\mathbf{s}_n \end{aligned}$$

Using this formula we find that by solving the n linear problems for the canonical basis vectors of the co-domain $A\mathbf{x} = \mathbf{e}_i$, we can extract a solution for **any** righthand side without solving additional linear systems. Since each of these linear systems involves the same matrix A, the row operations we use to solve each system will be identical. In order to avoid duplicating our efforts, then we can write an augmented system which solves each

of these systems simultaneously. If we order the canonical basis vectors numerically, this yields the following augmented matrix:

$$\left(\begin{array}{c|c} A & \mathbb{I} \end{array}\right)$$

By using row operations to systematically convert the matrix A into the identity matrix, we will find the resulting columns on the right hand side of this augmented system are exactly the individual solutions $\mathbf{s}_i$ found above. These columns together form the inverse matrix for the matrix A. In the computation below we demonstrate how to construct the general formula for the inverse of a two-by two matrix with arbitrary entries.

$$\begin{pmatrix} a & b & | & 1 & 0 \\ c & d & | & 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & \frac{b}{a} & | & \frac{1}{a} & 0 \\ c & d & | & 0 & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & \frac{b}{a} & | & \frac{1}{a} & 0 \\ 0 & \frac{ad-bc}{a} & | & -\frac{c}{a} & 1 \end{pmatrix}$$
$$\begin{pmatrix} 1 & \frac{b}{a} & | & \frac{1}{a} & 0 \\ 0 & 1 & | & -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$
$$\begin{pmatrix} 1 & 0 & | & \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ 0 & 1 & | & -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$
$$A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

As we might expect, we find that the determinant of this matrix must be non-zero for the inverse to be well defined.

Briefly summarizing the technique. To find the inverse matrix for a given square matrix, A set up an augmented system with A on the left and the appropriately sized identity matrix on the right:

$$\left(\begin{array}{c|c} A & \mathbb{I} \end{array}\right)$$

Perform row operations on the whole augmented system to convert the entries of A into the entries of the identity matrix. When this process is **complete** you will have an altered version of the same augmented system with the following structure:

$$\left(\begin{array}{c|c} \mathbb{I} & Q \end{array}\right)$$

By construction Q is the inverse matrix for the original matrix A, and more specifically, the columns of Q solve the individual problems:

$$A\mathbf{q}_{(:,j)} = \mathbf{e}_j$$

5. The Eigenvalue Problem

To solve the eigenvalue problem for an $n \times n$ matrix we must first find the eigenvalues by solving the algebraic equation:

$$\det(A - \lambda I) = 0$$

This is an n-th degree polynomial and may have anywhere between 0 and n distinct real roots. In order to completely solve the eigenvalue, a complete factorization of this characteristic equation must be obtained. Depending on context, complex eigenvalues may or may not be of interest. If the factorization involves multiple instances of the same eigenvalue (e.g. a term like $(\lambda - 1)^2$ or $(\lambda + 3)^3$) then the matrix has repeated eigenvalues. A matrix with repeated eigenvalues may or may not be 'defective.'

Once the eigenvalues are found, representative eigenvectors for each eigenvalue must be found. The eigenvector(s) for each eigenvalue, λ_j are found by selecting a basis for the null space of the matrix $A - \lambda_j I$. If the eigenvalue λ_j is a simple eigenvalue, then this null space will be one dimensional. If the eigenvalue is repeated m times in the factorization of the characteristic polynomial, then the null space may be up to m dimensional. In the case of repeated eigenvalues, anytime the null space corresponding to the repeated eigenvalue is less than m dimensional the matrix is termed 'defective.'

To illustrate this process we use several simple numerical examples

Example 1: Find the eigenvalues and eigenvectors of the matrix:

$$A = \begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}$$

To find the eigenvalues we set the determinant of $A - \lambda I$ equal to zero.

$$\begin{aligned} det(A - \lambda I) = \begin{vmatrix} 4 - \lambda & 3 \\ 1 & 2 - \lambda \end{vmatrix} &= 0 \\ (4 - \lambda)(2 - \lambda) - 3 &= 0 \\ \lambda^2 - 6\lambda + 8 - 3 &= 0 \\ (\lambda - 5)(\lambda - 1) &= 0 \end{aligned}$$

Solving this equation, we obtain two distinct, real eigenvalues.

$$\lambda_1 = 5, \quad \lambda_2 = 1$$

Now to find the corresponding eigenvectors we substitute each of these values into $A - \lambda I$ and search for vectors in the null space.

$\lambda = 5$

$$(A - \lambda I)\mathbf{x} = 0$$
$$\begin{pmatrix} 4-5 & 3 \\ 1 & 2-5 \end{pmatrix} \mathbf{x} = \begin{pmatrix} -1 & 3 \\ 1 & -3 \end{pmatrix} \mathbf{x} = 0$$

The eigenvalue makes $A - \lambda I$ a singular matrix, so we get an interesting null space. Any vector of the form:

$$x_1 - 3x_2 = 0, \quad k \begin{pmatrix} -3 \\ 1 \end{pmatrix},$$

will be an eigenvector for this eigenvalue.

Exercise 2.1: Repeat the same computations with λ_2 to find the other eigenvector.

Example 2:

Solve the eigenvalue problem for the following matrix:

$$\begin{pmatrix} 1 & 3 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

First we find the eigenvalues by solving $det(A - \lambda I) = 0$.

$$\begin{vmatrix} 1-\lambda & 3 & 0 \\ 3 & 1-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} = 0 - 0 + (2-\lambda) \begin{vmatrix} 1-\lambda & 3 \\ 3 & 1-\lambda \end{vmatrix}$$
$$(2-\lambda)[(1-\lambda)^2 - 9] = 0$$
$$(2-\lambda)[\lambda^2 - 2\lambda - 8] = 0$$
$$(2-\lambda)(\lambda - 4)(\lambda + 2) = 0$$

This yields three distinct eigenvalues. Take each eigenvalue, substitute into the matrix and solve for the resulting eigenvector(s) (which are vectors in the null space of $A - \lambda I$.

$\lambda = 2$

$$\begin{pmatrix} -1 & 3 & 0 \\ 3 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{x} = 0$$

Solving for the null space of this matrix we find the vector:

$$\mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

will span this null space.

$\lambda = 4$

$$\begin{pmatrix} -3 & 3 & 0 \\ 3 & -3 & 0 \\ 0 & 0 & 2 \end{pmatrix} \mathbf{x} = 0$$

Solving for the null space of this matrix we find the vector:

$$\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

will span this null space.

Exercise 2.2 Find the remaining eigenvector for this matrix.

Example 3 In the case of complex eigenvalues, the eigenvectors are found in an identical manner (however the vectors will be complex).

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

Solving for the eigenvalues of this matrix we get the following characteristic polynomial:

$$\lambda^2 - 4\lambda + 5 = 0$$

Using the quadratic formula on this equation yields:

$$\lambda = \frac{4 \pm \sqrt{16 - 4(5)}}{2}$$
$$\lambda = 2 \pm i$$

While these are unexpected, and don't have any easy interpretation in the context of linear algebra, they are very important in many applications. We can find the corresponding complex eigenvectors just as before: $\lambda = 2 - i$

$$\begin{pmatrix} i & -1 \\ 1 & i \end{pmatrix} \mathbf{x} = 0$$

These might look like completely different equations, but they turn out to be complex multiples of one another. (Multiply the first equation by $-i$ and you obtain exactly the second equation.) This means that any vector of the form:

$$ix_1 - x_2 = 0$$

Will satisfy both equations. We can take the complex vector:

$$\mathbf{x} = \begin{pmatrix} 1 \\ i \end{pmatrix}$$

Note: since the roots are complex conjugates, the corresponding eigenvectors must also be complex conjugates: $\lambda = 2 + i$

$$\mathbf{x} = \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

6. Diagonalization and the Jordon-Canonical Form

According to the Diagonalization theorem in Chapter 3, when given a matrix A possessing a full set of eigenvectors (i.e. a basis) the representation of the matrix A in the eigenbasis will be diagonal. A diagonal representation has many advantages, some computational, and some conceptual. In this section we give several explicit examples of the diagonalization process, and we show how the process may be modified in situations where the matrix is defective.

Example 1:

Consider the matrix below:

$$A = \begin{pmatrix} 1 & 2 & -2 \\ -1 & 3 & -1 \\ -1 & 0 & 2 \end{pmatrix}$$

If one solves the eigenvalue problem for this matrix, one obtains three distinct eigenvalues, and three corresponding eigenvectors.

$$\lambda_1 = 1, \qquad \lambda_2 = 2, \qquad \lambda_3 = 3,$$
$$\phi_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \phi_3 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix},$$

Since the eigenvectors form a basis for $\mathbb{R}^3$ it is legitimate to seek the representation of A under this basis. By playing with the order of the eigenvectors we can adjust the order of the eigenvalues in the diagonal representation. Let us define Φ_{123} as the following matrix:

$$\Phi_{123} \equiv \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Using matrix multiplication you can check for yourself that:

$$\Phi_{123}^{-1} A \Phi_{123} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

If we choose a different order for the eigenvectors, then the diagonal representation is altered as well. For example, define

$$\Phi_{132} \equiv \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Then the new representation of A becomes:

$$\Phi_{132}^{-1} A \Phi_{132} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Example 2:

Sometimes there are not enough true eigenvectors to form a basis. When this happens we call the underlying matrix 'defective.' Defective matricies do not have diagonal representations, but we can find a simple representation by using 'generalized eigenvectors.' we briefly illustrate the process with an example.

Consider the matrix:

$$D = \begin{pmatrix} -1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

If one tries to solve the eigenvalue problem for this matrix one obtains the following characteristic polynomial which determines the eigenvalues:

$$(1-\lambda)^2(-1-\lambda) = 0$$

left

This polynomial only possesses two different eigenvalues. When searching for the eigenvectors, we find one eigenvector for each of the eigenvalues. This gives us only two linearly independent eigenvectors which are not sufficient to form a basis for $\mathbb{R}^3$. We seek a third vector to complete this basis. We can find a generalized eigenvector by solving one additional problem.

The eigenvectors were obtained by solving the problem:

$$(D - \lambda_i I)\phi_i = 0$$

in order to complete our basis we seek a generalized eigenvector by using the true eigenvector we already have. In general if λ_j is a repeated eigenvalue which is missing one or more corresponding eigenvectors we find a generalized eigenvector by solving the problem:

$$(D - \lambda_j I)\mathbf{g} = \phi_j$$

That is we seek a vector which 'turns into' an eigenvector when multiplied by the matrix $(D - \lambda_j I)$. This choice is guaranteed to be linearly independent with the true eigenvector, and has many of the same computational advantages. Returning to the example at hand, we find the following true eigenvalue, eigenvector pairs:

$$\lambda_1 = 1, \qquad \lambda_2 = -1$$

$$\phi_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix},$$

Since λ_1 was the repeated eigenvalue, we use it's corresponding eigenvector to seek a generalized eigenvector. We need to solve the following problem:

$$(D - 1I)\mathbf{g} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} -2 & 0 & 0 \\ -2 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \mathbf{g} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

This problem has the general solution:

$$\mathbf{g} = \begin{pmatrix} 0 \\ 1 \\ k \end{pmatrix}$$

so, for example we can select $k = 0$, and we obtain a candidate for our generalized eigenvector. Using this in conjunction with our eigenvectors we can attempt a change of bases, to see how this will affect the representation of the matrix D. Let:

$$\Phi_{1g-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

Then we obtain:

$$\Phi_{1g-1}^{-1} D \Phi_{1g-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

This representation has the eigenvalues along the diagonal, and possesses a single off-diagonal entry corresponding to the generalized eigenvector.

Remark: This process generalizes quite naturally to situations where one has multiple missing eigenvectors. The use of generalized eigenvectors just guarantees the new representation of the matrix is as simple as possible.

Example:
A good non-trivial example of a defective matrix is given by:

$$D = \begin{pmatrix} 6 & 3 & 0 \\ 1 & 4 & 1 \\ -1 & -1 & 5 \end{pmatrix}$$

This matrix has the (algebraically simplified) characteristic polynomial:

$$(\lambda - 3)(\lambda - 6)^2 = 0$$

and the repeated eigenvalue only has a single corresponding eigenvector. (Note: if you analyze this matrix in Matlab, it incorrectly identifies three eigenvalues.)

We find the generalized eigenvector by solving:

$$(D - 6I)\mathbf{g} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \tag{1}$$

Two possible solutions are:

$$\mathbf{g} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}, \text{ or } \mathbf{g} = \begin{pmatrix} 2/3 \\ 1/3 \\ 0 \end{pmatrix}$$

Note that since $D - \lambda I$ is singular you will always get an infinite family of solutions. Picking one that results in the eigenvector for the repeated eigenvalue generates a linearly independent choice which can be used to complete the basis. Making the output equal to the chosen eigenvector is what produces the value of "1" on the off diagonal. Ordering the eigenvectors before the generalized eigenvectors places the off-diagonal entry above the diagonal where the eigenvalues appear.

Example 3: Computational advantages of diagonal representation.
The diagonal representation of a matrix has many computational advantages, but one of the easiest to illustrate is the advantage gained over repeated matrix multiplication.

The following result is always valid: If two matrices are similar, so that there exists a change of bases transforming one matrix into the other, then **all** powers of those matrices are similarly related.

$$\Phi^{-1}A\Phi = B, \rightarrow \Phi^{-1}A^n\Phi = B^n \quad \forall n \in \mathbb{N}$$

This is especially useful if B is diagonal.

Consider trying to compute the following matrix-vector multiplication problem:

$$A^n\mathbf{x}$$

If A is a large matrix, and n is a large integer then performing repeated matrix multiplication will require many computations, but If A has a diagonal representation, then we can reduce all of these calculations to a single change of bases computation, and a single matrix multiplication. First we change bases to the eigenbasis for A.

$$\Phi^{-1}A^n\Phi\hat{\mathbf{x}}$$

We know that $\Phi^{-1}A\Phi = \Lambda$, a diagonal matrix with the eigenvalues along the diagonal, so using the result above, we find:

$$\Phi^{-1}A^n\Phi\hat{\mathbf{x}} = \Lambda^n\hat{\mathbf{x}}$$

And direct computation tells us that:

$$\Lambda^n = \begin{pmatrix} \lambda_1^n & 0 & 0 & \dots \\ 0 & \lambda_2^n & 0 & \dots \\ 0 & 0 & \lambda_3^n & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

So the result of the computation, which might have seemed insurmountable, actually becomes very easy when expressed in the eigenbasis. If we know the formulae for the eigenvectors ϕ_i, in the eigenbasis we can even write down a formula for the result which is valid for each n.

$$A^n\mathbf{x} = \hat{x}_1\lambda_1^n\boldsymbol{\phi}_1 + \hat{x}_2\lambda_2^n\boldsymbol{\phi}_2 + \hat{x}_3\lambda_3^n\boldsymbol{\phi}_3 + \cdots + \hat{x}_m\lambda_m^n\boldsymbol{\phi}_m.$$

So by solving the eigenvalue problem for the matrix A, the linear system:

$$\Phi\hat{\mathbf{x}} = \mathbf{x},$$

to find the representation for the vector $\mathbf{x}$ in the eigenbasis, we can trivially solve **any** matrix multiplication problem with powers of A.

7. Singular Valued Decomposition and its Applications

We have seen in the previous section that not all matrices can be diagonalized. In this section we investigate a generalization of diagonalization which works for non-square matrices. This technique also yields a powerful conceptual decomposition of a matrix which yields the solution to certain optimization problems. We begin by outlining the procedure used to find the singular valued decomposition, then we explore some of the applications.

A singular valued decomposition of a matrix is a decomposition into three distinct matricies:

$$A = U\Sigma V^T$$

where Σ is a diagonal matrix with the same size as A, and U and V are both square orthogonal matricies.

To find the parts of this decomposition one first constructs the matrix A^TA, this square symmetric matrix is positive definite and possesses a full set of mutually orthogonal eigenvectors. One solves the eigenvalue problem for the matrix A^TA and records both the eigenvalues and the eigenvectors.

One then orders the eigenvalues from largest to smallest and constructs the singular values by taking the square roots of each eigenvalue. The matrix Σ is then constructed as follows:

$$\Sigma = \begin{pmatrix} \sqrt{\lambda_1} & 0 & 0 & \dots \\ 0 & \sqrt{\lambda_2} & 0 & \dots \\ 0 & 0 & \sqrt{\lambda_3} & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

again, the size of Σ should be identical to A. The matrix V is constructed from the normalized eigenvectors of A^TA. Since the eigenvectors of symmetric matrices are mutually orthogonal, by normalizing these vectors one obtains an orthogonal matrix, this is important computationally, because the V^T appearing in the decomposition is actually the inverse of V.

Once Σ and V are known, one can construct the columns of U by computing the following:

$$\mathbf{u}_i = \frac{A\mathbf{v}_i}{\|A\mathbf{v}_i\|}$$

thus each $\mathbf{u}_i$ is the normalized output direction for each of the columns of V.

Conceptually the singular valued decomposition introduces a coordinate system which is aligned with the directions of maximal and minimal stretching induced by the matrix A. We can gain a great deal of insight into the singular value decomposition by looking at its behavior geometrically.

Consider a generic, non-singular 2×2 matrix. If we consider the set of all unit vectors in $\mathbb{R}^2$ we can visualize this as the unit circle. By acting on all of these points with the matrix A and plotting the outputs, we will in general see an ellipse, centered at the origin, but not necessarily align with our traditional coordinate axes. See figure 1. The components of the singular value decomposition align with the different features of this graph. The columns of V (which are the rows of V^T) are the eigenvectors of A^TA, these are not the eigenvectors of A (unless A is symmetric), but these eigenvectors **do** correspond to the semi-major and semi-minor axes of the ellipse of outputs. Thus the eigenvectors for A^TA define an orthonormal coordinate system which is align with the directions which experience maximal and minimal stretching under the transformation induced by A. In Figure 2, the eigenvectors of A^TA are drawn in blue, note that they are in the same location in both figures, as we can think of this as a change of coordinate system for both the inputs and outputs. The singular values describe the length of the semi-major axis and semi minor axis of the ellipse of outputs.

The singular valued decomposition has powerful applications in error analysis, data reduction, visualization and data compression. Since the singular valued decomposition provides a coordinate system which captures and orders the stretching of a transformation, one can use it to obtain quick, globally valid, worst case scenario error estimates of measurements under linear mappings.

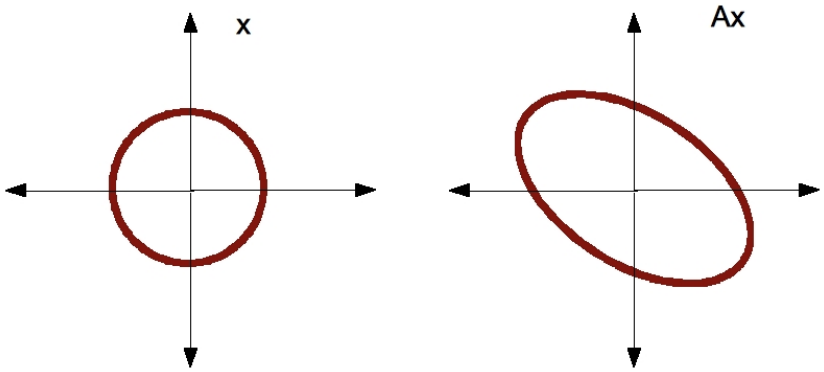

FIGURE 2. Input and Output Values under a non-singular matrix

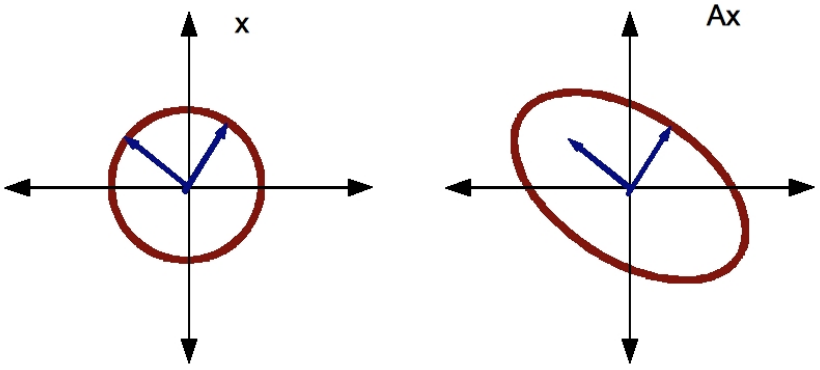

FIGURE 3. Coordinate system defined by the SVD