Basic Vector Space Methods in Signal and Systems Theory

By:

C. Sidney Burrus

Basic Vector Space Methods in Signal and Systems Theory

By:
C. Sidney Burrus

Online:
< http://cnx.org/content/col10636/1.5/ >

C O N N E X I O N S

Rice University, Houston, Texas

Collection structure revised: December 19, 2012
PDF generated: January 10, 2013
For copyright and attribution information for the modules contained in this collection, see p. 35.

Table of Contents

Chapter 1

Introduction[1]

1.1 Introduction

The tools, ideas, and insights from linear algebra, abstract algebra, and functional analysis can be extremely useful to signal processing and system theory in various areas of engineering, science, and social science. Indeed, many important ideas can be developed from the simple operator equation

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.1}$$

by considering it in a variety of ways. If $\mathbf{x}$ and $\mathbf{b}$ are vectors from the same or, perhaps, different vector spaces and $\mathbf{A}$ is an operator, there are three interesting questions that can be asked which provide a setting for a broad study.

1. Given $\mathbf{A}$ and $\mathbf{x}$, find $\mathbf{b}$. The analysis or operator problem or transform.
2. Given $\mathbf{A}$ and $\mathbf{b}$, find $\mathbf{x}$. The inverse or control problem or deconvolution or design.
3. Given $\mathbf{x}$ and $\mathbf{b}$, find $\mathbf{A}$. The synthesis or design problem or parameter identification.

Much can be learned by studying each of these problems in some detail. We will generally look at the finite dimensional problem where (1.1) can more easily be studied as a finite matrix multiplication [82], [84], [65], [87]

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & & \\ a_{31} & a_{32} & a_{33} & & \\ \vdots & & & & \vdots \\ a_{M1} & & & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{bmatrix} \tag{1.2}$$

but will also try to indicate what the infinite dimensional case might be [49], [94], [71], [68].

An application to signal theory is in [44], to optimization [61], and multiscale system theory [8]. The inverse problem (number 2 above) is the basis for a large study of pseudoinverses, approximation, optimization, filter design, and many applications. When used with the l_2 norm [59], [10] powerful results can be optained analytically but used with other norms such as l_∞, l_1, l_0 (a pseudonorm), an even larger set of problems can be posed and solved [2], [5].

A development of vector space ideas for the purpose of presenting wavelet representations is given in [30], [23]. An interesting idea of *unconditional bases* is given by Donoho [36].

[1] This content is available online at <http://cnx.org/content/m19560/1.4/>.

Linear regression analysis can be posed in the form of (1.1) and (1.2) where the $\mathbf{M}$ rows of $\mathbf{A}$ are the vectors of input data from $\mathbf{M}$ experiments, entries of $\mathbf{x}$ are the $\mathbf{N}$ weights for the $\mathbf{N}$ components of the inputs, and the $\mathbf{M}$ values of $\mathbf{b}$ are the outputs [2]. This can be used in machine learning problems [9], [52]. A problem similar to the design or synthesis problem is that of parameter identification where a model of some system is posed with unknown parameters. Then experiments with known inputs and measured outputs are run to identify these parameters. Linear regression is also an example of this [2], [9].

Dynamic systems are often modelled by ordinary differential equation where $\mathbf{b}$ is set to be the time derivative of $\mathbf{x}$ to give what are called the linear state equations:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{1.3}$$

or for difference equations and discrete-time or digital signals,

$$\mathbf{x}(n+1) = \mathbf{A}\,\mathbf{x}(n) \tag{1.4}$$

which are used in digital signal processing and the analysis of certain algorithms. State equations are useful in feedback control as well as in simulation of many dynamical systems and the eigenvalues and other properties of the square matix $\mathbf{A}$ are important indicators of the performance [97], [35].

The ideas of similarity transformations, diagonalization, the eigenvalue problem, Jordon normal form, singular value decomposition, etc. from linear algebra [82], [84], [53] are applicable to this problem.

Various areas in optimization and approximation use vector space math to great advantage [61], [59].

This booklet is intended to point out relationships, interpretations, and tools in linear algebra, matrix theory, and vector spaces that scientists and engineers might find useful. It is not a stand-alone linear algebra book. Details, definitions, and formal proofs can be found in the references. A very helpful source is Wikipedia.

There is a variety software systems to both pose and solve linear algebra problems. A particularly powerful one is Matlab [65] which is, in some ways, the gold standard since it started years ago a purely numerical matrix package. But there are others such as Octave, SciLab, LabVIEW, Mathematica, Maple, etc.

Chapter 2

A Matrix Times a Vector[1]

2.1 A Matrix Times a Vector

In this chapter we consider the first problem posed in the introduction

$$\mathbf{Ax} = \mathbf{b} \tag{2.1}$$

where the matrix $\mathbf{A}$ and vector $\mathbf{x}$ are given and we want to interpret and give structure to the calculation of the vector $\mathbf{b}$. Equation (2.1) has a variety of special cases. The matrix $\mathbf{A}$ may be square or may be rectangular. It may have full column or row rank or it may not. It may be symmetric or orthogonal or non-singular or many other characteristics which would be interesting properties as an operator. If we view the vectors as signals and the matrix as an operator or processor, there are two interesting interpretations.

- The operation (2.1) is a change of basis or coordinates for a fixed signal. The signal stays the same, the basis (or frame) changes.
- The operation (2.1) alters the characteristics of the signal (processes it) but within a fixed basis system. The basis stays the same, the signal changes.

An example of the first would be the discrete Fourier transform (DFT) where one calculates frequency components of a signal which are coordinates in a frequency space for a given signal. The definition of the DFT from [22] can be written as a matrix-vector operation by $\mathbf{c} = \mathbf{Wx}$ which, for $w = e^{-j2\pi/N}$ and $N = 4$, is

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} w^0 & w^0 & w^0 & w^0 \\ w^0 & w^1 & w^2 & w^3 \\ w^0 & w^2 & w^4 & w^6 \\ w^0 & w^3 & w^6 & w^9 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{2.2}$$

An example of the second might be convolution where you are processing or filtering a signal and staying in the same space or coordinate system.

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} h_0 & 0 & 0 & \cdots & 0 \\ h_1 & h_0 & 0 & & \\ h_2 & h_1 & h_0 & & \\ \vdots & & & & \vdots \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix}. \tag{2.3}$$

[1] This content is available online at <http://cnx.org/content/m19559/1.7/>.

A particularly powerful sequence of operations is to first change the basis for a signal, then process the signal in this new basis, and finally return to the original basis. For example, the discrete Fourier transform (DFT) of a signal is taken followed by setting some of the Fourier coefficients to zero followed by taking the inverse DFT.

Another application of (2.1) is made in linear regression where the input signals are rows of $\mathbf{A}$ and the unknown weights of the hypothesis are in $\mathbf{x}$ and the outputs are the elements of $\mathbf{b}$.

2.2 Change of Basis

Consider the two views:

1. The operation given in (2.1) can be viewed as $\mathbf{x}$ being a set of weights so that $\mathbf{b}$ is a weighted sum of the columns of $\mathbf{A}$. In other words, $\mathbf{b}$ will lie in the space spanned by the columns of $\mathbf{A}$ at a location determined by $\mathbf{x}$. This view is a composition of a signal from a set of weights as in (2.6) and (2.8) below. If the vector $\mathbf{a_i}$ is the i^{th} column of $\mathbf{A}$, it is illustrated by

$$\mathbf{Ax} = x_1 \begin{bmatrix} \vdots \\ \mathbf{a_1} \\ \vdots \end{bmatrix} + x_2 \begin{bmatrix} \vdots \\ \mathbf{a_2} \\ \vdots \end{bmatrix} + x_3 \begin{bmatrix} \vdots \\ \mathbf{a_3} \\ \vdots \end{bmatrix} = \mathbf{b}. \tag{2.4}$$

2. An alternative view has $\mathbf{x}$ being a signal vector and with $\mathbf{b}$ being a vector whose entries are inner products of $\mathbf{x}$ and the rows of $\mathbf{A}$. In other words, the elements of $\mathbf{b}$ are the projection coefficients of $\mathbf{x}$ onto the coordinates given by the rows of $\mathbf{A}$. The multiplication of a signal by this operator decomposes the signal and gives the coefficients of the decomposition. If $\bar{\mathbf{a}}_\mathbf{j}$ is the j^{th} row of $\mathbf{A}$ we have:

$$b_1 = \begin{bmatrix} \cdots \bar{\mathbf{a}}_\mathbf{1} \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{x} \\ \vdots \end{bmatrix} \qquad b_2 = \begin{bmatrix} \cdots \bar{\mathbf{a}}_\mathbf{2} \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{x} \\ \vdots \end{bmatrix} \qquad etc. \tag{2.5}$$

   Regression can be posed from this view with the input signal being the rows of $\mathbf{A}$.

These two views of the operation as a decomposition of a signal or the recomposition of the signal to or from a different basis system are extremely valuable in signal analysis. The ideas from linear algebra of subspaces, inner product, span, orthogonality, rank, etc. are all important here. The dimensions of the domain and range of the operators may or may not be the same. The matrices may or may not be square and may or may not be of full rank [50], [85].

2.2.1 A Basis and Dual Basis

A set of linearly independent vectors $\mathbf{x_n}$ forms a *basis* for a vector space if every vector $\mathbf{x}$ in the space can be uniquely written

$$\mathbf{x} = \sum_n a_n \, \mathbf{x_n} \tag{2.6}$$

and the *dual basis* is defined as a set vectors $\tilde{\mathbf{x}}_\mathbf{n}$ in that space allows a simple inner product (denoted by parenthesis: $(\mathbf{x}, \mathbf{y})$) to calculate the expansion coefficients as

$$a_n = \ (\mathbf{x}, \tilde{\mathbf{x}}_\mathbf{n}) \ = \mathbf{x}^\mathbf{T} \tilde{\mathbf{x}}_\mathbf{n} \tag{2.7}$$

A basis expansion has enough vectors but none extra. It is efficient in that no fewer expansion vectors will represent all the vectors in the space but is fragil in that losing one coefficient or one basis vector destroys the ability to exactly represent the signal by (2.6). The expansion (2.6) can be written as a matrix operation

$$\mathbf{F}\,\mathbf{a} = \mathbf{x} \tag{2.8}$$

where the columns of $\mathbf{F}$ are the basis vectors $\mathbf{x_n}$ and the vector $\mathbf{a}$ has the expansion coefficients a_n as entries. Equation (2.7) can also be written as a matrix operation

$$\tilde{\mathbf{F}}\,\mathbf{x} = \mathbf{a} \tag{2.9}$$

which has the dual basis vectors as rows of $\tilde{\mathbf{F}}$. From (2.8) and (2.9), we have

$$\mathbf{F}\tilde{\mathbf{F}}\,\mathbf{x} = \mathbf{x} \tag{2.10}$$

Since this is true for all $\mathbf{x}$,

$$\mathbf{F}\,\tilde{\mathbf{F}} = \mathbf{I} \tag{2.11}$$

or

$$\tilde{\mathbf{F}} = \mathbf{F^{-1}} \tag{2.12}$$

which states the dual basis vectors are the rows of the inverse of the matrix whose columns are the basis vectors (and vice versa). When the vector set is a basis, $\mathbf{F}$ is necessarily square and from (2.8) and (2.9), one can show

$$\mathbf{F}\,\tilde{\mathbf{F}} = \tilde{\mathbf{F}}\,\mathbf{F}. \tag{2.13}$$

Because this system requires two basis sets, the expansion basis and the dual basis, it is called *biorthogonal*.

2.2.2 Orthogonal Basis

If the basis vectors are not only independent but orthonormal, the basis set is its own dual and the inverse of $\mathbf{F}$ is simply its transpose.

$$\mathbf{F^{-1}} = \tilde{\mathbf{F}} = \mathbf{F^T} \tag{2.14}$$

When done in Hilbert spaces, this decomposition is sometimes called an abstract Fourier expansion [50], [48], [93].

2.2.3 Parseval's Theorem

Because many signals are digital representations of voltage, current, force, velocity, pressure, flow, etc., the inner product of the signal with itself (the norm squared) is a measure of the signal energy q.

$$q = (\mathbf{x}, \mathbf{x}) = ||\mathbf{x}||^{\mathbf{2}} = \mathbf{x^T x} = \sum_{n=0}^{N-1} x_n^2 \tag{2.15}$$

Parseval's theorem states that if the basis system is orthogonal, then the norm squared (or "energy") is invarient across a change in basis. If a change of basis is made with

$$\mathbf{c} = \mathbf{A}\mathbf{x} \tag{2.16}$$

then

$$q = (\mathbf{x}, \mathbf{x}) = ||\mathbf{x}||^{\mathbf{2}} = \mathbf{x}^{\mathbf{T}}\mathbf{x} = \sum_{n=0}^{N-1} x_n^2 = K\,(\mathbf{c}, \mathbf{c}) = K||\mathbf{c}||^{\mathbf{2}} = K\mathbf{c}^{\mathbf{T}}\mathbf{c} = K\sum_{k=0}^{N-1} c_k^2 \tag{2.17}$$

for some constant K which can be made unity by normalization if desired.

For the discrete Fourier transform (DFT) of x_n which is

$$c_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-j2\pi nk/N} \tag{2.18}$$

the energy calculated in the time domain: $q = \sum_n x_n^2$ is equal to the norm squared of the frequency coefficients: $q = \sum_k c_k^2$, within a multiplicative constant of $1/N$. This is because the basis functions of the Fourier transform are orthogonal: "the sum of the squares is the square of the sum" which means means the energy calculated in the time domain is the same as that calculated in the frequency domain. The energy of the signal (the square of the sum) is the sum of the energies at each frequency (the sum of the squares). Because of the orthogonal basis, the cross terms are zero. Although one seldom directly uses Parseval's theorem, its truth is what make sense in talking about frequency domain filtering of a time domain signal. A more general form is known as Plancherel theorem [29].

If a transformation is made on the signal with a non-orthogonal basis system, then Parseval's theorem does not hold and the concept of energy does not move back and forth between domains. We can get around some of these restrictions by using frames rather than bases.

2.2.4 Frames and Tight Frames

In order to look at a more general expansion system than a basis and to generalize the ideas of orthogonality and of energy being calculated in the original expansion system or the transformed system, the concept of *frame* is defined. A frame decomposition or representation is generally more robust and flexible than a basis decomposition or representation but it requires more computation and memory [1], [91], [29]. Sometimes a frame is called a redundant basis or representing an underdetermined or underspecified set of equations.

If a set of vectors, $\mathbf{f_k}$, span a vector space (or subspace) but are not necessarily independent nor orthogonal, bounds on the energy in the transform can still be defined. A set of vectors that span a vector space is called a *frame* if two constants, A and B exist such that

$$0 < A||\mathbf{x}||^2 \leq \sum_k |\,(\mathbf{f_k}, \mathbf{x})\,|^2 \leq B||\mathbf{x}||^2 < \infty \tag{2.19}$$

and the two constants are called the *frame bounds* for the system. This can be written

$$0 < A||\mathbf{x}||^2 \leq ||\mathbf{c}||^2 \leq B||\mathbf{x}||^2 < \infty \tag{2.20}$$

where

$$\mathbf{c} = \mathbf{F}\mathbf{x} \tag{2.21}$$

If the $\mathbf{f_k}$ are linearly independent but not orthogonal, then the frame is a non-orthogonal basis. If the $\mathbf{f_k}$ are not independent the frame is called redundant since there are more than the minimum number of expansion vectors that a basis would have. If the frame bounds are equal, $A = B$, the system is called a *tight frame* and it has many of features of an orthogonal basis. If the bounds are equal to each other and to one, $A = B = 1$, then the frame is a basis and is tight. It is, therefore, an orthogonal basis.

So a frame is a generalization of a basis and a tight frame is a generalization of an orthogonal basis. If , $A = B$, the frame is tight and we have a scaled Parseval's theorem:

$$A||\mathbf{x}||^2 = \sum_k |\,(\mathbf{f_k}, \mathbf{x})\,|^2 \tag{2.22}$$

If $A = B > 1$, then the number of expansion vectors are more than needed for a basis and A is a measure of the redundancy of the system (for normalized frame vectors). For example, if there are three frame vectors in a two dimensional vector space, $A = 3/2$.

A finite dimensional matrix version of the redundant case would have $\mathbf{F}$ in (2.8) with more columns than rows but with full row rank. For example

$$\begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{2.23}$$

has three frame vectors as the columns of $\mathbf{A}$ but in a two dimensional space.

The prototypical example is called the Mercedes-Benz tight frame where three frame vectors that are 120° apart are used in a two-dimensional plane and look like the Mercedes car hood ornament. These three frame vectors must be as far apart from each other as possible to be tight, hence the 120° separation. But, they can be rotated any amount and remain tight [91], [57] and, therefore, are not unique.

$$\begin{bmatrix} 1 & -0.5 & -0.5 \\ 0 & 0.866 & -0.866 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \tag{2.24}$$

In the next section, we will use the pseudo-inverse of $\mathbf{A}$ to find the optimal $\mathbf{x}$ for a given $\mathbf{b}$.

So the frame bounds A and B in (2.19) are an indication of the redundancy of the expansion system f_k and to how close they are to being orthogonal or tight. Indeed, (2.19) is a sort of approximate Parseval's theorem [95], [31], [73], [56], [29], [91], [42], [58].

The dual frame vectors are also not unique but a set can be found such that (2.9) and, therefore, (2.10) hold (but (2.13) does not). A set of dual frame vectors could be found by adding a set of arbitrary but independent rows to $\mathbf{F}$ until it is square, inverting it, then taking the first N columns to form $\tilde{\mathbf{F}}$ whose rows will be a set of dual frame vectors. This method of construction shows the non-uniqueness of the dual frame vectors. This non-uniqueness is often resolved by optimizing some other parameter of the system [31].

If the matrix operations are implementing a frame decomposition and the rows of $\mathbf{F}$ are orthonormal, then $\tilde{\mathbf{F}} = \mathbf{F}^\mathbf{T}$ and the vector set is a tight frame [95], [31]. If the frame vectors are normalized to $||\mathbf{x_k}|| = 1$, the decomposition in (2.6) becomes

$$\mathbf{x} = \frac{1}{A} \sum_n (\mathbf{x}, \tilde{\mathbf{x}}_\mathbf{n}) \ \mathbf{x_n} \tag{2.25}$$

where the constant A is a measure of the redundancy of the expansion which has more expansion vectors than necessary [31].

The matrix form is

$$\mathbf{x} = \frac{1}{A} \, \mathbf{F} \, \mathbf{F}^\mathbf{T} \, \mathbf{x} \tag{2.26}$$

where $\mathbf{F}$ has more columns than rows. Examples can be found in [24].

2.2.5 Sinc Expansion as a Tight Frame

The Shannon sampling theorem [20] can be viewied as an infinite dimensional signal expansion where the *sinc* functions are an orthogonal basis. The sampling theorem with critical sampling, i.e. at the Nyquist rate, is the expansion:

$$g(t) \;=\; \sum_n g(Tn) \, \frac{sin\left(\frac{\pi}{T}(t - Tn)\right)}{\frac{\pi}{T}(t - Tn)} \tag{2.27}$$

where the expansion coefficients are the samples and where the *sinc* functions are easily shown to be orthogonal.

Over sampling is an example of an infinite-dimensional tight frame [64], [24]. If a function is over-sampled but the sinc functions remains consistent with the upper spectral limit W, using A as the amount of over-sampling, the sampling theorem becomes:

$$AW = \frac{\pi}{T}, \qquad \text{for } A \geq 1 \tag{2.28}$$

and we have

$$g(t) = \frac{1}{A}\sum_n g(Tn)\,\frac{sin\left(\frac{\pi}{AT}(t-Tn)\right)}{\frac{\pi}{AT}(t-Tn)} \tag{2.29}$$

where the sinc functions are no longer orthogonal. In fact, they are no longer a basis as they are not independent. They are, however, a tight frame and, therefore, have some of the characteristics of an orthogonal basis but with a "redundancy" factor A as a multiplier in the formula [24] and a generalized Parseval's theorem. Here, moving from a basis to a frame (actually from an orthogonal basis to a tight frame) is almost invisible.

2.2.6 Frequency Response of an FIR Digital Filter

The discrete-time Fourier transform (DTFT) of the impulse response of an FIR digital filter $h(n)$ is its frequency response. The discrete Fourier transform (DFT) of $h(n)$ gives samples of the frequency response [20]. This is a powerful analysis tool in digital signal processing (DSP) and suggests that an inverse (or pseudoinverse) method could be useful for design [20].

2.2.7 Conclusions

Frames tend to be more robust than bases in tolerating errors and missing terms. They allow flexibility is designing wavelet systems [31] where frame expansions are often chosen.

In an infinite dimensional vector space, if basis vectors are chosen such that all expansions converge very rapidly, the basis is called an *unconditional basis* and is near optimal for a wide class of signal representation and processing problems. This is discussed by Donoho in [37].

Still another view of a matrix operator being a change of basis can be developed using the eigenvectors of an operator as the basis vectors. Then a signal can decomposed into its eigenvector components which are then simply multiplied by the scalar eigenvalues to accomplish the same task as a general matrix multiplication. This is an interesting idea but will not be developed here.

2.3 Change of Signal

If both $\mathbf{x}$ and $\mathbf{b}$ in (2.1) are considered to be signals in the same coordinate or basis system, the matrix operator $\mathbf{A}$ is generally square. It may or may not be of full rank and it may or may not have a variety of other properties, but both $\mathbf{x}$ and $\mathbf{b}$ are viewed in the same coordinate system and therefore are the same size.

One of the most ubiquitous of these is convolution where the input to a linear, shift invariant system with impulse response $h(n)$ is calculated by (2.1) if $\mathbf{A}$ is the convolution matrix and $\mathbf{x}$ is the input [20].

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} h_0 & 0 & 0 & \cdots & 0 \\ h_1 & h_0 & 0 & & \\ h_2 & h_1 & h_0 & & \\ \vdots & & & & \vdots \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix}. \tag{2.30}$$

It can also be calculated if $\mathbf{A}$ is the arrangement of the input and $\mathbf{x}$ is the the impulse response.

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} x_0 & 0 & 0 & \cdots & 0 \\ x_1 & x_0 & 0 & & \\ x_2 & x_1 & x_0 & & \\ \vdots & & & & \vdots \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \end{bmatrix}. \tag{2.31}$$

If the signal is periodic or if the DFT is being used, then what is called a *circulate* is used to represent cyclic convolution. An example for $N = 4$ is the Toeplitz system

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} h_0 & h_3 & h_2 & h_1 \\ h_1 & h_0 & h_3 & h_2 \\ h_2 & h_1 & h_0 & h_3 \\ h_3 & h_2 & h_1 & h_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{2.32}$$

One method of understanding and generating matrices of this sort is to construct them as a product of first a decomposition operator, then a modification operator in the new basis system, followed by a recomposition operator. For example, one could first multiply a signal by the DFT operator which will change it into the frequency domain. One (or more) of the frequency coefficients could be removed (set to zero) and the remainder multiplied by the inverse DFT operator to give a signal back in the time domain but changed by having a frequency component removed. That is a form of signal filtering and one can talk about removing the energy of a signal at a certain frequency (or many) because of Parseval's theorem.

It would be instructive for the reader to make sense out of the cryptic statement “the DFT diagonalizes the cyclic convolution matrix" to add to the ideas in this note.

2.4 Factoring the Matrix A

For insight, algorithm development, and/or computational efficiency, it is sometime worthwhile to factor $\mathbf{A}$ into a product of two or more matrices. For example, the *DFT* matrix [22] illustrated in (2.2) can be factored into a product of fairly sparce matrices. If fact, the fast Fourier transform (FFT) can be derived by factoring the DFT matrix into $Nlog\left(N\right)$ factors (if $N = 2^m$), each requiring order N multiplies. This is done in [22].

Using eigenvalue theory [85], a full rank square matrix can be factored into a product

$$\mathbf{AV} = \mathbf{V}\Lambda \tag{2.33}$$

where $\mathbf{V}$ is a matrix with columns of the eigenvectors of $\mathbf{A}$ and Λ is a diagonal matrix with the eigenvalues along the diagonal. The inverse is a method to “diagonalize" a matrix

$$\Lambda = \mathbf{V^{-1}AV} \tag{2.34}$$

If a matrix has “repeated eigenvalues", in other words, two or more of the N eigenvalues have the same value but less than N indepentant eigenvectors, it is not possible to diagonalize the matrix but an “almost" diagonal form called the *Jordan normal form* can be acheived. Those details can be found in most books on matrix theory [83].

A more general decompostion is the singular value decomposition (SVD) which is similar to the eigenvalue problem but allows rectangular matrices. It is particularly valuable for expressing the pseudoinverse in a simple form and in making numerical calculations [88].

2.5 State Equations

If our matrix multiplication equation is a vector differential equation (DE) of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{2.35}$$

or for difference equations and discrete-time signals or digital signals,

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) \tag{2.36}$$

an inverse or even pseudoinverse will not solve for $\mathbf{x}$. A different approach must be taken [41] and different properties and tools from linear algebra will be used. The solution of this first order vector DE is a coupled set of solutions of first order DEs. If a change of basis is made so that $\mathbf{A}$ is diagonal (or Jordan form), equation (2.35) becomes a set on uncoupled (or almost uncoupled in the Jordan form case) first order DEs and we know the solution of a first order DE is an exponential. This requires consideration of the eigenvalue problem, diagonalization, and solution of scalar first order DEs [41].

State equations are often used to model or describe a system such as a control system or a digital filter or a numerical algorithm [41], [98].

Chapter 3

General Solutions of Simultaneous Equations[1]

The second problem posed in the introduction is basically the solution of simultaneous linear equations [60], [3], [6] which is fundamental to linear algebra [54], [86], [66] and very important in diverse areas of applications in mathematics, numerical analysis, physical and social sciences, engineering, and business. Since a system of linear equations may be over or under determined in a variety of ways, or may be consistent but ill conditioned, a comprehensive theory turns out to be more complicated than it first appears. Indeed, there is a considerable literature on the subject of *generalized inverses* or *pseudo-inverses*. The careful statement and formulation of the general problem seems to have started with Moore [69] and Penrose [74], [75] and developed by many others. Because the generalized solution of simultaneous equations is often defined in terms of minimization of an equation error, the techniques are useful in a wide variety of approximation and optimization problems [11], [62] as well as signal processing.

The ideas are presented here in terms of finite dimensions using matrices. Many of the ideas extend to infinite dimensions using Banach and Hilbert spaces [77], [72], [96] in functional analysis.

3.1 The Problem

Given an M by N real matrix $\mathbf{A}$ and an M by 1 vector $\mathbf{b}$, find the N by 1 vector $\mathbf{x}$ when

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & & \\ a_{31} & a_{32} & a_{33} & & \\ \vdots & & & & \vdots \\ a_{M1} & & & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{bmatrix} \tag{3.1}$$

or, using matrix notation,

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.2}$$

If $\mathbf{b}$ does not lie in the range space of $\mathbf{A}$ (the space spanned by the columns of $\mathbf{A}$), there is no exact solution to (3.2), therefore, an approximation problem can be posed by minimizing an equation error defined by

$$\epsilon = \mathbf{A}\mathbf{x} - \mathbf{b}. \tag{3.3}$$

[1] This content is available online at <http://cnx.org/content/m19561/1.5/>.

A generalized solution (or an optimal approximate solution) to (3.2) is usually considered to be an $\mathbf{x}$ that minimizes some norm of ϵ. If that problem does not have a unique solution, further conditions, such as also minimizing the norm of $\mathbf{x}$, are imposed. The l_2 or root-mean-squared error or Euclidean norm is $\sqrt{\epsilon^{\mathbf{T}*}\epsilon}$ and minimization sometimes has an analytical solution. Minimization of other norms such as l_∞ (Chebyshev) or l_1 require iterative solutions. The general l_p norm is defined as q where

$$q = ||x||_p = \left(\sum_n |x(n)|^p\right)^{1/p} \tag{3.4}$$

for $1 < p < \infty$ and a "pseudonorm" (not convex) for $0 < p < 1$. These can sometimes be evaluated using IRLS (iterative reweighted least squares) algorithms [13], [17], [89], [46], [33].

If there is a non-zero solution of the homogeneous equation

$$\mathbf{A}\mathbf{x} = \mathbf{0}, \tag{3.5}$$

then (3.2) has infinitely many generalized solutions in the sense that any particular solution of (3.2) plus an arbitrary scalar times any non-zero solution of (3.5) will have the same error in (3.3) and, therefore, is also a generalized solution. The number of families of solutions is the dimension of the null space of $\mathbf{A}$.

This is analogous to the classical solution of linear, constant coefficient differential equations where the total solution consists of a particular solution plus arbitrary constants times the solutions to the homogeneous equation. The constants are determined from the initial (or other) conditions of the solution to the differential equation.

3.2 Ten Cases to Consider

Examination of the basic problem shows there are ten cases [60] listed in Figure 1 to be considered. These depend on the shape of the M by N real matrix $\mathbf{A}$, the rank r of $\mathbf{A}$, and whether $\mathbf{b}$ is in the span of the columns of $\mathbf{A}$.

- 1a. $M = N = r$: One solution with no error, ϵ.
- 1b. $M = N > r$: $\mathbf{b} \in span\{\mathbf{A}\}$: Many solutions with $\epsilon = \mathbf{0}$.
- 1c. $M = N > r$: $\mathbf{b}$ *not* $\in span\{\mathbf{A}\}$: Many solutions with the same minimum error.
- 2a. $M > N = r$: $\mathbf{b} \in span\{\mathbf{A}\}$: One solution $\epsilon = \mathbf{0}$.
- 2b. $M > N = r$: $\mathbf{b}$ *not* $\in span\{\mathbf{A}\}$: One solution with minimum error.
- 2c. $M > N > r$: $\mathbf{b} \in span\{\mathbf{A}\}$: Many solutions with $\epsilon = \mathbf{0}$.
- 2d. $M > N > r$: $\mathbf{b}$ *not* $\in span\{\mathbf{A}\}$: Many solutions with the same minimum error.
- 3a. $N > M = r$: Many solutions with $\epsilon = \mathbf{0}$.
- 3b. $N > M > r$: $\mathbf{b} \in span\{\mathbf{A}\}$: Many solutions with $\epsilon = \mathbf{0}$
- 3c. $N > M > r$: $\mathbf{b}$ *not* $\in span\{\mathbf{A}\}$: Many solutions with the same minimum error.

Figure 1. Ten Cases for the Pseudoinverse.

Here we have:

- case 1 has the same number of equations as unknowns ($\mathbf{A}$ is square, $M = N$),
- case 2 has more equations than unknowns, therefore, is over specified ($\mathbf{A}$ is taller than wide, $M > N$),
- case 3 has fewer equations than unknowns, therefore, is underspecified ($\mathbf{A}$ is wider than tall $N > M$).

This is a setting for frames and sparse representations.

In case 1a and 3a, $\mathbf{b}$ is necessarily in the span of $\mathbf{A}$. In addition to these classifications, the possible orthogonality of the columns or rows of the matrices gives special characteristics.

3.3 Examples

Case 1: Here we see a 3 x 3 square matrix which is an example of case 1 in Figure 1 and 2.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3.6}$$

If the matrix has rank 3, then the **b** vector will necessarily be in the space spanned by the columns of **A** which puts it in case 1a. This can be solved for **x** by inverting **A** or using some more robust method. If the matrix has rank 1 or 2, the **b** may or may not lie in the spanned subspace, so the classification will be 1b or 1c and minimization of $||x||_2^2$ yields a unique solution.

Case 2: If **A** is 4 x 3, then we have more equations than unknowns or the overspecified or overdetermined case.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \tag{3.7}$$

If this matrix has the maximum rank of 3, then we have case 2a or 2b depending on whether **b** is in the span of **A** or not. In either case, a unique solution **x** exists which can be found by (3.15) or (3.21). For case 2a, we have a single exact solution with no equation error, $\varepsilon = \mathbf{0}$ just as case 1a. For case 2b, we have a single optimal approximate solution with the least possible equation error. If the matrix has rank 1 or 2, the classification will be 2c or 2d and minimization of $||x||_2^2$ yelds a unique solution.

Case 3: If **A** is 3 x 4, then we have more unknowns than equations or the underspecified case.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3.8}$$

If this matrix has the maximum rank of 3, then we have case 3a and **b** must be in the span of **A** . For this case, many exact solutions **x** exist, all having zero equation error and a single one can be found with minimum solution norm $||\mathbf{x}||$ using (3.17) or (3.22). If the matrix has rank 1 or 2, the classification will be 3b or 3c.

3.4 Solutions

There are several assumptions or side conditions that could be used in order to define a useful unique solution of (3.2). The side conditions used to define the Moore-Penrose pseudo-inverse are that the l_2 norm squared of the equation error ϵ be minimized and, if there is ambiguity (several solutions with the same minimum error), the l_2 norm squared of **x** also be minimized. A useful alternative to minimizing the norm of **x** is to require certain entries in **x** to be zero (sparse) or fixed to some non-zero value (equality constraints).

In using sparsity in posing a signal processing problem (e.g. compressive sensing), an l_1 norm can be used (or even an l_0 "pseudo norm") to obtain solutions with zero components if possible [40], [78].

In addition to using side conditions to achieve a unique solution, side conditions are sometimes part of the original problem. One interesting case requires that certain of the equations be satisfied with no error and the approximation be achieved with the remaining equations.

3.5 Moore-Penrose Pseudo-Inverse

If the l_2 norm is used, a unique generalized solution to (3.2) always exists such that the norm squared of the equation error $\epsilon^{\mathbf{T}*}\epsilon$ and the norm squared of the solution $\mathbf{x}^{\mathbf{T}*}\mathbf{x}$ are both minimized. This solution is denoted by

$$\mathbf{x} = \mathbf{A}^{+}\mathbf{b} \tag{3.9}$$

where $\mathbf{A}^{+}$ is called the Moore-Penrose inverse [3] of $\mathbf{A}$ (and is also called the generalized inverse [6] and the pseudoinverse [3])

Roger Penrose [75] showed that for all $\mathbf{A}$, there exists a unique $\mathbf{A}^{+}$ satisfying the four conditions:

$$\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A} \tag{3.10}$$

$$\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+} \tag{3.11}$$

$$\left[\mathbf{A}\mathbf{A}^{+}\right]^{*} = \mathbf{A}\mathbf{A}^{+} \tag{3.12}$$

$$\left[\mathbf{A}^{+}\mathbf{A}\right]^{*} = \mathbf{A}^{+}\mathbf{A} \tag{3.13}$$

There is a large literature on this problem. Five useful books are [60], [3], [6], [25], [76]. The Moore-Penrose pseudo-inverse can be calculated in Matlab [67] by the `pinv(A,tol)` function which uses a singular value decomposition (SVD) to calculate it. There are a variety of other numerical methods given in the above references where each has some advantages and some disadvantages.

3.6 Properties

For cases 2a and 2b in Figure 1, the following N by N system of equations called the *normal equations*[3], [60] have a unique minimum squared equation error solution (minimum $\varepsilon^{T}\varepsilon$). Here we have the over specified case with more equations than unknowns. A derivation is outlined in "Derivations" (Section 3.8.1: Derivations), equation (3.28) below.

$$\mathbf{A}^{\mathbf{T}*}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathbf{T}*}\mathbf{b} \tag{3.14}$$

The solution to this equation is often used in least squares approximation problems. For these two cases $\mathbf{A}^{T}\mathbf{A}$ is non-singular and the N by M pseudo-inverse is simply,

$$\mathbf{A}^{+} = \left[\mathbf{A}^{\mathbf{T}*}\mathbf{A}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}*}. \tag{3.15}$$

A more general problem can be solved by minimizing the weighted equation error, $\varepsilon^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\varepsilon$ where $\mathbf{W}$ is a positive semi-definite diagonal matrix of the error weights. The solution to that problem [6] is

$$\mathbf{A}^{+} = \left[\mathbf{A}^{\mathbf{T}*}\mathbf{W}^{\mathbf{T}*}\mathbf{W}\mathbf{A}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}*}\mathbf{W}^{\mathbf{T}*}\mathbf{W}. \tag{3.16}$$

For the case 3a in Figure 1 with more unknowns than equations, $\mathbf{A}\mathbf{A}^{T}$ is non-singular and has a unique minimum norm solution, $||\mathbf{x}||$. The N by M pseudoinverse is simply,

$$\mathbf{A}^{+} = \mathbf{A}^{\mathbf{T}*}\left[\mathbf{A}\mathbf{A}^{\mathbf{T}*}\right]^{-\mathbf{1}}. \tag{3.17}$$

with the formula for the minimum weighted solution norm $||x||$ is

$$\mathbf{A}^{+} = \left[\mathbf{W}^{\mathbf{T}}\mathbf{W}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}\left[\mathbf{A}\left[\mathbf{W}^{\mathbf{T}}\mathbf{W}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}\right]^{-\mathbf{1}}. \tag{3.18}$$

For these three cases, either (3.15) or (3.17) can be directly calculated, but not both. However, they are equal so you simply use the one with the non-singular matrix to be inverted. The equality can be shown from an equivalent definition [3] of the pseudo-inverse given in terms of a limit by

$$\mathbf{A}^{+} = \underset{\delta\to 0}{lim}\left[\mathbf{A}^{\mathbf{T}*}\mathbf{A} + \delta^{2}\mathbf{I}\right]^{-1}\mathbf{A}^{\mathbf{T}*} = \underset{\delta\to 0}{lim}\mathbf{A}^{\mathbf{T}*}\left[\mathbf{A}\mathbf{A}^{\mathbf{T}*} + \delta^{2}\mathbf{I}\right]^{-1}. \tag{3.19}$$

For the other 6 cases, SVD or other approaches must be used. Some properties [3], [25] are:

- $[\mathbf{A}^{+}]^{+} = \mathbf{A}$
- $[\mathbf{A}^{+}]^{*} = [\mathbf{A}^{*}]^{+}$
- $[\mathbf{A}^{*}\mathbf{A}]^{+} = \mathbf{A}^{+}\mathbf{A}^{*+}$
- $\lambda^{+} = 1/\lambda$ for $\lambda \neq 0$ else $\lambda^{+} = 0$
- $\mathbf{A}^{+} = [\mathbf{A}^{*}\mathbf{A}]^{+}\mathbf{A}^{*} = \mathbf{A}^{*}[\mathbf{A}\mathbf{A}^{*}]^{+}$
- $\mathbf{A}^{*} = \mathbf{A}^{*}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}\mathbf{A}\mathbf{A}^{*}$

It is informative to consider the range and null spaces [25] of $\mathbf{A}$ and $\mathbf{A}^{+}$

- $R(\mathbf{A}) = R(\mathbf{A}\mathbf{A}^{+}) = R(\mathbf{A}\mathbf{A}^{*})$
- $R(\mathbf{A}^{+}) = R(\mathbf{A}^{*}) = R(\mathbf{A}^{+}\mathbf{A}) = R(\mathbf{A}^{*}\mathbf{A})$
- $R(I - \mathbf{A}\mathbf{A}^{+}) = N(\mathbf{A}\mathbf{A}^{+}) = N(\mathbf{A}^{*}) = N(\mathbf{A}^{+}) = R(\mathbf{A})^{\perp}$
- $R(I - \mathbf{A}^{+}\mathbf{A}) = N(\mathbf{A}^{+}\mathbf{A}) = N(\mathbf{A}) = R(\mathbf{A}^{*})^{\perp}$

3.7 The Cases with Analytical Soluctions

The four Penrose equations in (3.11) are remarkable in defining a unique pseudoinverse for any $\mathbf{A}$ with any shape, any rank, for any of the ten cases listed in Figure 1. However, only four cases of the ten have analytical solutions (actually, all do if you use SVD).

- If $\mathbf{A}$ is case 1a, (square and nonsingular), then

$$\mathbf{A}^{+} = \mathbf{A}^{-1} \tag{3.20}$$

- If $\mathbf{A}$ is case 2a or 2b, (over specified) then

$$\mathbf{A}^{+} = \left[\mathbf{A}^{\mathbf{T}}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathbf{T}} \tag{3.21}$$

- If $\mathbf{A}$ is case 3a, (under specified) then

$$\mathbf{A}^{+} = \mathbf{A}^{\mathbf{T}}\left[\mathbf{A}\mathbf{A}^{\mathbf{T}}\right]^{-1} \tag{3.22}$$

Figure 2. Four Cases with Analytical Solutions

Fortunately, most practical cases are one of these four but even then, it is generally faster and less error prone to use special techniques on the normal equations rather than directly calculating the inverse matrix. Note the matrices to be inverted above are all r by r (r is the rank) and nonsingular. In the other six cases from the ten in Figure 1, these would be singular, so alternate methods such as SVD must be used [60], [3], [6].

In addition to these four cases with "analytical" solutions, we can pose a more general problem by asking for an optimal approximation with a weighted norm [6] to emphasize or de-emphasize certain components or range of equations.

- If $\mathbf{A}$ is case 2a or 2b, (over specified) then the weighted error pseudoinverse is

$$\mathbf{A}^{+} = \left[\mathbf{A}^{\mathbf{T}*}\mathbf{W}^{\mathbf{T}*}\mathbf{W}\mathbf{A}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}*}\mathbf{W}^{\mathbf{T}*}\mathbf{W} \tag{3.23}$$

- If $\mathbf{A}$ is case 3a, (under specified) then the weighted norm pseudoinverse is

$$\mathbf{A}^{+} = \left[\mathbf{W}^{\mathbf{T}}\mathbf{W}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}\left[\mathbf{A}\left[\mathbf{W}^{\mathbf{T}}\mathbf{W}\right]^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}\right]^{-\mathbf{1}} \tag{3.24}$$

Figure 3. Three Cases with Analytical Solutions and Weights

These solutions to the weighted approxomation problem are useful in their own right but also serve as the foundation to the Iterative Reweighted Least Squares (IRLS) algorithm developed in the next chapter.

3.8 Geometric interpretation and Least Squares Approximation

A particularly useful application of the pseudo-inverse of a matrix is to various least squared error approximations [60], [11]. A geometric view of the derivation of the normal equations can be helpful. If $\mathbf{b}$ does not lie in the range space of $\mathbf{A}$, an error vector is defined as the difference between $\mathbf{Ax}$ and $\mathbf{b}$. A geometric picture of this vector makes it clear that for the length of ϵ to be minimum, it must be orthogonal to the space spanned by the columns of $\mathbf{A}$. This means that $\mathbf{A}^{*}\epsilon = \mathbf{0}$. If both sides of (3.2) are multiplied by $\mathbf{A}^{*}$, it is easy to see that the normal equations of (3.14) result in the error being orthogonal to the columns of $\mathbf{A}$ and, therefore its being minimal length. If $\mathbf{b}$ does lie in the range space of $\mathbf{A}$, the solution of the normal equations gives the exact solution of (3.2) with no error.

For cases 1b, 1c, 2c, 2d, 3a, 3b, and 3c, the homogeneous equation (3.5) has non-zero solutions. Any vector in the space spanned by these solutions (the null space of $\mathbf{A}$) does not contribute to the equation error ϵ defined in (3.3) and, therefore, can be added to any particular generalized solution of (3.2) to give a family of solutions with the same approximation error. If the dimension of the null space of $\mathbf{A}$ is d, it is possible to find a unique generalized solution of (3.2) with d zero elements. The non-unique solution for these seven cases can be written in the form [6].

$$\mathbf{x} = \mathbf{A}^{+}\mathbf{b} + \left[\mathbf{I} - \mathbf{A}^{+}\mathbf{A}\right]\mathbf{y} \tag{3.25}$$

where $\mathbf{y}$ is an arbitrary vector. The first term is the minimum norm solution given by the Moore-Penrose pseudo-inverse $\mathbf{A}^{+}$ and the second is a contribution in the null space of $\mathbf{A}$. For the minimum $||x||$, the vector $\mathbf{y} = 0$.

3.8.1 Derivations

To derive the necessary conditions for minimizing q in the overspecified case, we differentiate $q = \varepsilon^{\mathbf{T}}\varepsilon$ with respect to $\mathbf{x}$ and set that to zero. Starting with the error

$$q = \varepsilon^{\mathbf{T}}\varepsilon = [\mathbf{Ax} - \mathbf{b}]^{\mathbf{T}}[\mathbf{Ax} - \mathbf{b}] = \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{Ax} - \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{b} - \mathbf{b}^{\mathbf{T}}\mathbf{Ax} + \mathbf{b}^{\mathbf{T}}\mathbf{b} \tag{3.26}$$

$$q = \mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{Ax} - \mathbf{2x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{b} + \mathbf{b}^{\mathbf{T}}\mathbf{b} \tag{3.27}$$

and taking the gradient or derivative gives

$$\nabla_{\mathbf{x}}q = 2\mathbf{A}^{\mathbf{T}}\mathbf{Ax} - 2\mathbf{A}^{\mathbf{T}}\mathbf{b} = \mathbf{0} \tag{3.28}$$

which are the normal equations in (3.14) and the pseudoinverse in (3.15) and (3.21).

If we start with the weighted error problem

$$q = \varepsilon^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}\varepsilon = [\mathbf{Ax} - \mathbf{b}]^{\mathbf{T}}\mathbf{W}^{\mathbf{T}}\mathbf{W}[\mathbf{Ax} - \mathbf{b}] \tag{3.29}$$

using the same steps as before gives the normal equations for the minimum weighted squared error as

$$\mathbf{A^T W^T W A x} = \mathbf{A^T W^T W b} \tag{3.30}$$

and the pseudoinverse as

$$\mathbf{x} = \left[\mathbf{A^T W^T W A}\right]^{-\mathbf{1}} \mathbf{A^T W^T W b} \tag{3.31}$$

To derive the necessary conditions for minimizing the Euclidian norm $||x||_2$ when there are few equations and many solutions to (3.1), we define a Lagrangian

$$\mathcal{L}(\mathbf{x}, \mu) = ||\mathbf{W x}||_2^2 + \mu^{\mathbf{T}} (\mathbf{A x} - \mathbf{b}) \tag{3.32}$$

take the derivatives in respect to both $\mathbf{x}$ and μ and set them to zero.

$$\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{2 W^T W x} + \mathbf{A^T} \mu = \mathbf{0} \tag{3.33}$$

and

$$\nabla_{\mu} \mathcal{L} = \mathbf{A x} - \mathbf{b} = \mathbf{0} \tag{3.34}$$

Solve these two equation simultaneously for $\mathbf{x}$ eliminating μ gives the pseudoinverse in (3.17) and (3.22) result.

$$\mathbf{x} = \left[\mathbf{W^T W}\right]^{-\mathbf{1}} \mathbf{A^T} \left[\mathbf{A} \left[\mathbf{W^T W}\right]^{-\mathbf{1}} \mathbf{A^T}\right]^{-\mathbf{1}} \mathbf{b} \tag{3.35}$$

Because the weighting matrices $\mathbf{W}$ are diagonal and real, multiplication and inversion is simple. These equations are used in the Iteratively Reweighted Least Squares (IRLS) algorithm described in the next chapter.

3.9 Regularization

To deal with measurement error and data noise, a process called "regularization" is sometimes used [45], [11], [70].

3.10 Least Squares Approximation with Constraints

The solution of the overdetermined simultaneous equations is generally a least squared error approximation problem. A particularly interesting and useful variation on this problem adds inequality and/or equality constraints. This formulation has proven very powerful in solving the constrained least squares approximation part of FIR filter design [80]. The equality constraints can be taken into account by using Lagrange multipliers and the inequality constraints can use the Kuhn-Tucker conditions [43], [86], [63]. The iterative reweighted least squares (IRLS) algorithm described in the next chapter can be modified to give results which are an optimal constrained least p-power solution [15], [19], [17].

3.11 Conclusions

There is remarkable structure and subtlety in the apparently simple problem of solving simultaneous equations and considerable insight can be gained from these finite dimensional problems. These notes have emphasized the l_2 norm but some other such as l_∞ and l_1 are also interesting. The use of sparsity [78] is particularly interesting as applied in Compressive Sensing [4], [38] and in the sparse FFT [51]. There are also interesting and important applications in infinite dimensions. One of particular interest is in signal analysis using wavelet basis functions [32]. The use of weighted error and weighted norm pseudoinverses provide a base for iterative reweighted least squares (IRLS) algorithms.

Chapter 4

Approximation with Other Norms and Error Measures[1]

4.1 Approximation with Other Norms and Error Measures

Most of the discussion about the approximate solutions to $\mathbf{Ax} = \mathbf{b}$ are about the result of minimizing the l_2 equation error $||Ax - b||_2$ and/or the l_2 norm of the solution $||\mathbf{x}||_2$ because in some cases that can be done by analytic formulas and also because the l_2 norm has a energy interpretation. However, both the l_1 and the l_∞[28] have well known applications that are important [34], [21] and the more general l_p error is remarkably flexible [14], [18]. Donoho has shown [39] that l_1 optimization gives essentially the same sparsity as the true sparsity measure in l_0.

In some cases, one uses a different norm for the minimization of the equation error than the one for minimization of the solution norm. And in other cases, one minimizes a weighted error to emphasize some equations relative to others [7]. A modification allows minimizing according to one norm for one set of equations and another for a different set. A more general error measure than a norm can be used which used a polynomial error [18] which does not satisfy the scaling requirement of a norm, but is convex. One could even use the so-called l_p norm for $1 > p > 0$ which is not even convex but is an interesting tool for obtaining sparse solutions.

[1] This content is available online at <http://cnx.org/content/m45576/1.1/>.

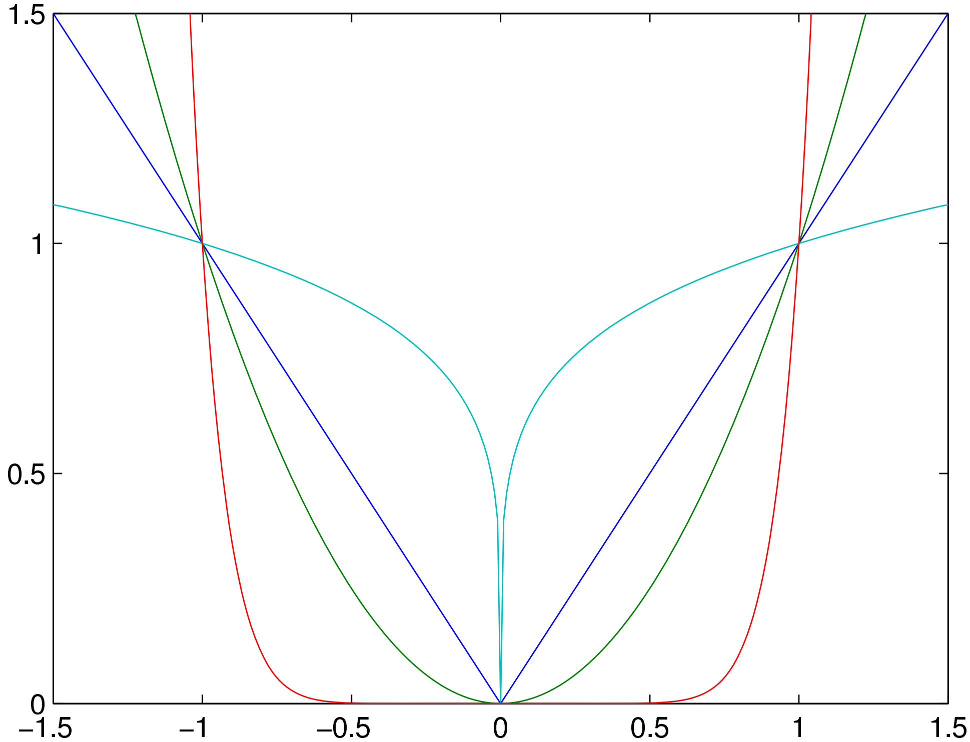

Figure 4.1: Different l_p norms: p = .2, 1, 2, 10.

Note from the figure how the l_{10} norm puts a large penalty on large errors. This gives a Chebyshev-like solution. The $l_{0.2}$ norm puts a large penalty on small errors making them tend to zero. This (and the l_1 norm) give a sparse solution.

4.2 The L_p Norm Approximation

The **IRLS** (iterative reweighted least squares) algorithm allows an iterative algorithm to be built from the analytical solutions of the weighted least squares with an iterative reweighting to converge to the optimal l_p approximation [12].

4.2.1 The Overdetermined System with more Equations than Unknowns

If one poses the l_p approximation problem in solving an overdetermined set of equations (case 2 from Chapter 3), it comes from defining the equation error vector

$$\mathbf{e} = \mathbf{A}\mathbf{x} - \mathbf{b} \tag{4.1}$$

and minimizing the p-norm

$$||\mathbf{e}||_p = \left(\sum_n |e_n|^p \right)^{1/p} \tag{4.2}$$

or

$$||\mathbf{e}||_\mathbf{p}^\mathbf{p} = \sum_\mathbf{n} |\mathbf{e_n}|^\mathbf{p} \tag{4.3}$$

neither of which can we minimize easily. However, we do have formulas [7] to find the minimum of the weighted squared error

$$||\mathbf{We}||_\mathbf{2}^\mathbf{2} = \sum_\mathbf{n} \mathbf{w_n^2} |\mathbf{e_n}|^\mathbf{2} \tag{4.4}$$

one of which is derived in , equation and is

$$\mathbf{x} = \left[\mathbf{A^T W^T W A}\right]^{-1} \mathbf{A^T W^T W b} \tag{4.5}$$

where $\mathbf{W}$ is a diagonal matrix of the error weights, w_n. From this, we propose the iterative reweighted least squared (IRLS) error algorithm which starts with unity weighting, $\mathbf{W} = \mathbf{I}$, solves for an initial $\mathbf{x}$ with (4.5), calculates a new error from(4.1), which is then used to set a new weighting matrix $\mathbf{W}$

$$\mathbf{W} = diag(w_n)^{(p-2)/2} \tag{4.6}$$

to be used in the next iteration of (4.5). Using this, we find a new solution $\mathbf{x}$ and repeat until convergence (if it happens!).

This core idea has been repeatedly proposed and developed in different application areas over the past 50 years with a variety of success [12]. Used in this basic form, it reliably converges for $2 < p < 3$. In 1990, a modification was made to partially update the solution each iteration with

$$\mathbf{x}(k) = q\,\hat{\mathbf{x}}(k) + (1-q)\,\mathbf{x}(k-1) \tag{4.7}$$

where $\hat{\mathbf{x}}$ is the new weighted least squares solution of which is used to partially update the previous value $\mathbf{x}(k-1)$ using a convergence up-date factor $0 < q < 1$ which gave convergence over a larger range of around $1.5 < p < 5$ but but it was slower.

A second improvement showed that a specific up-date factor of

$$q = \frac{1}{p-1} \tag{4.8}$$

significantly increased the speed of convergence. With this particular factor, the algorithm becomes a form of Newton's method which has quadratic convergence.

A third modification applied homotopy [16], , , [81] by starting with a value for p which is equal to 2 and increasing it each iteration (or each few iterations) until it reached the desired value, or, in the case of $p < 2$, decrease it. This made a significant increase in both the range of p that allowed convergence and in the speed of calculations. Some of the history and details can be found applied to digital filter design in [14], [18].

A Matlab program that implements these ideas applied to our pseudoinverse problem with more equations than unknowns (case 2a) is:

```
%~m-file~IRLS1.m~to~find~the~optimal~solution~to~Ax=b

%~~minimizing~the~L_p~norm~||Ax-b||_p,~using~IRLS.

%~~Newton~iterative~update~of~solution,~x,~for~~M~>~N.

%~~For~2<p<infty,~use~homotopy~parameter~K~=~1.01~to~2
```

```
%~~For~0<p<2,~use~K~=~approx~0.7~-~0.9

%~~csb~10/20/2012

function~x~=~IRLS1(A,b,p,K,KK)

if~nargin~<~5,~KK=10;~~end;

if~nargin~<~4,~K~=~2;~~end;

if~nargin~<~3,~p~=~10;~end;

pk~=~2;~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Initial~homotopy~value

x~~=~pinv(A)*b;~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Initial~L_2~solution

E~=~[];

for~k~=~1:KK~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Iterate

~~~if~p~>=~2,~pk~=~min([p,~K*pk]);~~~~~~~~~~~%~Homotopy~change~of~p

~~~~~~else~pk~=~max([p,~K*pk]);~end

~~~e~~=~A*x~-~b;~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Error~vector

~~~w~~=~abs(e).^((pk-2)/2);~~~~~~~~~~~~~~~~~~%~Error~weights~for~IRLS

~~~W~~=~diag(w/sum(w));~~~~~~~~~~~~~~~~~~~~~~%~Normalize~weight~matrix

~~~WA~=~W*A;~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~apply~weights

~~~x1~~=~(WA'*WA)\(WA'*W)*b;~~~~~~~~~~~~~~~~~%~weighted~L_2~sol.

~~~q~~=~1/(pk-1);~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Newton's~parameter

~~~if~p~>~2,~x~=~q*x1~+~(1-q)*x;~nn=p;~~~~~~~%~partial~update~for~p>2

~~~~~~else~x~=~x1;~nn=2;~end~~~~~~~~~~~~~~~~~%~no~partial~update~for~p<2

~~~ee~=~norm(e,nn);~~~E~=~[E~ee];~~~~~~~~~~~~%~Error~at~each~iteration

end

plot(E)
```

This can be modified to use different p's in different bands of equations or to use weighting only when the error exceeds a certain threshold to achieve a constrained LS approximation [14], [18], [90]. Our work was originally done in the context of filter design but others have done similar things in sparsity analysis [47], [34], [92].

This is presented as applied to the overdetermined system (Case 2a and 2b) but can also be applied to other cases. A particularly important application of this section is to the design of digital filters.

4.2.2 The Underdetermined System with more Unknowns than Equations

If one poses the l_p approximation problem in solving an underdetermined set of equations (case 3 from Chapter 3), it comes from defining the solution norm as

$$||x||_p = \left(\sum_n |x\left(n\right)|^p\right)^{1/p} \tag{4.9}$$

and finding $\mathbf{x}$ to minimizing this p-norm while satisfying $\mathbf{Ax} = \mathbf{b}$.

It has been shown this is equivalent to solving a least weighted norm problem for specific weights.

$$||x||_p = \left(\sum_n w(n)^2 |x\left(n\right)|^2\right)^{1/2} \tag{4.10}$$

The development follows the same arguments as in the previous section but using the formula [79], [7] derived in

$$\mathbf{x} = \left[\mathbf{W^T W}\right]^{-\mathbf{1}} \mathbf{A^T} \left[\mathbf{A}\left[\mathbf{W^T W}\right]^{-\mathbf{1}} \mathbf{A^T}\right]^{-\mathbf{1}} \mathbf{b} \tag{4.11}$$

with the weights, $w\left(n\right)$, being the diagonal of the matrix, $\mathbf{W}$, in the iterative algorithm to give the minimum weighted solution norm in the same way as (4.5) gives the minimum weighted equation error.

A Matlab program that implements these ideas applied to our pseudoinverse problem with more unknowns than equations (case 3a) is:

```
%~m-file~IRLS2.m~to~find~the~optimal~solution~to~Ax=b

%~~minimizing~the~L_p~norm~||x||_p,~using~IRLS.

%~~Newton~iterative~update~of~solution,~x,~for~~M~<~N.

%~~For~2<p<infty,~use~homotopy~parameter~K~=~1.01~to~2

%~~For~0<p<2,~use~K~=~approx~0.7~to~0.9

%~~csb~10/20/2012

function~x~=~IRLS2(A,b,p,K,KK)

if~nargin~<~5,~KK=~10;~~end;

if~nargin~<~4,~K~=~.8;~~end;

if~nargin~<~3,~p~=~1.1;~end;

pk~=~2;~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~%~Initial~homotopy~value

x~~=~pinv(A)*b;~~~~~~~~~~~~~~~~~~~~~~~~~%~Initial~L_2~solution

E~=~[];
```

```
for~k~=~1:KK

~~~if~p~>=~2,~pk~=~min([p,~K*pk]);~~~~~~%~Homotopy~update~of~p

~~~~~~else~pk~=~max([p,~K*pk]);~end

~~~W~~=~diag(abs(x).^((2-pk)/2)+0.00001);~~%~norm~weights~for~IRLS

~~~AW~=~A*W;~~~~~~~~~~~~~~~~~~~~~~~~~~~%~applying~new~weights

~~~x1~=~W*AW'*((AW*AW')\b);~~~~~~~~~~~~~%~Weighted~L_2~solution

~~~q~~=~1/(pk-1);~~~~~~~~~~~~~~~~~~~~~~%~Newton's~parameter

~~~if~p~>=~2,~x~=~q*x1~+~(1-q)*x;~nn=p;~%~Newton's~partial~update~for~p>2

~~~~~~else~x~=~x1;~nn=1;~end~~~~~~~~~~~%~no~Newton's~partial~update~for~p<2

~~~ee~=~norm(x,nn);~~E~=~[E~ee];~~~~~~~~%~norm~at~each~iteration

end;

plot(E)
```

This approach is useful in sparse signal processing and for frame representation.

4.3 The Chebyshev, Minimax, or L_∞ Appriximation

The **Chebyshev** optimization problem minimizes the maximum error:

$$\varepsilon_m = \max_n |\varepsilon(n)| \tag{4.12}$$

This is particularly important in filter design. The Remez exchange algorithm applied to filter design as the Parks-McClellan algorithm is very efficient [21]. An interesting result is the limit of an $||\mathbf{x}||_p$ optimization as $p \to \infty$ is the Chebyshev optimal solution. So, the Chebyshev optimal, the minimax optimal, and the L_∞ optimal are all the same [28], [21].

A particularly powerful theorem which characterizes a solution to $\mathbf{Ax} = \mathbf{b}$ is given by Cheney [28] in Chapter 2 of his book:

- **A Characterization Theorem:** *For an M by N real matrix, $\mathbf{A}$ with $M > N$, every minimax solution $\mathbf{x}$ is a minimax solution of an appropriate $N+1$ subsystem of the M equations. This optimal minimax solution will have at least $N+1$ equal magnitude errors and they will be larger than any of the errors of the other equations.*

This is a powerful statement saying an optimal minimax solution will have out of M, at least $N+1$ maximum magnitude errors and they are the minimum size possible. What this theorem doesn't state is which of the M equations are the $N+1$ appropriate ones. Cheney develops an algorithm based on this theorem which finds these equations and exactly calculates this optimal solution in a finite number of steps. He shows how this can be combined with the minimum $||\mathbf{e}||_p$ using a large p, to make an efficient solver for a minimax or Chebyshev solution.

This theorem is similar to the Alternation Theorem [21] but more general and, therefore, somewhat more difficult to implement.

4.4 The L_1 Approximation and Sparsity

The **sparsity** optimization is to minimize the number of non-zero terms in a vector. A "pseudonorm", $||\mathbf{x}||_0$, is sometimes used to denote a measure of sparsity. This is not convex, so is not really a norm but the convex (in the limit) norm $||\mathbf{x}||_1$ is close enough to the $||\mathbf{x}||_0$ to give the same sparsity of solution [39]. Finding a sparse solution is not easy but interative reweighted least squares (IRLS) [18], [90], weighted norms [47], [34], and a somewhat recent result is called Basis Pursuit [26], [27] are possibilities.

This approximation is often used with an underdetermined set of equations (Case 3a) to obtain a sparse solution $\mathbf{x}$.

Using the IRLS algorithm to minimize the l_p equation error often gives a sparse error if one exists. Using the algorithm in the illustrated Matlab program with $p = 1.1$ on the problem in Cheney [28] gives a zero error in equation 4 while using no larger p gives any zeros.

Chapter 5

Constructing the Operator (Design)[1]

5.1 Constructing the Operator (unfinished)

Solving the third problem posed in the introduction to these notes is rather different from the other two. Here we want to find an operator or matrix that when multiplied by $\mathbf{x}$ gives $\mathbf{b}$. Clearly a solution to this problem would not be unique as stated. In order to pose a better defined problem, we generally give a set or family of inputs $\mathbf{x}$ and the corresponding outputs $\mathbf{b}$. If these families are independent, and if the number of them is the same as the size of the matrix, a unique matrix is defined and can be found by solving simultaneous equations. If a smaller number is given, the remaining degrees of freedom can be used to satisfy some other criterion. If a larger number is given, there is probably no exact solution and some approximation will be necessary.

If the unknown operator matrix is of dimension M by N, then we take N inputs $\mathbf{x_k}$ for $k = 1, 2, \cdots, N$, each of dimension N and the corresponding N outputs $\mathbf{b_k}$, each of dimension M and form the matrix equation:

$$\mathbf{AX} = \mathbf{B} \tag{5.1}$$

where $\mathbf{A}$ is the M by N unknown operator, $\mathbf{X}$ is the N by N input matrix with N columns which are the inputs $\mathbf{x_k}$ and $\mathbf{B}$ is the M by N output matrix with columns $\mathbf{b_k}$. The operator matrix is then determined by:

$$\mathbf{A} = \mathbf{B}\mathbf{X}^{\mathbf{-1}} \tag{5.2}$$

if the inputs are independent which means $\mathbf{X}$ is nonsingular.

This problem can be posed so that there are more (perhaps many more) inputs and outputs than N with a resulting equation error which can be minimized with some form of pseudoinverse.

Linear regression can be put in this form. If our matrix equation is

$$\mathbf{Ax} = \mathbf{b} \tag{5.3}$$

where $\mathbf{A}$ is a row vector of unknown weights and $\mathbf{x}$ is a column vector of known inputs, then $\mathbf{b}$ is a scaler inter product. If a seond experiment gives a second scaler inner product from a second column vector of known inputs, then we augment $\mathbf{X}$ to have two rows and $\mathbf{b}$ to be a length-2 row vector. This is continued for N experiment to give (5.3) as a 1 by N row vector times an M by N matrix which equals a 1 by M row vector. It this equation is transposed, it is in the form of (5.3) which can be approximately solved by the pesuedo inverse to give the unknown weights for the regression.

Alternatively, the matrix may be constrained by structure to have less than N^2 degrees of freedom. It may be a cyclic convolution, a non cyclic convolution, a Toeplitz, a Hankel, or a Toeplitz plus Hankel matrix.

[1] This content is available online at <http://cnx.org/content/m19562/1.4/>.

A problem of this sort came up in research on designing efficient prime length fast Fourier transform (FFT) algorithms where $\mathbf{x}$ is the data and $\mathbf{b}$ is the FFT of $\mathbf{x}$. The problem was to derive an operator that would make this calculation using the least amount of arithmetic. We solved it using a special formulation [55] and Matlab.

Bibliography

[1] *Wavelets, Frames and Operator Theory.* American Mathematical Society, 2004.

[2] Arthur Albert. *Regression and the Moore-Penrose Pseudoinverse.* Academic Press, New York, 1972.

[3] Arthur Albert. *Regression and the Moore-Penrose Pseudoinverse.* Academic Press, New York, 1972.

[4] Richard G. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):1188211;124, July 2007. also: http://dsp.rice.edu/cs.

[5] Adi Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Wiley and Sons, New York, 1974. Second edition, Springer, 2003.

[6] Adi Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Wiley and Sons, New York, 1974. Second edition, Springer, 2003.

[7] Adi Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* Wiley and Sons, New York, 1974. Second edition, Springer, 2003.

[8] A. Benveniste, R. Nikoukhah, and A. S. Willsky. Multiscale system theory. *IEEE Transactions on Circuits and Systems, I*, 41(1):28211;15, January 1994.

[9] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[10] [U+FFFD]e Bjrck. *Numerical Methods for Least Squares Problems.* Blaisdell, Dover, SIAM, Philadelphia, 1996.

[11] [U+FFFD]e Bjrck. *Numerical Methods for Least Squares Problems.* Blaisdell, Dover, SIAM, Philadelphia, 1996.

[12] [U+FFFD]e Bjrck. *Numerical Methods for Least Squares Problems.* Blaisdell, Dover, SIAM, Philadelphia, 1996.

[13] C. S. Burrus and J. A. Barreto. Least -power error design of fir filters. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, volume 2, page 5458211;548, ISCAS-92, San Diego, CA, May 1992.

[14] C. S. Burrus and J. A. Barreto. Least -power error design of fir filters. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, volume 2, page 5458211;548, ISCAS-92, San Diego, CA, May 1992.

[15] C. S. Burrus, J. A. Barreto, and I. W. Selesnick. Reweighted least squares design of fir filters. In *Paper Summaries for the IEEE Signal Processing Society's Fifth DSP Workshop*, page 3.1.1, Starved Rock Lodge, Utica, IL, September 138211;16 1992.

[16] C. S. Burrus, J. A. Barreto, and I. W. Selesnick. Reweighted least squares design of fir filters. In *Paper Summaries for the IEEE Signal Processing Society's Fifth DSP Workshop*, page 3.1.1, Starved Rock Lodge, Utica, IL, September 138211;16 1992.

[17] C. S. Burrus, J. A. Barreto, and I. W. Selesnick. Iterative reweighted least squares design of fir filters. *IEEE Transactions on Signal Processing*, 42(11):29268211;2936, November 1994.

[18] C. S. Burrus, J. A. Barreto, and I. W. Selesnick. Iterative reweighted least squares design of fir filters. *IEEE Transactions on Signal Processing*, 42(11):29268211;2936, November 1994.

[19] C. Sidney Burrus. Constrained least squares design of fir filters using iterative reweighted least squares. In *Proceedings of EUSIPCO-98*, page 2818211;282, Rhodes, Greece, September 8-11 1998.

[20] C. Sidney Burrus. *Digital Signal Processing and Digital Filter Design.* Connexions, cnx.org, 2008. http://cnx.org/content/col10598/latest/.

[21] C. Sidney Burrus. *Digital Signal Processing and Digital Filter Design.* Connexions, cnx.org, 2008. http://cnx.org/content/col10598/latest/.

[22] C. Sidney Burrus. *Fast Fourier Transforms.* Connexions, cnx.org, 2008. http://cnx.org/content/col10550/latest/.

[23] C. Sidney Burrus, Ramesh A. Gopinath, and Haitao Guo. *Introduction to Wavelets and the Wavelet Transform.* Prentice Hall, Upper Saddle River, NJ, 1998. to appear on the web in Connexions: cnx.org.

[24] C. Sidney Burrus, Ramesh A. Gopinath, and Haitao Guo. *Introduction to Wavelets and the Wavelet Transform.* Prentice Hall, Upper Saddle River, NJ, 1998. to appear on the web in Connexions: cnx.org.

[25] S. L. Campbell and C. D. Meyer, Jr. *Generalized Inverses of Linear Transformations.* Pitman, London, 1979. Reprint by Dover in 1991.

[26] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Sci. Compt.*, 20:33–61, 1998.

[27] Scott S. Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):1298211;159, March 2001. http://www-stat.stanford.edu/donoho/reports.html.

[28] E. W. Cheney. *Introduction to Approximation Theory.* McGraw-Hill, New York, 1966. Second edition, AMS Chelsea, 2000.

[29] Ole Christensen. *An Introduction to Frames and Riesz Bases.* Birkh[U+FFFD]er, 2002.

[30] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, PA, 1992. Notes from the 1990 CBMS-NSF Conference on Wavelets and Applications at Lowell, MA.

[31] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, PA, 1992. Notes from the 1990 CBMS-NSF Conference on Wavelets and Applications at Lowell, MA.

[32] Ingrid Daubechies. *Ten Lectures on Wavelets.* SIAM, Philadelphia, PA, 1992. Notes from the 1990 CBMS-NSF Conference on Wavelets and Applications at Lowell, MA.

[33] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Gunturk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):18211;38, January 2010.

[34] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Gunturk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):18211;38, January 2010.

[35] Paul M. DeRusso, Rob J. Roy, and Charles M. Close. *State Variables for Engineers.* Wiley, 1965. Second edition 1997.

[36] David L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):1008211;115, December 1993. Also Stanford Statistics Dept. Report TR-410, Nov. 1992.

[37] David L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):1008211;115, December 1993. Also Stanford Statistics Dept. Report TR-410, Nov. 1992.

[38] David L. Donoho. Compressed sensing. Technical report, Statistics Department, Stanford University, September 2004. http://www-stat.stanford.edu/donoho/ Reports/2004/CompressedSensing091604.pdf.

[39] David L. Donoho. For most large underdetermined systems of linear equations the minimal -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):7978211;829, June 2006. http://stats.stanford.edu/donoho/Reports/2004/l1l0EquivCorrected.pdf.

[40] David L. Donoho and Michael Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via minimization. Technical report, Statistics Department, Stanford University, 2002.

[41] Richard C. Dorf. *Time-Domain Analysis and Design of Control Systems*. Addison-Wesley, 1965.

[42] Paulo J. S. G. Ferreira. Mathematics for multimedia signal processing ii: Discrete finite frames and signal reconstruction. *Signal Processing for Multimedia*, page 358211;54, 1999. J. S. Byrnes (editor), IOS Press.

[43] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, second edition, 1987.

[44] Lewis Franks. *Signal Theory*. Prentice8211;Hall, Englewood Cliffs, NJ, 1969.

[45] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, MD, 1996. 3rd edition, 4th edition is forthcoming.

[46] Irina F. Gorodnitsky and Bhaskar D. Rao. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3), March 1997.

[47] Irina F. Gorodnitsky and Bhaskar D. Rao. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3), March 1997.

[48] Paul R. Halmos. *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*. Chelsea, New York, 1951. second edition 1957.

[49] Paul R. Halmos. *Finite-Dimensional Vector Spaces*. Van Nostrand, Princeton, NJ, 1958. Springer 1974.

[50] Paul R. Halmos. *Finite-Dimensional Vector Spaces*. Van Nostrand, Princeton, NJ, 1958. Springer 1974.

[51] Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. Nearly optimal sparse fourier transform. *arXiv:1201.2501v1 [cs.DS] 12 Jan 2012*, 2012.

[52] Simon O. Haykin. *Neural Networks and Learning Machines*. Prentice Hall, 2008.

[53] Jim Hefferon. *Linear Algebra*. Virginia Commonwealth Univeristy Mathematics Textbook Series, 2011. Copyright: cc-by-sa, URL:joshua.smcvt.edu.

[54] Jim Hefferon. *Linear Algebra*. Virginia Commonwealth Univeristy Mathematics Textbook Series, 2011. Copyright: cc-by-sa, URL:joshua.smcvt.edu.

[55] Howard W. Johnson and C. S. Burrus. On the structure of efficient dft algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(1):2488211;254, February 1985.

[56] Jelena Kovacevic and Amina Chebira. Life beyond bases: The advent of frames (part i). *IEEE Signal Processing Magazine*, 24(4):868211;104, July 2007.

[57] Jelena Kovacevic and Amina Chebira. Life beyond bases: The advent of frames (part ii). *IEEE Signal Processing Magazine*, 24(5):1158211;125, September 2007.

[58] Jelena Kovacevic, Vivek K. Goyal, and Martin Vetterli. *Signal Processing: Foundations*. On Line publication by fourierandwavelets.org, March 2012. available online: http://fourierandwavelets.org/.

[59] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Inglewood Cliffs, NJ, 1974. Second edition by SIAM in 1987.

[60] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Inglewood Cliffs, NJ, 1974. Second edition by SIAM in 1987.

[61] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969, 1997.

[62] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, New York, 1969, 1997.

[63] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Springer, third edition, 2008.

[64] R. J. Marks II. *Introduction to Shannon Sampling and Interpolation Theory*. Springer-Verlag, New York, 1991.

[65] Cleve Moler. *Numerical Computing with MATLAB*. The MathWorks, Inc., South Natick, MA, 2008. available: http://www.mathworks.com/moler/.

[66] Cleve Moler. *Numerical Computing with MATLAB*. The MathWorks, Inc., South Natick, MA, 2008. available: http://www.mathworks.com/moler/.

[67] Cleve Moler, John Little, and Steve Bangert. *Matlab User's Guide*. The MathWorks, Inc., South Natick, MA, 1989.

[68] Todd K. Moon and Wynn C. Stirling. *Mathematical Methods and Algorithms for Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2000.

[69] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the AMS*, 26:3948211;395, 1920.

[70] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Reiview*, 40:6368211;666, 1998. available: http://www.mat.univie.ac.at/neum/.

[71] J. Tinsley Oden and Leszek F. Demkowicz. *Applied Functional Analysis*. CRC Press, Boca Raton, 1996.

[72] J. Tinsley Oden and Leszek F. Demkowicz. *Applied Functional Analysis*. CRC Press, Boca Raton, 1996.

[73] Soo-Chang Pei and Min-Hung Yeh. An introduction to discrete finite frames. *IEEE Signal Processing Magazine*, 14(6):848211;96, November 1997.

[74] R. Penrose. A generalized inverse for matrices. *Proc. Cambridge Phil. Soc.*, 51:4068211;413, 1955.

[75] R. Penrose. On best approximate solutions of linear matrix equations. *Proc. Cambridge Phil. Soc.*, 52:178211;19, 1955.

[76] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, New York, 1971.

[77] Frigyes Riesz and B[U+FFFD] Sz.8211;Nagy. *Functional Analysis*. Dover, New York, 1955.

[78] Ivan Selesnick. Introduction to sparsity in signal processing. *Connexions Web Site*, May 2012. Available: http://cnx.org/content/m43545/latest/.

[79] Ivan W. Selesnick. Least squares solutions to linear system of equations, 2009. to be published in Connexions.

[80] Ivan W. Selesnick, Markus Lang, and C. Sidney Burrus. Constrained least square design of fir filters without explicitly specified transition bands. *IEEE Transactions on Signal Processing*, 44(8):18798211;1892, August 1996.

[81] Allan J. Sieradski. *An Introduction to Topology and Homotopy*. PWS8211;Kent, Boston, 1992.

[82] Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press, New York, 1976. 4th Edition, Brooks Cole, 2005.

[83] Gilbert Strang. *Linear Algebra and Its Applications*. Academic Press, New York, 1976. 4th Edition, Brooks Cole, 2005.

[84] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge, New York, 1986. 4th Edition, 2009.

[85] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge, New York, 1986. 4th Edition, 2009.

[86] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge, New York, 1986. 4th Edition, 2009.

[87] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[88] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[89] Ricardo Vargas and C. Sidney Burrus. Iterative design of digital filters. *arXiv*, 2012. arXiv:1207.4526v1 [cs.IT] July 19, 2012.

[90] Ricardo Vargas and C. Sidney Burrus. Iterative design of digital filters. *arXiv*, 2012. arXiv:1207.4526v1 [cs.IT] July 19, 2012.

[91] Shayne F. D. Waldron. *An Introduction to Finite Tight Frames, Draft*. Springer, 2010. http://www.math.auckland.ac.nz/waldron/Harmonic-frames/Tuan-Shayne/book.pdf.

[92] Andrew E. Yagle. Non-iterative reweighted-norm least-squares local minimization for sparse solutions to underdetermined linear systems of equations, 2008. preprint at http://web.eecs.umich.edu/aey/sparse/sparse11.pdf.

[93] N. Young. *An Introduction to Hilbert Space*. Cambridge Press, 1988.

[94] R. M. Young. *An Introduction to Nonharmonic Fourier Series*. Academic Press, New York, 1980.

[95] R. M. Young. *An Introduction to Nonharmonic Fourier Series*. Academic Press, New York, 1980.

[96] R. M. Young. *An Introduction to Nonharmonic Fourier Series*. Academic Press, New York, 1980.

[97] Lotfi A. Zadeh and Charles A. Desoer. *Linear System Theory: The State Space Approach*. Dover, 1963, 2008.

[98] Lotfi A. Zadeh and Charles A. Desoer. *Linear System Theory: The State Space Approach*. Dover, 1963, 2008.

Index of Keywords and Terms

Keywords are listed by the section with that keyword (page numbers are in parentheses). Keywords do not necessarily appear in the text of the page. They are merely associated with that section. *Ex.* apples, § 1.1 (1) **Terms** are referenced by the page they appear on. *Ex.* apples, 1

Attributions

Collection: *Basic Vector Space Methods in Signal and Systems Theory*
Edited by: C. Sidney Burrus
URL: http://cnx.org/content/col10636/1.5/
License: http://creativecommons.org/licenses/by/3.0/

Module: "Introduction"
By: C. Sidney Burrus
URL: http://cnx.org/content/m19560/1.4/
Pages: 1-2
Copyright: C. Sidney Burrus
License: http://creativecommons.org/licenses/by/3.0/

Module: "A Matrix Times a Vector"
By: C. Sidney Burrus
URL: http://cnx.org/content/m19559/1.7/
Pages: 3-10
Copyright: C. Sidney Burrus
License: http://creativecommons.org/licenses/by/3.0/

Module: "General Solutions of Simultaneous Equations"
By: C. Sidney Burrus
URL: http://cnx.org/content/m19561/1.5/
Pages: 11-17
Copyright: C. Sidney Burrus
License: http://creativecommons.org/licenses/by/3.0/

Module: "Approximation with Other Norms and Error Measures"
By: C. Sidney Burrus
URL: http://cnx.org/content/m45576/1.1/
Pages: 19-25
Copyright: C. Sidney Burrus
License: http://creativecommons.org/licenses/by/3.0/

Module: "Constructing the Operator (Design)"
By: C. Sidney Burrus
URL: http://cnx.org/content/m19562/1.4/
Pages: 27-28
Copyright: C. Sidney Burrus
License: http://creativecommons.org/licenses/by/3.0/

Basic Vector Space Methods in Signal and Systems Theory
Linear algebra, vector space methods, and functional analysis are a powerful setting for many topics in engineering, science (including social sciences), and business. This collection starts with the simple idea of a matrix times a vector and develops tools and interpretations for many signal processing and system analysis and design methods.