FFTs for the rest of us

John Kerl
Lockheed Martin Management and Data Systems

October 4, 2000

Contents

1 Introduction

Fast Fourier transforms are of central importance in digital signal processing. People working in DSP fields stand to benefit from understanding FFTs. Yet, many of us understand FFTs poorly, or not at all. Why is this? What a Fourier transform does is very, very simple; straightforward implementations of Fourier transforms are also easy to understand. Unfortunately, these straightforward implementations run slowly. A fast Fourier transform is a particular implementation of a Fourier transform which resorts to great trickiness in order to significantly decrease execution time. Unfortunately, the tricky implementation details that make a fast Fourier transform fast often obscure the underlying simplicity -it's the first F in FFT that tends to throw most of us off track. (Even then, as we'll see, that first F is really not so horrible to understand either!)

This document will describe what Fourier transforms do, how they do it, and lastly, how the fast Fourier transform makes it all happen efficiently. You might ask, why another document on FFTs? Any signal-processing or numerical-analysis text shows a derivation of the FFT algorithm; in fact, there are entire books devoted to FFTs. Answer: Yes, there are many rigorous and thorough descriptions of FFTs. However, I've never found a simple, intuitive description of how they work — every source I've read on FFTs seems to make certain assumptions about what the reader already knows.

This document fills that void by not assuming that the reader already understands its subject. I will only assume that your background includes high-school algebra and a semester's worth of calculus (for the very few integrals I'll need to discuss). I hope to provide an intuitive understanding of Fourier transforms, such that you will be able to know what results to expect from a computation, understand what others are talking about, or be able to implement an FFT in software — that is, to have a practical mastery of the subject which you can use at will.

In that sense, this document is something more than other sources. However, it is also something less than other sources — I will not prove mathematical theorems and will make no attempt at mathematical rigor. I will glibly declare that Fourier transforms have certain properties, and I will provide lots of examples to support those claims. Mathematicians use (indeed, require) proofs to convince each other of the correctness of their arguments — in fact, no self-respecting mathematician would accept an argument not backed up by a solid proof — but for non-mathematicians, proofs often confuse more than convince. This document is for the rest of us.

8

2

What is spectral analysis?

Fourier transforms do spectral analysis. What does that mean? It means that a Fourier transform splits a signal into its component frequencies — into a spectrum. What is a ”signal”? As we'll see, it can be pretty much anything — starlight, sound, photographs, earthquake tremors, etc.

Paint

Spectral analysis sounds like an intimidating term, so here is an example. Suppose you had some red paint, some yellow paint and some blue paint. Suppose you took 3/4 cup of blue, 1/4 cup of yellow, and just a teaspoon of red, and mixed them together — and only then showed the mixture to me. By looking at the mixture's blue-green hue, I could guess that blue was the predominant color, but I couldn't tell you with any accuracy what proportions of colors you'd used. However, if I had a magical paint unmixer, I could take this mixture of a little more than a cup, then separate it back to 3/4 cup blue paint, 1/4 cup yellow and a teaspoon of red. Then I could exactly recover the amount of each original color. A paint unmixer would be a very magical device!

Light

Here is another example. We all know that if you pass a ray of light through a prism, the ray gets spread out into a rainbow of its component colors. If you look up at the constellation Orion at night, you can tell that Betelgeuse is redder than most stars, and Rigel is bluer. That's about all you can tell. But if an astronomer trains a telescope on one of those stars, then splits the light up into its component colors, then they can obtain a great deal of information -they can even tell what gases comprise the outer surface of the star, from light years away, even though under the greatest magnification the star is nothing more than a point! So a prism is a powerful thing.

Sound

Here is a third example. In the Beatles' song *Twist and Shout* , at one point all four Beatles stop singing, then John sings “laaaa” with a C, then four beats later Paul comes in a second later singing “laaaa” with an E, then George with a G — a nice little chord so far — then Ringo comes in with a funky-sounding B-flat. Then they all scream for a while, the song picks up again, and they all go back to singing “Shake it up baby, now.” We know that sound is simply compression waves in the air between the source and our ear. We could plot the compression waves making up John's C, versus time. This would give us a nice sine wave (figure 1). 1

We could plot Paul's E, all by itself, too (figure 2).

George's G, all by itself, is figure 3; Ringo's B-flat, all by itself, is figure 4.

But in *Twist and Shout*, these voices aren't alone. First there's John's C (figure 1), then John's C and Paul's E together (figure 5), then George's G gets added in (figure 6), and lastly Ringo adds his B-flat (figure 7),

Note that there is only one atmosphere between the radio and your ear; all these sound waves must be happening all at the same time, vibrating the same air molecules. It's kind of like mixing paint as I mentioned above. Well, we nonetheless hear all four voices, so somehow our ear is able to separate, or unmix, those separate frequencies.

1 The human voice is much, much richer than a simple sine wave. I'm oversimplifying for ease of presentation.

10

How does the ear (or the mind, rather) do that? I have no idea. But I'll show what a Fourier transform does with it. (In this section I'm just talking about what a Fourier transform does; I'll reserve the how for the next section.) ...

11

Figure 6: John's C, Paul's E and George's G

41 I I I I I I I I I

21

01

-2

-L

0

0.1

0.2

0.3

0.4

0.5

0.6

0.7

0.8

0.9

1

Fourier transforms of each of the singers' voices individually are shown in figures 8 through 11. A Fourier transform of John's C and Paul's E is in figure 12; a Fourier transform of after George's G is added in is in figure 13. Lastly a Fourier transform of all four of them is in figure 14.

In these transform plots, you can see the different frequencies all split out -much as a prism would do. Take a look at figure 1. Count the number of peaks in the sine wave plot. You'll see that there are 12 of them. Similarly for figures 2 through 4: 15, 18 and 21 peaks. Now look at figures 8 through 14, noting where on the horizontal axis the spikes are located: 12, 15, 18 and 21. This is easier to see in the close-up in figure 15.

So this is what a Fourier transform does: It takes some periodic waveform (sound, light, etc.) and shows what different frequencies, in what different strengths, comprise that waveform.

...

12

Figure 10: Fourier transform of George's G

2.4

Two-dimensional transforms

Here is a fourth example of spectral analysis. We alread.' looked at splitting light waveforms into colors, and splitting sound waveforms into frequencies (or pitches). In both cases, we

13

Figure 13: Fourier transform of John's a, Paul's E and George's G

300! I I I I I

200

100

-150

-100

-50

a

50

100

150

were looking at light or sound amplitude as a function of time. But there's no reason we couldn't plot something as a function of distance instead -for example, if you look at the rolling hills in western Kansas, you would notice-that there are highs and lows every several hundred feet. Or you could look at a snapshot of waves in the sea. Now, what kind of frequency is involved here? It is a spatial frequency -for the prairie, one hill per every few hundred feet; for the sea, one wave per every few dozen feet. In both cases we've got two dimensions we're looking at -north/south and east/west. In our light and sound examples, there was only one dimension, which was time. It turns out that the Fourier transform can be easily and straightforwardly extended to two dimensions -one does a one-dimensional transform of every row of the two-dimensional input, then does a one-dimensionsal transform of every column of that intermediate result. (Or, one can do columns, then rows; it turns out that it works out the same.) The examples in this section will be two-dimensional.

Suppose you are looking at some two-dimensional data which repeats itself in the horizontal direction (figure 16). Its two-dimensional Fourier transform is figure 17. Note that the waveform has four peaks in the horizontal direction; the transform has a spot at coordinates (4,0).

15

Figure 17: Fourier transform of two-dimensional waveform with horizontal frequency -15 -10 -5 0 5 10 15 Now suppose you are looking at some two-dimensional data which repeats itself in the vertical direction -figure 18. Its two-dimensional Fourier transform is figure 19. Note that the waveform has three peaks in the vertical direction; the transform has a spot at coordinates (0,3). Thirdly, suppose you have two-dimensional data whiclrtepeats itself in a diagonal direction 16 Figure 19: Fourier transform of two-dimensional waveform with vertical frequency

-15

-10

-5

0

5

figure 20, with Fourier transform in figure 21

10

15

Here, if you count peaks up the side of the graph, you'll see 5 of them; if you count peaks along the front of the graph, you'll also 5 of them. Note that the transform plot has a spot at (5, 5).

Lastly, suppose you superimpose the first three data set

17

figure 22, with Fourier transform

Figure 21: Fourier transform of two-dimensional waveform with diagonal frequency

-15

-10

in figure 23.

-5

0

5

10

15

Note how the three different spatial frequencies have been split out. Again, as we saw above, this is like unmixing paint to find its component colors, or using a prism to separate light into its colors. The Fourier transform is that magical unmixer, that mathematical prism, which recovers frequency information from a signal. In the next few sections, we'll see how it works.

18

Figure 23: Fourier transform of two-dimensional waveform with superposition of frequencies

-15

-10

-5

0

5

10

3

Complex numbers and related terms

15

First, before we define what a Fourier transform is, I want to take a few paragraphs to define some terms having to do with complex numbers, so that we're all on the same page. You're probably already familiar with complex numbers, so I really don't need to start at the very beginning -but the beginning is the easiest place to start.

19

3.1

Definitions; rectangular coordinates

We define i to be the positive square root of-l, that is, we declare that i. i = -1. 2 (The number -i has this same property: -i. -i = -1.) i times any real number is called an *imaginary* number for example, -3i, or 1.27i.

If we add a real number and an imaginary number, we get a complex number -for example, 2 + 3i. The part without the i is called the real part; the part with the i is called the imaginary part. Note that a real number, such as 5, is also a complex number (5 + Oi) -it simply has zero imaginary part. By the same token, a purely imaginary number, e.g. 2i, is also a complex number (0 + 2i) -it has a zero real part.

There are several ways to think of complex numbers. There is the algebraic notation we've been using, e.g. 2 + 3i. But we can also think of complex numbers as a pair of coordinates on a plane -we might write 2 + 3i as (2,3) (see figure 24). We're all familiar with plane graphs from our school days. But complex numbers have additional properties that ordinary coordinate pairs don't -namely, we can multiply and divide them.

3.1.1

Arithmetic operations on complex numbers

Addition and subtraction are simple: One adds (or subtracts) the real parts, then adds (or subtracts) the imaginary parts. For example, (2+3i)+(7+11i) = (2+7)+(3+11)i = 9+14i. Multiplication also is clear from the algebraic notation, remembering that i. i is -1. (a + bi) .(c + di) = (a. c) + (b .c. i) + (a. d. i) + (b .d .i. i) = (ac- bd) + (bc + ad)i. This seems like an awkward multiplication rule, but it falls right out of the definition of i.

(Complex numbers may also be divided by one another, but we'll have no need of complex division in this document.)

3.1.2

Negation and conjugation

If a complex number w is a + bi, then we say that its negative, -w, is -(a + bi), or -a -bi, meaning that the negative of a complex number is its mirror image straight through the origin.

If a complex number w is a + bi, then we say that the complex conjugate of w is a -bi; we write[1] this as w or w* (See figure 24).

[1] Engineering types tend to refer to i as j. I have no idea why the cultural difference comes about. The meaning is the same; only the name is different.

The conjugate of a number is its reflection through the real axis.

The value w*w is occasionally referred to as the power of w. (This is a term from physics -power as in watts, not power as in exponentiation.) If w is a + bi, the quantity w*w is (a + bi)(a -bi) = a2 + b2. Note that for any complex number w, w*w is always real, and always non-negative.

Polar coordinates

Just as we can think of coordinate pairs in either rectangular or polar coordinates, we can think of complex numbers in either rectangular or polar coordinates (see figure 24.)

3.2.1

Magnitude

The magnitude of a complex number is simply the straight-line distance from the origin (the complex number O+Oi) to the number. What’s the formula for this? The Pythagorean theorem comes in handy. The magnitude r of the complex number a + bi is va2 + b2. So, the magnitude of 2 + 3i is ../13, or a little less than four .

Phase

The phase of a complex number is simply the angle from the positive real axis of the complex plane, to a line connecting the origin and the number. We typically measure this angle in radians, although degrees would work too. One problem is that there is more than one way to measure an angle- is i’s phase 7r /2 (900), or -37r /2 (−270°), or 57r /2 (450°)? All of these are correct. Nonetheless we tend to adopt certain conventions, usually that phases are between -7r and 7r .(People differ as to whether the complex number -1 is said to have phase -7r , or 7r .This is just a matter of convention; it won’t be important in this document.)

The property that a complex number with phase 4¿ is the exact same as a complex number with phase ifJ + 27r (or 4¿ plus or minus any integer multiple of 27r , for that matter) is called the modularity property of phase. In figure 24 on page 21, note that the angle in polar coordinates could as well be as + 27r .

What’s the actual formula to calculate the phase of a complex number? In figure 24, we want to find the angle 0 given the length a and height b of the right triangle. The defintion of the tangent of 0 is b/a, that is, tan(O) = b/a, so..we might say that the phase of the complex number a + bi is tan-l(b/a). Let’s try that idea out, and see why it doesn’t work.

22

Using this rule, tan-l(b/a), what’s the phase of2+2i? It’s tan-l(2/2), which is 7r/4 (45) as we expect. But what about the phase of -2- 2i? We would expect -37r/4 (-135), but our formula tells us that the phase of -2- 2i is tan-l(-2/ -2), or 7r/4. Worse yet, what’s the phase of i? We’d expect 7r /2 (900) but our formula tells us tan-l(l/O); the division by zero makes this meaningless. For these reasons, what one really uses to calculate phase in a computer program is a two-argument arctangent function, namely, atan2(b. a). This function (created just for this purpose) can distinguish between 2 + 2i and -2- 2i, and can handle the case were the real part of a complex number is zero.

3.2.3

Polar to rectangular conversion

Converting the other way, from polar to rectangular coordinates, is straightforward if we look at figure 24. From trigonometry we remember that a = r cos 0 and b = r sin 0. That is, if a complex number has magnitude r and phase 0, then we can write it down in rectangular coordinates as rcosO+irsinO, or r(cosO+isinO). In the 1700's a fellow named Leonhard Euler found out that cos 0 + i sin 0 can be more compactly written down as ei9. (That's the familiar e from calculus, the base of the natural logarithm.) So, if a complex number has magnitude r and phase 0, we can write it down compactly as rei9.

Arithmetic operations in polar coordinates

What's the benefit of using polar coordinates? It turns out that multiplication of complex numbers is very simple in polar coordinates. If one complex number is rei , and another is qei4 , then their product is rei(qei4 , which is rqei .That is, we multiply the magnitudes and add the phases.

Negation and conjugation in polar coordinates

How do we negate a complex number in polar coordinates? If a complex number is rei/J , then we can negate it by writing down -rei/J .But we like to think of magnitudes as being non-negative. Alternatively, we could think of the negative of rei/J as being rei(/J+71) -that is, the same magnitude, but rotated around the origin by 11 radians (1800):

rein

rei()+7r)

Taking the conjugate of a complex number is also easy in polar coordinates. If w is a + bi in rectangular coordinates and reiD in polar coordinates, tilen w's magnitude is r = ,fa'i+ij'I; for its phase, we can say that tan(O) = b/a.

23

The magnitude of w's conjugate, w*, is r = va2 + (-b)2, or r = , which is the same as the magnitude of w. For the phase 1 of w*, we can say that tan1 = -b/a. So the phase of w* is the opposite of w's phase.

This means that if w = reifJ , then w* = re-ifJ .That is, to conjugate a complex number in polar coordinates, we simply negate the phase:

(rei9)* = re-i9

The unit circle

The unit circle on the complex plane is all the complex numbers whose magnitude is I. This means they can be written down as ei8 for some 0. If we multiply two numbers on the unit circle, then we get another number on the unit circle: ei8 .ei4 = ei(8+4 . Multiplying two numbers on the unit circle is simply a matter of adding their phases (see figure 25).

Let's beat this point into the ground (because it's so important later on when we talk about Fourier transforms) by looking at a few numbers on the unit circle. Remember that phase is measured starting from the positive real axis, and that 360° around a circle is the same as 27f radians. So, what is eiO? It's the number 1. ei21r is also the number

1. ei1r is halfway around the unit circle, so it's -1. ei1r /2 is 90° from the positive real axis, so it's i.

To summarize: eiO = I; ei11/2 = i; ei1l = -I; e-i11/2 = -i.

WN

Complex numbers on the unit circle are used so much in Fourier transforms that people use some special notations for them. We will use the terminology

WN = ei27r/N

What does this mean? It means that WN is the complex number one Nth of the way around the unit circle. For example, W2 is 1/2 of the way around the unit circle, which is -1; W4 is 1/4 of the way around the unit circle, which is i.

This w notation frees us from writing down e, i, and...7r , allowing us to focus simply on fractions of a cycle, where by cycle I mean one trip around the unit circle.

3.4.1

Powers of W N

Remember that to multiply two numbers on the unit circle, we just add their phases. So, since exponentiation is repeated multiplication, a number on the unit circle, eiO raised to the mth power, (eiO)m, is eimO. So,

wN = ei21rm/N

This is a complex number m/N of the way around the unit circle. Here is what wf, wJ and Wllook like (figure 26).

If we multiply two different powers of W N , again we just add their phases:

wm .wl - Wm+l N N- N

Note that we can think of the m and N in WN as a fraction m / N , since WN = ei27rm/ N .

In particular, we can simplify this “fraction” if we like. For example, WJ is the same as

wl, since 2/6 is the same as 1/3. (See figure 27).

3.4.2

Modularity of powers of WN

What is wl? It’s 1/4 of the way around the unit circle, so it’s i. What is wi? It’s a full trip around the unit circle, plus a fourth of another trip, so it’s also i. So we see that:

wmmodN -Wm

N -N

(1)

where by m mod N I mean the remainder, or modulus, when m is divided by N.

This means that there are only N distinct powers of W N .This fact will turn out to be useful later when we look at fast Fourier transforms. (See also figure 26).

3.4.3

Negation and conjugation using W N

There are some special properties of WN which will come in handy later on.

We saw in section 3.2.5 that to negate a complex numbdf in polar coordinates, we just add 7r to its phase. How do we negate a power of W N ? Adding 7r to the phase of a complex

26

number is the same as multiplying the number by ei1r : eiO .ei1r = ei(O+1r)

So:

= Wm+N/2

N

The conjugate of WN is of course simply WNm:

(W7l)* = wNm

See figure 28.

The cis function

.Also ei7r isWN /2

, N .

So far, we've talked about complex numbers as having fixed values, e.g. 2 + 3i or ei1T/8.

But we can also have complex-valued functions, e.g. 28 + 3t2i, or ei21Tax. The latter is particularly important in discussing Fourier transforms.

Frequency

What does the function ei27rax look like? How can we plot it? If x can take on any real value, then ei27rax can take on any value on the unit circle. Specifically, suppose x starts at 0, then increases. ei27rax starts on the unit circle at the positive real axis, moving around counterclockwise. As x increases constantly, ei27rax rotates at a constant speed. That speed is called frequency. If a is a small number, then as x increases, ei27rax will rotate slowly. If a is a big number, then as x increases, ei27rax will rotate quickly.

To be specific, let's let x go from, say, O to 1, and look at what the function ei21Tax does. If a is 1, then as x goes from O to 1, ei21Tax will go once around the unit circle, ending up where it started. If a is 2, then as x goes from O to 1, ei21Tax will go twice around the unit circle, ending up where it started. If a is 1/8, then as x goes from O to 1, ei21Tax will go only 1/8 of the way around the unit circle.

Remembering that ei27rux is the same as cos(27rax) +isin(27rax), we can plot x vs. the real and imaginary parts of ei27rux (figure 29).

Negative frequency

What about negative frequencies -what if a is negative? This means that as x increases, ei21rux goes around the unit circle backwards -clockwise, rather than counterclockwise. We

remember from trigonometry that the cosine is an even function, that is, cos(-x) = cos(x), and that the sine is an odd function, that is, sin(-x) = -sin(x). So, e-i21rax = cos(27rax) - i sin(27rax) .

Graphically, the difference between a positive and a negative frequency is that the imaginary part is upside down, while the real part is the same (figure 30).

Figure 30: ei27rax with a = -2.25. Cosine part same as above; sine part inverted.

.I I , I I I , I I

1

0.5

0

-0.5

0

0.1

0.2

0.3

0.4

0.5

0.6

0.7

0.8

0.9

1

To summarize, I define the frequency, a, of the complex-valued function f(x) = ei27rQx to be the number of counter-clockwise trips that the function's value f(x) makes around the unit circle, as x traverses the unit interval [0,1]. What do we call this e ix function? It is sometimes called a complex exponential, or a complex sinusoid; another name, from cosx + isinx, is the cis function.

31

3.5.3

Addition of frequencies

In section 3.3, we saw that when we multiply two complex numbers on the unit circle, we simply add their phases. Likewise, when we multiply two complex-valued cis functions together, we simply add their frequencies:

ei27rax .ei27rbx = ei27r(a+b)x

3.6

Why use complex numbers?

Lastly, why do we use complex numbers at all? What's the physical significance of an imaginary number anyway? Aren't imaginary numbers merely. ..imaginary? You might think that when we talk about, say, sound waves, we should use real numbers. After all, sound is air being repeatedly compressed and decompressed; there's only one parameter , pressure, to work with. However, when we look at things which vary sinusoidally, it's often mathematically useful to think of magnitude (peak absolute displacement) and phase, which are easy to manipulate using complex numbers.

32

4

Vector spaces

At this point, we're almost ready to see how Fourier transforms work. But I need to define a few more terms. It may seem that I'm going astray -but in fact, these definitions will lead us directly into the heart of Fourier transforms.

4.1

Vectors and inner products in Rn

We're all familiar with vectors in Cartesian n-space, that is, ordered n-tuples of reals. For example, vectors in R3 are simply ordered triples of real numbers, such as (3, 4, 5) or

(0, 0, 0).

The elements a, b, and c in a vector u = (a, b, c) are called scalars. For n-tuples of reals, scalars are reals; for n-tuples of complexes, scalars are complexes; etc.

For concreteness, let's continue to focus on R3 for a while.

4.1.1

R 3 as a vector space

The set of all vectors (a, b, c), that is, for any real numbers a, b and c, forms a vector space which is all of R3. (Below, we'll see that there are more restrictive subsets of Rn which are also vector spaces.

We are accustomed to doing certain operations on vectors:

- We can add or subtract two vectors by adding or subtracting corresponding elements, and this gives us another vector. For example, (1,2,3)+(4,7,9) = (1+4, 2+7, 3+9) = (5, 9, 12).
- We can multiply a vector by a scalar by multiplying each element of the vector by the scalar , and this gives us another vector. For example, (1, 2, 3) .4 = (1 .4, 2 .4, 3 .4) =

  (4,8, 12).
- Addition and multiplication are associative, commutative, left- and right-distributive,

etc.

(Any set of objects which satisfies these criteria, or axioms — not just Cartesian n-tuples — can also be called a vector space.)

4.1.2

R3 as an inner product space

Additionally, we can take the dot product or inner product of two vectors u = (a, b, c) and v = (d, e, I) by summing the products of corresponding elements: u .v = a. d + b. e + c .f .

This gives us a scalar. For example, (1,2,3) .(4,7,9) =1.4+2.7+3.9=45.

The inner product has some special properties:

- Two vectors are said to be perpendicular, or orthogonal, if their inner product is zero.
- The length, or norm, of a vector u can be defined as lul = . (This makes sense since if u = (a, b, c), then ? is just v'a2 + b2 + c2, which is the usual Euclidean distance.)
- The norm of a vector is zero only if the vector is the zero vector, that is, (0,0,0). If a vector u is non-zero, then its norm is positive. That is, ˘ is always zero or positive, never negative, for any vector u. The norm is said to be positive definite.

(Any vector space which has an inner product satisifying these criteria is called an inner product space.)

4.1.3

Orthonormality

We can say that a set of vectors is orthonormal if all the vectors in the set are orthogonal to each other, and if their norms are all 1. Since two vectors are orthogonal when their dot product is zero, we can say say more compactly that a set of vectors $\{el, e2, e3\}$ is orthonormal if the dot product ej .ek is 1 when j = k, and O when $j \neq k$.

4.1.4

Linear combinations, dimensions and bases

A linear combination of a set of vectors is the sum of each of the vectors, each multiplied by a scalar .For example, au + (3v + 'YW is a linear combination of u, v and w.

We can make a vector space of dimension m by making a set of all linear combinations of m orthonormal vectors. The number m is called the dimension of the space, and the set of orthonormal vectors is called an orthonormal basis for the space. Some more jargon: The basis vectors are also said to *span* the space.

For example, the vectors i = (1,0,0), j = (0, 1,0) and...k = (0,0, I) are the most familiar basis for R3; they are called the standard basis. We can write any vector u = (a, b, c) as a linear combination ofi,j and k: u = ai+bj+ck = a(I,O,O)+b(O, 1,0)+c(0,0, I) = (a,b,c).

34

We can think of basis vectors as building blocks with which to build any vector in the space they span: For example, we take a certain amount of i, plus a certain amount of j, plus a certain amount of k, and we get the vector we want.

4.1.5

Coordinates

The coefficients on the basis vectors are called coordinates for that basis. For example, the vector (3,4,5) is 3i + 4j + 5k; the coordinates of u for i, j, and k are 3, 4, and 5.

How do we actually calculate the coordinates for a vector? We simply dot the vector with each basis vector. The results of those dot products are the coordinates. That is, if R 3 has basis vectors el, e2, e3 , the coordinates Ck of a vector u with respect to el, e2, e3 are:

Ck

ekoU

(2)

Using this formula for the standard basis is trivial, and seems pointless. For example, given the vector u = (3,4,5), equation 2 tells us that u's coordinates with respect to i, j and k

are:

Cl

C2

C3

i.u

j.u

k.u

(1,0,0)

(0, 1,0)

(0,0,1)

(3, 4, 5)

(3, 4, 5)

(3, 4, 5)

3

4

5

(3)

Not too informative! But remember that any set of m orthonormal vectors is a basis for an m-dimensional vector space, and that any set of 3 orthonormal vectors is a basis for R3. In the next section, and in following sections, we'll use equation 2 for something a little less trivial.

4.1.6

Coordinates in alternate bases

I said above that there are more restrictive subsets of R n which are also vector spaces. For example, i and k are a basis for a two-dimensional vector space, namely, the xz plane.

An alternate basis for R3, which I'll be using for several examples, is the following:

·,

1

·,

J

k'

(../3/2,

(-1/2,

(0,

35

1/2, 0)

../3/2, 't)

0, 1)

(4)

Using non-standard bases demonstrates the usefulness of equation 2. The coordinates Cl, C2 and C3 of the vector u = (3,4,5) with respect to it, jt and kt are as follows:

4.1.7

Cl = it. u

C2 = jt .U

C3 = kt. u

-

(V3/2,

(-1/2,

(0,

1/2,

J3/2,

0,

0)

0)

1)

(3,

(3,

(3,

4,

4,

4,

5) =

5) =

5) =

Coordinate transformation as matrix multiplication

2fl:t..!

2

5

(5)

Equations 2 and 3 are sequences of dot products. But multiplying a matrix by a vector is also just a sequence of dot products: Element 1 of the product is the dot product of row 1 of the matrix with the vector; element 2 of the product is the dot product of row 2 of the matrix with the vector; etc. So equation 3 can also be written as:

[:]=[? ? ?]

[:]

(6)

The matrix-by-vector product is nothing more than several dot products, stacked on top of one another. Similarly, equation 5 can be written as:

2

5

-

J3/2

-1/2

O

1/2 O

J3/2 o

0 1

[:]

(7)

That is, to transform a vector u from from the standard basis to an alternate basis, we can simply multiply u by a matrix whose rows are the vectors in the alternate basis. If the basis vectors are ek, then we might call the matrix E.

4.1.8

The orthonormality condition as matrix multiplication

We saw in section 4.1.3 that a set of basis vectors ek is orthonormal if ej .ek is 1 when j = k, and O when $j \neq k$. Using the matrix notation from the previous section, we can write these dot products more compactly as follows: We will make one matrix whose rows are the basis vectors; we'll multiply it by a second matrix whose columns are the basis vectors:

36

J3/2

-1/2

O

1/2 O

J3/2 o

0 1

V3/2

1/2

O

-1/2 O

J3/2 o

0 1

[lOO)0 1 0

0 0 1

(8)

When we multiply one matrix by the other, the kth row and lth column of the product matrix is simply the dot product of the kth row of the first matrix with the lth column of the second matrix. If the vectors are orthonormal, those dot products will be 1 when k = l, that is, along the diagonal of the product matrix; the dot products will be O when $k \neq l$, that is, off the diagonal.

For example, from the previous section, el is (-13/2,1/2,0), which can be found in the top row of the left-hand matrix in equation 7; e2 is (-1/2, -13/2,0), which can be found in the middle column of the right- hand matrix in equation 7.

The first matrix is simply E from the previous section; the second matrix is E's transpose, ET. We can state the orthonormality condition very compactly as follows: If E is a real matrix whose rows are the vectors ek, ET is the transpose of that matrix, and I is the identity matrix (ones on the diagonal, zeros elsewhere), then the vectors are orthonormal when

4.2

E .ET = I

(9)

Vectors, inner products and orthonormality in cn

Vectors over the complex numbers are very similar to vectors over the reals, with one exception. To make the inner product positive definite, we take the complex conjugate of the first operand in the dot product:

u .v = a* .d+ b* .e + c* .f

For example, (i,i,i) .(i,i,i) = -i. i +

i.i+-i.i=1+1+1=3.

4.2.1

Coordinate transformations in Cn

As I discussed in section 4.1.5 for real vectors, we find the coordinates of a complex vector u with respect to a standard basis ek by dotting u with each basis vector: