

Appendix A

Størmer-Cowell and Nyström Integration Methods

The differential equations of classical mechanics often involve only second derivatives with no first derivatives present. In this case it is possible to work directly with the second order equations instead of converting them into a first order set of twice the dimensionality. The result can be a saving in computer time and an increase in accuracy. We shall describe methods due to *Størmer* and *Cowell* and *Nyström*. See Chapter 2 for notation.

A.1 Preliminary Derivation of Størmer-Cowell Method

Consider a set of second-order equations of the form

$$\ddot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}, t). \quad (\text{A.1.1})$$

[We remark that if a Hamiltonian is of the form $H = p \cdot p/2 + V(q, t)$, it leads to differential equations of the form (1.1).] Then, using arguments similar to those of Chapter 2, we have the integration formulas

$$\nabla^2 \mathbf{y}^{n+1} = \nabla^2 D^{-2} \mathbf{f}^{n+1}, \quad (\text{A.1.2})$$

$$\nabla^2 \mathbf{y}^{n+1} = \nabla^2 (1 - \nabla)^{-1} D^{-2} \mathbf{f}^n. \quad (\text{A.1.3})$$

By expanding (1.2) and (1.3) and using (2.4.13), we may rewrite our results as

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 [\nabla / \log(1 - \nabla)]^2 \mathbf{f}^{n+1}, \quad (\text{A.1.4})$$

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 (1 - \nabla)^{-1} [\nabla / \log(1 - \nabla)]^2 \mathbf{f}^n. \quad (\text{A.1.5})$$

As in Chapter 2, we interpret the right sides (1.4) and (1.5) in terms of power series. After truncation we obtain the predictor and corrector formulas

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 \sum_{k=0}^N \alpha_k \nabla^k \mathbf{f}^{n+1}, \quad (\text{corrector}) \quad (\text{A.1.6})$$

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 \sum_{k=0}^N \beta_k \nabla^k \mathbf{f}^n, \quad (\text{predictor}) \quad (\text{A.1.7})$$

or the expanded versions

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 \sum_{k=0}^N \tilde{\alpha}_k^N \mathbf{f}^{n+1-k}, \quad (\text{corrector}) \quad (\text{A.1.8})$$

$$\mathbf{y}^{n+1} = 2\mathbf{y}^n - \mathbf{y}^{n-1} + h^2 \sum_{k=0}^N \tilde{\beta}_k^N \mathbf{f}^{n-k}. \quad (\text{predictor}) \quad (\text{A.1.9})$$

The corrector and predictor truncation errors associated with (1.6) and (1.7) may be estimated using arguments similar to the Adams case. The result is

$$\mathbf{y}_{\text{true}}^{n+1} - \mathbf{y}_{\text{corr}}^{n+1} \approx h^{N+3} \alpha_{N+1} (d^{N+3} \mathbf{y} / dt^{N+3})|_{t=t^n}, \quad (\text{A.1.10})$$

$$\mathbf{y}_{\text{true}}^{n+1} - \mathbf{y}_{\text{pred}}^{n+1} \approx h^{N+3} \beta_{N+1} (d^{N+3} \mathbf{y} / dt^{N+3})|_{t=t^n}. \quad (\text{A.1.11})$$

Note that the predictor and corrector are one order higher accurate than their Adams counterparts. See (2.4.37) and (2.4.38). This increase in accuracy arises from the fact that the original differential equations being integrated are second order with no first derivatives present.

The coefficients α_k, β_k are listed in Table 1 below, and the associated coefficients $\tilde{\alpha}_k^N, \tilde{\beta}_k^N$ are listed in Tables 2 and 3.

Table 1

k	0	1	2	3	4	5	6	7	8	9
α_k	1	-1	$\frac{1}{12}$	0	$\frac{-1}{240}$	$\frac{-1}{240}$	$\frac{-221}{60480}$	$\frac{-19}{6048}$	$\frac{-9829}{3628800}$	$\frac{-407}{172800}$
β_k	1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{3}{40}$	$\frac{863}{12096}$	$\frac{275}{4032}$	$\frac{33953}{518400}$	$\frac{8183}{129600}$
$ \beta_k/\alpha_k $	1	0	1	∞	19	18	~ 20	~ 22	~ 24	~ 27

Table 2

The Størmer-Cowell Corrector Coefficients $\tilde{\alpha}_k^N$.

$k \ N$	2	3	4	5	6	7	8	9
0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{18}{240}$	$\frac{4315}{60480}$	$\frac{4125}{60480}$	$\frac{237671}{3628800}$	$\frac{229124}{3628800}$
1	10	10	204	209	53994	55324	3398072	3474995
2	1	1	14	4	-2307	-6297	-653032	-960724
3		0	4	14	7948	14598	1426304	2144252
4			-1	-6	-4827	-11477	-1376650	-2453572
5				1	1578	5568	884504	1961426
6					-221	-1551	-368272	-1086220
7						190	90032	397724
8							-9829	-86752
9								8547

The denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column.

Table 3
The Størmer-Cowell Predictor Coefficients $\tilde{\beta}_k^N$.

$k \ N$	2	3	4	5	6	7	8	9
0	$\frac{13}{12}$	$\frac{14}{12}$	$\frac{299}{240}$	$\frac{317}{240}$	$\frac{168398}{120960}$	$\frac{176648}{120960}$	$\frac{5537111}{3628800}$	$\frac{5766235}{3628800}$
1	-2	-5	-176	-266	-185844	-243594	-9209188	-11271304
2	1	4	194	374	317946	491196	21390668	29639132
3		-1	-96	-276	-311704	-600454	-31323196	-50569612
4			19	109	184386	473136	30831050	59700674
5				-18	-60852	-234102	-20331636	-49202260
6					8630	66380	8646188	27892604
7						-8250	-2148868	-10397332
8							237671	2299787
9								-229124

The denominator of each of the coefficients of the first line is to be repeated for all the coefficients of the corresponding column. Note that the entries for $N = 6$ and $N = 7$ are not reduced to lowest terms. Both numerator and denominator should be divided by two.

Exercises

A.1.1. Make a study of the $\tilde{\alpha}$'s and $\tilde{\beta}$'s similar to that made in Exercise 2.4.4 for the \tilde{a} 's and \tilde{b} 's.

A.2 Summed Formulation

In principle the integration formulas (1.8) and (1.9) can be used as they stand. However in practice it has been found that a so called *summed* formulation has better performance with respect to round-off errors. It also reduces the truncation error by an additional factor of h without requiring additional starting values. Because the derivation of the summed formulation is a bit involved, we shall first state the procedure, and then provide the derivation.

A.2.1 Procedure

Suppose we know the values $\mathbf{y}^0 \dots \mathbf{y}^N$ and $\mathbf{f}^0 \dots \mathbf{f}^N$ from some starting routine such as Runge-Kutta. [Note that to use standard Runge-Kutta, one must first convert the set (1.1) into a first-order set. There are variants of Runge-Kutta due to *Nystrom* that work directly with (1.1). See Section 5 at the end of this appendix.] We use the starting values to make some preparatory calculations. Define vectors \mathbf{G}^n for $n = -1, 0, \dots, N$ by the rule

$$\mathbf{G}^{-1} = 0, \quad (\text{A.2.1})$$

$$\mathbf{G}^n = h \sum_{m=0}^n \mathbf{f}^m \text{ for } n \in [0, N].$$

Next define a vector $\boldsymbol{\sigma}$ using

$$\boldsymbol{\sigma} = h^{-1}(\mathbf{y}^N - \mathbf{y}^{N-1}) - \sum_{k=0}^{N+1} \tilde{\alpha}_k^{N+1} \mathbf{G}^{N-k}. \quad (\text{A.2.2})$$

Finally, define vectors \mathbf{g}^n for $n = -1, 0, \dots, N$ by writing

$$\mathbf{g}^n = \mathbf{G}^n + \boldsymbol{\sigma}. \quad (\text{A.2.3})$$

This completes the preparatory calculations.

The integration routine itself is given by the rules

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h \sum_{k=0}^{N+1} \tilde{\alpha}_k^{N+1} \mathbf{g}^{n+1-k}, \quad (\text{corrector}) \quad (\text{A.2.4})$$

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h \sum_{k=0}^{N+1} \tilde{\beta}_k^{N+1} \mathbf{g}^{n-k}, \quad (\text{predictor}) \quad (\text{A.2.5})$$

$$\mathbf{g}^{n+1} = \mathbf{g}^n + h \mathbf{f}^{n+1}. \quad (\text{A.2.6})$$

Their truncation errors have the estimates

$$\mathbf{y}_{\text{true}}^{n+1} - \mathbf{y}_{\text{corr}}^{n+1} \approx h^{N+4} \alpha_{N+2} (d^{N+4} \mathbf{y} / dt^{N+4})|_{t=t^n}, \quad (\text{A.2.7})$$

$$\mathbf{y}_{\text{true}}^{n+1} - \mathbf{y}_{\text{pred}}^{n+1} \approx h^{N+3} \beta_{N+1} (d^{N+3} \mathbf{y} / dt^{N+3})|_{t=t^n}. \quad (\text{A.2.8})$$

Note that the corrector error is a factor of h smaller than that of the predictor. Whether or not this improvement in accuracy is realized in practice depends upon how many times the corrector is iterated. It can be shown that the simplest sequence *PEC* is insufficient. For this reason, and to check the convergence of successive iterations, it is better to use the sequences *PECEC* or *PECECE*.

A.2.2 Derivation

We now present the derivation for this procedure. We begin by rewriting (1.6) and (1.7) with an upper summation limit of $N + 1$:

$$\nabla^2 \mathbf{y}^{n+1} = h^2 \sum_{k=0}^{N+1} \alpha_k \nabla^k \mathbf{f}^{n+1}, \quad (\text{corrector}) \quad (\text{A.2.9})$$

$$\nabla^2 \mathbf{y}^{n+1} = h^2 \sum_{k=0}^{N+1} \beta_k \nabla^k \mathbf{f}^n. \quad (\text{predictor}) \quad (\text{A.2.10})$$

We observe that the vectors \mathbf{g}^n obey the rules

$$\mathbf{g}^{-1} = \boldsymbol{\sigma}, \quad (\text{A.2.11})$$

$$\mathbf{g}^n = h \sum_{m=0}^n \mathbf{f}^m + \boldsymbol{\sigma} \quad \text{for } n \geq 0.$$

It is easily checked that

$$\nabla \mathbf{g}^n = h \mathbf{f}^n \quad \text{for } n \geq 0. \quad (\text{A.2.12})$$

Insert (2.12) into (2.9). The result is

$$\nabla^2 \mathbf{y}^{n+1} = h \nabla \sum_{k=0}^{N+1} \alpha_k \nabla^k \mathbf{g}^{n+1}. \quad (\text{A.2.13})$$

Suppose we could peel off a ∇ from each side of (2.13). The result would be

$$\nabla \mathbf{y}^{n+1} = h \sum_{k=0}^{N+1} \alpha_k \nabla^k \mathbf{g}^{n+1}. \quad (\text{A.2.14})$$

Normally this operation is not justified since, by (2.4.5), the two sides of (2.14) could differ by a *constant* vector. However, in our case we assert that the value of $\boldsymbol{\sigma}$ was cleverly defined in (2.2) to insure that (2.14) would be correct. To check this claim, set $n+1 = N$ in (2.14). The result, using (2.3), is

$$\nabla \mathbf{y}^N = h \alpha_o \boldsymbol{\sigma} + h \sum_{k=0}^{N+1} \alpha_k \nabla^k \mathbf{G}^N. \quad (\text{A.2.15})$$

From Table 1 we find $\alpha_o = 1$, and, after expansion, we see that (2.15) is equivalent to (2.2). Thus (2.14) is correct. Upon expansion it gives the corrector (2.4).

Let us now work on the predictor (2.10). Use of (2.12) gives

$$\nabla^2 \mathbf{y}^{n+1} = h \nabla \sum_{k=0}^{N+1} \beta_k \nabla^k \mathbf{g}^n. \quad (\text{A.2.16})$$

Again we peel off a ∇ from each side, but this time we must insert a constant vector \mathbf{c} . The result is

$$\nabla \mathbf{y}^{n+1} = h \sum_{k=0}^{N+1} \beta_k \nabla^k \mathbf{g}^n + \mathbf{c}. \quad (\text{A.2.17})$$

What should \mathbf{c} be? We shall see that to within sufficient accuracy it can be set to zero. Assuming this to be the case, the expansion of (2.17) gives the predictor (2.5).

We now estimate the size of \mathbf{c} . From the analog of (2.4.20) for \mathbf{g} we find

$$\mathbf{g}^n = (1 - \nabla) \mathbf{g}^{n+1} \quad (\text{A.2.18})$$

so that (2.17) can also be written as

$$\nabla \mathbf{y}^{n+1} = h(1 - \nabla) \sum_{k=0}^{N+1} \beta_k \nabla^k \mathbf{g}^{n+1} + \mathbf{c}. \quad (\text{A.2.19})$$

Subtract (2.14) from (2.19). The result is

$$\mathbf{c} = h \left[\sum_{k=0}^{N+1} \alpha_k \nabla^k - (1 - \nabla) \sum_{k=0}^{N+1} \beta_k \nabla^k \right] \mathbf{g}^{n+1}. \quad (\text{A.2.20})$$

From their definition in (1.4) and (1.5), the coefficients α_k and β_k satisfy the identity

$$(1 - z) \sum_0^\infty \beta_k z^k = \sum_0^\infty \alpha_k z^k. \quad (\text{A.2.21})$$

It follows that

$$\sum_{k=0}^{N+1} \alpha_k \nabla^k - (1 - \nabla) \sum_{k=0}^{N+1} \beta_k \nabla^k = \beta_{N+1} \nabla^{N+2}. \quad (\text{A.2.22})$$

Using (2.12) and (2.22), the expression for \mathbf{c} can be simplified to give

$$\mathbf{c} = h^2 \beta_{N+1} \nabla^{N+1} \mathbf{f}^{n+1}. \quad (\text{A.2.23})$$

Finally, remembering that $\mathbf{f} = \ddot{\mathbf{y}}$ and using (2.4.12) we find

$$\mathbf{c} \approx h^{N+3} \beta_{N+1} (d^{N+3} \mathbf{y} / dt^{N+3})|_{t=t^n}. \quad (\text{A.2.24})$$

Since (2.17) is equivalent to (2.10), we know it has the intrinsic error given in (1.11) with N replaced by $N+1$. This error is of order h smaller than \mathbf{c} . Thus the error made in dropping \mathbf{c} is the dominant predictor error, and (2.8) is correct.

Exercises

A.2.1. Verify (2.21).

A.3 Computation of First Derivative

It often happens that values of $\dot{\mathbf{y}}$ are required at various points of a trajectory. For example, from time to time we may need the velocity to compute the energy. If the trajectory is being integrated with the Adams method, the velocity is available at each step. However, with Størmer-Cowell only the coordinates \mathbf{y}^n are computed.

This apparent defect can be overcome with numerical differentiation. We observe that by using (2.4.13) the equation

$$\dot{\mathbf{y}}^{n+1} = D\mathbf{y}^{n+1} \quad (\text{A.3.1})$$

can be written in the form

$$\dot{\mathbf{y}}^{n+1} = -h^{-1} \log(1 - \nabla) \mathbf{y}^{n+1}. \quad (\text{A.3.2})$$

To use (3.2) as it stands requires the storage of previous \mathbf{y} 's. However, we may rewrite it in the form

$$\dot{\mathbf{y}}^{n+1} = -h^{-1} [\log(1 - \nabla)/\nabla] \nabla \mathbf{y}^{n+1}, \quad (\text{A.3.3})$$

or using (2.14),

$$\dot{\mathbf{y}}^{n+1} = -[\log(1 - \nabla)/\nabla] \sum_0^{N+1} \alpha_k \nabla^k \mathbf{g}^{n+1}. \quad (\text{A.3.4})$$

From the definition of the coefficients a_k and α_k [see (2.4.23), (1.4), and (1.6)] we learn that

$$-[\log(1 - z)/z] \sum_0^\infty \alpha_k z^k = \sum_0^\infty a_k z^k. \quad (\text{A.3.5})$$

Consequently, we also may write to within sufficient accuracy

$$\dot{\mathbf{y}}^{n+1} = \sum_{k=0}^{N+2} a_k \nabla^k \mathbf{g}^{n+1}, \quad (\text{A.3.6})$$

or expanding out,

$$\dot{\mathbf{y}}^{n+1} = \sum_{k=0}^{N+2} \tilde{a}_k^{N+2} \mathbf{g}^{n+1-k}. \quad (\text{A.3.7})$$

We conclude that $\dot{\mathbf{y}}$ can be computed in terms of the stored \mathbf{g} 's any time it is required. [If the reader is wondering about the upper summation limit of $N + 2$ in (3.6) and (3.7), it is not a misprint. He or she should see Exercise 4.2 at the end of Section A.4.]

Exercises

A.3.1. Verify (3.5).

A.4 Example Program and Numerical Results

We show below, with some associated subroutines, a summed Størmer-Cowell program.

A.4.1 Program

The program is written to solve (2.2.5) with the initial conditions (2.2.6). We have set $N = 3$ and $h = 1/10$, and the solution is initiated with the Runge-Kutta routine rk3 using a step size of $h/20$.

```

c This is the main program for illustrating a Stormer-Cowell
c method for numerical integration.
c
      implicit double precision (a-h,o-z)
c
c Print heading.
c
      write(6,100)
100 format
& (1h , 'time', 4x, 'ycomp', 10x, 'ydcomp', 10x, 'ytrue',
& 10x, 'ydtrue', /)
c
c Set up initial conditions and parameters. n is the number of integration
c steps we wish to make.
c
      t=0.d0
      h=.1d0
      n=15
      y=0.d0
      ydot=1.d0
c
      call sc(t,h,n,y,ydot)
c
      end
c
c This is a sixth order Stormer-Cowell integration subroutine.
c
      subroutine sc(t,h,n,y,ydot)
      implicit double precision (a-h,o-z)
      dimension g(5)
c
      write(6,*) 'Starting with Runge-Kutta integration'
c
c Set up initial g values.
c
      g(1)=0.d0
      call evalsc(y,t,f)
      g(2)=h*f
      call prints(t,y,ydot,y1true(t),y2true(t),0)
      do 10 i=2,4
      call rk3(t,h/20.d0,20,y,ydot)
      call evalsc(y,t,f)
      g(i+1)=g(i)+h*f
      if(i .eq. 3) yb=y
      call prints(t,y,ydot,y1true(t),y2true(t),0)
10  continue
      sigma=(y-yb)/h-(1.d0/240.d0)*
& (19.d0*g(5)+204.d0*g(4)+14.d0*g(3)+4.d0*g(2))
      do 20 i=1,5
20  g(i)=g(i)+sigma
      hdiv=h/240.d0
      n=n-3
      tint=t
      write (6,*) 'Continuing with Stormer-Cowell integration'

```

```

c
c Printing and integration loop.
c
c      do 100 i=1,n
c
c Predictor step.
c
t=t+h
p=y+hdiv*(299.d0*g(5)-176.d0*g(4)+194.d0*g(3)
& -96.d0*g(2)+19.d0*g(1)
call evalsc(p,t,f)
g6=g(5)+h*f
call dif(g,g6,ydot)
call prints(t,p,ydot,y1true(t),y2true(t),0)
c
c Corrector steps.
c
do 50 j=1,3
c=y+hdiv*(19.d0*g6+204.d0*g(5)+14.d0*g(4)
& +4.d0*g(3)-1.d0*g(2))
call evalsc(c,t,f)
g6=g(5)+h*f
call dif(g,g6,ydot)
call prints(t,c,ydot,y1true(t),y2true(t),1)
50 continue
c
c Update gs
c
do 60 j=1,4
60 g(j)=g(j+1)
g(5)=g6
y=c
t=tint+float(i)*h
100 continue
c
      return
      end

c This subroutine computes ydot from the g values.
c
subroutine dif(g,g6,ydot)
implicit double precision (a-h,o-z)
dimension g(5)
c
ydot=(1.d0/1440.d0)*(475.d0*g6+1427.d0*g(5)-798.d0*g(4)
& +482.d0*g(3)-173.d0*g(2)+27.d0*g(1))
c
      return
      end

c
c This subroutine evaluates f, the right side of the
c differential equation for the second order set.
c
subroutine evalsc(y,t,f)

```

```

    implicit double precision (a-h,o-z)
c
f=2.d0*t-y
c
return
end

```

A.4.2 Numerical Results

Below are the results of running this program. The format of the column *ycomp* is the same as that of *y1comp* in Example 2.4.1. The column *ydcomp* contains values of \dot{y} computed using (A.42). We observe that the solution is accurate to essentially eight significant figures.

time	ycomp	ydcomp	ytrue	ydtrue
Starting with Runge-Kutta integration				
0.0000	0.00000000E+00	0.10000000E+01	0.00000000E+00	0.10000000E+01
0.1000	0.10016658E+00	0.10049958E+01	0.10016658E+00	0.10049958E+01
0.2000	0.20133067E+00	0.10199334E+01	0.20133067E+00	0.10199334E+01
0.3000	0.30447979E+00	0.10446635E+01	0.30447979E+00	0.10446635E+01
Continuing with Stormer-Cowell integration				
0.4000	0.41058164E+00	0.10789390E+01	0.41058166E+00	0.10789390E+01
	0.41058166E+00	0.10789390E+01		
	0.41058166E+00	0.10789390E+01		
	0.41058166E+00	0.10789390E+01		
0.5000	0.52057444E+00	0.11224174E+01	0.52057446E+00	0.11224174E+01
	0.52057446E+00	0.11224174E+01		
	0.52057446E+00	0.11224174E+01		
	0.52057446E+00	0.11224174E+01		
0.6000	0.63535750E+00	0.11746644E+01	0.63535753E+00	0.11746644E+01
	0.63535753E+00	0.11746644E+01		
	0.63535753E+00	0.11746644E+01		
	0.63535753E+00	0.11746644E+01		
0.7000	0.75578228E+00	0.12351578E+01	0.75578231E+00	0.12351578E+01
	0.75578232E+00	0.12351578E+01		
	0.75578232E+00	0.12351578E+01		
	0.75578232E+00	0.12351578E+01		
0.8000	0.88264387E+00	0.13032933E+01	0.88264391E+00	0.13032933E+01
	0.88264392E+00	0.13032933E+01		
	0.88264392E+00	0.13032933E+01		
	0.88264392E+00	0.13032933E+01		
0.9000	0.10166730E+01	0.13783900E+01	0.10166731E+01	0.13783900E+01
	0.10166731E+01	0.13783900E+01		
	0.10166731E+01	0.13783900E+01		
	0.10166731E+01	0.13783900E+01		
1.0000	0.11585290E+01	0.14596977E+01	0.11585290E+01	0.14596977E+01
	0.11585290E+01	0.14596977E+01		
	0.11585290E+01	0.14596977E+01		
	0.11585290E+01	0.14596977E+01		
1.1000	0.13087926E+01	0.15464039E+01	0.13087926E+01	0.15464039E+01
	0.13087927E+01	0.15464039E+01		
	0.13087927E+01	0.15464039E+01		
	0.13087927E+01	0.15464039E+01		

1.2000	0.14679609E+01	0.16376422E+01	0.14679609E+01	0.16376422E+01
	0.14679609E+01	0.16376422E+01		
	0.14679609E+01	0.16376422E+01		
	0.14679609E+01	0.16376422E+01		
1.3000	0.16364418E+01	0.17325012E+01	0.16364418E+01	0.17325012E+01
	0.16364418E+01	0.17325012E+01		
	0.16364418E+01	0.17325012E+01		
	0.16364418E+01	0.17325012E+01		
1.4000	0.18145502E+01	0.18300328E+01	0.18145503E+01	0.18300329E+01
	0.18145503E+01	0.18300328E+01		
	0.18145503E+01	0.18300328E+01		
	0.18145503E+01	0.18300328E+01		
1.5000	0.20025050E+01	0.19292628E+01	0.20025050E+01	0.19292628E+01
	0.20025050E+01	0.19292628E+01		
	0.20025050E+01	0.19292628E+01		
	0.20025050E+01	0.19292628E+01		

Exercises

A.4.1. Compare the error estimate (2.7) with the actual error made in the example above.

A.4.2. Check the derivation of (3.7) and find a formula similar to (2.7) for the expected error in $\dot{\mathbf{y}}$. Compare with the error in the example above. Suppose the sum in (3.6) were terminated at $N + 1$. Show that the error in its expanded form, the analog of (3.7), would then be larger.

A.5 Nyström Runge-Kutta Methods

We close this appendix with a brief description of Nyström Runge-Kutta (NRK) methods. They are analogous to ordinary Runge-Kutta methods, but are designed to work directly with second-order equations of the form (1.1). We will present methods that are analogous to the Runge-Kutta methods RK3 given by (2.3.2), (2.3.3) and RK4 given by (2.3.4), (2.3.5). Unlike Størmer-Cowell methods, we will need integration formulas for both \mathbf{y} and $\dot{\mathbf{y}}$.

The method that is analogous to RK3 is given by

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h\dot{\mathbf{y}}^n + (h^2/6)(\mathbf{a} + 2\mathbf{b}), \quad (\text{A.5.1})$$

$$\dot{\mathbf{y}}^{n+1} = \dot{\mathbf{y}}^n + (h/6)(\mathbf{a} + 4\mathbf{b} + \mathbf{c}), \quad (\text{A.5.2})$$

where at each step

$$\mathbf{a} = \mathbf{f}(\mathbf{y}^n, t^n), \quad (\text{A.5.3})$$

$$\mathbf{b} = \mathbf{f}[\mathbf{y}^n + (h/2)\dot{\mathbf{y}}^n + (h^2/8)\mathbf{a}, t^n + h/2], \quad (\text{A.5.4})$$

$$\mathbf{c} = \mathbf{f}[\mathbf{y}^n + h\dot{\mathbf{y}}^n + (h^2/2)\mathbf{b}, t^n + h]. \quad (\text{A.5.5})$$

This is a three-stage fourth-order method. That is, it is locally correct through order h^4 , and makes local errors of order h^5 . Note that this method is one order higher in accuracy than its counterpart RK3. Accordingly, we will call it NRK4. This increase in accuracy

again arises from the fact that the original differential equations being integrated are second order with no first derivatives present.

The method that is analogous to RK4 is given by

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h\dot{\mathbf{y}}^n + (h^2/192)(23\mathbf{a} + 75\mathbf{b} - 27\mathbf{c} + 25\mathbf{d}), \quad (\text{A.5.6})$$

$$\dot{\mathbf{y}}^{n+1} = \dot{\mathbf{y}}^n + (h/192)(23\mathbf{a} + 125\mathbf{b} - 81\mathbf{c} + 125\mathbf{d}), \quad (\text{A.5.7})$$

where at each step

$$\mathbf{a} = \mathbf{f}(\mathbf{y}^n, t^n), \quad (\text{A.5.8})$$

$$\mathbf{b} = \mathbf{f}[\mathbf{y}^n + (2/5)h\dot{\mathbf{y}}^n + (2/25)h^2\mathbf{a}, t^n + (2/5)h], \quad (\text{A.5.9})$$

$$\mathbf{c} = \mathbf{f}[\mathbf{y}^n + (2/3)h\dot{\mathbf{y}}^n + (2/9)h^2\mathbf{a}, t^n + (2/3)h], \quad (\text{A.5.10})$$

$$\mathbf{d} = \mathbf{f}[\mathbf{y}^n + (4/5)h\dot{\mathbf{y}}^n + (4/25)h^2(\mathbf{a} + \mathbf{b}), t^n + (4/5)h]. \quad (\text{A.5.11})$$

This is a four-stage fifth-order method, and is again one order higher in accuracy than its counterpart RK4. Accordingly, we will call it NRK5.

Nyström Runge-Kutta methods can also be described in terms of Butcher tableaux. Let $\bar{\mathbf{b}}$, \mathbf{b} , and \mathbf{c} be s -dimensional vectors with real entries, and let $\bar{\mathbf{a}}$ be an $s \times s$ matrix with real entries. Consider stepping formulas of the form

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h\dot{\mathbf{y}}^n + h^2 \sum_{i=1}^s \bar{b}_i \mathbf{k}_i, \quad (\text{A.5.12})$$

$$\dot{\mathbf{y}}^{n+1} = \dot{\mathbf{y}}^n + h \sum_{i=1}^s b_i \mathbf{k}_i, \quad (\text{A.5.13})$$

where at each step

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}^n + hc_i \dot{\mathbf{y}}^n + h^2 \sum_{j=1}^s \bar{a}_{ij} \mathbf{k}_j, t^n + c_i h). \quad (\text{A.5.14})$$

Evidently the procedures (5.1) through (5.5) and (5.6) through (5.11) are of this form.

In terms of the notation just introduced, the general problem now is to impose various conditions on the vectors $\bar{\mathbf{b}}$, \mathbf{b} , and \mathbf{c} and the matrix $\bar{\mathbf{a}}$ so that the integration method will be of some particular order m , and perhaps have some other desirable properties. For this purpose, it is convenient to arrange the vectors $\bar{\mathbf{b}}$, \mathbf{b} , and \mathbf{c} and the matrix $\bar{\mathbf{a}}$ into a tableau (again called a Butcher tableau) of the form

$$\begin{array}{c|ccc} c_1 & \bar{a}_{11} & \cdots & \bar{a}_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & \bar{a}_{s1} & \cdots & \bar{a}_{ss} \\ \hline & \bar{b}_1 & \cdots & \bar{b}_s \\ \hline & b_1 & \cdots & b_s \end{array}. \quad (\text{A.5.15})$$

The Butcher tableau for NRK4, the method (5.1) through (5.5), is

$$\begin{array}{c|ccc}
 0 & 0 & 0 & 0 \\
 1/2 & 1/8 & 0 & 0 \\
 1 & 0 & 1/2 & 0 \\
 \hline
 & 1/6 & 2/6 & 0 \\
 & 1/6 & 4/6 & 1/6
 \end{array} . \quad (\text{A.5.16})$$

The Butcher tableau for NRK5, the method (5.6) through (5.11), is

$$\begin{array}{c|ccccc}
 0 & 0 & 0 & 0 & 0 \\
 2/5 & 2/25 & 0 & 0 & 0 \\
 2/3 & 2/9 & 0 & 0 & 0 \\
 4/5 & 4/25 & 4/25 & 0 & 0 \\
 \hline
 & 23/192 & 75/192 & -27/192 & 25/192 \\
 & 23/192 & 125/192 & -81/192 & 125/192
 \end{array} . \quad (\text{A.5.17})$$

At this point we observe that, as in the case of ordinary Runge-Kutta, it is sometimes useful to rewrite the relations (5.12) through (5.14) in a somewhat different form. At each step introduce intermediate times t_i and coordinates \mathbf{y}_i by the rules

$$t_i = t^n + c_i h, \quad (\text{A.5.18})$$

$$\mathbf{y}_i = \mathbf{y}^n + h c_i \dot{\mathbf{y}}^n + h^2 \sum_{j=1}^s \bar{a}_{ij} \mathbf{k}_j. \quad (\text{A.5.19})$$

With this convention (5.14) can be rewritten in the form

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}_i, t_i). \quad (\text{A.5.20})$$

Finally we copy (5.12) and (5.13) and place them last,

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h \dot{\mathbf{y}}^n + h^2 \sum_{i=1}^s \bar{b}_i \mathbf{k}_i, \quad (\text{A.5.21})$$

$$\dot{\mathbf{y}}^{n+1} = \dot{\mathbf{y}}^n + h \sum_{i=1}^s b_i \mathbf{k}_i, \quad (\text{A.5.22})$$

Evidently the relations (5.18) through (5.22) are equivalent to the relations (5.12) through (5.14), but in this expanded form it is clear that the \mathbf{k}_i are the values of \mathbf{f} at the intermediate points t_i , \mathbf{y}_i .

Finally, we remark that there are Nyström counterparts to embedded Runge-Kutta pairs so that it is possible to develop Nyström procedures with adaptive step-size control. Also, as described in Section 12.2, there are symplectic Nyström-Runge-Kutta procedures. See the books of Hairer *et al.* and Sanz-Serna and Calvo listed in the bibliography at the end of this appendix.

Exercises

A.5.1. Apply one step of the fourth-order Nyström method (5.1) through (5.5) to the differential equation

$$\ddot{y} = t^2 \quad (\text{A.5.23})$$

with the initial conditions

$$y(0) = 0 \text{ and } \dot{y}(0) = 0. \quad (\text{A.5.24})$$

That is, use (5.1) and (5.2) to compute $y(h)$ and $\dot{y}(h)$. Verify that your result has the advertised accuracy. Repeat the calculation for the case $\ddot{y} = t^3$, and show that, as expected, the result is not exact.

Bibliography

- [1] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*. (John Wiley 1962) QA 372.H48. *Error Propagation for Difference Methods*. (John Wiley 1963) QA 431.H44.
- [2] J.F. Frankena, “Størmer-Cowell: straight, summed and split. An overview”, *J. Computational and Applied Mathematics* **62**, p. 129-154 (1995).
- [3] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Non-stiff Problems*, Springer (1993).
- [4] J.M. Sanz-Serna and M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall (1994) or Dover (1994).

Appendix B

Computer Programs for Numerical Integration

In this appendix we list computer programs that are more efficient versions of those described in Chapter 2. The programs are all subroutines, and need to be supplemented by a main calling program and input and output statements of the reader's own design. For the most part, the programs are self-explanatory. All the integration programs call the subroutine *eval* (or *feval*), which computes the vector \mathbf{f} appearing in the differential equation $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t)$. To change from one set of differential equations to another, it is only necessary to change *eval*. The integration routines themselves are general purpose. One only need specify *ne*, the number of equations to be simultaneously integrated. As listed, the routines are set up to integrate the equation used in the Examples in Chapter 2, *i.e.* $\ddot{x} + x = 2t$.

The routines given in this appendix are suitable for *semi-serious* use. Readers who wish to pursue numerical integration in a serious way may wish to use a canned integration package. Much work has gone into writing some such routines. However, the reader should be sure to understand how the package he or she has selected actually works. Some produce unpleasant surprises. See the references at the end of Chapter 2.

B.1 A 3rd Order Runge-Kutta Routine

B.1.1 Butcher Tableau for *RK3*

The Butcher tableau for *RK3* is given in (2.3.9).

B.1.2 The Routine *RK3*

```

subroutine rk3 (h,ns,nf,t,y)
c This is a Runge Kutta routine that makes local errors of order
c h**4. h is the step size, ns is the number of steps, t is the
c time, and y is the dependent variable array. Finally, nf is a
c flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne)
c yt is a temporary storage array and f is ydot.
c a, b, and c are used in integration.
      tint=t
c tint is the initial time.
      do 100 i=1,ns
      call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
10   a(j)=h*f(j)
      do 20 j=1,ne
20   yt(j)=y(j)+.5*a(j)
      tt=t+.5*h
c tt is a temporary time
      call eval (tt,yt,f)
      do 30 j=1,ne
30   b(j)=h*f(j)
      do 40 j=1,ne
40   yt(j)=y(j)+2.*b(j)-a(j)
      tt=t+h
      call eval (tt,yt,f)
      do 50 j=1,ne
50   c(j)=h*f(j)
      do 60 j=1,ne
60   y(j)=y(j)+(a(j)+4.*b(j)+c(j))/6.
      t=tint+float(i)*h
100 continue
      return
      end

```

Note: The *flag* *nf* is not actually used. It is incorporated to make the program easier to modify.

B.2 A 4th Order Runge-Kutta Routine

B.2.1 Butcher Tableau for *RK4*

The Butcher tableau for *RK4* is given in (2.3.10).

B.2.2 The Routine *RK4*

```

subroutine rk4 (h,ns,nf,t,y)
c This is a Runge Kutta routine that makes local errors of order
c h**5. h is the step size, ns is the number of steps, t is the
c time, and y is the dependent variable array. Finally, nf is a
c flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),d(ne)
c yt is a temporary storage array and f is ydot.
c a, b, c, and d are used in integration.
      tint=t
c tint is the initial time.
      do 100 i=1,ns
         call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
10   a(j)=h*f(j)
      do 20 j=1,ne
20   yt(j)=y(j)+.5*a(j)
      tt=t+.5*h
c tt is a temporary time
      call eval (tt,yt,f)
      do 30 j=1,ne
30   b(j)=h*f(j)
      do 40 j=1,ne
40   yt(j)=y(j)+.5*b(j)
      call eval (tt,yt,f)
      do 50 j=1,ne
50   c(j)=h*f(j)
      do 60 j=1,ne
60   yt(j)=y(j)+c(j)
      tt=t+h
      call eval (tt,yt,f)
      do 70 j=1,ne
70   d(j)=h*f(j)
      do 80 j=1,ne
80   y(j)=y(j)+(a(j)+2.*b(j)+2.*c(j)+d(j))/6.
      t=tint+float(i)*h
100 continue
      return
      end

```

Note: The *flag nf* is not actually used. It is incorporated to make the program easier to modify.

B.3 A Subroutine to Compute f

```
subroutine eval (t,y,f)
c This is a subroutine that evaluates ydot.
dimension y(*),f(*)
c In the following lines put in the expressions for the f(i).
f(1)=y(2)
f(2)=2.*t-y(1)
return
end
```

B.4 A Partial Double-Precision Version of RK3

```

subroutine rk3pdp (h,ns,nf,t,y)
c This is a Runge Kutta routine that works in partial double precision
c and makes local errors of order h**4. h is the step size, ns is the
c number of steps, t is the time, and y is the dependent variable array.
c Finally, nf is a flag to control whatever.
c In the next line, put in after ne the number of equations to be
c integrated.
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),yd(ne)
      double precision yd
c yt is a temporary storage array and f is ydot.
c yd is a double precision storage array.
c a, b, and c are used in integration.
      tint=t
c tint is the initial time.
      do 2 j=1,ne
         2 yd(j)=dble(y(j))
c The input array y is transferred into the double precision array yd.
c Beginning of rk3 loop.
      do 100 i=1,ns
         call eval (t,y,f)
c eval is a subroutine that evaluates ydot.
      do 10 j=1,ne
         10 a(j)=h*f(j)
         do 20 k=1,ne
            20 yt(j)=y(j)+.5*a(j)
            tt=tt+.5*h
c tt is a temporary time
         call eval (tt,yt,f)
         do 30 j=1,ne
            30 b(j)=h*f(j)
            do 40 j=1,ne
               40 yt(j)=y(j)+2.*b(j)-a(j)
               tt=t+h
               call eval (tt,yt,f)
               do 50 j=1,ne
                  50 c(j)=h*f(j)
                  do 60 j=1,ne
                     60 yd(j)=yd(j)+dble((a(j)+4.*b(j)+c(j))/6.)
c The array yd is incremented in double precision.
                     t=tint+float(i)*h
                     do 70 j=1,ne
                        70 y(j)=sngl(yd(j))
c Preparation of y for transfer out or the next run through the loop.
100 continue
      return
      end

```

The *error curve* for this routine is exactly the same as that given in Figure (2.3.1) when the step size h is .05 or larger. However, for smaller h the error curve is much better. The error continues to decrease as h^3 until h reaches a value a little less than 10^{-7} and then remains approximately constant as h is decreased further. This is because the only serious round-off error occurs in the statement $y(j) = sngl(yd(j))$, and this error is independent of the number of steps. We see that partial double precision is worthwhile if good accuracy is required.

B.5 A 6th Order 8 Stage Runge-Kutta Routine

This section describes a sixth-order eight-stage Runge-Kutta routine. Note that, according to Table 2.3.1, it should be possible to achieve sixth-order accuracy using seven stages. Therefore the routine in this section, while workable, is not optimal with regard to employing only the minimum number of required stages. On the other hand, again according to Table 2.3.1, there is no eight-stage method that has an order higher than six.

B.5.1 Butcher Tableau for $RK6$

The Butcher tableau for RK6 is

0	0	0	0	0	0	0	0	0
1/2	1/2	0	0	0	0	0	0	0
1/2	0	1/2	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	1	0	0	0	0	0
	1/6	2/6	2/6	1/6	*	*	*	*

(B.5.1)

B.5.2 The Routine $RK6$

```

subroutine rk6(h,ns,t,y)
c Written by Rob Ryne, Spring 1986, based on a routine of
c J. Milutinovic.
c For a reference, see page 76 of F. Ceschino and J. Kuntzmann,
c Numerical Solution of Initial Value Problems, Prentice Hall 1966.
c This integration routine makes local truncation errors at each
c step of order h**7.
c That is, it is locally correct through terms of order h**6.
c Each step requires 8 function evaluations: The method has
c 8 stages.
c
      implicit double precision (a-h,o-z)
c
      parameter (ne=2)
      dimension y(ne),yt(ne),f(ne),a(ne),b(ne),c(ne),d(ne),
# e(ne),g(ne),o(ne),p(ne)
c
      tint=t
      do 200 i=1,ns
      call feval(t,y,f)
      do 10 j=1,ne
10    a(j)=h*f(j)
      do 20 j=1,ne
20    yt(j)=y(j)+a(j)/9.d+0
      tt=t+h/9.d+0
      call feval(tt,yt,f)

```

```

      do 30 j=1,ne
30 b(j)=h*f(j)
      do 40 j=1,ne
40 yt(j)=y(j) + (a(j) + 3.d+0*b(j))/24.d+0
      tt=t+h/6.d+0
      call feval(tt,yt,f)
      do 50 j=1,ne
50 c(j)=h*f(j)
      do 60 j=1,ne
60 yt(j)=y(j)+(a(j)-3.d+0*b(j)+4.d+0*c(j))/6.d+0
      tt=t+h/3.d+0
      call feval(tt,yt,f)
      do 70 j=1,ne
70 d(j)=h*f(j)
      do 80 j=1,ne
80 yt(j)=y(j) + (-5.d+0*a(j) + 27.d+0*b(j) -
# 24.d+0*c(j) + 6.d+0*d(j))/8.d+0
      tt=t+.5d+0*h
      call feval(tt,yt,f)
      do 90 j=1,ne
90 e(j)=h*f(j)
      do 100 j=1,ne
100 yt(j)=y(j) + (221.d+0*a(j) - 981.d+0*b(j) +
# 867.d+0*c(j)- 102.d+0*d(j) + e(j))/9.d+0
      tt = t+2.d+0*h/3.d+0
      call feval(tt,yt,f)
      do 110 j=1,ne
110 g(j)=h*f(j)
      do 120 j=1,ne
120 yt(j) = y(j)+(-183.d+0*a(j)+678.d+0*b(j)-472.d+0*c(j)-
# 66.d+0*d(j)+80.d+0*e(j) + 3.d+0*g(j))/48.d+0
      tt = t + 5.d+0*h/6.d+0
      call feval(tt,yt,f)
      do 130 j=1,ne
130 o(j)=h*f(j)
      do 140 j=1,ne
140 yt(j) = y(j)+(716.d+0*a(j)-2079.d+0*b(j)+1002.d+0*c(j) +
# 834.d+0*d(j)-454.d+0*e(j)-9.d+0*g(j)+72.d+0*o(j))/82.d+0
      tt = t + h
      call feval(tt,yt,f)
      do 150 j=1,ne
150 p(j)=h*f(j)
      do 160 j=1,ne
160 y(j) = y(j)+(41.d+0*a(j)+216.d+0*c(j)+27.d+0*d(j) +
# 272.d+0*e(j)+27.d+0*g(j)+216.d+0*o(j)+41.d+0*p(j))/840.d+0
      t=tint+i*h
200 continue
      return
      end

```

Note: This program calls the subroutine `feval`, which is another name for `eval`.

B.6 Embedded Runge-Kutta Pairs

In this section we will describe the construction of Runge-Kutta pairs and provide two examples. Our discussion here is meant to provide background, but not actual code. As in the case of other adaptive codes, much time has been spent by professional mathematicians and numerical analysts writing optimal code for embedded Runge-Kutta procedures. Readers are advised not to try writing such code on their own without first exploring existing programs and without being prepared to expend considerable time and effort.

B.6.1 Preliminaries

Section 2.5.1 sketched the possibility of pairs of Runge-Kutta methods whose orders differ (usually by one) and that, in making one integration step, share many or all intermediate evaluation points. By subtracting the higher-order result from the lower-order result, one can estimate the error in the lower-order result, and adjust the step size accordingly. In this section we will describe two examples of Runge-Kutta pairs. Each example has the feature that both methods of each pair employ the same \mathbf{k} vectors. Thus both methods can be carried out simultaneously with little added expense.

Equations (2.3.6) through (2.3.8) described the general Runge-Kutta method and its characterization by an associated Butcher table. If there are two methods that utilize the same \mathbf{k} vectors, we may make definitions of the form

$$\mathbf{y}^{n+1} = \mathbf{y}^n + h \sum_{i=1}^s b_i \mathbf{k}_i, \quad \text{stepping formula} \quad (\text{B.6.1})$$

$$\hat{\mathbf{y}}^{n+1} = \mathbf{y}^n + h \sum_{i=1}^s \hat{b}_i \mathbf{k}_i, \quad \text{error estimator} \quad (\text{B.6.2})$$

where at each step

$$\mathbf{k}_i = \mathbf{f}(\mathbf{y}^n + h \sum_{j=1}^s a_{ij} \mathbf{k}_j, t^n + c_i h). \quad (\text{B.6.3})$$

This pair of methods may be described by an extended Butcher tableau of the form

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \\ \hline & \hat{b}_1 & \cdots & \hat{b}_s \end{array} \quad (\text{B.6.4})$$

As the annotation is intended to indicate, the relation (6.1) is to be used as a stepping formula to propagate the solution, and the relation (6.2) is to be used in conjunction with (6.1) to estimate and control the local error in making a given step.

B.6.2 Fehlberg 4(5) Pair

Fehlberg was the first to propose and develop embedded pairs. One such pair, called Fehlberg 4(5), is that described by the (extended) Butcher tableau below:

Butcher Tableau for Fehlberg 4(5)

0	0	0	0	0	0	0
$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0	0
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$	0	0	0	0
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$	0	0	0
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$	0	0
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	0
<hr/>						
	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
<hr/>						
	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

The procedure has $s = 6$ stages. The stepping formula is locally exact through terms of order h^4 . The error estimator is locally exact through terms of order h^5 . This procedure is therefore referred to as a 4(5) procedure. Because the stepping formula used to propagate the solution is of order 4, the whole procedure itself is locally exact through terms of order $m = 4$. Note that, according to Table 2.3.1, the highest order a 6-stage method can have is $m = 5$. Although the order 4 is relatively low in view of the number of stages involved, there is some freedom in selecting the entries in the matrix a and the vectors b and \hat{b} ; and Fehlberg selected them to minimize the size of the order h^5 error terms in the stepping formula.

Error Estimation

Since \mathbf{y}^{n+1} is locally exact through order 4 and $\hat{\mathbf{y}}^{n+1}$ is locally exact through order 5, we may define a local error vector Δ^n by the rule

$$\Delta^n = \mathbf{y}^{n+1} - \hat{\mathbf{y}}^{n+1} = h \sum_{i=1}^6 d_i \mathbf{k}_i \quad (\text{B.6.6})$$

where

$$d_i = b_i - \hat{b}_i. \quad (\text{B.6.7})$$

Note that in this approach only \mathbf{y}^{n+1} is computed using (6.1), $\hat{\mathbf{y}}^{n+1}$ is not computed, and Δ^n is computed using the far right side of (6.6). So doing minimizes work and round-off error.

Control of Step Size

We now wish to use $\|\Delta^n\|$ to control the subsequent step size h_{n+1} or to specify how to repeat, if necessary, the current step with a smaller step size.¹ Exactly how to do so is an art based on considerable experience with various possibilities, and is best left to professional numerical analysts. Typical programs require the user to specify some desired *tolerance* Tol and perhaps some initial step size h_0 , and the program automatically adjusts the step size as the integration proceeds based on this information.² As is the case with adaptive predictor-corrector and extrapolation methods, the prospective user is advised to first try and understand programs professionally written before attempting to write any of his/her own.

Interpolation-Dense Output

When employing a fixed step size method it is easy to integrate from some initial time t^0 to the final time $t^0 + T$ simply by requiring the relation

$$Nh = T \quad (\text{B.6.8})$$

between the step size h and the interval duration T . Recall Section 2.1. However, when the time step is variable, the time $t^0 + T$ is generally not among the times t^n at which the \mathbf{y}^n are computed, and so the quantity $\mathbf{y}(t^0 + T)$ is not among the \mathbf{y}^n .

In the case of a Runge-Kutta method, since it is a single-step method, this problem of finding an accurate $\mathbf{y}(t^0 + T)$ is fairly easy to solve. First, over the course of integration, monitor the times t^n and find the first such time, call it t^{n^*} , for which

$$t^{n^*} \geq t^0 + T. \quad (\text{B.6.9})$$

Next, define a step size h^* by the rule

$$h^* = t^0 + T - t^{n^*}. \quad (\text{B.6.10})$$

Note that $h^* \leq 0$. Finally, execute one step of the Runge-Kutta method in use with the time step h^* and the initial condition \mathbf{y}^{n^*} . So doing determines $\mathbf{y}(t^0 + T)$ to the accuracy of the integration method. In effect, this method integrates backward from the time t^{n^*} to the time $t^0 + T$.

There are some situations, for example when graphical output is needed, in which one desires an accurate and efficient method for finding $\mathbf{y}(t^n + \theta h_n)$ for any $\theta \in [0, 1]$. There are procedures that prepare, at each integration step, polynomials in θ for this purpose, and these procedures utilize the \mathbf{k} vectors computed in the course of a Runge-Kutta step. See, for example, the book of Hairer, Nørsett, and Wanner cited at the end of this appendix.

¹For computational efficiency, in this application it is convenient to define the vector norm $\| * \|$ to be the component moduli sum norm (3.7.20).

²There are also procedures for determining h_0 automatically so that only Tol need be specified.

B.6.3 Dormand-Prince 5(4) Pair

Inspired by the work of Fehlberg, Dormand and Prince and others developed procedures for which the stepping formula is higher order than the error estimator, and the stepping formula is optimized to minimize its still higher-order error terms.³ One procedure of Dormand and Prince is specified by the following Butcher tableau:

Butcher Tableau for Dormand-Prince 5(4)

0	0	0	0	0	0	0	0
$\frac{1}{5}$	$\frac{1}{5}$	0	0	0	0	0	0
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$	0	0	0	0	0
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$	0	0	0	0
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$	0	0	0
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$	0	0
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

The procedure has 7 stages. But the method for the stepping formula is FSAL, and therefore only requires the work of a 6-stage method after the first integration step. See the material at the end of Subsection 2.3.4. The stepping formula is locally exact through terms of order h^5 . The error estimator is locally exact through terms of order h^4 . This method is therefore referred to as a 5(4) method.⁴ Because the stepping formula is of order 5, and is used to propagate the solution, the whole procedure itself is locally exact through terms of order $m = 5$. Note that, according to Table 2.3.1, the highest order a 6-stage method can have is $m = 5$. Thus the order is optimum in view of the effective number of stages involved. Moreover, there is still some freedom in selecting the entries in the matrix a and the vectors b and \hat{b} . As mentioned at the beginning of this subsection, Dormand and Prince selected them to minimize the size of the order h^6 error terms in the stepping formula.⁵

³Procedures that use the higher-order formula for stepping are called *local extrapolation* procedures. Note that here the word *extrapolation* has a different meaning than in Section 2.6.

⁴Some authors use the notation (4)5. However, whatever the notation, it is always the order of the stepping formula that is not in parentheses, and the order of the error estimator that is in parentheses.

⁵Why choose, for the stepping formula, a seven-stage FSAL method rather than a six-stage non-FSAL method? Both choices could yield a fifth-order stepping formula and a fourth-order error estimator. Dor-

Error Estimation

Since \mathbf{y}^{n+1} is locally exact through order 5 and $\hat{\mathbf{y}}^{n+1}$ is locally exact through order 4, we may now define a local error vector Δ^n by the rule

$$\Delta^n = \hat{\mathbf{y}}^{n+1} - \mathbf{y}^{n+1} = -h \sum_{i=1}^7 d_i \mathbf{k}_i. \quad (\text{B.6.12})$$

Note that in carrying out the sum (6.12) the $i = 2$ term may be omitted since, according to (6.11), both b_2 and \hat{b}_2 vanish, and therefore $d_2 = 0$.

Control of Step Size

We again wish to use $\|\Delta^n\|$ to control the step size. In this case it should be understood that what is now being controlled is the error in the lower order error estimator with the hope that, since it is of one order higher, the error in the stepping formula will be even better controlled. Again, there are several possible procedures, and again the prospective user is advised to first try and understand programs professionally written before attempting to write any of his/her own.

Interpolation-Dense Output

The same methods and considerations apply here as in Subsection 6.2.

mand and Prince found that by so doing they were better able to choose the entries in the matrix a and the vectors b and \hat{b} to minimize the order h^6 error terms in the stepping formula.

B.7 A 5th Order PECEC Adams Routine

```

subroutine adams5 (h,ns,nf,t,y)
c This is an N=4 PECEC Adams routine that is locally correct
c through terms of order h**5 and makes local errors of
c order h**6.
c h is the step size, and ns is the number of steps. t is the time, and
c y is the dependent variable array. nf is a flag that controls the mode
c of entry. If nf = 'start', the trajectory is started with Runge Kutta.
c If nf = 'cont', the solution is continued using previous "f" values.
c ns must exceed 4 when Adams is called with nf = 'start'.
c In the next line, put in after ne the number of equations to be
c integrated.
parameter (ne=2)
dimension y(ne),yp(ne),yc(ne),f1(ne),f2(ne),f3(ne),
& f4(ne),f5(ne),f6(ne)
c yp and yc are corrector arrays. f1 through f6 form the array of
c "f" values.
dimension a(5),am(5),b(5),bm(5)
c a,am and b,bm are coefficients used in the corrector and predictor,
c respectively.
data (a(i), i=1,5) /-19.,106.,-264.,646.,251./
data (b(i), i=1,5) /251.,-1274.,2616.,-2774.,1901./
save am,bm,f1,f2,f3,f4,f5
nsa=ns
c nsa is the number of steps to be made by Adams.
if (nf .eq. 'cont') go to 20
c When nf = 'cont', the integration has already been started earlier,
c and "f" values are assumed to exist. Otherwise, start with Runge Kutta.
c Set up the initial f array using Runge Kutta.
call eval (t,y,f1)
c eval is a subroutine that evaluates ydot.
call rk3 (h/5.,5,0,t,y)
call eval (t,y,f2)
call rk3 (h/5.,5,0,t,y)
call eval (t,y,f3)
call rk3 (h/5.,5,0,t,y)
call eval (t,y,f4)
call rk3 (h/5.,5,0,t,y)
call eval (t,y,f5)
c Now go into the finite difference procedure.
nsa=ns-4
c nsa is the number of steps to be made by Adams. If the integration
c began with Runge Kutta, Adams has 4 fewer steps to make.
hdiv=h/720.
do 10 i=1,5
  am(i)=hdiv*a(i)
10 bm(i)=hdiv*b(i)
c am and bm are used in the corrector and predictor.
20 tint=t
c tint is the initial time for Adams.
  do 100 i=1,nsa
c Begin with predictor.

```

```

      do 30 j=1,ne
30  yp(j)=y(j)+bm(1)*f1(j)+bm(2)*f2(j)+bm(3)*f3(j)
     & +bm(4)*f4(j)+bm(5)*f5(j)
c First evaluation.
      call eval (t+h,yp,f6)
c First use of corrector. Here we use yp as a storage array.
      do 40 j=1,ne
         yp(j)=y(j)+am(1)*f2(j)+am(2)*f3(j)+am(3)*f4(j)
     & +am(4)*f5(j)
40  yc(j)=yp(j)+am(5)*f6(j)
c Second evaluation.
      call eval (t+h,yc,f6)
c Second use of corrector.
      do 50 j=1,ne
50  y(j)=yp(j)+am(5)*f6(j)
c Update table of f values.
      do 60 j=1,ne
         f1(j)=f2(j)
         f2(j)=f3(j)
         f3(j)=f4(j)
         f4(j)=f5(j)
60  f5(j)=f6(j)
      t=tint+float(i)*h
100 continue
      return
      end

```

Note: Here the *flag nf* controls the mode of entry.

B.8 A 10^{th} Order PECEC Adams Routine

```

subroutine adams10(h,ns,nf,t,y)
c Written by Rob Ryne, Spring 1986, based on a routine of Alex Dragt.
c This N=9 Adams integration routine makes local truncation errors
c at each step of order  $h^{**11}$ . That is, it is locally correct through
c order  $h^{**10}$ . Due to round off errors, its true precision is
c realized only when more than 64 bits are used.
c Warning: because this is a high-order method, the step size must be
c correspondingly small to achieve stability. For example, for the simple
c harmonic oscillator with unit frequency ( $xdoubleprime+x=0$ ), at least
c 50 steps per oscillation are required to safely achieve stability and
c for the error analysis based on finite-difference considerations
c to be relevant.
      implicit double precision (a-h,o-z)
c
      character*6 nf
      parameter (ne=2)
c
      dimension y(ne),yp(ne),yc(ne),f1(ne),f2(ne),f3(ne),f4(ne),
# f5(ne),f6(ne),f7(ne),f8(ne),f9(ne),f10(ne),f11(ne)
c
      dimension a(10),am(10),b(10),bm(10)
c
      data (a(i),i=1,10)/57281.d0,-583435.d0,2687864.d0,
# -7394032.d0,13510082.d0,-17283646.d0,16002320.d0,
# -11271304.d0,9449717.d0,2082753.d0/
      data (b(i),i=1,10)/-2082753.d0,20884811.d0,-94307320.d0,
# 252618224.d0,-444772162.d0,538363838.d0,-454661776.d0,
# 265932680.d0,-104995189.d0,30277247.d0/
c
      nsa=ns
      if (nf.eq.'cont') go to 20
c
c rk start
      iqt=5
      qt=float(iqt)
      hqt=h/qt
      call feval(t,y,f1)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f2)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f3)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f4)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f5)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f6)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f7)
      call rk78ii(hqt,iqt,t,y)
      call feval(t,y,f8)

```

```

call rk78ii(hqt,iqt,t,y)
call feval(t,y,f9)
call rk78ii(hqt,iqt,t,y)
call feval(t,y,f10)
nsa=ns-9
hdiv=h/7257600.0d+00
do 10 i=1,10
am(i)=hdiv*a(i)
10 bm(i)=hdiv*b(i)
c
c Adams routine
c
20 tint=t
do 100 i=1,nsa
do 30 j=1,ne
yp(j)=y(j)+bm(1)*f1(j)+bm(2)*f2(j)+bm(3)*f3(j)
# +bm(4)*f4(j)+bm(5)*f5(j)+bm(6)*f6(j)+bm(7)*f7(j)
# +bm(8)*f8(j)+bm(9)*f9(j)+bm(10)*f10(j)
30 continue
call feval(t+h,yp,f11)
do 40 j=1,ne
yp(j)=y(j)+am(1)*f2(j)+am(2)*f3(j)+am(3)*f4(j)+am(4)*f5(j)
# +am(5)*f6(j)+am(6)*f7(j)+am(7)*f8(j)+am(8)*f9(j)+am(9)*f10(j)
40 yc(j)=yp(j)+am(10)*f11(j)
41 call feval(t+h,yc,f11)
do 50 j=1,ne
50 y(j)=yp(j)+am(10)*f11(j)
do 60 j=1,ne
f1(j)=f2(j)
f2(j)=f3(j)
f3(j)=f4(j)
f4(j)=f5(j)
f5(j)=f6(j)
f6(j)=f7(j)
f7(j)=f8(j)
f8(j)=f9(j)
f9(j)=f10(j)
60 f10(j)=f11(j)
t=tint+i*h
100 continue
return
end

```

Notes: This program calls the subroutine **feval**, which is another name for **eval**. Here the flag **nf** controls the mode of entry.

Bibliography

- [1] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I. Non-stiff Problems*, Springer (1993).
- [2] J.C. Butcher, *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*, John Wiley, (1987).
- [3] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, First Edition, John Wiley (2003).
- [4] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, Second Edition, John Wiley (2008). <http://www.math.auckland.ac.nz/~butcher/ODE-book-2008/>.
- [5] J. Dormand and P. Prince, “A family of embedded Runge-Kutta formulae”, *J. Comp. Appl. Math.* **6**, 19-26 (1980).

Appendix C

Baker-Campbell-Hausdorff and Zassenhaus Formulas, Bases, and Paths

The purpose of this appendix is to describe the Lie-algebraic results of Henry Frederick Baker (1866-1956), John Edward Campbell (1862-1924), and Felix Hausdorff (1868-1942), and the related results of Hans Zassenhaus (1912-1991). We also discuss differentiating the exponential function, bases for Lie algebras, and paths in Lie groups and Lie algebras.

C.1 Differentiating the Exponential Function

C.2 The Baker-Campbell-Hausdorff Formula

C.3 The Baker-Campbell-Hausdorff Series

$$\begin{aligned} \log(e^y e^x) = & x + y - \frac{1}{2}[x, y] + \frac{1}{12}[x, [x, y]] + \frac{1}{12}[[x, y], y] - \frac{1}{24}[x, [[x, y], y]] \\ & - \frac{1}{720}[x, [x, [x, [x, y]]]] + \frac{1}{180}[x, [x, [[x, y], y]]] + \frac{1}{180}[x, [[[x, y], y], y]] \\ & + \frac{1}{120}[[x, y], [[x, y], y]] + \frac{1}{360}[[x, [x, y]], [x, y]] - \frac{1}{720}[[[[x, y], y], y]y] \\ & + \frac{1}{1440}[x, [x, [x, [[x, y], y]]]] - \frac{1}{360}[x, [x, [[x, y], y], y]] - \frac{1}{240}[x, [[x, y], [[x, y], y]]] \\ & - \frac{1}{720}[x, [[x, [x, y]], [x, y]]] + \frac{1}{1440}[x, [[[x, y], y], y], y]] + \frac{1}{30240}[x, [x, [x, [x, [x, [x, y]]]]]] \\ & - \frac{1}{5040}[x, [x, [x, [x, [[x, y], y]]]]] + \frac{1}{3780}[x, [x, [x, [[[x, y], y], y]]]] \\ & + \frac{1}{1680}[x, [x, [[x, y], [[x, y], y]]]] + \frac{1}{10080}[x, [x, [[x, [x, y]], [x, y]]]] \\ & + \frac{1}{3780}[x, [x, [[[x, y], y], y], y]] + \frac{13}{15120}[x, [[x, y], [[[x, y], y], y]]] \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{1260} [x, [[x, [[x, y], y]], [x, y]]] - \frac{1}{5040} [x, [[[x, y], y], y], y] \\
& + \frac{1}{1260} [[x, y], [[x, y], [[x, y], y]]] - \frac{1}{2016} [[x, y], [[[x, y], y], y], y] \\
& + \frac{1}{2016} [[x, [x, y]], [x, [[x, y], y]]] - \frac{1}{5040} [[[x, y], y], [[x, y], y], y] \\
& + \frac{1}{10080} [[x, [x, [x, y]]], [x, [x, y]]] + \frac{1}{10080} [[x, [[x, y], y]], [[x, y], y]] \\
& - \frac{1}{1512} [[x, [[[x, y], y], y]], [x, y]] - \frac{1}{5040} [[[x, [x, y]], [x, y]], [x, y]] \\
& + \frac{1}{30240} [[[[[x, y], y], y], y], y] - \frac{1}{60480} [x, [x, [x, [x, [[x, y], y]]]]] \\
& + \frac{1}{10080} [x, [x, [x, [x, [[[x, y], y], y]]]]] + \frac{1}{20160} [x, [x, [x, [[x, y], [[x, y], y]]]]] \\
& + \frac{1}{15120} [x, [x, [x, [x, [[x, [x, y]], [x, y]]]]] - \frac{23}{120960} [x, [x, [x, [[[x, y], y], y], y]]] \\
& - \frac{13}{30240} [x, [x, [[x, y], [[[x, y], y], y]]]] - \frac{1}{2520} [x, [x, [[x, [x, y]], y], [x, y]]] \\
& + \frac{1}{10080} [x, [x, [[[x, y], y], [[[x, y], y], y]]]] - \frac{1}{2520} [x, [[x, y], [[x, y], [[x, y], y]]]] \\
& + \frac{1}{4032} [x, [[x, y], [[[x, y], y], y], y]] - \frac{1}{4032} [x, [[x, [x, y]], [x, [[x, y], y]]]] \\
& + \frac{1}{10080} [x, [[[x, y], y], [[[x, y], y], y]]] - \frac{1}{20160} [x, [[x, [x, y]], [x, [x, y]]]] \\
& - \frac{1}{20160} [x, [[x, [[x, y], y]], [[x, y], y]]] + \frac{1}{3024} [x, [[x, [[[x, y], y], y]], [x, y]]] \\
& + \frac{1}{10080} [x, [[[x, [x, y]], [x, y]], [x, y]]] - \frac{1}{60480} [x, [[[[[x, y], y], y], y], y]] \\
& - \frac{1}{1209600} [x, [x, [x, [x, [x, [x, [x, y]]]]]]] + \frac{1}{151200} [x, [x, [x, [x, [x, [x, [[x, y], y]]]]]]] \\
& - \frac{1}{56700} [x, [x, [x, [x, [x, [[[x, y], y], y]]]]] - \frac{1}{43200} [x, [x, [x, [x, [[x, y], [[x, y], y]]]]]] \\
& - \frac{1}{100800} [x, [x, [x, [x, [x, [[x, [x, y]], [x, y]]]]]]] + \frac{1}{75600} [x, [x, [x, [x, [[[x, y], y], y], y]]]] \\
& + \frac{11}{302400} [x, [x, [x, [[x, y], [[[x, y], y], y]]]]] + \frac{1}{75600} [x, [x, [x, [[x, [[x, y], y]], [x, y]]]]] \\
& + \frac{1}{75600} [x, [x, [x, [[[[x, y], y], y], y], y]]] + \frac{1}{100800} [x, [x, [[x, y], [[x, y], [[x, y], y]]]]] \\
& + \frac{1}{20160} [x, [x, [[x, y], [[[x, y], y], y], y]]] + \frac{1}{67200} [x, [x, [[x, [x, y]], [x, [[x, y], y]]]]] \\
& + \frac{1}{30240} [x, [x, [[[[x, y], y], [[[x, y], y], y]]]]] + \frac{11}{201600} [x, [x, [[x, [[x, y], y]], [[x, y], y]]]] \\
& + \frac{11}{151200} [x, [x, [[x, [[[x, y], y], y]], [x, y]]]] + \frac{1}{43200} [x, [x, [[[x, [x, y]], [x, y]], [x, y]]]] \\
& - \frac{1}{56700} [x, [x, [[[[[x, y], y], y], y], y], y]] + \frac{1}{6048} [x, [[x, y], [[x, y], [[[x, y], y], y]]]] \\
& - \frac{23}{302400} [x, [[x, y], [[[[[x, y], y], y], y], y], y]] + \frac{23}{302400} [x, [[x, [x, y]], [x, [[[x, y], y], y]]]] \\
& + \frac{1}{37800} [x, [[x, [x, y]], [[x, [x, y]], [x, y]]]] - \frac{11}{120960} [x, [[[x, y], y], [[[x, y], y], y], y]]
\end{aligned}$$

$$\begin{aligned}
& + \frac{1}{25200} [x, [[x, [x, [[x, y], y]]], [x, [x, y]]]] - \frac{1}{15120} [x, [[x, [[x, y], y], y]], [[x, y], y]]] \\
& - \frac{1}{20160} [x, [[[x, y], [[x, y], y]], [[x, y], y]]] + \frac{1}{33600} [x, [[x, [[x, y], [[x, y], y]]], [x, y]]]] \\
& - \frac{1}{7560} [x, [[x, [[[x, y], y], y], y]], [x, y]]] - \frac{17}{100800} [x, [[[x, [[x, y], y]], [x, y]], [x, y]]]] \\
& + \frac{1}{151200} [x, [[[[[[x, y], y], y], y], y], y]]] + \frac{1}{10080} [[x, y], [[x, y], [[x, y], [[x, y], y]]]]] \\
& - \frac{1}{7560} [[x, y], [[x, y], [[[x, y], y], y], y]]] - \frac{1}{15120} [[x, y], [[[x, y], y], [[[x, y], y], y]]]] \\
& + \frac{1}{43200} [[x, y], [[[[[x, y], y], y], y], y]]] + \frac{1}{25200} [[x, [x, y]], [x, [[x, y], [[x, y], y]]]]] \\
& - \frac{1}{17280} [[x, [x, y]], [x, [[[x, y], y], y], y]]] - \frac{1}{8400} [[x, [x, y]], [[x, [[x, y], y]], [x, y]]]] \\
& + \frac{1}{33600} [[[x, y], y], [[[x, y], y], y], y]]] + \frac{1}{43200} [[x, [x, [x, y]], [x, [x, [[x, y], y]]]]] \\
& + \frac{1}{60480} [[x, [[x, y], y]], [x, [[[x, y], y], y], y]]] + \frac{1}{20160} [[x, [[x, y], y]], [[x, y], [[x, y], y]]]] \\
& + \frac{1}{60480} [[[x, y], y], [[[x, y], y], y], y]]] + \frac{1}{302400} [[x, [x, [x, [x, y]]], [x, [x, [x, y]]]]]] \\
& + \frac{1}{67200} [[x, [x, [[x, y], y]]], [x, [[x, y], y]]]] + \frac{1}{90720} [[x, [[x, y], y], y], [[[x, y], y], y]]] \\
& - \frac{1}{25200} [[[x, [x, y]], [x, y]], [x, [[x, y], y], y]]] - \frac{1}{21600} [[x, [x, [[x, y], y], y]], [x, [x, y]]]] \\
& - \frac{1}{30240} [[x, [[x, [x, y]], [x, y]]], [x, [x, y]]]] + \frac{1}{60480} [[x, [[[x, y], y], y], y], [[x, y], y]]] \\
& + \frac{1}{15120} [[[x, y], [[[x, y], y], y], y]], [[x, y], y]]] - \frac{1}{10080} [[x, [[x, y], [[[x, y], y], y]], [x, y]]]] \\
& + \frac{1}{21600} [[x, [[[x, y], y], y], y], y], [x, y]]] + \frac{1}{20160} [[[x, [x, y], y]], [[x, y], y], [x, y]]] \\
& + \frac{1}{10080} [[[x, [[x, y], y], y]], [x, y], [x, y]]] + \frac{1}{50400} [[[x, [x, y]], [x, y]], [x, y], [x, y]]] \\
& - \frac{1}{1209600} [[[[[[[x, y], y], y], y], y], y], y]]] + \frac{1}{2419200} [x, [x, [x, [x, [x, [x, [[x, y], y]]]]]]]] \\
& - \frac{1}{302400} [x, [x, [x, [x, [x, [x, [[[x, y], y], y]]]]]]]] + \frac{1}{604800} [x, [x, [x, [x, [x, [[x, y], [[x, y], y]]]]]]]] \\
& - \frac{1}{403200} [x, [x, [x, [x, [x, [[x, [x, y]], [x, y]]]]]]]] + \frac{37}{3628800} [x, [x, [x, [x, [x, [[[x, y], y], y], y]]]]]] \\
& + \frac{1}{56700} [x, [x, [x, [x, [[x, y], [[[x, y], y], y]]]]]]] + \frac{1}{37800} [x, [x, [x, [x, [[x, [x, y], y]], [x, y]]]]]] \\
& - \frac{1}{67200} [x, [x, [x, [x, [[[x, y], y], y], y], y]]]] + \frac{17}{604800} [x, [x, [x, [[x, y], [[x, y], [[x, y], y]]]]]]] \\
& - \frac{11}{241920} [x, [x, [x, [[x, y], [[[x, y], y], y], y]]]]] + \frac{1}{75600} [x, [x, [x, [[x, [x, y]], [x, [[x, y], y]]]]]]] \\
& - \frac{1}{40320} [x, [x, [x, [[[x, y], y], [[[x, y], y], y]]]]]] + \frac{1}{241920} [x, [x, [x, [[x, [x, [x, y]]], [x, [x, y]]]]]]] \\
& - \frac{1}{43200} [x, [x, [x, [[x, [[x, y], y]], [[x, y], y]]]]]] - \frac{29}{453600} [x, [x, [x, [[x, [[[x, y], y], y]], [x, y]]]]]] \\
& - \frac{1}{50400} [x, [x, [x, [[[x, [x, y]], [x, y]], [x, y]], [x, y]]]]] + \frac{37}{3628800} [x, [x, [x, [[[[[x, y], y], y], y], y], y]]]]
\end{aligned}$$

$$\begin{aligned}
& - \frac{1}{12096} [x, [x, [[x, y], [[x, y], [[[x, y], y], y]]]]] + \frac{23}{604800} [x, [x, [[x, y], [[[x, y], y], y], y], y]]] \\
& - \frac{23}{604800} [x, [x, [[x, [x, y]], [x, [[[x, y], y], y]]]]] - \frac{1}{75600} [x, [x, [[x, [x, y]], [[x, [x, y]], [x, y]]]]] \\
& + \frac{11}{241920} [x, [x, [[[[x, y], y], [[[x, y], y], y], y]]]] - \frac{1}{50400} [x, [x, [[x, [x, [[x, y], y]]], [x, [x, y]]]]] \\
& + \frac{1}{30240} [x, [x, [[x, [[[x, y], y], y]], [[x, y]y]]]] + \frac{1}{40320} [x, [x, [[[[x, y], [[x, y], y]], [[x, y]y]]]]] \\
& - \frac{1}{67200} [x, [x, [[x, [[x, y], [[x, y], y]]], [x, y]]]] + \frac{1}{15120} [x, [x, [[x, [[[x, y], y], y], y], [x, y]]]] \\
& + \frac{17}{201600} [x, [x, [[[[x, [[x, y], y]], [x, y]], [x, y]]]]] - \frac{1}{302400} [x, [x, [[[[[[x, y], y], y], y], y], y], y]]] \\
& - \frac{1}{20160} [x, [[x, y], [[x, y], [[x, y], [[x, y], y]]]]] + \frac{1}{15120} [x, [[x, y], [[x, y], [[[x, y], y], y], y]]] \\
& + \frac{1}{30240} [x, [[x, y], [[[x, y], y], [[x, y], y], y]]] - \frac{1}{86400} [x, [[x, y], [[[x, y], y], y], y], y]]] \\
& - \frac{1}{50400} [x, [[x, [x, y]], [x, [[x, y], [[x, y], y]]]]] + \frac{1}{34560} [x, [[x, [x, y]], [x, [[[x, y], y], y], y]]] \\
& + \frac{1}{16800} [x, [[x, [x, y]], [[x, [x, y]], [x, y]]]] - \frac{1}{67200} [x, [[[[x, y], y], [[[x, y], y], y], y], y]]] \\
& - \frac{1}{86400} [x, [[x, [x, [x, y]]], [x, [x, [[x, y], y]]]]] - \frac{1}{120960} [x, [[x, [[x, y], y]], [x, [[[x, y], y], y]]]] \\
& - \frac{1}{40320} [x, [[x, [[x, y], y]], [[x, y], [[x, y], y]]]] - \frac{1}{120960} [x, [[[[x, y], y], y], [[[x, y], y], y], y]]] \\
& - \frac{1}{604800} [x, [[x, [x, [x, [x, y]]]], [x, [x, [x, y]]]]] - \frac{1}{134400} [x, [[x, [x, [[x, y], y]]], [x, [[x, y], y]]]] \\
& - \frac{1}{181440} [x, [[x, [[x, [[[x, y], y], y], y]], [[x, y], [[x, y], y], y]]]] + \frac{1}{50400} [x, [[[[x, x, y]], [x, y]], [x, [[x, y], y]]]] \\
& + \frac{1}{43200} [x, [[x, [x, [[[x, y], y], y], y]], [x, [x, y]]]] + \frac{1}{60480} [x, [[x, [[x, [x, y]], [x, y]], [x, [x, y]]]], [x, [x, y]]]] \\
& - \frac{1}{120960} [x, [[x, [[[x, y], y], y], y], [[x, y], y]]] - \frac{1}{30240} [x, [[[[x, y], y], [[x, y], y], y], [x, y]]]] \\
& + \frac{1}{20160} [x, [[x, [[x, y], [[[x, y], y], y], y]], [x, y]]] - \frac{1}{43200} [x, [[x, [[[[x, y], y], y], y], y], [x, y]]]] \\
& - \frac{1}{40320} [x, [[[x, [[x, y], y]], [[x, y], y], [x, y]]], [x, y]]] - \frac{1}{20160} [x, [[[x, [[[[x, y], y], y], y], [x, y]], [x, y]]], [x, y]]] \\
& - \frac{1}{100800} [x, [[[x, [x, y]], [x, y]], [x, y], [x, y]]] + \frac{1}{2419200} [x, [[[[[[x, y], y], y], y], y], y], y]]] \\
& + \dots
\end{aligned}$$

This result is taken from the Thesis of P. V. Koseleff. See the references at the end of this appendix.

Exercises

C.3.1. According to the BCH theorem the *exponential* function has the remarkable property that the quantity C in the relation

$$\exp(A) \exp(B) = \exp(C) \quad (\text{C.3.1})$$

depends only on elements in the Lie algebra generated by A and B . See Section 3.7.3 and the BCH series in Section C.3 above. The purpose of this exercise is to study the properties of two other functions.

Begin with the truncated exponential function t1exp defined by

$$\text{t1exp}(A) = I + A. \quad (\text{C.3.2})$$

See (4.1.22). Show that

$$\text{t1exp}(A)\text{t1exp}(B) = \text{t1exp}(C) \quad (\text{C.3.3})$$

with

$$C = A + B + AB. \quad (\text{C.3.4})$$

Evidently the degree 1 term $A + B$ is in the Lie algebra generated by A and B , but the degree 2 term AB is not.

Next consider the truncated exponential function t2exp defined by

$$\text{t2exp}(A) = I + A + A^2/2!. \quad (\text{C.3.5})$$

Show that in this case

$$\text{t2exp}(A)\text{t2exp}(B) = \text{t2exp}(C) \quad (\text{C.3.6})$$

with

$$C = A + B + (1/2)(AB - BA) + \dots. \quad (\text{C.3.7})$$

Evidently the terms explicitly displayed on the right side of (C.3.7) are in the Lie algebra generated by A and B . Show that the remaining terms, which are of degree 3 and higher, are not.

C.3.2. Exercise on BCH like relation for the Cayley function.

C.4 Zassenhaus Formulas

C.5 Bases

C.6 Paths

C.6.1 Paths in the Group Yield Paths in the Lie Algebra

C.6.2 Paths in the Lie Algebra Yield Paths in the Group

C.6.3 Differential Equations

Bibliography

Baker-Campbell-Hausdorff Formula/Series

- [1] P.V. Koseleff, Formal calculus for Lie methods in Hamiltonian mechanics, Ph.D. Thesis, École Polytechnique (1993). The BCH series presented in Section 3 is taken from this Thesis.
- [2] A. Arnal, F. Casas, and C. Chiralt, “A note on the Baker-Campbell-Hausdorff series in terms of right-nested commutators”, <https://arxiv.org/abs/2006.15869v1> (2020).
- [3] E.B. Dynkin, “On the representation by means of commutators of the series $\log(e^x e^y)$ for noncommutative x and y ”, *Mat. Sbornik (N.S.)* **25**(67), 155-162 (1949).
- [4] Karl Goldberg, “The Formal Power Series for Log $e^x e^y$ ”, *Duke Journal of Mathematics* **23**, 13 (1956).
- [5] Erik Eriksen, “Properties of Higher-Order Commutator Products and the Baker-Hausdorff Formula”, *Journal of Mathematical Physics* **9**, 790 (1968).
- [6] J.B. Kogut, *Rev. Mod. Phys.* **51**, 4 (1979).
- [7] J.A. Oteo, “The Baker-Campbell-Hausdorff formula and nested commutator identities”, *J. Math. Phys.* **32**, 419-424 (1991).
- [8] A. Bonfiglioli and R. Fulci, *Topics in Noncommutative Algebra: The Theorem of Campbell, Baker, Hausdorff, and Dynkin*, Lecture Notes in Mathematics 2034, Springer (2012).
- [9] F. Casas and A. Murua, “An efficient algorithm for computing the Baker-Campbell-Hausdorff series and some of its applications”, *J. Math. Phys.* **50**, 033513 (2009).

Zassenhaus Formula

- [10] F. Casas, A. Murua, and M. Nadinic, “Efficient computation of the Zassenhaus formula”, <https://arxiv.org/abs/1204.0389v2> (2012).
- [11] R.M. Wilcox, “Exponential Operators and Parameter Differentiation in Quantum Physics”, *J. Math. Phys.* **8**, p. 962 (1967).
- [12] F. Bayen, “On the convergence of the Zassenhaus formula”, *Lett. Math. Phys.* **3**, 161-167 (1979).

Appendix D

Canonical Transformations

Appendix E

Mathematica Notebooks

Appendix F

Properties of Harmonic Functions, Analyticity, Aberration Expansions, and Smoothing

F.1 The Static Case

According to Poincaré's Theorem 1.3.3, trajectories will be analytic functions of the initial conditions in some domain if the right sides of the equations of motion (1.3.4) are analytic. Correspondingly, according to the results of Section 26.2, the Taylor map (7.5.5) will converge in some domain about the origin. For problems of particular interest to us the Hamiltonians, and hence the equations of motion, will involve the scalar and vector potentials ψ and \mathbf{A} as in, for example, (1.5.29), (1.6.16), and (1.6.17). Consequently, we are interested in knowing the analytic properties of ψ and \mathbf{A} .

In the static case these potentials are determined in terms of the charge density $\rho(\mathbf{r})$ and the current density $\mathbf{j}(\mathbf{r})$ by the Poisson equations

$$\nabla^2\psi = -4\pi\rho, \quad (\text{F.1.1})$$

$$\nabla^2\mathbf{A} = -4\pi\mathbf{j}/c, \quad (\text{F.1.2})$$

which have the solutions

$$\psi(\mathbf{r}) = \int d^3\mathbf{r}' \rho(\mathbf{r}') / \| \mathbf{r}' - \mathbf{r} \|, \quad (\text{F.1.3})$$

$$\mathbf{A}(\mathbf{r}) = (1/c) \int d^3\mathbf{r}' \mathbf{j}(\mathbf{r}') / \| \mathbf{r}' - \mathbf{r} \| . \quad (\text{F.1.4})$$

Here the notation $\| \cdot \|$ indicates the Euclidean norm,

$$\| \mathbf{r}' - \mathbf{r} \| = [(x' - x)^2 + (y' - y)^2 + (z' - z)^2]^{1/2}. \quad (\text{F.1.5})$$

It follows immediately from (1.1) that if $\psi(\mathbf{r})$ is analytic in the components x, y, z of \mathbf{r} at some point \mathbf{r}^0 , then $\rho(\mathbf{r})$ must also be analytic at \mathbf{r}^0 . The purpose of this appendix is to show the converse: if $\rho(\mathbf{r})$ is analytic at some point \mathbf{r}^0 , then $\psi(\mathbf{r})$ will also be analytic at \mathbf{r}^0 . We note, with some surprise, that this is a *local* statement. Although, according to (1.3),

the value of $\psi(\mathbf{r})$ at the point \mathbf{r} is determined by the value of $\rho(\mathbf{r}')$ at *all* points \mathbf{r}' , the quantity $\rho(\mathbf{r}')$ need not be analytic everywhere, but only at \mathbf{r}^0 , for $\psi(\mathbf{r})$ to be analytic at \mathbf{r}^0 . Finally, since (1.4) is analogous to (1.3), our proof will also show that if all components of $\mathbf{j}(\mathbf{r})$ are analytic at \mathbf{r}^0 , then all components of $\mathbf{A}(\mathbf{r})$ will also be analytic at \mathbf{r}^0 .

Before proceeding further, and so as to not raise the reader's expectations too high, we confess that the analyticity we will prove is analyticity in the vicinity of *real* points. That is, at any real point (x^0, y^0, z^0) , we will be able to prove analyticity in the complex variables $(x^0 + i\tilde{x}, y^0 + i\tilde{y}, z^0 + i\tilde{z})$ for $\tilde{x}, \tilde{y}, \tilde{z}$ finite but possibly small.

Why should we be interested in this question? First, we note that if $\rho(\mathbf{r})$ is zero in some region, then it is automatically analytic in that region. Therefore, as a particular case, we will find that vacuum solutions to Poisson's equation (solutions to Laplace's equation) must be analytic. This particular case is in fact the most common case since we are usually interested in orbits that remain within evacuated beam pipes. However, in some cases we are interested in the behavior of orbits that pass through regions of nonzero charge and/or current densities. Examples that come to mind include plasma lenses, electron cloud lenses, lithium lenses, beam-beam effects, and space-charge effects. In these cases we may still expect to have convergent aberration expansions provided $\rho(\mathbf{r})$ and $\mathbf{j}(\mathbf{r})$ are analytic in the region *traversed* by all orbits of interest. We emphasize again that no analyticity assumptions need be made about the behavior of $\rho(\mathbf{r})$ and $\mathbf{j}(\mathbf{r})$ in regions not traversed by orbits. From a mathematical perspective, the discussion that follows will be somewhat discursive; but it will have the advantage of obtaining several interesting results along the way. For a more direct approach, see Exercises 17 and 18.

By a suitable translation, and without loss of generality, we may take \mathbf{r}^0 to be the origin. Then, by the theory of Section 26.2, the assumption that ρ is analytic implies that it has an expansion of the form

$$\rho(\mathbf{r}) = \sum_{ijk} c_{ijk} x^i y^j z^k \quad (\text{F.1.6})$$

that converges in some polydisc \mathcal{D} given by inequalities of the form

$$|x| < R_x, |y| < R_y, |z| < R_z. \quad (\text{F.1.7})$$

Here we are treating x, y, z as three *complex* variables. For further discussion, it is convenient to work within a smaller domain \mathcal{R} contained within \mathcal{D} . Let ϵ be some fixed small positive number. Define a quantity R by the rule

$$R = \text{minimum of } (R_x - \epsilon), (R_y - \epsilon), (R_z - \epsilon). \quad (\text{F.1.8})$$

Then we define \mathcal{R} to be the closed set

$$|x| \leq R, |y| \leq R, |z| \leq R. \quad (\text{F.1.9})$$

We are now in a position to obtain a bound on the Taylor coefficients c_{ijk} . From the Cauchy formula

$$c_{jkl} = \frac{1}{(2\pi i)^3} \oint_{|x|=R} \oint_{|y|=R} \oint_{|z|=R} dx dy dz \frac{\rho(x, y, z)}{x^{j+1} y^{k+1} z^{\ell+1}} \quad (\text{F.1.10})$$

we get the result

$$|c_{jk\ell}| \leq KR^{-(j+k+\ell)} \quad (\text{F.1.11})$$

where the constant K is defined by the equation

$$K = \max |\rho(x, y, z)| \text{ for } |x| = |y| = |z| = R. \quad (\text{F.1.12})$$

Suppose the terms in the series (1.6) are grouped together according to their total degree. Doing so gives the result

$$\rho(\mathbf{r}) = \sum_{D=0}^{\infty} \sum_{\alpha=1}^{N(D,3)} d_{D\alpha} h_D^\alpha(\mathbf{r}). \quad (\text{F.1.13})$$

Here $h_D^\alpha(\mathbf{r})$ denotes a monomial of the form $x^i y^j z^k$ and of degree D (that is, $i + j + k = D$) and labelled by an index α . From (6.3.36) we know that $N(D, 3)$, the number of monomials of degree D in 3 variables, is given by the relation

$$N(D, 3) = \frac{(D+2)!}{D!2!} = (D+2)(D+1)/2. \quad (\text{F.1.14})$$

Hence, we may label the monomials of degree D in such a way that the index α ranges over the values of 1 to $N(D, 3)$, as indicated in (1.13). We note that that (1.13) is an ordering and grouping of (1.6), and hence is a permissible operation that cannot change its value for $\mathbf{r} \in \mathcal{R}$. See Sections 26.1 through 26.3.

We next study the relation between the monomials h_D^α , harmonic polynomials, and spherical harmonics. For the moment let x , y , and z be real. Introduce, in the standard way, spherical coordinates by the relations

$$x = r \sin \theta \cos \phi, \quad (\text{F.1.15})$$

$$y = r \sin \theta \sin \phi, \quad (\text{F.1.16})$$

$$z = r \cos \theta, \quad (\text{F.1.17})$$

with r given by the relation

$$r = (x^2 + y^2 + z^2)^{1/2}. \quad (\text{F.1.18})$$

Now consider the functions $H_\ell^m(\mathbf{r})$ defined by the relation

$$H_\ell^m(\mathbf{r}) = r^\ell Y_\ell^m(\theta, \phi) \quad (\text{F.1.19})$$

where the Y_ℓ^m are the usual *spherical harmonics*. We observe that the H_ℓ^m are *homogeneous polynomials* of degree ℓ in the variables x , y , z . In fact, for H_ℓ^m we have the relation

$$\begin{aligned} H_\ell^m(\mathbf{r}) &= r^\ell Y_\ell^m(\theta, \phi) = (2\ell+1)^{1/2} \{4\pi[(2\ell)!]\}^{-1/2} r^\ell P_\ell^\ell(\cos \theta) \\ &= [(-1)^\ell / (2^\ell \ell!)] \{[(2\ell+1)/(4\pi)][(2\ell)!]\}^{1/2} r^\ell (\sin \theta)^\ell e^{i\ell\phi} \\ &= [(-1)^\ell / (2^\ell \ell!)] \{[(2\ell+1)/(4\pi)][(2\ell)!]\}^{1/2} (x + iy)^\ell. \end{aligned} \quad (\text{F.1.20})$$

To see that the remaining H_ℓ^m are also homogeneous polynomials of degree ℓ we define, in the usual way, the angular momentum operator \mathcal{L} by the rule

$$\mathcal{L} = -ir \times \partial. \quad (\text{F.1.21})$$

By convention and construction the Y_ℓ^m have the property

$$\mathcal{L}_- Y_\ell^m = [(\ell + m)(\ell - m + 1)]^{1/2} Y_\ell^{m-1} \quad (\text{F.1.22})$$

where \mathcal{L}_- is defined by the rule

$$\mathcal{L}_- = \mathcal{L}_x - i\mathcal{L}_y. \quad (\text{F.1.23})$$

It is easily verified that \mathcal{L} commutes with r , and hence we also have the relation

$$\mathcal{L}_- H_\ell^m = [(\ell + m)(\ell - m + 1)]^{1/2} H_\ell^{m-1}. \quad (\text{F.1.24})$$

Moreover, we see from (1.21) and (1.23) that \mathcal{L}_- maps homogeneous polynomials into homogeneous polynomials, and leaves their degrees unchanged. It follows that all the H_ℓ^m are homogeneous polynomials of degree ℓ . Finally, we see from (1.19) that the H_ℓ^m satisfy Laplace's equation,

$$\nabla^2 H_\ell^m = 0, \quad (\text{F.1.25})$$

and therefore are entitled to be called *harmonic* polynomials.

Consider now the functions $H_\ell^{ms}(\mathbf{r})$ defined by the relations

$$H_\ell^{ms}(\mathbf{r}) = r^{2s} r^\ell Y_\ell^m = r^{2s} H_\ell^m(\mathbf{r}), \quad s = 0, 1, 2, \dots. \quad (\text{F.1.26})$$

They are evidently homogeneous polynomials in x, y, z of degree D , with D given by the relation

$$D = \ell + 2s. \quad (\text{F.1.27})$$

See (1.18). They are also linearly independent. Let us count their number for fixed D . The cases of even and odd D need to be treated separately. For even D we have the polynomials

$$r^D Y_D^m, r^2 r^{D-2} Y_{D-2}^m, \dots r^{D-2} r^2 Y_2^m, r^D, \quad (\text{F.1.28})$$

and their total number is

$$\begin{aligned} \text{number} &= [2D + 1] + [2(D - 2) + 1] + \dots + [2(2) + 1] + 1 \\ &= (D + 2)(D + 1)/2 = N(D, 3). \end{aligned} \quad (\text{F.1.29})$$

For odd D we have the polynomials

$$r^D Y_D^m, r^2 r^{D-2} Y_{D-2}^m, \dots r^{D-1} r Y_1^m, \quad (\text{F.1.30})$$

and their total number is

$$\begin{aligned} \text{number} &= [2D + 1] + [2(D - 2) + 1] + \dots + [2(1) + 1] \\ &= (D + 2)(D + 1)/2 = N(D, 3). \end{aligned} \quad (\text{F.1.31})$$

Comparison of (1.14), (1.29), and (1.31) shows that the number of homogeneous polynomials H_ℓ^{ms} with fixed degree D equals the number of monomials h_D^α with fixed D . Consequently, there exist expansion coefficients a and relations of the form

$$\begin{aligned} x^i y^j z^k &= h_D^\alpha(\mathbf{r}) = \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} a_{m\ell s}^{D\alpha} r^{2s} r^\ell Y_\ell^m \\ &= \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} a_{m\ell s}^{D\alpha} H_\ell^{ms}(\mathbf{r}). \end{aligned} \quad (\text{F.1.32})$$

With the aid of (1.32), the series (1.13) can also be written in the form

$$\begin{aligned}\rho(\mathbf{r}) &= \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} r^{2s} r^{\ell} Y_{\ell}^m \\ &= \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} r^{2s} H_{\ell}^m(\mathbf{r}).\end{aligned}\quad (\text{F.1.33})$$

In this form we see, as a consequence of analyticity, that spherical harmonics Y_{ℓ}^m always occur in conjunction with powers of r of the form $r^{\ell+2s}$ with $s = 0, 1, 2, \dots$.

The transition from the Taylor series (1.13) to what we will call the *harmonic* series (1.33) is not simply a different ordering, and therefore questions of convergence have to be examined anew. We begin by finding bounds for the polynomials $H_{\ell}^m(\mathbf{r})$. For this purpose it is convenient to use the familiar expansion

$$\frac{1}{\|\mathbf{r}' - \mathbf{r}\|} = 4\pi \sum_{\ell, m} (2\ell + 1)^{-1} r^{\ell} Y_{\ell}^m(\Omega) (r')^{-\ell-1} \bar{Y}_{\ell}^m(\Omega') \text{ for } r < r'. \quad (\text{F.1.34})$$

For real angles we note the bound

$$|Y_{\ell}^m(\Omega)| \leq [(2\ell + 1)/4\pi]^{1/2}. \quad (\text{F.1.35})$$

It follows that the expansion (1.34) is absolutely convergent for \mathbf{r} and \mathbf{r}' real and $r < r'$. Now multiply both sides of (1.34) by $Y_{\ell'}^{m'}(\Omega')$, integrate over Ω' , and use the orthogonality of the Y_{ℓ}^m to get the integral representation

$$H_{\ell}^m(\mathbf{r}) = r^{\ell} Y_{\ell}^m(\Omega) = (4\pi)^{-1} (2\ell + 1) (r')^{\ell+1} \int d\Omega' Y_{\ell}^m(\Omega') / \|\mathbf{r}' - \mathbf{r}\|. \quad (\text{F.1.36})$$

As it stands, the representation (1.36) holds for \mathbf{r} real and satisfying $r < r'$. We will now analytically continue it to possibly complex \mathbf{r} while keeping \mathbf{r}' real. There is no difficulty in extending the left side of (1.36) to complex \mathbf{r} since it is a polynomial. The extension of the right side is also straight forward. Moreover, when the left and right sides of (1.36) are extended to complex \mathbf{r} , they will continue to agree. That is, the integral representation (1.36) is also valid for complex \mathbf{r} . See Exercise 9. Introduce the unit vector

$$\mathbf{e}(\Omega') = \mathbf{r}'/r' = \mathbf{e}_x \sin \theta' \cos \phi' + \mathbf{e}_y \sin \theta' \sin \phi' + \mathbf{e}_z \cos \theta'. \quad (\text{F.1.37})$$

Also introduce real vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ and a complex vector $\boldsymbol{\zeta}$ by the relation

$$\mathbf{r}/r' = \boldsymbol{\xi} + i\boldsymbol{\eta} = \boldsymbol{\zeta}. \quad (\text{F.1.38})$$

With the aid of these definitions (1.36) can be recast in the form

$$H_{\ell}^m(\mathbf{r}) = (4\pi)^{-1} (2\ell + 1) (r')^{\ell} \int d\Omega' Y_{\ell}^m(\Omega') / \|\mathbf{e}(\Omega') - \boldsymbol{\zeta}\|. \quad (\text{F.1.39})$$

Let us examine the denominator $\|\mathbf{e} - \boldsymbol{\zeta}\|$. It has the form

$$\|\mathbf{e} - \boldsymbol{\zeta}\| = [(\mathbf{e} - \boldsymbol{\zeta}) \cdot (\mathbf{e} - \boldsymbol{\zeta})]^{1/2}. \quad (\text{F.1.40})$$

For the dot product appearing in (1.40) we find the expression

$$\begin{aligned} (\mathbf{e} - \boldsymbol{\zeta}) \cdot (\mathbf{e} - \boldsymbol{\zeta}) &= \mathbf{e} \cdot \mathbf{e} - 2\mathbf{e} \cdot \boldsymbol{\zeta} + \boldsymbol{\zeta} \cdot \boldsymbol{\zeta} \\ &= 1 - 2\mathbf{e} \cdot \boldsymbol{\zeta} + \boldsymbol{\zeta} \cdot \boldsymbol{\zeta} = (1 - 2\mathbf{e} \cdot \boldsymbol{\xi} + \boldsymbol{\xi} \cdot \boldsymbol{\xi} - \boldsymbol{\eta} \cdot \boldsymbol{\eta}) \\ &\quad + 2i(\boldsymbol{\xi} \cdot \boldsymbol{\eta} - \mathbf{e} \cdot \boldsymbol{\eta}). \end{aligned} \quad (\text{F.1.41})$$

Suppose the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are restricted in length to satisfy the relations

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq 1/16, \quad \boldsymbol{\eta} \cdot \boldsymbol{\eta} \leq 1/16. \quad (\text{F.1.42})$$

Then the quantities appearing in the real part of (1.41) obey the inequalities

$$|\mathbf{e} \cdot \boldsymbol{\xi}| \leq \|\mathbf{e}\| \|\boldsymbol{\xi}\| \leq 1/4, \quad (\text{F.1.43})$$

$$-1/16 \leq \boldsymbol{\xi} \cdot \boldsymbol{\xi} - \boldsymbol{\eta} \cdot \boldsymbol{\eta} \leq 1/16, \quad (\text{F.1.44})$$

and the real part itself satisfies the inequality

$$|1 - 2\mathbf{e} \cdot \boldsymbol{\xi} + \boldsymbol{\xi} \cdot \boldsymbol{\xi} - \boldsymbol{\eta} \cdot \boldsymbol{\eta}| \geq 7/16. \quad (\text{F.1.45})$$

From (1.41) and (1.45) it follows that $\|\mathbf{e} - \boldsymbol{\zeta}\|$ satisfies the inequality

$$|\|\mathbf{e} - \boldsymbol{\zeta}\|| = |[(\mathbf{e} - \boldsymbol{\zeta}) \cdot (\mathbf{e} - \boldsymbol{\zeta})]^{1/2}| \geq \sqrt{7}/4. \quad (\text{F.1.46})$$

Now use (1.35) and (1.46) in (1.39) to get the bound

$$|H_\ell^m(\mathbf{r})| \leq (16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2}(r')^\ell. \quad (\text{F.1.47})$$

Suppose \mathbf{r} is in the closed polydisc \mathcal{R} given by (1.9). Then we have the relations

$$\xi_x = \operatorname{Re}(x)/r' \leq R/r', \text{ etc.}; \quad (\text{F.1.48})$$

$$\eta_x = \operatorname{Im}(x)/r' \leq R/r', \text{ etc.} \quad (\text{F.1.49})$$

From these relations it follows that

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq 3(R/r')^2, \quad (\text{F.1.50})$$

$$\boldsymbol{\eta} \cdot \boldsymbol{\eta} \leq 3(R/r')^2. \quad (\text{F.1.51})$$

Finally set r' to the value

$$r' = 4\sqrt{3}R. \quad (\text{F.1.52})$$

Then the inequalities (1.42) are satisfied and (1.47) takes the final form

$$|H_\ell^m(\mathbf{r})| \leq (16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2}(4\sqrt{3})^\ell R^\ell \text{ for } \mathbf{r} \in \mathcal{R}. \quad (\text{F.1.53})$$

Inspection of (1.33) shows that what we really require are bounds on the polynomials $r^{2s} H_\ell^m(\mathbf{r})$. From the definition (1.18) and (1.9) we find the result

$$|r^2| = |x^2 + y^2 + z^2| \leq |x^2| + |y^2| + |z^2| \leq 3R^2. \quad (\text{F.1.54})$$

Consequently we find for the quantity r^{2s} the bound

$$|r^{2s}| = |r^2|^s \leq 3^s R^{2s} \leq (\sqrt{3})^{2s} R^{2s} \leq (4\sqrt{3})^{2s} R^{2s}. \quad (\text{F.1.55})$$

It follows that the polynomials $r^{2s} H_\ell^m(\mathbf{r})$ have the bounds

$$|r^{2s} H_\ell^m(\mathbf{r})| \leq (16\pi/\sqrt{7})[(2\ell+1)/4\pi]^{3/2} (4\sqrt{3})^{\ell+2s} R^{\ell+2s}. \quad (\text{F.1.56})$$

We next find bounds on the coefficients $b_{m\ell s}$. For each degree D we have the relation

$$\sum_{i+j+k=D} c_{ijk} x^i y^j z^k = \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} r^{\ell+2s} Y_\ell^m(\theta, \phi). \quad (\text{F.1.57})$$

Multiply both sides of (1.57) by $\bar{Y}_{\ell'}^{m'}$ and integrate over solid angle to obtain the relation

$$b_{m'\ell's'} r^D = r^D \int d\Omega \sum_{i+j+k=D} c_{ijk} (\sin \theta \cos \phi)^i (\sin \theta \sin \phi)^j (\cos \theta)^k \bar{Y}_{\ell'}^{m'}(\theta, \phi) \quad (\text{F.1.58})$$

where s' satisfies the condition

$$\ell' + 2s' = D. \quad (\text{F.1.59})$$

Here we have also used (1.15) through (1.17). It follows from (1.35) and (1.58) that we have the inequality

$$|b_{m'\ell's'}| \leq 4\pi[(2\ell+1)/4\pi]^{1/2} \sum_{i+j+k=D} |c_{ijk}|. \quad (\text{F.1.60})$$

Now use (1.11), and the fact that the number of terms appearing in the sum (1.60) is $N(D, 3)$, to get the result

$$|b_{m'\ell's'}| \leq 2\pi K[(2\ell+1)/4\pi]^{1/2} (D+2)(D+1) R^{-D}. \quad (\text{F.1.61})$$

Let us use the results (1.56) and (1.61) to examine the convergence of the series (1.33). Suppose \mathbf{r} is in a polydisc of the form (1.9) with R replaced by some value R'' yet to be selected. Consider the series

$$\begin{aligned} & \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} 2\pi K[(2\ell+1)/4\pi]^{1/2} (D+2)(D+1) R^{-D} |r^{2s} H_\ell^m(\mathbf{r})| \\ & \leq K(2/\sqrt{7}) \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} (D+2)(D+1)(2\ell+1)^2 (4\sqrt{3})^D (R''/R)^D \\ & \leq K(2/\sqrt{7}) \sum_{D=0}^{\infty} (D+2)(D+1)(4\sqrt{3})^D (R''/R)^D \sum_{\ell+2s=D} (2\ell+1)^3. \end{aligned} \quad (\text{F.1.62})$$

Here we have used (1.56) with R replaced by R'' . Evidently the series (1.62) is convergent if R'' satisfies the inequality

$$R'' < (4\sqrt{3})^{-1} R. \quad (\text{F.1.63})$$

We conclude that the series (1.33) converges absolutely in the polydisc

$$|x| \leq R'', |y| \leq R'', |z| \leq R'' \quad (\text{F.1.64})$$

with R'' given by (1.63).

With these matters concerning the series (1.6) and (1.33) for $\rho(\mathbf{r})$ behind us, we turn to the behavior of $\psi(\mathbf{r})$. Break up the integration required by (1.3) into two regions by rewriting it in the form

$$\psi(\mathbf{r}) = \psi_<(\mathbf{r}) + \psi_>(\mathbf{r}) \quad (\text{F.1.65})$$

where

$$\psi_<(\mathbf{r}) = \int_{\|\mathbf{r}'\| \leq R'} d^3\mathbf{r}' \rho(\mathbf{r}') / \|\mathbf{r}' - \mathbf{r}\|, \quad (\text{F.1.66})$$

$$\psi_>(\mathbf{r}) = \int_{\|\mathbf{r}'\| \geq R'} d^3\mathbf{r}' \rho(\mathbf{r}') / \|\mathbf{r}' - \mathbf{r}\|. \quad (\text{F.1.67})$$

We will examine each of the functions $\psi_<(\mathbf{r})$ and $\psi_>(\mathbf{r})$ separately.

The behavior of $\psi_>(\mathbf{r})$ near the origin is relatively easy to discern. By analysis similar to that of equations (1.40) through (1.46), we see that $(1/\|\mathbf{r}' - \mathbf{r}\|)$ with $\|\mathbf{r}'\| \geq R'$ is analytic in the components of \mathbf{r} in a small neighborhood of the origin. Consequently, $\psi_>(\mathbf{r})$ is also analytic in the components of \mathbf{r} in a small neighborhood of the origin, and this conclusion is independent of the nature of $\rho(\mathbf{r}')$ for $\|\mathbf{r}'\| \geq R'$ save for some mild distribution theoretic or integrability conditions such as, for example, that the integral (1.67) be absolutely convergent. We also observe that $(1/\|\mathbf{r}' - \mathbf{r}\|)$ with $\|\mathbf{r}'\| \geq R'$ is harmonic in the variables \mathbf{r} in a small neighborhood of the origin,

$$\nabla^2(1/\|\mathbf{r}' - \mathbf{r}\|) = (\partial_x^2 + \partial_y^2 + \partial_z^2)(1/\|\mathbf{r}' - \mathbf{r}\|) = 0 \text{ for } \|\mathbf{r}'\| \geq R' \text{ and } \mathbf{r} \text{ near 0.} \quad (\text{F.1.68})$$

Consequently, $\psi_>(\mathbf{r})$ is also harmonic around the origin,

$$\nabla^2\psi_>(\mathbf{r}) = 0 \text{ for } \mathbf{r} \text{ near 0.} \quad (\text{F.1.69})$$

Finding the behavior of $\psi_<(\mathbf{r})$ near the origin requires more work. We know that an expansion for $\rho(\mathbf{r}')$ of the form (1.33) converges for \mathbf{r}' in some sufficiently small polydisc. Select the R' in (1.66) and (1.67) such that the region \mathbf{r}' real and $\|\mathbf{r}'\| \leq R'$ lies within this polydisc. Then we may use the expansion (1.33) for $\rho(\mathbf{r}')$ in the integral (1.66) to compute $\psi_<(\mathbf{r})$.

Let us do this one term at a time. Define functions $\mathcal{X}_{m\ell s}(\mathbf{r})$ by the rule

$$\mathcal{X}_{m\ell s}(\mathbf{r}) = \int_{\|\mathbf{r}'\| \leq R'} d^3\mathbf{r}' (r')^{2s+\ell} Y_\ell^m(\Omega') / \|\mathbf{r}' - \mathbf{r}\|. \quad (\text{F.1.70})$$

Then, $\psi_<(\mathbf{r})$ will have the expansion

$$\psi_<(\mathbf{r}) = \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} \mathcal{X}_{m\ell s}(\mathbf{r}). \quad (\text{F.1.71})$$

The integral (1.70) can be written in the iterated form

$$\mathcal{X}_{m\ell s}(\mathbf{r}) = \int_0^{R'} dr' (r')^2 (r')^{2s+\ell} \int d\Omega' Y_\ell^m(\Omega') / \| \mathbf{r}' - \mathbf{r} \| . \quad (\text{F.1.72})$$

The second integral in (1.72) has the value

$$\int d\Omega' Y_\ell^m(\Omega') / \| \mathbf{r}' - \mathbf{r} \| = [4\pi/(2\ell+1)](r'_</r'_>^\ell) Y_\ell^m(\Omega), \quad (\text{F.1.73})$$

where $r_<$ and $r_>$ are defined by the equations

$$r_< = \text{the lesser of } r, r', \quad (\text{F.1.74})$$

$$r_> = \text{the greater of } r, r'. \quad (\text{F.1.75})$$

This value can be used in (1.72) to yield the result

$$\mathcal{X}_{m\ell s}(\mathbf{r}) = \mathcal{X}_{m\ell s}^1(\mathbf{r}) + \mathcal{X}_{m\ell s}^2(\mathbf{r}), \quad (\text{F.1.76})$$

where

$$\mathcal{X}_{m\ell s}^1(\mathbf{r}) = -4\pi[(2\ell+1)+(2s+2)]^{-1}(2s+2)^{-1}r^{2s+2}r^\ell Y_\ell^m(\theta, \phi), \quad (\text{F.1.77})$$

$$\mathcal{X}_{m\ell s}^2(\mathbf{r}) = 4\pi[(2\ell+1)(2s+2)]^{-1}(R')^{2s+2}r^\ell Y_\ell^m(\theta, \phi). \quad (\text{F.1.78})$$

We see that $\mathcal{X}_{m\ell s}(\mathbf{r})$ is a linear combination of the functions $r^{2s+2}r^\ell Y_\ell^m$ and $r^\ell Y_\ell^m$. From our earlier work we know that both these functions are polynomials in the components of \mathbf{r} , and are therefore entire analytic functions of the variables x, y, z .

The next thing to check is that the series (1.71) for $\psi_<(\mathbf{r})$ converges. Following the decomposition (1.76), we will write $\psi_<(\mathbf{r})$ in the form

$$\psi_<(\mathbf{r}) = \psi_<^1(\mathbf{r}) + \psi_<^2(\mathbf{r}) \quad (\text{F.1.79})$$

where

$$\psi_<^1(\mathbf{r}) = -4\pi \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} [(2\ell+1)+(2s+2)]^{-1}(2s+2)^{-1}r^{2s+2}r^\ell Y_\ell^m(\theta, \phi), \quad (\text{F.1.80})$$

$$\psi_<^2(\mathbf{r}) = 4\pi \sum_{D=0}^{\infty} \sum_{\ell+2s=D} \sum_{m=-\ell}^{\ell} b_{m\ell s} [(2\ell+1)(2s+2)]^{-1}(R')^{2s+2}r^\ell Y_\ell^m(\theta, \phi). \quad (\text{F.1.81})$$

It is easily verified, using the bounds (1.56) and (1.61), that both the series $\psi_<^1(\mathbf{r})$ and $\psi_<^2(\mathbf{r})$ converge, and converge absolutely, for \mathbf{r} sufficiently near the origin and R' sufficiently small. Consequently, $\psi_<(\mathbf{r})$ itself is analytic in a neighborhood of the origin. We now know that both $\psi_>$ and $\psi_<$ are analytic around the origin, and hence their sum ψ is analytic about the origin, which is what we have wanted to prove.

At this point we make some observations about the functions $\psi_<^1(\mathbf{r})$ and $\psi_<^2(\mathbf{r})$. We claim that they satisfy the equations

$$\nabla^2 \psi_<^1(\mathbf{r}) = -4\pi\rho(\mathbf{r}), \quad (\text{F.1.82})$$

$$\nabla^2 \psi_<^2(\mathbf{r}) = 0. \quad (\text{F.1.83})$$

Since the series for $\psi_<^1$ and $\psi_<^2$ converge absolutely, the operator ∇^2 can be taken under the summation signs and allowed to act term by term. The claim (1.83) then follows immediately from (1.25). To verify (1.82) we note that the operator ∇^2 has the spherical decomposition

$$\nabla^2 = r^{-1} \partial_r^2 r - \mathcal{L}^2/r^2, \quad (\text{F.1.84})$$

and the spherical harmonics have the property

$$\mathcal{L}^2 Y_\ell^m = \ell(\ell+1) Y_\ell^m. \quad (\text{F.1.85})$$

It follows that

$$\begin{aligned} \nabla^2(r^{2s+2} r^\ell Y_\ell^m) &= \{[(r^{-1} \partial_r^2 r) - \ell(\ell+1)/r^2] r^{\ell+2s+2}\} Y_\ell^m \\ &= [(\ell+2s+3)(\ell+2s+2) - \ell(\ell+1)] r^{\ell+2s} Y_\ell^m \\ &= [(2\ell+1) + (2s+2)](2s+2) r^{2s} r^\ell Y_\ell^m. \end{aligned} \quad (\text{F.1.86})$$

We see that the ℓ and s dependent multiplicative factors in (1.86) cancel like factors in the denominators appearing in (1.80), and comparison of the resulting expression for $\nabla^2 \psi_<^1$ with that given in (1.33) for ρ shows that the assertion (1.82) is also correct.

Exercises

F.1.1. This section has been devoted to the rather laborious task of showing that if $\rho(\mathbf{r})$ is analytic in the components of \mathbf{r} at some point \mathbf{r}^0 , then the same is true for $\psi(\mathbf{r})$. By contrast, the converse is easy to prove. Show that if $\psi(\mathbf{r})$ is analytic in the components of \mathbf{r} at some point \mathbf{r}^0 , then the same is true for $\rho(\mathbf{r})$.

F.1.2. Consider, as examples, three possible forms for $\rho(\mathbf{r})$ as shown below. In each case find the corresponding $\psi(\mathbf{r})$, and discuss its analytic properties.

$$\begin{aligned} \rho(\mathbf{r}) &= \text{constant for } \|\mathbf{r}\| \leq R, \\ &= 0 \text{ for } \|\mathbf{r}\| > R; \end{aligned} \quad (\text{F.1.87})$$

$$\rho(\mathbf{r}) = a \exp(-br^2); \quad (\text{F.1.88})$$

$$\rho(\mathbf{r}) = a \exp(-br). \quad (\text{F.1.89})$$

F.1.3. Show that the electron charge density for any hydrogen atom energy eigenstate,

$$\rho(\mathbf{r}) = \bar{\mathcal{X}}(\mathbf{r}) \mathcal{X}(\mathbf{r}), \quad (\text{F.1.90})$$

is *not* analytic at the origin. Relate this “singular” behavior to the fact that the energy eigenstate wave function $\mathcal{X}(\mathbf{r})$ satisfies the Schrödinger equation

$$-\left[\hbar^2/(2m)\right]\nabla^2 \mathcal{X} - (e^2/r)\mathcal{X} = E\mathcal{X}. \quad (\text{F.1.91})$$

F.1.4. Suppose that $f(D, \mathbf{r})$ is a homogeneous polynomial of degree D in the components of \mathbf{r} . According to Exercise 1.5.12 it obeys the relation

$$(\mathbf{r} \cdot \partial) f(D, \mathbf{r}) = D f(D, \mathbf{r}). \quad (\text{F.1.92})$$

Show that the operators $(\mathbf{r} \cdot \partial)$ and \mathcal{L} commute. You have provided another demonstration that \mathcal{L}_- maps homogeneous polynomials into themselves and leaves their degrees unchanged.

F.1.5. Verify (1.25) directly for H_ℓ^ℓ as given by (1.20). Show that ∇^2 commutes with \mathcal{L} , and hence (1.25) holds for all H_ℓ^m .

F.1.6. Verify the sums (1.29) and (1.31).

F.1.7. Verify the bound (1.35).

F.1.8. Verify (1.36) through (1.53). Verify that H_ℓ^ℓ as given by (1.20) satisfies the bound (1.53).

F.1.9. The conditions (1.42) were chosen to simplify analysis. Show that the analysis can be modified to improve the bound (1.53) and the requirement (1.63) by replacing $(4\sqrt{3})$ by a smaller factor. Verify that H_ℓ^ℓ as given by (1.20) satisfies your improved bound. Hint: Let δ be a real number in the open interval $0 < \delta < 1$. Show that if $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ satisfy the conditions

$$\boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq (1 - \delta)^2/4, \quad \boldsymbol{\eta} \cdot \boldsymbol{\eta} \leq (1 - \delta)^2/4, \quad (\text{F.1.93})$$

then one has the inequality

$$|[\|\mathbf{e} - \boldsymbol{\zeta}\|]| \geq \delta^{1/2}. \quad (\text{F.1.94})$$

F.1.10. Since H_ℓ^m is a homogeneous polynomial of degree ℓ , show that both sides of (1.39) can be divided by $(r')^\ell$ so that, with the aid of (1.38), it can be rewritten in the form

$$H_\ell^m(\boldsymbol{\zeta}) = (4\pi)^{-1}(2\ell + 1) \int d\Omega' Y_\ell^m(\Omega') / \|\mathbf{e}(\Omega') - \boldsymbol{\zeta}\|. \quad (\text{F.1.95})$$

Verify that the right side of (1.95) is analytic in the components of $\boldsymbol{\zeta}$ for $\boldsymbol{\zeta}$ sufficiently near 0, and that it has a Taylor expansion about 0 that converges absolutely for $\boldsymbol{\zeta}$ in a sufficiently small polydisc about the origin. (Here you may assume that the composition of two analytic functions is again analytic. See Section 38.2.) Show that the coefficients of this Taylor expansion can be determined from a knowledge of the values of the right side of (1.95) when $\boldsymbol{\zeta}$ is *real* and near 0. We know that both sides of (1.95) are equal when $\boldsymbol{\zeta}$ is real and near 0. (Such a set is an example of what is called a *real environment*. See Exercise 38.2.8.) It follows that both sides of (1.95) have identical Taylor coefficients. Consequently, both sides of (1.95) must also be equal when $\boldsymbol{\zeta}$ is complex and in a sufficiently small polydisc about the origin. Show, in fact, that they must be equal in the domain (1.93).

F.1.11. Verify that (1.61) is a consequence of (1.11).

F.1.12. Verify that the series (1.62) is convergent if (1.64) is satisfied.

F.1.13. Verify that if (in the static case) ρ vanishes in some region, then ψ is analytic in this region.

F.1.14. Verify (1.77) and (1.78). Show that the series (1.80) and (1.81) converge.

F.1.15. Verify (1.86) and (1.82).

F.1.16. Our discussion so far of how ψ inherits the analytic properties of ρ has taken a rather circuitous path through the territory of Taylor and harmonic series. Is there an approach that displays the inheritance directly? There is. Consider $\psi_<$ as given by (1.66). Introduce the variable Δ by the definition

$$\mathbf{r}' = \mathbf{r} + \Delta, \quad (\text{F.1.96})$$

and show that (1.66) can also be written in the form

$$\psi_<(\mathbf{r}) = \int_{\|\mathbf{r}+\Delta\| \leq R'} d^3\Delta \rho(\mathbf{r} + \Delta) / \|\Delta\|. \quad (\text{F.1.97})$$

Next introduce polar coordinates for Δ by the relation

$$\Delta = \Delta \mathbf{e}(\Omega). \quad (\text{F.1.98})$$

Show that (1.97) can be rewritten in the form

$$\psi_<(\mathbf{r}) = \int d\Omega \int_0^{\tilde{\Delta}} d\Delta \Delta \rho(\mathbf{r} + \Delta), \quad (\text{F.1.99})$$

where $\tilde{\Delta}$ is given by the expression

$$\tilde{\Delta}(\mathbf{r}, \Omega, R') = -\mathbf{r} \cdot \mathbf{e}(\Omega) + \{(R')^2 + [\mathbf{r} \cdot \mathbf{e}(\Omega)]^2 - r^2\}^{1/2}. \quad (\text{F.1.100})$$

Finally, introduce the variable τ by writing

$$\Delta = \tau \tilde{\Delta}, \quad (\text{F.1.101})$$

and show that (1.99) can be rewritten as

$$\psi_<(\mathbf{r}) = \int d\Omega \int_0^1 d\tau \tau \tilde{\Delta}^2 \rho[\mathbf{r} + \tau \tilde{\Delta} \mathbf{e}(\Omega)]. \quad (\text{F.1.102})$$

As it stands, (1.102) may be viewed as an integral representation for $\psi_<(\mathbf{r})$ that is valid for small real \mathbf{r} . Now consider making \mathbf{r} complex. Verify that $\tilde{\Delta}$ as given by (1.100) is analytic in \mathbf{r} for \mathbf{r} contained in a sufficiently small polydisc about 0. By hypothesis, $\rho(\mathbf{r})$ is also analytic in some such polydisc. Verify that the same is true for the function $\rho[\mathbf{r} + \tau \tilde{\Delta} \mathbf{e}(\Omega)]$. Finally, show that $\psi_<(\mathbf{r})$ as given by (1.102) must be analytic in some such polydisc.

F.2 The Time Dependent Case

We will now sketch how the results obtained so far can be extended to the time dependent case. In the time dependent case the scalar potential satisfies the inhomogeneous wave equation

$$[\nabla^2 - (1/c^2)\partial_t^2]\psi(\mathbf{r}, t) = -4\pi\rho(\mathbf{r}, t). \quad (\text{F.2.1})$$

Let us assume that ρ , although time dependent, has a *bounded* Fourier spectrum. That is, we assume that $\rho(\mathbf{r}, t)$ can be written in the form

$$\rho(\mathbf{r}, t) = (1/2\pi) \int_{-\omega_{\max}}^{\omega_{\max}} d\omega \tilde{\rho}(\mathbf{r}, \omega) \exp(-i\omega t) \quad (\text{F.2.2})$$

where ω_{\max} is some finite frequency cutoff. Then $\psi(\mathbf{r}, t)$ will also have a bounded Fourier spectrum,

$$\psi(\mathbf{r}, t) = (1/2\pi) \int_{-\omega_{\max}}^{\omega_{\max}} d\omega \tilde{\psi}(\mathbf{r}, \omega) \exp(-i\omega t). \quad (\text{F.2.3})$$

We know from (2.1) that the Fourier transform $\tilde{\psi}$ satisfies the inhomogeneous Helmholtz equation

$$(\nabla^2 + k^2)\tilde{\psi}(\mathbf{r}, \omega) = -4\pi\tilde{\rho}(\mathbf{r}, \omega) \quad (\text{F.2.4})$$

where

$$k = \omega/c. \quad (\text{F.2.5})$$

With the assumption of an outgoing wave boundary condition, equation (2.4) has the solution

$$\tilde{\psi}(\mathbf{r}, \omega) = \int d^3\mathbf{r}' \tilde{\rho}(\mathbf{r}', \omega) [\exp(ik \parallel \mathbf{r}' - \mathbf{r} \parallel)] / \parallel \mathbf{r}' - \mathbf{r} \parallel. \quad (\text{F.2.6})$$

We now assume that $\tilde{\rho}(\mathbf{r}, \omega)$, for all ω in the closed interval $[-\omega_{\max}, \omega_{\max}]$, is analytic in the components of \mathbf{r} at some point \mathbf{r}^0 . Then it can be shown that $\tilde{\psi}(\mathbf{r}, \omega)$ will also be analytic at \mathbf{r}^0 . It follows from (2.3) that $\psi(\mathbf{r}, t)$ will also be analytic in the components of \mathbf{r} at \mathbf{r}^0 for all t . Finally, again from (2.3), we see that $\psi(\mathbf{r}, t)$ will also be an analytic function of t for all t .

We now outline the proof of this assertion. As before, we take \mathbf{r}^0 to be the origin and break up the region of integration in (2.6) to write

$$\tilde{\psi}(\mathbf{r}, \omega) = \tilde{\psi}_<(\mathbf{r}, \omega) + \tilde{\psi}_>(\mathbf{r}, \omega) \quad (\text{F.2.7})$$

where

$$\tilde{\psi}_<(\mathbf{r}, \omega) = \int_{\parallel \mathbf{r}' \parallel \leq R'} d^3\mathbf{r}' \tilde{\rho}(\mathbf{r}', \omega) [\exp(ik \parallel \mathbf{r}' - \mathbf{r} \parallel)] / \parallel \mathbf{r}' - \mathbf{r} \parallel, \quad (\text{F.2.8})$$

$$\tilde{\psi}_>(\mathbf{r}, \omega) = \int_{\parallel \mathbf{r}' \parallel \geq R'} d^3\mathbf{r}' \tilde{\rho}(\mathbf{r}', \omega) [\exp(ik \parallel \mathbf{r}' - \mathbf{r} \parallel)] / \parallel \mathbf{r}' - \mathbf{r} \parallel. \quad (\text{F.2.9})$$

It follows, from arguments similar to those made earlier, that $\tilde{\psi}_>(\mathbf{r}, \omega)$ is analytic in the components of \mathbf{r} in a small neighborhood of the origin, and satisfies the Helmholtz equation

$$[\nabla^2 + k^2]\tilde{\psi}_>(\mathbf{r}, \omega) = 0 \text{ for } \mathbf{r} \text{ near } 0. \quad (\text{F.2.10})$$

To study the behavior of $\tilde{\psi}_<$ we assume, as before, that R' is sufficiently small that $\tilde{\rho}(\mathbf{r}', \omega)$ has a convergent expansion of the form (1.33) where the coefficients $b_{m\ell s}(\omega)$ are now ω dependent. Again consider each term at a time and, in analogy to (1.78), examine the integrals

$$\tilde{\mathcal{X}}_{m\ell s}(\mathbf{r}, \omega) = \int_0^{R'} dr' (r')^2 (r')^{2s+\ell} \int d\Omega' Y_\ell^m(\Omega') [\exp(ik \|\mathbf{r}' - \mathbf{r}\|)] / \|\mathbf{r}' - \mathbf{r}\|. \quad (\text{F.2.11})$$

The second integral in (2.11) has the value

$$\int d\Omega' Y_\ell^m(\Omega') [\exp(ik \|\mathbf{r}' - \mathbf{r}\|)] / \|\mathbf{r}' - \mathbf{r}\| = 4\pi ik j_\ell(kr_<) h_\ell^1(kr_>) Y_\ell^m(\Omega). \quad (\text{F.2.12})$$

Consequently our problem is reduced to studying the behavior of the integral

$$\int_0^{R'} dr' (r')^{\ell+2s+2} 4\pi ik j_\ell(kr_<) h_\ell^1(kr_>). \quad (\text{F.2.13})$$

It can be shown that this integral has the same analytic behavior as the related integral

$$\int_0^{R'} dr' (r')^{\ell+2s+2} [4\pi/(2\ell+1)] r_<^\ell / r_>^{\ell+1} \quad (\text{F.2.14})$$

for the analogous time independent case. In particular, $\tilde{\mathcal{X}}_{m\ell s}$ can be written in the form

$$\tilde{\mathcal{X}}_{m\ell s}(\mathbf{r}, \omega) = \tilde{\mathcal{X}}_{m\ell s}^1(\mathbf{r}, \omega) + \tilde{\mathcal{X}}_{m\ell s}^2(\mathbf{r}, \omega) \quad (\text{F.2.15})$$

where both $\tilde{\mathcal{X}}^1$ and $\tilde{\mathcal{X}}^2$ are entire analytic functions of the variables x, y, z . Correspondingly, $\tilde{\psi}_<(\mathbf{r}, \omega)$ can be written in the form

$$\tilde{\psi}_<(\mathbf{r}, \omega) = \tilde{\psi}_<^1(\mathbf{r}, \omega) + \tilde{\psi}_<^2(\mathbf{r}, \omega) \quad (\text{F.2.16})$$

where both $\tilde{\psi}_<^1$ and $\tilde{\psi}_<^2$ are analytic in a neighborhood about the origin, and satisfy the equations

$$(\nabla^2 + k^2) \tilde{\psi}_<^1(\mathbf{r}, \omega) = -4\pi \tilde{\rho}(\mathbf{r}, \omega), \quad (\text{F.2.17})$$

$$(\nabla^2 + k^2) \tilde{\psi}_<^2(\mathbf{r}, \omega) = 0. \quad (\text{F.2.18})$$

Exercises

F.2.1. Review Exercise 1.13. Now consider the time dependent case. Show that the frequency cutoff in (2.2) is essential to the argument. Show that if all frequencies are allowed, then the effect of singularities in ρ can *propagate*. That is, a singularity in ρ at some point \mathbf{r}' , and some time, can produce a singularity in ψ at some other point \mathbf{r} at some later time even though ρ may be analytic at \mathbf{r} .

F.2.2. Evaluate the integral (2.13), and complete the analysis of the time dependent case.

F.2.3. Extend the method of Exercise 1.16 to the time-dependent case (2.6).

F.3 Smoothing Properties of the Laplacian Kernel

In Section 22.2 we encountered the relation

$$\mathbf{H}(\mathbf{r}) = [1/(4\pi)] \int_V d^3\mathbf{r}' \mathbf{F}(\mathbf{r}') G(\mathbf{r}, \mathbf{r}'). \quad (\text{F.3.1})$$

See (22.2.92). It follows from the work at the beginning of this appendix that if $\mathbf{F}(\mathbf{r}')$ is analytic, then $\mathbf{H}(\mathbf{r})$ will also be analytic. Now we will assume only that $\mathbf{F}(\mathbf{r}')$ is smooth, and then study what can be said about the properties of $\mathbf{H}(\mathbf{r})$. By *smooth*, we mean that

....

Bibliography

- [1] J.D. Jackson, *Classical Electrodynamics*, John Wiley (1999).
- [2] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Chapt. IV, Interscience (1962).
- [3] E.H. Lieb and M. Loss, *Analysis*, American Mathematical Society (1997). Suppose $\rho(\mathbf{r})$ is not analytic but does have n derivatives. Then, roughly speaking, it can be shown that $\psi(\mathbf{r})$ will have $n + 2$ derivatives. For precise results, see Chapt. 10 of this book.
- [4] A.R. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press (1960).
- [5] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press (1975).
- [6] N.M. Giunter, *Potential Theory and its Application to Basic Problems in Mathematical Physics*, F. Ungar, New York (1967).
- [7] A. Erdelyi et al., edit., *Higher Transcendental Functions*, Bateman Manuscript Project, Vol. 2, McGraw-Hill (1953).
- [8] The method of Exercise 1.16 is due to Robert Warnock. I am grateful to him for this and many other helpful comments.

Appendix G

Specification of $m \geq 1$ Current Filaments/Windings

Finding Level Lines

However, we must confess at this point that the level-line Ansatz gives us little and perhaps even confusing information about how to produce suitable current windings.

Thus determining the level lines of H provides a recipe for placing current windings with all windings having the *same* current.

In general integrating the equations of motion pair (3.66) and (3.67) numerically is likely easier than finding the level lines of H . But we have agreed to treat the various possible cases separately.

The Normal $m \geq 1$ Cases

Suppose V is of the form

$$V(\phi, z) = \sin(m\phi)f_{ms}(z), \quad (\text{G.0.1})$$

and therefore H is of the form

$$H(Q, P) = -(1/a)\sin(mP)f_{ms}(Qa). \quad (\text{G.0.2})$$

In this case finding level lines of H is elementary since (3.72) can be rewritten in the form

$$\sin(mP) = -aH/f_{ms}(Qa). \quad (\text{G.0.3})$$

Once some constant value of H has been specified and a set of Q values has been selected, then the corresponding values of P can be found by solving (3.73) for P .

For example, suppose that $f_{1,s}(z)$ has support only in the interval $z \in [0, L]$ and that in this interval it has the value

$$f_{1,s}(z) = \lambda[1 - \cos(\pi z/L)]. \quad (\text{G.0.4})$$

Figure * illustrates the behavior of $f_{1,s}(z)$ for the case

Appendix H

Harmonic Functions

Section 13.2 provided cylindrical harmonic expansions for the harmonic function ψ . This appendix does the same for the gradients of ψ . It also studies the *range* of the transverse gradient operators when acting on the space of harmonic functions. In particular, given a harmonic function χ , it shows that there exists a harmonic function ψ such that $\partial_x \psi = \chi$ or a ψ such that $\partial_y \psi = \chi$. Finally, it provides representations for harmonic functions in two variables.

H.1 Representation of Gradients

We know that the harmonic function $\psi(x, y, z)$ has the representation (13.2.37). We would like to find similar representations for $\partial_x \psi(x, y, z)$, $\partial_y \psi(x, y, z)$, and $\partial_z \psi(x, y, z)$ which, of course, are also harmonic functions.

H.1.1 Low-Order Results

We will begin by finding low-order results, and then work out results to all orders. Refer to (13.2.37) to write ψ in the form

$$\psi = \psi_0 + \psi_c + \psi_s \quad (\text{H.1.1})$$

where

$$\psi_0(x, y, z) = \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell}, \quad (\text{H.1.2})$$

$$\begin{aligned} \psi_c(x, y, z) &= \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell+m} \\ &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x + iy)^m, \end{aligned} \quad (\text{H.1.3})$$

$$\begin{aligned}\psi_s(x, y, z) &= \sum_{m=1}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell+m} \\ &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x + iy)^m.\end{aligned}\quad (\text{H.1.4})$$

Let us expand ψ through terms of third order. We can then differentiate this expansion to find the first few terms in the expansions of its gradients. Through terms of degree 3, we find for the constituents of ψ the expansions

$$\psi_0 = C_0^{[0]}(z) - (1/4)(x^2 + y^2)C_0^{[2]}(z) + \dots, \quad (\text{H.1.5})$$

$$\begin{aligned}\psi_c &= \Re(x + iy)C_{1,c}^{[0]}(z) + \Re(x + iy)^2C_{2,c}^{[0]}(z) + \Re(x + iy)^3C_{3,c}^{[0]}(z) \\ &\quad - (1/8)(x^2 + y^2)\Re(x + iy)C_{1,c}^{[2]}(z) + \dots \\ &= xC_{1,c}^{[0]}(z) + (x^2 - y^2)C_{2,c}^{[0]}(z) + (x^3 - 3xy^2)C_{3,c}^{[0]}(z) - (1/8)(x^2 + y^2)x C_{1,c}^{[2]}(z) + \dots, \\ &= xC_{1,c}^{[0]}(z) + (x^2 - y^2)C_{2,c}^{[0]}(z) \\ &\quad + x^3[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] - xy^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \dots,\end{aligned}\quad (\text{H.1.6})$$

$$\begin{aligned}\psi_s &= \Im(x + iy)C_{1,s}^{[0]}(z) + \Im(x + iy)^2C_{2,s}^{[0]}(z) + \Im(x + iy)^3C_{3,s}^{[0]}(z) \\ &\quad - (1/8)(x^2 + y^2)\Im(x + iy)C_{1,s}^{[2]}(z) + \dots \\ &= yC_{1,s}^{[0]}(z) + 2xyC_{2,s}^{[0]}(z) + (-y^3 + 3x^2y)C_{3,s}^{[0]}(z) - (1/8)(x^2 + y^2)y C_{1,s}^{[2]}(z) + \dots, \\ &= yC_{1,s}^{[0]}(z) + 2xyC_{2,s}^{[0]}(z) \\ &\quad - y^3[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)] + x^2y[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \dots.\end{aligned}\quad (\text{H.1.7})$$

Differentiating these expansions, and retaining terms through second order, give the results

$$\partial_z \psi_0 = C_0^{[1]}(z) - (1/4)(x^2 + y^2)C_0^{[3]}(z) + \dots, \quad (\text{H.1.8})$$

$$\partial_z \psi_c = xC_{1,c}^{[1]}(z) + (x^2 - y^2)C_{2,c}^{[1]}(z) + \dots, \quad (\text{H.1.9})$$

$$\partial_z \psi_s = yC_{1,s}^{[1]}(z) + 2xyC_{2,s}^{[1]}(z) + \dots, \quad (\text{H.1.10})$$

$$\partial_x \psi_0 = -(1/2)x C_0^{[2]}(z) + \dots, \quad (\text{H.1.11})$$

$$\begin{aligned}\partial_x \psi_c &= C_{1,c}^{[0]}(z) + 2xC_{2,c}^{[0]}(z) + 3x^2[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] \\ &\quad - y^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \dots,\end{aligned}\quad (\text{H.1.12})$$

$$\partial_x \psi_s = 2yC_{2,s}^{[0]}(z) + 2xy[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \dots, \quad (\text{H.1.13})$$

$$\partial_y \psi_0 = -(1/2)yC_0^{[2]}(z) + \dots, \quad (\text{H.1.14})$$

$$\partial_y \psi_c = -2yC_{2,c}^{[0]}(z) - 2xy[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \dots, \quad (\text{H.1.15})$$

$$\begin{aligned} \partial_y \psi_s &= C_{1,s}^{[0]}(z) + 2xC_{2,s}^{[0]}(z) - 3y^2[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)] \\ &\quad + x^2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] + \dots. \end{aligned} \quad (\text{H.1.16})$$

Finally, upon employing the decomposition (1.1), we find the results

$$\begin{aligned} \partial_z \psi &= C_0^{[1]}(z) + xC_{1,c}^{[1]}(z) + yC_{1,s}^{[1]}(z) \\ &\quad + (x^2 - y^2)C_{2,c}^{[1]}(z) + 2xyC_{2,s}^{[1]}(z) - (1/4)(x^2 + y^2)C_0^{[3]}(z) + \dots \\ &= C_0^{[1]}(z) + xC_{1,c}^{[1]}(z) + yC_{1,s}^{[1]}(z) \\ &\quad + x^2[C_{2,c}^{[1]}(z) - (1/4)C_0^{[3]}(z)] - y^2[C_{2,c}^{[1]}(z) + (1/4)C_0^{[3]}(z)] + 2xyC_{2,s}^{[1]}(z) + \dots, \end{aligned} \quad (\text{H.1.17})$$

$$\begin{aligned} \partial_x \psi &= C_{1,c}^{[0]}(z) + x[2C_{2,c}^{[0]}(z) - (1/2)C_0^{[2]}(z)] + 2yC_{2,s}^{[0]}(z) \\ &\quad + 3x^2[C_{3,c}^{[0]}(z) - (1/8)C_{1,c}^{[2]}(z)] - y^2[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] \\ &\quad + 2xy[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] \dots, \end{aligned} \quad (\text{H.1.18})$$

$$\begin{aligned} \partial_y \psi &= C_{1,s}^{[0]}(z) - y[2C_{2,c}^{[0]}(z) + (1/2)C_0^{[2]}(z)] + 2xC_{2,s}^{[0]}(z) \\ &\quad - 3y^2[C_{3,s}^{[0]}(z) + (1/8)C_{1,s}^{[2]}(z)] + x^2[3C_{3,s}^{[0]}(z) - (1/8)C_{1,s}^{[2]}(z)] \\ &\quad - 2xy[3C_{3,c}^{[0]}(z) + (1/8)C_{1,c}^{[2]}(z)] + \dots. \end{aligned} \quad (\text{H.1.19})$$

H.1.2 Results to All Orders

Let us study the effect of the operators ∂_x , ∂_y , and ∂_z on each of the terms in (1.1). Finding the effect of ∂_z is easy because it acts only on the $C_{m,\alpha}^{[n]}(z)$ to raise the value of n by 1. The effects of ∂_x and ∂_y are more complicated. Observe that ψ_0 , ψ_c , and ψ_s are sums over the “basis” functions $\rho^{2\ell}$, $\rho^{2\ell}\Re(x+iy)^m$, and $\rho^{2\ell}\Im(x+iy)^m$, respectively. We will first compute ∂_x and ∂_y of these basis functions; then compute ∂_x and ∂_y of ψ_0 , ψ_c , and ψ_s ; and finally compute $\partial_x \psi$ and $\partial_y \psi$ and also $\partial_z \psi$.

Transverse Gradients of $\rho^{2\ell}$

Let us begin by calculating ∂_x and ∂_y of $\rho^{2\ell}$. We find that

$$\partial_x \rho^{2\ell} = \partial_x (x^2 + y^2)^\ell = \ell(x^2 + y^2)^{\ell-1} 2x = 2\ell \rho^{2\ell-2} \Re(x+iy) \quad (\text{H.1.20})$$

and

$$\partial_y \rho^{2\ell} = \partial_y (x^2 + y^2)^\ell = \ell(x^2 + y^2)^{\ell-1} 2y = 2\ell \rho^{2\ell-2} \Im(x+iy). \quad (\text{H.1.21})$$

Transverse Gradients of $\rho^{2\ell}\Re(x+iy)^m$ and $\rho^{2\ell}\Im(x+iy)^m$

Determining the transverse gradients of $\rho^{2\ell}\Re(x+iy)^m$ and $\rho^{2\ell}\Im(x+iy)^m$ requires somewhat more work. Note that for these calculations we have $m \geq 1$. We find that

$$\begin{aligned}
\partial_x[\rho^{2\ell}\Re(x+iy)^m] &= \partial_x\{\rho^{2\ell}[(x+iy)^m + (x-iy)^m]/2\} \\
&= [\partial_x\rho^{2\ell}][(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}\partial_x[(x+iy)^m + (x-iy)^m]/2 \\
&= 2\ell\rho^{2\ell-2}\Re(x+iy)[(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}m[(x+iy)^{m-1} + (x-iy)^{m-1}]/2 \\
&= \ell\rho^{2\ell-2}[(x+iy) + (x-iy)][(x+iy)^m + (x-iy)^m]/2 + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} + (x-iy)^{m+1}] + \rho^2[(x+iy)^{m-1} + (x-iy)^{m-1}]\}/2 \\
&\quad + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Re(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1},
\end{aligned} \tag{H.1.22}$$

$$\begin{aligned}
\partial_x[\rho^{2\ell}\Im(x+iy)^m] &= \partial_x\{\rho^{2\ell}[(x+iy)^m - (x-iy)^m]/(2i)\} \\
&= [\partial_x\rho^{2\ell}][(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}\partial_x[(x+iy)^m - (x-iy)^m]/(2i) \\
&= 2\ell\rho^{2\ell-2}\Re(x+iy)[(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}m[(x+iy)^{m-1} - (x-iy)^{m-1}]/(2i) \\
&= \ell\rho^{2\ell-2}[(x+iy) + (x-iy)][(x+iy)^m - (x-iy)^m]/(2i) + m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} - (x-iy)^{m+1}] + \rho^2[(x+iy)^{m-1} - (x-iy)^{m-1}]\}/(2i) \\
&\quad + m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Im(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1},
\end{aligned} \tag{H.1.23}$$

$$\begin{aligned}
\partial_y[\rho^{2\ell}\Re(x+iy)^m] &= \partial_y\{\rho^{2\ell}[(x+iy)^m + (x-iy)^m]/2\} \\
&= [\partial_y\rho^{2\ell}][(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}\partial_y[(x+iy)^m + (x-iy)^m]/2 \\
&= 2\ell\rho^{2\ell-2}\Im(x+iy)[(x+iy)^m + (x-iy)^m]/2 + \rho^{2\ell}im[(x+iy)^{m-1} - (x-iy)^{m-1}]/2 \\
&= \ell\rho^{2\ell-2}[(x+iy) - (x-iy)][(x+iy)^m + (x-iy)^m]/(2i) - m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\{[(x+iy)^{m+1} - (x-iy)^{m+1}] - \rho^2[(x+iy)^{m-1} - (x-iy)^{m-1}]\}/(2i) \\
&\quad - m\rho^{2\ell}\Im(x+iy)^{m-1} \\
&= \ell\rho^{2\ell-2}\Im(x+iy)^{m+1} - (\ell+m)\rho^{2\ell}\Im(x+iy)^{m-1},
\end{aligned} \tag{H.1.24}$$

$$\begin{aligned}
\partial_y[\rho^{2\ell}\Im(x+iy)^m] &= \partial_y\{\rho^{2\ell}[(x+iy)^m - (x-iy)^m]/(2i)\} \\
&= [\partial_y\rho^{2\ell}][(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}\partial_y[(x+iy)^m - (x-iy)^m]/(2i) \\
&= 2\ell\rho^{2\ell-2}\Im(x+iy)[(x+iy)^m - (x-iy)^m]/(2i) + \rho^{2\ell}im[(x+iy)^{m-1} + (x-iy)^{m-1}]/(2i) \\
&= -\ell\rho^{2\ell-2}[(x+iy) - (x-iy)][(x+iy)^m - (x-iy)^m]/2 + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= -\ell\rho^{2\ell-2}\{[(x+iy)^{m+1} + (x-iy)^{m+1}] - \rho^2[(x+iy)^{m-1} + (x-iy)^{m-1}]\}/2 \\
&\quad + m\rho^{2\ell}\Re(x+iy)^{m-1} \\
&= -\ell\rho^{2\ell-2}\Re(x+iy)^{m+1} + (\ell+m)\rho^{2\ell}\Re(x+iy)^{m-1}.
\end{aligned} \tag{H.1.25}$$

Transverse Gradients of ψ_0

We are now prepared to compute the transverse gradients of ψ_0 . Based on (1.2), (1.20), and (1.21), we find that

$$\begin{aligned}
\partial_x \psi_0 &= \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2\ell}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell-2} \Re(x+iy) \\
&= \sum_{\ell=1}^{\infty} (-1)^\ell \frac{2\ell}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell-2} \Re(x+iy) \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{(2\ell+2)}{2^{2\ell+2} (\ell+1)! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy) \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{2}{2^{2\ell+2} \ell! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)
\end{aligned} \tag{H.1.26}$$

and

$$\begin{aligned}
\partial_y \psi_0 &= \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2\ell}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell-2} \Im(x+iy) \\
&= \sum_{\ell=1}^{\infty} (-1)^\ell \frac{2\ell}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell-2} \Im(x+iy) \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{(2\ell+2)}{2^{2\ell+2} (\ell+1)! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy) \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{2}{2^{2\ell+2} \ell! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)
\end{aligned} \tag{H.1.27}$$

Transverse Gradients of ψ_c and ψ_s

We are also ready to compute the transverse gradients of ψ_c and ψ_s . These results are more lengthy. Based on (1.3), (1.4), and (1.22) through (1.25), we find the following results.

Result for $\partial_x \psi_c$

$$\begin{aligned}
\partial_x \psi_c &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) [\ell \rho^{2\ell-2} \Re(x+iy)^{m+1} + (\ell+m) \rho^{2\ell} \Re(x+iy)^{m-1}] \\
&= \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} (-1)^{\ell} \frac{m! \ell}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell-2} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m! (\ell+1)}{2^{2\ell+2} (\ell+1)! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,c}^{[2\ell]}(z) - [1/(4m)] C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m.
\end{aligned}$$

(H.1.28)

Result for $\partial_x \psi_s$

$$\begin{aligned}
\partial_x \psi_s &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) [\ell \rho^{2\ell-2} \Im(x+iy)^{m+1} + (\ell+m) \rho^{2\ell} \Im(x+iy)^{m-1}] \\
&= \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} (-1)^{\ell} \frac{m! \ell}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell-2} \Im(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m! (\ell+1)}{2^{2\ell+2} (\ell+1)! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1}. \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1}. \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,s}^{[2\ell]}(z) - [1/(4m)] C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m.
\end{aligned} \tag{H.1.29}$$

Result for $\partial_y \psi_c$

$$\begin{aligned}
\partial_y \psi_c &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) [\ell \rho^{2\ell-2} \Im(x+iy)^{m+1} - (\ell+m) \rho^{2\ell} \Im(x+iy)^{m-1}] \\
&= \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} (-1)^{\ell} \frac{m! \ell}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell-2} \Im(x+iy)^{m+1} \\
&\quad - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m! (\ell+1)}{2^{2\ell+2} (\ell+1)! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad - \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad - \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= - \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,c}^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy)^{m+1} \\
&\quad - \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy)^{m-1} \\
&= - \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad - \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,c}^{[2\ell]}(z) + [1/(4m)] C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m.
\end{aligned} \tag{H.1.30}$$

Result for $\partial_y \psi_s$

$$\begin{aligned}
\partial_y \psi_s &= \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) [-\ell \rho^{2\ell-2} \Re(x+iy)^{m+1} + (\ell+m) \rho^{2\ell} \Re(x+iy)^{m-1}] \\
&= - \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} (-1)^{\ell} \frac{m! \ell}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell-2} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m! (\ell+1)}{2^{2\ell+2} (\ell+1)! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m! (\ell+m)}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad - \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{m!}{2^{2\ell+2} \ell! (\ell+m+1)!} C_{m,s}^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy)^{m+1} \\
&\quad + \sum_{m=3}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m-1)!} C_{m,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy)^{m-1} \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,s}^{[2\ell]}(z) + [1/(4m)] C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m.
\end{aligned} \tag{H.1.31}$$

Gradients of ψ

The last step is to put all the previous results together using (1.1). So doing gives the following final results.

Result for $\partial_x \psi$

$$\begin{aligned}
\partial_x \psi(x, y, z) &= \partial_x \psi_0 + \partial_x \psi_c + \partial_x \psi_s \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{2}{2^{2\ell+2} \ell! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^\ell \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_0^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m \\
&= \sum_{\ell=0}^{\infty} (-1)^\ell \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) - (1/4)C_0^{[2\ell+2]}(z)] \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^\ell \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m.
\end{aligned} \tag{H.1.32}$$

Result for $\partial_y \psi$

$$\begin{aligned}
\partial_y \psi(x, y, z) &= \partial_y \psi_0 + \partial_y \psi_c + \partial_y \psi_s \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell+1} \frac{2}{2^{2\ell+2} \ell! (\ell+1)!} C_0^{[2\ell+2]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad - \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,c}^{[2\ell]}(z) \rho^{2\ell} \Im(x+iy) \\
&\quad - \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,c}^{[2\ell]}(z) + [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m \\
&\quad + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) + [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell} \Re(x+iy) \\
&\quad + \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) + [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell} \Re(x+iy)^m \\
&\quad - \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) + (1/4)C_0^{[2\ell+2]}] \rho^{2\ell} \Im(x+iy) \\
&\quad - \sum_{m=2}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,c}^{[2\ell]}(z) + [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell} \Im(x+iy)^m.
\end{aligned} \tag{H.1.33}$$

Result for $\partial_z \psi$

$$\begin{aligned}
\partial_z \psi(x, y, z) &= \partial_z \psi_0 + \partial_z \psi_c + \partial_z \psi_s \\
&= \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_0^{[2\ell+1]}(z) \rho^{2\ell} \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,c}^{[2\ell+1]}(z) \rho^{2\ell} \Re(x+iy)^m \\
&\quad + \sum_{m=1}^{\infty} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} C_{m,s}^{[2\ell+1]}(z) \rho^{2\ell} \Im(x+iy)^m.
\end{aligned} \tag{H.1.34}$$

Exercises

H.1.1. Verify (1.22) through (1.22) for selected example values of ℓ and m .

H.1.2. Verify (1.26) and (1.27).

H.1.3. Verify (1.28) through (1.31).

H.1.4. Verify (1.32) through (1.34). Show that, when expanded to low order, they yield results identical to (1.17) through (1.19).

H.1.5. Write

$$\mathbf{B} = \nabla\psi \quad (\text{H.1.35})$$

to find the cylindrical coordinate results

$$B_\rho = \partial\psi/\partial\rho, \quad (\text{H.1.36})$$

$$B_\phi = (1/\rho)\partial\psi/\partial\phi, \quad (\text{H.1.37})$$

$$B_z = \partial\psi/\partial z. \quad (\text{H.1.38})$$

Also invoke the relations

$$B_x = (\cos\phi)B_\rho - (\sin\phi)B_\phi \quad (\text{H.1.39})$$

$$B_y = (\sin\phi)B_\rho + (\cos\phi)B_\phi. \quad (\text{H.1.40})$$

Use these results to derive (1.32) through (1.34).

H.2 Range of Transverse Gradient Operators

We next enquire about the *range* of the operators ∂_x and ∂_y . Consider ∂_x . Suppose χ is some given (real) harmonic function. Can we find a (real) harmonic ψ such that either

$$\partial_x\psi = \chi \quad (\text{H.2.1})$$

or

$$\partial_y\psi = \chi? \quad (\text{H.2.2})$$

We will verify, by construction, that the answer is *yes*.

H.2.1 Solution of $\partial_x\psi = \chi$

Since χ is assumed (real) harmonic, it has a representation of the form

$$\begin{aligned} \chi(x, y, z) &= \sum_{\ell=0}^{\infty} (-1)^\ell \frac{1}{2^{2\ell}\ell!(\ell!)} B_0^{[2\ell]}(z) \rho^{2\ell} \\ &+ \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m} \\ &+ \sum_{m=1}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^\ell \frac{m!}{2^{2\ell}\ell!(\ell+m)!} B_{m,s}^{[2\ell]}(z) \rho^{2\ell+m}. \end{aligned} \quad (\text{H.2.3})$$

Evidently, we must compare (1.32) and (2.3). We must try to satisfy the pair of equations

$$\begin{aligned}
& \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell} + \\
& + \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) - (1/4) C_0^{[2\ell+2]}(z)] \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,c}^{[2\ell]}(z) - [1/(4m)] C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\
& \stackrel{?}{=} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} B_0^{[2\ell]}(z) \rho^{2\ell} + \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m}
\end{aligned} \tag{H.2.4}$$

and

$$\begin{aligned}
& \sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,s}^{[2\ell]}(z) - [1/(4m)] C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\
& \stackrel{?}{=} \sum_{m=1}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,s}^{[2\ell]}(z) \rho^{2\ell+m}.
\end{aligned} \tag{H.2.5}$$

Here we have used (13.2.7) and (13.2.8).

Let us first work on question (2.4). Suppose we set

$$C_{1,c}^{[0]}(z) = B_0^{[0]}(z). \tag{H.2.6}$$

Then (2.4) becomes the question

$$\begin{aligned}
& + \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) - (1/4) C_0^{[2\ell+2]}(z)] \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,c}^{[2\ell]}(z) - [1/(4m)] C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\
& \stackrel{?}{=} \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m}.
\end{aligned} \tag{H.2.7}$$

Note that the right side of (2.7) can be written in the form

$$\begin{aligned}
& \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m} \\
& = \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} B_{1,c}^{[2\ell]}(z) \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m}.
\end{aligned} \tag{H.2.8}$$

Therefore, upon equating like terms, we see that (2.7) is equivalent to the two questions

$$2[C_{2,c}^{[2\ell]}(z) - (1/4)C_0^{[2\ell+2]}(z)] \stackrel{?}{=} B_{1,c}^{[2\ell]}(z) \tag{H.2.9}$$

and

$$(m+1)C_{m+1,c}^{[2\ell]}(z) - [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z) = B_{m,c}^{[2\ell]}(z) \text{? for } m \geq 2. \tag{H.2.10}$$

We will solve (2.9) by making the stipulation

$$C_0^{[0]} = 0, \tag{H.2.11}$$

in which case (2.9) has the solution

$$C_{2,c}^{[0]}(z) = (1/2)B_{1,c}^{[0]}(z). \tag{H.2.12}$$

Let us next rewrite (2.10) in the recursive form

$$C_{m+1,c}^{[2\ell]}(z) = [1/(m+1)]B_{m,c}^{[2\ell]}(z) + \{1/[(4m)(m+1)]\}C_{m-1,c}^{[2\ell+2]}(z) \text{? for } m \geq 2. \tag{H.2.13}$$

In view of (2.6) and (2.12), this is a well-defined recursion relation. For example, putting $m = 2$ in (2.13) gives the result

$$C_{3,c}^{[0]}(z) = (1/3)B_{2,c}^{[0]}(z) + \{1/[(8)(3)]\}C_{1,c}^{[2]}(z), \tag{H.2.14}$$

which, using (2.6), can be rewritten as

$$C_{3,c}^{[0]}(z) = (1/3)B_{2,c}^{[0]}(z) + (1/24)B_0^{[2]}(z). \tag{H.2.15}$$

Now put $m = 3$. Doing so gives the result

$$C_{4,c}^{[0]}(z) = (1/4)B_{3,c}^{[0]}(z) + \{1/[(12)(4)]\}C_{2,c}^{[2]}(z), \tag{H.2.16}$$

which, using (2.12), can be rewritten as

$$C_{4,c}^{[0]}(z) = (1/4)B_{3,c}^{[0]}(z) + (1/96)B_{1,c}^{[2]}(z). \tag{H.2.17}$$

Finally, put $m = 4$, The result is

$$C_{5,c}^{[0]}(z) = (1/5)B_{4,c}^{[0]}(z) + \{1/[(16)(5)]\}C_{3,c}^{[2]}(z), \quad (\text{H.2.18})$$

which, using (2.15), can be rewritten as

$$C_{5,c}^{[0]}(z) = (1/5)B_{4,c}^{[0]}(z) + (1/240)B_{2,c}^{[2]}(z) + (1/1920)B_0^{[4]}(z). \quad (\text{H.2.19})$$

The pattern is now clear. All the $C_{m,c}^{[0]}(z)$ are determined in terms of the $B_{m,c}^{[n]}(z)$, and (2.4) is satisfied.

Move on to look at (2.5), which can also be written in the form

$$\begin{aligned} & \sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell+1} \\ & + \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\ & \stackrel{?}{=} \sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} B_{1,s}^{[2\ell]}(z) \rho^{2\ell+1} \\ & + \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,s}^{[2\ell]}(z) \rho^{2\ell+m}. \end{aligned} \quad (\text{H.2.20})$$

Upon equating like terms in (2.20) we see that it is equivalent to the two relations

$$2C_{2,s}^{[2\ell]}(z) \stackrel{?}{=} B_{1,s}^{[2\ell]}(z) \quad (\text{H.2.21})$$

and

$$(m+1)C_{m+1,s}^{[2\ell]}(z) - [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z) \stackrel{?}{=} B_{m,s}^{[2\ell]}(z) \text{ for } m \geq 2. \quad (\text{H.2.22})$$

The relation (2.21) has the solution

$$C_{2,s}^{[0]}(z) = (1/2)B_{1,s}^{[0]}(z), \quad (\text{H.2.23})$$

and (2.22) can be written in the recursive form

$$C_{m+1,s}^{[2\ell]}(z) \stackrel{?}{=} [1/(m+1)]B_{m,s}^{[2\ell]}(z) + \{1/[(4m)(m+1)]\}C_{m-1,s}^{[2\ell+2]}(z) \text{ for } m \geq 2. \quad (\text{H.2.24})$$

Now make the stipulation

$$C_{1,s}^{[0]}(z) = 0. \quad (\text{H.2.25})$$

Then, for $m = 2$, we get the result

$$C_{3,s}^{[0]}(z) = (1/3)B_{2,s}^{[0]}(z). \quad (\text{H.2.26})$$

Now put $m = 3$ to get the result

$$C_{4,s}^{[0]}(z) = (1/4)B_{3,s}^{[0]}(z) + \{1/[(12)(4)]\}C_{2,s}^{[2]}(z), \quad (\text{H.2.27})$$

which, when combined with (2.23), gives the result

$$C_{4,s}^{[0]}(z) = (1/4)B_{3,s}^{[0]}(z) + (1/96)B_{1,s}^{[2]}(z). \quad (\text{H.2.28})$$

To continue the calculation, put $m = 4$ to find

$$C_{5,s}^{[0]}(z) = (1/5)B_{4,s}^{[0]}(z) + \{1/[(16)(5)]\}C_{3,s}^{[2]}(z), \quad (\text{H.2.29})$$

which, when combined with (2.26), gives the result

$$C_{5,s}^{[0]}(z) = (1/5)B_{4,s}^{[0]}(z) + (1/240)B_{2,s}^{[2]}(z). \quad (\text{H.2.30})$$

The pattern becomes clear when $m = 5$. In this case we find that

$$C_{6,s}^{[0]}(z) = (1/6)B_{5,s}^{[0]}(z) + \{1/[(20)(6)]\}C_{4,s}^{[2]}(z), \quad (\text{H.2.31})$$

which, when combined with (2.28), gives the result

$$C_{6,s}^{[0]}(z) = (1/6)B_{5,s}^{[0]}(z) + (1/480)B_{3,s}^{[2]}(z) + (1/11520)B_{1,s}^{[4]}(z). \quad (\text{H.2.32})$$

We see that all the $C_{m,s}^{[0]}(z)$ are determined in terms of the $B_{m,s}^{[n]}(z)$, and (2.5) is satisfied. Thus, both (2.4) and (2.5) have been satisfied, and therefore our goal (2.1) has been met.

At this point we should comment on the stipulations (2.11) and (2.25). Evidently the stipulation (2.11) specifies that all terms of the form

$$\sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_0^{[2\ell]}(z) \rho^{2\ell} \quad (\text{H.2.33})$$

are omitted from ψ . And the stipulation (2.25) specifies that all terms of the form

$$\sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell+1} \quad (\text{H.2.34})$$

are omitted from ψ . We have seen that these terms are not needed to meet the goal (2.1), and their omission simplifies the recursion relations that specify ψ in terms of χ .

H.2.2 Solution of $\partial_y \psi = \chi$

We next address the question of whether there is a ψ such that (2.2) is satisfied. By symmetry we know that there must be such a ψ , but it is instructive to work out the details. Evidently we must compare (1.33) and (2.3). We must try to satisfy the pair of equations

$$\begin{aligned} & \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} C_{1,s}^{[2\ell]}(z) \rho^{2\ell} + \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell+1} \\ & + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,s}^{[2\ell]}(z) + [1/(4m)]C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\ & \stackrel{?}{=} \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! \ell!} B_0^{[2\ell]}(z) \rho^{2\ell} + \sum_{m=1}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m} \end{aligned} \quad (\text{H.2.35})$$

and

$$\begin{aligned}
& -\sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) + (1/4) C_0^{[2\ell+2]}] \rho^{2\ell+1} \\
& - \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,c}^{[2\ell]}(z) + [1/(4m)] C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell+2} \\
& \stackrel{?}{=} \sum_{m=1}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,s}^{[2\ell]}(z) \rho^{2\ell+m}.
\end{aligned} \tag{H.2.36}$$

Let us first work on question (2.35). Begin by setting

$$C_{1,s}^{[0]}(z) = B_0^{[0]}(z). \tag{H.2.37}$$

When this is done, (2.35) becomes

$$\begin{aligned}
& \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} C_{2,s}^{[2\ell]}(z) \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1) C_{m+1,s}^{[2\ell]}(z) + [1/(4m)] C_{m-1,s}^{[2\ell+2]}(z)\} \rho^{2\ell+m} \\
& \stackrel{?}{=} \cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} B_{1,c}^{[2\ell]}(z) \rho^{2\ell+1} \\
& + \sum_{m=2}^{\infty} \cos(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,c}^{[2\ell]}(z) \rho^{2\ell+m},
\end{aligned} \tag{H.2.38}$$

which is equivalent to the questions

$$2C_{2,s}^{[0]}(z) \stackrel{?}{=} B_{1,c}^{[0]}(z) \tag{H.2.39}$$

and

$$(m+1) C_{m+1,s}^{[0]}(z) + [1/(4m)] C_{m-1,s}^{[2]}(z) \stackrel{?}{=} B_{m,c}^{[0]}(z) \text{ for } m \geq 2. \tag{H.2.40}$$

The question (2.39) has the answer

$$C_{2,s}^{[0]}(z) = (1/2) B_{1,c}^{[0]}(z), \tag{H.2.41}$$

and (2.40) can be written in the recursive form

$$C_{m+1,s}^{[0]}(z) \stackrel{?}{=} [1/(m+1)] B_{m,c}^{[0]}(z) - \{1/[(4m)(m+1)]\} C_{m-1,s}^{[2]}(z). \tag{H.2.42}$$

We see from (2.37) and (2.41) that this recursion relation has a unique solution. Setting $m = 2$ in (2.42) gives the result

$$C_{3,s}^{[0]}(z) = (1/3) B_{2,c}^{[0]}(z) - \{1/[(8)(3)]\} C_{1,s}^{[2]}(z), \tag{H.2.43}$$

which, in view of (2.37), becomes

$$C_{3,s}^{[0]}(z) = (1/3)B_{2,c}^{[0]}(z) - (1/24)B_0^{[2]}(z). \quad (\text{H.2.44})$$

Setting $m = 3$ gives

$$C_{4,s}^{[0]}(z) = (1/4)B_{3,c}^{[0]}(z) - \{1/[(12)(4)]\}C_{2,s}^{[2]}(z), \quad (\text{H.2.45})$$

which, in view of (2.41), becomes

$$C_{4,s}^{[0]}(z) = (1/4)B_{3,c}^{[0]}(z) - (1/96)B_{1,c}^{[2]}(z). \quad (\text{H.2.46})$$

Setting $m = 4$ gives

$$C_{5,s}^{[0]}(z) = (1/5)B_{3,c}^{[0]}(z) - \{1/[(16)(5)]\}C_{3,s}^{[2]}(z), \quad (\text{H.2.47})$$

which, in view of (2.44), becomes

$$C_{5,s}^{[0]}(z) = (1/5)B_{4,c}^{[0]}(z) - (1/96)B_{2,c}^{[2]}(z) - (1/96)B_0^{[4]}(z). \quad (\text{H.2.48})$$

Evidently we are able to find all the $C_{m,s}^{[0]}(z)$ in terms of the $B_{m,c}^{[n]}(z)$, and therefore (2.35) can be satisfied.

Now turn to satisfying (2.36), which is equivalent to the question

$$\begin{aligned} & -\sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{2}{2^{2\ell} \ell! (\ell+1)!} [C_{2,c}^{[2\ell]}(z) + (1/4)C_0^{[2\ell+2]}] \rho^{2\ell+1} \\ & - \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} \{(m+1)C_{m+1,c}^{[2\ell]}(z) + [1/(4m)]C_{m-1,c}^{[2\ell+2]}(z)\} \rho^{2\ell+2} \\ & \stackrel{?}{=} \sin(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} B_{1,s}^{[2\ell]}(z) \rho^{2\ell+1} \\ & + \sum_{m=2}^{\infty} \sin(m\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{m!}{2^{2\ell} \ell! (\ell+m)!} B_{m,s}^{[2\ell]}(z) \rho^{2\ell+m}. \end{aligned} \quad (\text{H.2.49})$$

Upon equating like terms, we find the questions

$$2[C_{2,c}^{[0]}(z) + (1/4)C_0^{[2]}] \stackrel{?}{=} -B_{1,s}^{[0]}(z) \quad (\text{H.2.50})$$

and

$$(m+1)C_{m+1,c}^{[0]}(z) + [1/(4m)]C_{m-1,c}^{[2]}(z) \stackrel{?}{=} -B_{m,s}^{[0]}(z). \quad (\text{H.2.51})$$

We again make the stipulation (2.11) so that (2.50) has the answer

$$C_{2,c}^{[0]}(z) = -(1/2)B_{1,s}^{[0]}(z). \quad (\text{H.2.52})$$

Moreover, (2.51) is equivalent to the recursion relation

$$C_{m+1,c}^{[0]}(z) \stackrel{?}{=} -[1/(m+1)]B_{m,s}^{[0]}(z) - \{1/[(4m)(m+1)]\}C_{m-1,c}^{[2]}(z). \quad (\text{H.2.53})$$

We now add the further stipulation

$$C_{1,c}^{[0]}(z) = 0. \quad (\text{H.2.54})$$

In view of (2.52) and (2.54), the recursion relation (2.53) now has a unique solution. Setting $m = 2$ and using (2.54) give the result

$$C_{3,c}^{[0]}(z) = -(1/3)B_{2,s}^{[0]}(z). \quad (\text{H.2.55})$$

Next set $m = 3$ to find the result

$$C_{4,c}^{[0]}(z) = -(1/4)B_{3,s}^{[0]}(z) - \{1/[(12)(4)]\}C_{2,c}^{[2]}(z), \quad (\text{H.2.56})$$

which, in view of (2.52), becomes the relation

$$C_{4,c}^{[0]}(z) = -(1/4)B_{3,s}^{[0]}(z) + (1/96)B_{1,s}^{[2]}(z). \quad (\text{H.2.57})$$

Now set $m = 4$ to find the result

$$C_{5,c}^{[0]}(z) = -(1/5)B_{4,s}^{[0]}(z) - \{1/[(16)(5)]\}C_{3,c}^{[2]}(z), \quad (\text{H.2.58})$$

which, in view of (2.55), becomes the relation

$$C_{5,c}^{[0]}(z) = -(1/5)B_{4,s}^{[0]}(z) + (1/240)B_{2,s}^{[2]}(z). \quad (\text{H.2.59})$$

Set $m = 5$ to find the result

$$C_{6,c}^{[0]}(z) = -(1/6)B_{5,s}^{[0]}(z) - \{1/[(20)(6)]\}C_{4,c}^{[2]}(z), \quad (\text{H.2.60})$$

which, in view of (2.57), becomes the relation

$$C_{6,c}^{[0]}(z) = -(1/6)B_{5,s}^{[0]}(z) + (1/480)B_{3,s}^{[2]}(z) - (1/11520)B_{1,s}^{[4]}(z). \quad (\text{H.2.61})$$

Evidently we are able to find all the $C_{m,c}^{[0]}(z)$ in terms of the $B_{m,s}^{[n]}(z)$, and therefore (2.36) can be satisfied. Since both (2.35) and (2.36) have been satisfied, our goal (2.2) has been met.

We note that the stipulation (2.54) specifies that all terms of the form

$$\cos(\phi) \sum_{\ell=0}^{\infty} (-1)^{\ell} \frac{1}{2^{2\ell} \ell! (\ell+1)!} C_{1,c}^{[2\ell]}(z) \rho^{2\ell+1} \quad (\text{H.2.62})$$

are omitted from ψ . We have seen that these terms [and those of (2.33)] are not needed to meet the goal (2.2), and their omission simplifies the recursion relations that specify ψ in terms of χ .

H.3 Harmonic Functions in Two Variables and Their Associated Fields

It is also useful to have representations for harmonic functions in two variables. We will consider the two cases of harmonic functions in the variable pairs x, z and y, z .

H.3.1 Harmonic Functions in x, z

Suppose $\psi(x, z)$ is a harmonic function in the two variables x, z . That is, we have the relation

$$[(\partial_x)^2 + (\partial_z)^2]\psi(x, z) = 0. \quad (\text{H.3.1})$$

Decompose ψ into *even* and *odd* parts with respect to its x dependence,

$$\psi = \psi_{ev} + \psi_{od}. \quad (\text{H.3.2})$$

Then, because the operation $x \rightarrow -x$ commutes with $[(\partial_x)^2 + (\partial_z)^2]$, each separate part of ψ must be harmonic,

$$[(\partial_x)^2 + (\partial_z)^2]\psi_{ev}(x, z) = 0, \quad (\text{H.3.3})$$

$$[(\partial_x)^2 + (\partial_z)^2]\psi_{od}(x, z) = 0. \quad (\text{H.3.4})$$

Series Representation

For the even part let us make the Ansatz

$$\begin{aligned} \psi_{ev}(x, z) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n)!] x^{2n} E^{[2n]}(z) \\ &= E^{[0]}(z) - (1/2)x^2 E^{[2]}(z) + (1/24)x^4 E^{[4]}(z) + \dots, \end{aligned} \quad (\text{H.3.5})$$

where the functions $E^{[n]}(z)$ and the meaning of the $[n]$ notation are yet to be determined. For such a ψ_{ev} to be harmonic there must be the relation

$$\begin{aligned} 0 &= \nabla^2 \psi_{ev}(x, z) = \partial_x^2 \psi_{ev}(x, z) + \partial_z^2 \psi_{ev}(x, z) \\ &= [\partial_z^2 E^{[0]}(z) - E^{[2]}(z)] - (1/2)x^2 [\partial_z^2 E^{[2]}(z) - E^{[4]}(z)] + \dots, \end{aligned} \quad (\text{H.3.6})$$

from which we conclude that

$$E^{[n+2]}(z) = \partial_z^2 E^{[n]}(z). \quad (\text{H.3.7})$$

Similarly, for the odd part we make the Ansatz

$$\begin{aligned} \psi_{od}(x, z) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n+1)!] x^{2n+1} O^{[2n]}(z) \\ &= x O^{[0]}(z) - (1/6)x^3 O^{[2]}(z) + (1/120)x^5 O^{[4]}(z) + \dots. \end{aligned} \quad (\text{H.3.8})$$

Now the harmonic requirement yields the result

$$\begin{aligned} 0 &= \nabla^2 \psi_{od}(x, z) = \partial_x^2 \psi_{od}(x, z) + \partial_z^2 \psi_{od}(x, z) \\ &= x[\partial_z^2 O^{[0]}(z) - O^{[2]}(z)] - (1/6)x^3 [\partial_z^2 O^{[2]}(z) - O^{[4]}(z)] + \dots, \end{aligned} \quad (\text{H.3.9})$$

from which we conclude that

$$O^{[n+2]}(z) = \partial_z^2 O^{[n]}(z). \quad (\text{H.3.10})$$

We see that in both cases the $[n]$ notation is the usual one. Thus, ψ is specified by the two functions $E^{[0]}(z)$ and $O^{[0]}(z)$. These functions may, in principle, be chosen arbitrarily. Often they are required to go to zero as $|z| \rightarrow \infty$.

Let us compute the “fields” associated with ψ_{ev} and ψ_{od} , call them \mathbf{B}^{ev} and \mathbf{B}^{od} . We find the results

$$B_x^{ev} = \partial_x \psi_{ev} = -xE^{[2]}(z) + (1/6)x^3E^{[4]}(z) - (1/120)x^5E^{[6]}(z) + \dots, \quad (\text{H.3.11})$$

$$B_y^{ev} = \partial_y \psi_{ev} = 0, \quad (\text{H.3.12})$$

$$B_z^{ev} = \partial_z \psi_{ev} = E^{[1]}(z) - (1/2)x^2E^{[3]}(z) + (1/24)x^4E^{[5]}(z) + \dots; \quad (\text{H.3.13})$$

$$B_x^{od} = \partial_x \psi_{od} = O^{[0]}(z) - (1/2)x^2O^{[2]}(z) + (1/24)x^4O^{[4]}(z) + \dots, \quad (\text{H.3.14})$$

$$B_y^{od} = \partial_y \psi_{od} = 0, \quad (\text{H.3.15})$$

$$B_z^{od} = \partial_z \psi_{od} = xO^{[1]}(z) - (1/6)x^3O^{[3]}(z) + (1/120)x^5O^{[5]}(z) + \dots. \quad (\text{H.3.16})$$

Note that, because the fields are the gradients of harmonic functions, the Cartesian components of \mathbf{B}^{ev} and \mathbf{B}^{od} must also be harmonic functions.

Explicit Construction from Analytic Functions

As is well known, there is an intimate connection between harmonic functions in two variables and analytic functions of a complex variable. This connection facilitates obtaining closed-form expressions for harmonic functions rather than dealing solely with power series. Let w be a complex variable written in the form

$$w = u + iv. \quad (\text{H.3.17})$$

Suppose $f(w)$ is a *real-analytic* function. That is, f is defined, analytic, and real for w on the real axis. Such a function can be extended into the complex plane by analytic continuation. For complex arguments, decompose f into real and imaginary parts by writing

$$f(u + iv) = f_r(u, v) + if_i(u, v). \quad (\text{H.3.18})$$

For f_r and f_i we find, by the chain rule, the result

$$[(\partial_u)^2 + (\partial_v)^2][f_r(u, v) + if_i(u, v)] = f''(u + iv)(1 + i^2) = 0. \quad (\text{H.3.19})$$

Thus, upon equating real and imaginary parts in (H.19), we see that both f_r and f_i are harmonic.

Let us expand f as a power series in the quantity iv . From Taylor’s theorem we find the result

$$\begin{aligned} f(u + iv) &= \sum_{n=0}^{\infty} f^{[n]}(u)(iv)^n/n! \\ &= \sum_{n=0}^{\infty} (-1)^n [1/(2n)!] v^{2n} f^{[2n]}(u) \\ &\quad + i \sum_{n=0}^{\infty} (-1)^n [1/(2n+1)!] v^{2n+1} f^{[2n+1]}(u). \end{aligned} \quad (\text{H.3.20})$$

Upon comparing (H.18) and (H.20), we see that there are the relations

$$\begin{aligned} f_r(u, v) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n)!] v^{2n} f^{[2n]}(u) \\ &= f^{[0]}(u) - (1/2)v^2 f^{[2]}(u) + (1/24)v^4 f^{[4]}(u) + \dots \end{aligned} \quad (\text{H.3.21})$$

and

$$\begin{aligned} f_i(u, v) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n+1)!] v^{2n+1} f^{[2n+1]}(u) \\ &= vf^{[1]}(u) - (1/6)v^3 f^{[3]}(u) + (1/120)v^5 f^{[5]}(u) + \dots . \end{aligned} \quad (\text{H.3.22})$$

Observe the resemblance between the pair (H.5) and (H.8) and the pair (H.21) and (H.22). Upon making the identifications

$$z \leftrightarrow u, \quad (\text{H.3.23})$$

$$x \leftrightarrow v, \quad (\text{H.3.24})$$

$$E^{[0]}(z) \leftrightarrow f^{[0]}(u), \quad (\text{H.3.25})$$

and

$$O^{[0]}(z) \leftrightarrow f^{[1]}(u), \quad (\text{H.3.26})$$

we see that these two pairs are the same. Thus, with these identifications, we have the relations

$$\psi_{ev}(x, z) = f_r(z, x), \quad (\text{H.3.27})$$

$$\psi_{od}(x, z) = f_i(z, x). \quad (\text{H.3.28})$$

Note that the identifications (H.25) and (H.26) require the restrictive relation

$$O^{[0]} = E^{[1]}. \quad (\text{H.3.29})$$

Of course, in general, ψ_{ev} and ψ_{od} need not be the real and imaginary parts of the *same* real-analytic function.

Since f_r and f_i are the real and imaginary parts of the analytic function f , they must satisfy the Cauchy-Riemann relations

$$\partial_z f_r = \partial_x f_i, \quad (\text{H.3.30})$$

$$\partial_x f_r = -\partial_z f_i. \quad (\text{H.3.31})$$

Consequently, if ψ_{ev} and ψ_{od} are the real and imaginary parts of the *same* real-analytic function, we have the relations

$$B_x^{ev} = \partial_x \psi_{ev} = \partial_x f_r = -\partial_z f_i = -\partial_z \psi_{od} = -B_z^{od}, \quad (\text{H.3.32})$$

$$B_z^{ev} = \partial_z \psi_{ev} = \partial_z f_r = \partial_x f_i = \partial_x \psi_{od} = B_x^{od}. \quad (\text{H.3.33})$$

These relations also follow from the representations (H.11) through (H.16) and the relation (H.29).

There is another application of the relation between analytic and harmonic functions that is useful. We will formulate and apply it in the subsection after the next where we discuss harmonic functions in the y, z variables. From that discussion the reader can easily infer the analogous results for harmonic functions in the variables x, z .

H.3.2 Harmonic Functions in y, z

The case of harmonic functions in the variables y, z is analogous to the x, z case. It is only necessary to make the substitution $x \rightarrow y$ in the relations found above.

Suppose $\psi(y, z)$ is a harmonic function in the two variables y, z . That is, we have the relation

$$[(\partial_y)^2 + (\partial_z)^2]\psi(y, z) = 0. \quad (\text{H.3.34})$$

Decompose ψ into *even* and *odd* parts with respect to its y dependence,

$$\psi = \psi_{ev} + \psi_{od}. \quad (\text{H.3.35})$$

Then, because the operation $y \rightarrow -y$ commutes with $[(\partial_y)^2 + (\partial_z)^2]$, each separate part of ψ must be harmonic,

$$[(\partial_y)^2 + (\partial_z)^2]\psi_{ev}(y, z) = 0, \quad (\text{H.3.36})$$

$$[(\partial_y)^2 + (\partial_z)^2]\psi_{od}(y, z) = 0. \quad (\text{H.3.37})$$

Series Representation

For the even part there is the representation

$$\begin{aligned} \psi_{ev}(y, z) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n)!] y^{2n} E^{[2n]}(z) \\ &= E^{[0]}(z) - (1/2)y^2 E^{[2]}(z) + (1/24)y^4 E^{[4]}(z) + \dots . \end{aligned} \quad (\text{H.3.38})$$

For the odd part there is the representation

$$\begin{aligned} \psi_{od}(y, z) &= \sum_{n=0}^{\infty} (-1)^n [1/(2n+1)!] y^{2n+1} O^{[2n]}(z) \\ &= yO^{[0]}(z) - (1/6)y^3 O^{[2]}(z) + (1/120)y^5 O^{[4]}(z) + \dots . \end{aligned} \quad (\text{H.3.39})$$

Thus, ψ is specified by the two functions $E^{[0]}(z)$ and $O^{[0]}(z)$. These functions may, in principle, be chosen arbitrarily. Often they are required to go to zero as $|z| \rightarrow \infty$.

For the “fields” associated with ψ_{ev} and ψ_{od} , call them \mathbf{B}^{ev} and \mathbf{B}^{od} , there are the results

$$B_x^{ev} = \partial_x \psi_{ev} = 0, \quad (\text{H.3.40})$$

$$B_y^{ev} = \partial_y \psi_{ev} = -yE^{[2]}(z) + (1/6)y^3 E^{[4]}(z) - (1/120)y^5 E^{[6]}(z) + \dots , \quad (\text{H.3.41})$$

$$B_z^{ev} = \partial_z \psi_{ev} = E^{[1]}(z) - (1/2)y^2 E^{[3]}(z) + (1/24)y^4 E^{[5]}(z) + \dots ; \quad (\text{H.3.42})$$

$$B_x^{od} = \partial_x \psi_{od} = 0, \quad (\text{H.3.43})$$

$$B_y^{od} = \partial_y \psi_{od} = O^{[0]}(z) - (1/2)y^2 O^{[2]}(z) + (1/24)y^4 O^{[4]}(z) + \dots , \quad (\text{H.3.44})$$

$$B_z^{od} = \partial_z \psi_{od} = yO^{[1]}(z) - (1/6)y^3 O^{[3]}(z) + (1/120)y^5 O^{[5]}(z) + \dots . \quad (\text{H.3.45})$$

Again, because the fields are the gradients of harmonic functions, the Cartesian components of \mathbf{B}^{ev} and \mathbf{B}^{od} must also be harmonic functions.

Explicit Construction from Analytic Functions

Suppose the functions $\psi_{ev}(y, z)$ and $\psi_{od}(y, z)$ are related to the real and imaginary parts of a real-analytic function f . Make the identifications

$$z \leftrightarrow u, \quad (\text{H.3.46})$$

$$y \leftrightarrow v, \quad (\text{H.3.47})$$

$$E^{[0]}(z) \leftrightarrow f^{[0]}(u), \quad (\text{H.3.48})$$

and

$$O^{[0]}(z) \leftrightarrow f^{[1]}(u). \quad (\text{H.3.49})$$

With these identifications, we have the relations

$$\psi_{ev}(y, z) = f_r(y, x), \quad (\text{H.3.50})$$

$$\psi_{od}(y, z) = f_i(y, x). \quad (\text{H.3.51})$$

The identifications (H.48) and (H.49) again require the restrictive relation

$$O^{[0]} = E^{[1]}. \quad (\text{H.3.52})$$

But, since f_r and f_i are the real and imaginary parts of the analytic function f , they must satisfy the Cauchy-Riemann relations

$$\partial_z f_r = \partial_x f_i, \quad (\text{H.3.53})$$

$$\partial_x f_r = -\partial_z f_i. \quad (\text{H.3.54})$$

Thus, we have the relations

$$B_y^{ev} = -B_z^{od}, \quad (\text{H.3.55})$$

$$B_z^{ev} = B_y^{od}. \quad (\text{H.3.56})$$

These relations also follow from the representations (H.40) through (H.45) and the relation (H.52).

H.3.3 More About $\mathbf{B}^{od}(y, z)$ and Another Application of Analytic Function Theory

To keep the promise made in Subsection H.3.1, we now discuss another application of the relation between analytic and harmonic functions. Consider the field $\mathbf{B}^{od}(y, z)$ given by (H.43) through (H.45). It is a candidate magnetic field for an infinitely wide (in x) parallel-faced dipole, see Figures 1.6.1 and 1.6.2, with symmetry about the midplane $y = 0$. Given a real-analytic function $f(w)$, define a related function $g(w)$ by the rule

$$g(w) = f^{[1]}(w). \quad (\text{H.3.57})$$

Evidently g will also be real analytic. As before, employ the relation (H.17) and decompose g into real and imaginary parts by writing

$$g(u + iv) = g_r(u, v) + g_i(u, v). \quad (\text{H.3.58})$$

In analogy to what was done for f , expand $g(u + iv)$ as a Taylor series in iv to get the relation

$$\begin{aligned} g(u + iv) &= \sum_{n=0}^{\infty} g^{[n]}(u)(iv)^n/n! \\ &= [g^{[0]}(u) - (1/2)v^2g^{[2]}(u) + (1/24)v^4g^{[4]}(u) + \dots] \\ &\quad + i[vg^{[1]}(u) - (1/6)v^3g^{[3]}(u) + (1/120)v^5g^{[5]}(u) + \dots]. \end{aligned} \quad (\text{H.3.59})$$

We see that

$$g_r(u, v) = g^{[0]}(u) - (1/2)v^2g^{[2]}(u) + (1/24)v^4g^{[4]}(u) + \dots \quad (\text{H.3.60})$$

and

$$g_i(u, v) = vg^{[1]}(u) - (1/6)v^3g^{[3]}(u) + (1/120)v^5g^{[5]}(u) + \dots. \quad (\text{H.3.61})$$

Make the identifications (H.46) and (H.47) and compare the series on the right side of (H.59) with the series (H.44) and (H.45) for B_y^{od} and B_z^{od} . We see that they are analogous if we make the identification

$$O^{[0]} = g^{[0]}. \quad (\text{H.3.62})$$

In particular, we have the relations

$$B_y^{od}(y, z) = g_r(z, y) = g^{[0]}(z) - (1/2)y^2g^{[2]}(z) + (1/24)y^4g^{[4]}(z) + \dots, \quad (\text{H.3.63})$$

$$B_z^{od}(y, z) = g_i(z, y) = yg^{[1]}(z) - (1/6)y^3g^{[3]}(z) + (1/120)y^5g^{[5]}(z) + \dots. \quad (\text{H.3.64})$$

The Cauchy-Riemann relations again hold for g_r and g_i ,

$$\partial_u g_r(u, v) = \partial_v g_i(u, v), \quad (\text{H.3.65})$$

$$\partial_v g_r(u, v) = -\partial_u g_i(u, v). \quad (\text{H.3.66})$$

In this case, because of (H.63) and (H.64), they have the consequence

$$\partial_z B_y^{od}(y, z) = \partial_z g_r(z, y) = \partial_y g_i(y, z) = \partial_y B_z^{od}(y, z), \quad (\text{H.3.67})$$

$$\partial_y B_y^{od}(y, z) = \partial_y g_r(z, y) = -\partial_z g_i(y, z) = -\partial_z B_z^{od}(y, z). \quad (\text{H.3.68})$$

These relations can also be verified directly from the representations (H.63) and (H.64).

We already know that $\nabla \times \mathbf{B}^{od} = 0$ because \mathbf{B}^{od} is the gradient of a scalar field. See (H.43) through (H.45). What can be said about $\nabla \cdot \mathbf{B}^{od}$? From the second Cauchy-Riemann relation (H.68) we see that

$$\nabla \cdot \mathbf{B}^{od} = \partial_y B_y^{od} + \partial_z B_z^{od} = 0, \quad (\text{H.3.69})$$

as expected and required.

Also observe that, in complex notation, the relations (H.58), (H.63), and (H.64) can be expressed in the compact form

$$B_y^{odd}(y, z) + iB_z^{odd}(y, z) = g(z + iy). \quad (\text{H.3.70})$$

Thus, in this application, the selection of one real-analytic function $g(w)$ specifies both components of the field for an infinitely wide parallel-faced dipole. Indeed, the application is even broader. It would also apply to an infinitely wide parallel-faced wiggler. In this case $g(u)$ would be roughly oscillatory in u . All that is required in both applications is that $g(u)$ vanish as $u \rightarrow \pm\infty$, in which case $\mathbf{B}^{od}(y, z)$ will vanish as $z \rightarrow \pm\infty$. Suppose, for example, that we set

$$g(u) = B \text{ bump}(u, c\ell, L) \quad (\text{H.3.71})$$

where $\text{bump}(u, c\ell, L)$ is one of the bump functions defined in Section 11.11. These functions are real analytic, and therefore the representation (H.70) can be implemented.

There is one last item to be discussed. Namely, it would be good to have a vector potential \mathbf{A}^{od} from which \mathbf{B}^{od} could be derived. Consider the following Ansatz,

$$A_x^{od}(x, y, z) = 0, \quad (\text{H.3.72})$$

$$A_y^{od}(x, y, z) = xB_z^{od}(y, z), \quad (\text{H.3.73})$$

$$A_z^{od}(x, y, z) = -xB_y^{od}(y, z). \quad (\text{H.3.74})$$

(Evidently, this vector potential is horizontal free.) It is easily verified that

$$(\nabla \times \mathbf{A}^{od})_x = \partial_y A_z^{od} - \partial_z A_y^{od} = x[-\partial_y B_z^{od}(y, z) - \partial_z B_y^{od}(y, z)] = 0 = B_x^{od}(y, z), \quad (\text{H.3.75})$$

$$(\nabla \times \mathbf{A}^{od})_y = \partial_z A_x^{od} - \partial_x A_z^{od} = -\partial_x A_z^{od} = B_y^{od}(y, z), \quad (\text{H.3.76})$$

$$(\nabla \times \mathbf{A}^{od})_z = \partial_x A_y^{od} - \partial_y A_x^{od} = \partial_x A_y^{od} = B_z^{od}(y, z), \quad (\text{H.3.77})$$

as desired. Also, we find that

$$\nabla \cdot \mathbf{A}^{od} = \partial_y A_z^{od} + \partial_z A_y^{od} = x[\partial_y B_z^{od}(y, z) - \partial_z B_y^{od}(y, z)] = 0. \quad (\text{H.3.78})$$

Here we have used the first Cauchy-Riemann relation (H.67). Thus, in addition to being horizontal free, the vector potential $\mathbf{A}^{od}(x, y, z)$ is in the Coulomb gauge. Since $\mathbf{A}^{od}(x, y, z)$ is in the Coulomb gauge, it follows, as is also readily verified from (H.72) through (H.74), that its Cartesian components are harmonic functions.

Note also, from its definition (H.72) through (H.74), that $\mathbf{A}^{od}(x, y, z)$ vanishes as $z \rightarrow \pm\infty$ if $\mathbf{B}^{od}(y, z)$ does so. This feature is desirable because we would like canonical and mechanical momenta to be equal in field-free regions. We also note the convenient feature that $A_y^{od}(x, y, z)$ vanishes in the midplane,

$$A_y^{od}(x, y = 0, z) = xB_z^{od}(y = 0, z) = 0. \quad (\text{H.3.79})$$

See (H.64). Thus, for the design orbit which lies in the midplane, there is no difference between mechanical and canonical for the x and y components of the momentum. Finally, as in the case of the vector potentials found in Exercises 13.3.4 and 13.5.5, the vector potential is primarily in the z direction except in the fringe-field regions.

Bibliography

- [1] The Ansatz (H.72) through (H.74) for $\mathbf{A}^{od}(x, y, z)$ is due to Peter Walstrom.

Appendix I

Poisson Bracket Relations

I.1 Poisson Brackets

$$z_a J_{aa'} z_{a'} = 0 \quad (\text{I.1.1})$$

$$[f_m, z_a] = (\partial_b f_m) J_{bb'} \partial_{b'} z_a = (\partial_b f_m) J_{bb'} \delta_{b'a} = (\partial_b f_m) J_{ba} \quad (\text{I.1.2})$$

$$\begin{aligned} [f_m, z_a] J_{aa'} z_{a'} &= (\partial_b f_m) J_{ba} J_{aa'} z_{a'} = -(\partial_b f_m) \delta_{ba'} z_{a'} \\ &= -z_b (\partial_b f_m) = -m f_m \end{aligned} \quad (\text{I.1.3})$$

$$[f_m, [f_n, z_a]] = [f_m, (\partial_b f_n)] J_{ba} = (\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) J_{ba} \quad (\text{I.1.4})$$

$$\begin{aligned} [f_m, [f_n, z_a]] J_{aa'} z_{a'} &= (\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) J_{ba} J_{aa'} z_{a'} = -(\partial_c f_m) J_{cc'} (\partial_{c'} \partial_b f_n) \delta_{ba'} z_{a'} \\ &= -(\partial_c f_m) J_{cc'} z_b (\partial_{c'} \partial_b f_n) \end{aligned} \quad (\text{I.1.5})$$

$$z_b (\partial_{c'} \partial_b f_n) = \partial_{c'} (z_b \partial_b f_n) - \delta_{c'b} (\partial_b f_n) = (n-1) (\partial_{c'} f_n) \quad (\text{I.1.6})$$

$$[f_m, [f_n, z_a]] J_{aa'} z_{a'} = -(n-1) (\partial_c f_m) J_{cc'} (\partial_{c'} f_n) = -(n-1) [f_m, f_n] \quad (\text{I.1.7})$$

$$\begin{aligned} [f_\ell, [f_m, [f_n, z_a]]] &= [f_\ell, (\partial_c f_m) (\partial_{c'} \partial_b f_n)] J_{cc'} J_{ba} \\ &= (\partial_d f_\ell) \{ \partial_{d'} [(\partial_c f_m) (\partial_{c'} \partial_b f_n)] \} J_{dd'} J_{cc'} J_{ba} \\ &= (\partial_d f_\ell) [(\partial_{d'} \partial_c f_m) (\partial_{c'} \partial_b f_n) + (\partial_c f_m) (\partial_{d'} \partial_{c'} \partial_b f_n)] J_{dd'} J_{cc'} J_{ba} \end{aligned} \quad (\text{I.1.8})$$

$$\begin{aligned}
& [f_\ell, [f_m, [f_n, z_a]]] J_{aa'} z_{a'} \\
= & (\partial_d f_\ell) (\partial_{d'} \partial_c f_m) (\partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} J_{ba} J_{aa'} z_{a'} + \\
& (\partial_d f_\ell) (\partial_c f_m) (\partial_{d'} \partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} J_{ba} J_{aa'} z_{a'} \\
= & -(\partial_d f_\ell) (\partial_{d'} \partial_c f_m) (\partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} \delta_{ba'} z_{a'} - \\
& (\partial_d f_\ell) (\partial_c f_m) (\partial_{d'} \partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} \delta_{ba'} z_{a'} \\
= & -(\partial_d f_\ell) (\partial_{d'} \partial_c f_m) z_b (\partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} - \\
& (\partial_d f_\ell) (\partial_c f_m) z_b (\partial_{d'} \partial_{c'} \partial_b f_n) J_{dd'} J_{cc'}
\end{aligned} \tag{I.1.9}$$

$$\begin{aligned}
& \partial_{d'} \partial_{c'} (z_b \partial_b f_n) = \partial_{d'} [\delta_{c'b} \partial_b f_n + z_b \partial_{c'} \partial_b f_n] \\
= & \partial_{d'} [\partial_{c'} f_n + z_b \partial_{c'} \partial_b f_n] \\
= & \partial_{d'} \partial_{c'} f_n + \partial_{d'} (z_b \partial_{c'} \partial_b f_n) \\
= & \partial_{d'} \partial_{c'} f_n + \delta_{d'b} \partial_{c'} \partial_b f_n + z_b \partial_{d'} \partial_{c'} \partial_b f_n \\
= & \partial_{d'} \partial_{c'} f_n + \partial_{c'} \partial_{d'} f_n + z_b \partial_{d'} \partial_{c'} \partial_b f_n \\
= & 2\partial_{d'} \partial_{c'} f_n + z_b \partial_{d'} \partial_{c'} \partial_b f_n
\end{aligned} \tag{I.1.10}$$

$$z_b \partial_{d'} \partial_{c'} \partial_b f_n = (n-2) \partial_{d'} \partial_{c'} f_n \tag{I.1.11}$$

$$\begin{aligned}
& [f_\ell, [f_m, [f_n, z_a]]] J_{aa'} z_{a'} \\
= & -(\partial_d f_\ell) (\partial_{d'} \partial_c f_m) z_b (\partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} - \\
& (\partial_d f_\ell) (\partial_c f_m) z_b (\partial_{d'} \partial_{c'} \partial_b f_n) J_{dd'} J_{cc'} \\
= & -(\partial_d f_\ell) (\partial_{d'} \partial_c f_m) (n-1) (\partial_{c'} f_n) J_{dd'} J_{cc'} - \\
& (\partial_d f_\ell) (\partial_c f_m) (n-2) (\partial_{d'} \partial_{c'} f_n) J_{dd'} J_{cc'} \\
= & -(n-1) (\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_c f_m) J_{cc'} (\partial_{c'} f_n) - \\
& (n-2) (\partial_d f_\ell) (\partial_c f_m) J_{dd'} J_{cc'} (\partial_{d'} \partial_{c'} f_n) \\
= & -(n-1) (\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_c f_m) J_{cc'} (\partial_{c'} f_n) + \\
& (n-2) (\partial_d f_\ell) (\partial_c f_m) J_{dd'} J_{c'c} (\partial_{d'} \partial_{c'} f_n) \\
= & -(n-1) (\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_c f_m) J_{cc'} (\partial_{c'} f_n) + \\
& (n-2) (\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_{c'} f_n) J_{c'c} (\partial_c f_m) \\
= & -(\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_c f_m) J_{cc'} (\partial_{c'} f_n)
\end{aligned} \tag{I.1.12}$$

$$\begin{aligned}
& [f_k, [f_\ell, [f_m, [f_n, z_a]]]] = [f_k, (\partial_d f_\ell) \{(\partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n) + (\partial_c f_m)(\partial_{d'} \partial_{c'} \partial_b f_n)\}] J_{dd'} J_{cc'} J_{ba} \\
= & [f_k, (\partial_d f_\ell)(\partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n)] J_{dd'} J_{cc'} J_{ba} + [f_k, (\partial_d f_\ell)(\partial_c f_m)(\partial_{d'} \partial_{c'} \partial_b f_n)] J_{dd'} J_{cc'} J_{ba}
\end{aligned} \tag{I.1.13}$$

$$\begin{aligned}
& [f_k, (\partial_d f_\ell)(\partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n)] = (\partial_e f_k) J_{ee'} \{ \partial_{e'} [(\partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n)] \} \\
= & (\partial_e f_k) J_{ee'} [(\partial_{e'} \partial_{d'} \partial_c f_m)(\partial_{c'} \partial_b f_n) +]
\end{aligned} \tag{I.1.14}$$

$$\begin{aligned}
& [f_k, (\partial_d f_\ell)(\partial_c f_m)(\partial_{d'} \partial_{c'} \partial_b f_n)] \\
= &
\end{aligned} \tag{I.1.15}$$

I.2 Preparatory Results

$$(z, Jz) = z_a J_{aa'} z_{a'} = 0 \tag{I.2.1}$$

$$(: f_n : z, Jz) = ([f_n, z], Jz) = [f_n, z_a] J_{aa'} z_{a'} = -n f_n \tag{I.2.2}$$

$$(: f_m :: f_n : z, Jz) = ([f_m, [f_n, z]], Jz) = [f_m, [f_n, z_a]] J_{aa'} z_{a'} = -(n-1) [f_m, f_n] \tag{I.2.3}$$

$$\begin{aligned}
(: f_\ell :: f_m :: f_n : z, Jz) &= ([f_\ell, [f_m, [f_n, z]]], Jz) = [f_\ell, [f_m, [f_n, z_a]]] J_{aa'} z_{a'} \\
&= (\partial_d f_\ell) J_{dd'} (\partial_{d'} \partial_c f_m) J_{cc'} (\partial_{c'} f_n) \\
&= (\partial f_\ell, JS(f_m) J \partial f_n).
\end{aligned} \tag{I.2.4}$$

Here $S(f_m)$ is the Hessian of f_m ,

$$S_{ab}(f_m) = \partial_a \partial_b f_m. \tag{I.2.5}$$

$$\begin{aligned}
& [f_k, [f_\ell, [f_m, [f_n, z_a]]]] J_{aa'} z_{a'} =
\end{aligned} \tag{I.2.6}$$

I.3 Application

Suppose $t^i = 0$ and $t^f = 1$ in (1.2.57) and (6.6.57). Suppose also that we confine our interest to the case where only h_3 through h_6 are possibly nonzero. Our task will be to find the Poincaré generating function F_+ corresponding to \mathcal{M} .

From (6.6.57) we have the result

$$F(z) = - \sum_m (m-2) h_m(z). \quad (\text{I.3.1})$$

And from (6.6.37) we know that

$$F_+(z) = [F + (Z, Jz)]/2. \quad (\text{I.3.2})$$

Also, by definition,

$$Z = \mathcal{M}z. \quad (\text{I.3.3})$$

Upon combining () through () we conclude that F_+ is given by the relation

$$F_+(z) = (1/2)[(\mathcal{M}z, Jz) - \sum_m (m-2) h_m(z)]. \quad (\text{I.3.4})$$

Our task will be to work out the implications of (). What we will find will be a homogeneous polynomial expansion for F_+ .

Evidently the hard part is to find an expansion for $(\mathcal{M}z, Jz)$. Let us write

$$\mathcal{M} = \exp(- : H :) = I - : H : + : H :^2 / 2! - : H :^3 / 3! + : H :^4 / 4! + \dots \quad (\text{I.3.5})$$

where the terms retained are sufficient to compute F_+ through terms of degree 6. Let $Z^{(n)}$ be the contribution to Z made by the term $: -H :^n / n!$ in \mathcal{M} and let $Y^{(n)}$ be the contribution that it makes to F_+ . That is, we make the definitions

$$Z^{(n)} = [: -H :^n / n!] z \quad (\text{I.3.6})$$

and

$$Y^{(n)}(z) = (Z^{(n)}, Jz) = ([: -H :^n / n!] z, Jz). \quad (\text{I.3.7})$$

From the relations listed in Section 23.1 above, we easily find the results

$$Y^{(0)}(z) = (Z^{(0)}, Jz) = (z, Jz) = 0, \quad (\text{I.3.8})$$

$$\begin{aligned} Y^{(1)}(z) &= (Z^{(1)}, Jz) = (: -H : z, Jz) = - \sum_m (: h_m : z, Jz) \\ &= \sum_m m h_m(z). \end{aligned} \quad (\text{I.3.9})$$

It follows from the work done so far that

$$\begin{aligned} F_+(z) &= (1/2)[(\mathcal{M}z, Jz) - \sum_m (m-2) h_m(z)] \\ &= (1/2)\{[\sum_m m h_m(z)] - [\sum_m (m-2) h_m(z)] + \dots\} = H(z) + \dots. \end{aligned} \quad (\text{I.3.10})$$

As already stated, we ultimately desire to have an expansion of F_+ in homogeneous polynomials. Let F_+^m denote the term in F_+ that is homogeneous of degree n . Then, based on the work done so far, we have the result

$$F_+^m = h_m + \dots . \quad (\text{I.3.11})$$

The next term we need is $Y^{(2)}(z)$. We have the result

$$Y^{(2)}(z) = (Z^{(2)}, Jz) = (: H :^2 z, Jz)/2!. \quad (\text{I.3.12})$$

The term $: H :^2$ has the expansion

$$\begin{aligned} : H :^2 &= \sum_m \sum_n : h_m :: h_n : \\ &= : h_3 :^2 + : h_3 :: h_4 : + : h_4 :: h_3 : \\ &\quad + : h_4 :^2 + : h_3 :: h_5 : + : h_5 :: h_3 : + \dots \end{aligned} \quad (\text{I.3.13})$$

where we have displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(2)}$ the result

$$\begin{aligned} Y^{(2)}(z) &= (1/2!)[(: h_3 :^2 z, Jz) \\ &\quad + (: h_3 :: h_4 : z, Jz) + (: h_4 :: h_3 : z, Jz) \\ &\quad + (: h_4 :^2 z, Jz) \\ &\quad + (: h_3 :: h_5 : z, Jz) + (: h_5 :: h_3 : z, Jz)] \\ &= (1/2!)([h_3, h_3] + [h_4, h_4] \\ &\quad + [h_3, h_4] + [h_4, h_3] + [h_3, h_5] + [h_5, h_3]) \\ &= (1/2)([h_3, h_4] + [h_3, h_5]). \end{aligned} \quad (\text{I.3.14})$$

We are now able to conclude that

$$F_+^3 = h_3, \quad (\text{I.3.15})$$

$$F_+^4 = h_4, \quad (\text{I.3.16})$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + \dots, \quad (\text{I.3.17})$$

$$F_+^6 = h_6 + (1/4)[h_3, h_5] + \dots. \quad (\text{I.3.18})$$

Let us move on to the term

$$Y^{(3)}(z) = (Z^{(3)}, Jz) = -(: H :^3 z, Jz)/3!. \quad (\text{I.3.19})$$

The term $: H :^3$ has the expansion

$$\begin{aligned} : H :^3 &= \sum_{\ell} \sum_m \sum_n : h_{\ell} :: h_m :: h_n : \\ &= : h_3 :^3 + : h_3 :^2 : h_4 : + : h_3 :: h_4 :: h_3 : + : h_4 :: h_3 :^2 + \dots \end{aligned} \quad (\text{I.3.20})$$

where we have displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(3)}$ the result

$$\begin{aligned} Y^{(3)}(z) &= (1/2!)[(: h_3 :^3 z, Jz) \\ &\quad + (: h_3 :^2 : h_4 : z, Jz) + (: h_3 :: h_4 :: h_3; z, Jz) + (: h_4 :: h_3 :^2 z, Jz)] \\ &= \end{aligned} \quad (\text{I.3.21})$$

We are now able to conclude that

$$F_+^3 = h_3, \quad (\text{I.3.22})$$

$$F_+^4 = h_4, \quad (\text{I.3.23})$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + (\partial h_3, JS(h_3)J\partial h_3) \quad (\text{I.3.24})$$

$$F_+^6 = h_6 + [h_3, h_5] + (\partial h_3, JS(h_3)J\partial h_4) + (\partial h_3, JS(h_4)J\partial h_3) + \dots \quad (\text{I.3.25})$$

Finally, as we will see, we need the term

$$Y^{(4)}(z) = (Z^{(4)}, Jz) = -(: H :^4 z, Jz)/3!. \quad (\text{I.3.26})$$

The term $: H :^4$ has the expansion

$$\begin{aligned} : H :^4 &= \sum_k \sum_\ell \sum_m \sum_n : h_k :: h_\ell :: h_m :: h_n : \\ &= : h_3 :^4 + \dots \end{aligned} \quad (\text{I.3.27})$$

where we have again displayed only the terms that will contribute to the F_+^m for $m \leq 6$. Correspondingly, we find for $Y^{(3)}$ the result

$$Y^{(4)}(z) = (1/2!)[(: h_3 :^4 z, Jz) = \quad (\text{I.3.28})$$

We are now able to conclude that

$$F_+^3 = h_3, \quad (\text{I.3.29})$$

$$F_+^4 = h_4, \quad (\text{I.3.30})$$

$$F_+^5 = h_5 + (1/4)[h_3, h_4] + (\partial h_3, JS(h_3)J\partial h_3), \quad (\text{I.3.31})$$

$$F_+^6 = h_6 + [h_3, h_5] + (\partial h_3, JS(h_3)J\partial h_4) + (\partial h_3, JS(h_4)J\partial h_3) + . \quad (\text{I.3.32})$$

Appendix J

Feigenbaum Cascade Denied/Achieved

Section 1.2.1 described Feigenbaum infinite period doubling cascades and mentioned that, for some maps in some parameter ranges, period doubling cascades begin but do not continue to completion. The purpose of this appendix is to provide a simple example of both incomplete and complete cascades.

J.1 Simple Map and Its Initial Bifurcations

Consider the simple one-dimensional map given by the relation

$$x_{n+1} = \mathcal{M}x_n = f(a, b; x_n) = a + bx_n/[1 + (x_n)^2] \quad (\text{J.1.1})$$

where a and b are parameters. Figure 1.1 shows the curves $y = f(a, b; x)$ for $b = 11.5$ and selected values of a . As is evident from (1.1) and from the figure, these curves are all vertical displacements of each other.

Also shown is the line $y = x$. Any intersection of the line $y = x$ and the curve $y = f(a, b; x)$ corresponds to a fixed point of \mathcal{M} . Observe that for sufficiently negative values of a there is only one intersection, and hence only one fixed point. This is the fixed point whose path is shown as a function of a in the lower left portion of the bifurcation diagram provided by Figure 1.2. Its path extends forever to the left, and has the asymptotic form

$$x_\infty \simeq a \text{ as } a \rightarrow -\infty. \quad (\text{J.1.2})$$

Further computation shows that this fixed point is stable.

However, as a is increased slightly beyond -5 , Figure 1.1 shows that there are two more intersections of the curve $y = f(a, b; x)$ with the line $y = x$. When these two intersections first occur (when the curve and line are tangent), a pair of fixed points is born together in a blue-sky bifurcation. These are the two fixed points that appear out of the blue at $a \simeq -4.8$ and $x_\infty \simeq +1$ in Figure 1.2, and move along separate paths as a is further increased. One of these fixed points (the one on the lower of the two paths) is unstable, and the other (the one on the very top path) is stable.

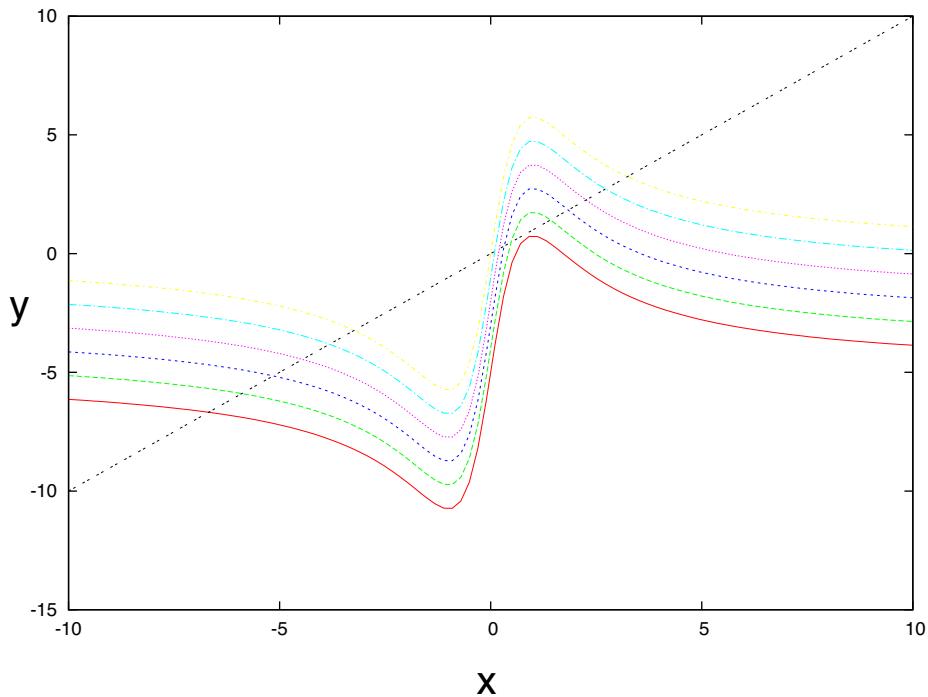


Figure J.1.1: The curves $y = f(a, b; x)$ for $b = 11.5$ and various values of $a \in [-5, 0]$. Also shown is the line $y = x$. Intersections of the line and the curve correspond to fixed points.

J.2 Complete Cascade Denied

As a is increased still further, the fixed point on the very top path begins a period doubling cascade, which we will call the *upper* cascade, at $a \simeq -4.3$ and $x \simeq +1.5$. Again see Figure 1.2. However, as is evident from the figure, the period doubling cascade does not run to completion. Instead it ceases and then begins to undo itself by successive mergers that begin at $a \simeq -3.2$ so that for $a \simeq -.8$ there is again a single fixed point. Its path extends forever to the right, and has the asymptotic form

$$x_\infty \simeq a \text{ as } a \rightarrow +\infty. \quad (\text{J.2.1})$$

Further computation shows that this fixed point is stable.

We began our discussion with the fixed point whose path appears in the lower left side of Figure 1.2 and has the asymptotic form (1.2). Let us now follow its history as a is increased. As indicated in Figure 1.2, it too begins a period doubling cascade, which we will call the *lower* cascade, and this cascade begins at $a \simeq +.8$ and $q_\infty \simeq -2.8$. And, like the upper period doubling cascade, it also does not run to completion. Instead it stops and then undoes itself by successive mergers so that for $a \simeq +4.3$ there is again a single fixed point. This fixed point is stable. Inspection of Figure 1.2 shows that, as a is increased still further, this fixed point blue-sky merges with the unstable fixed point that came out of the blue-sky bifurcation at $a \simeq -4.8$ and $q_\infty \simeq +1$ so that they are mutually annihilated at $q_\infty \simeq -1$ when $a \simeq +4.8$.

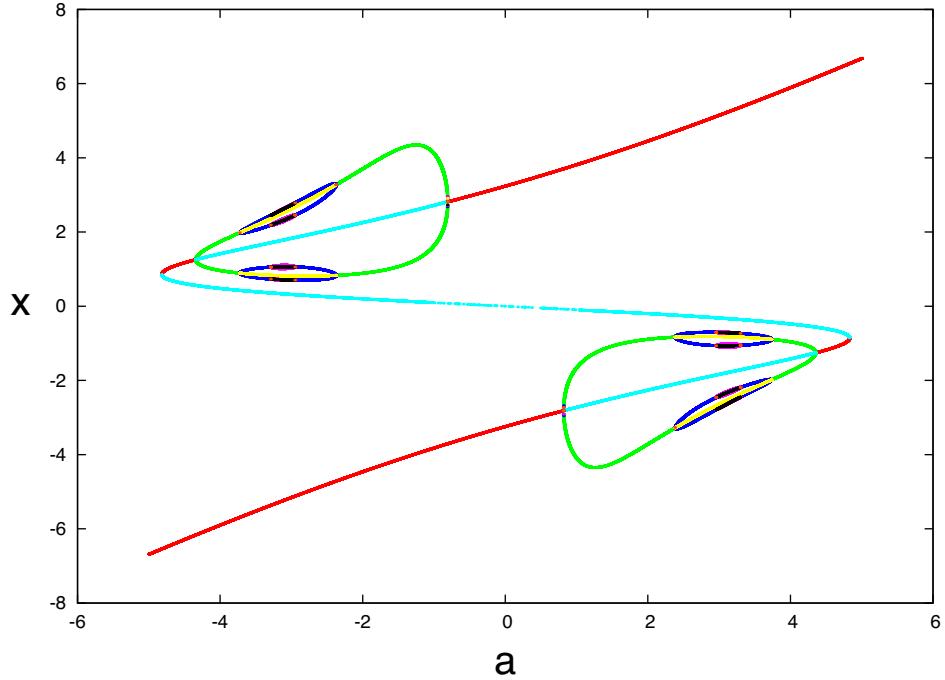


Figure J.1.2: Bifurcation diagram showing x_∞ as a function of a for the map (1.1) with $b = 11.5$ and $a \in [-5, 5]$. For $a = -5$, there is only one fixed point, and it is stable. As a is increased from this value, a blue-sky bifurcation occurs at $x_\infty \simeq +1$ when $a = \simeq -4.8$. Here a pair of fixed points, one stable and one unstable, is born. Now there are three fixed points. The one that bifurcates to larger values of x_∞ is stable, and the one that bifurcates to smaller values of x_∞ is unstable. The original fixed point persists, and remains stable. At $x_\infty \simeq +1$ and $a = \simeq +4.8$ a blue-sky merger occurs where two fixed points, one stable and the other unstable, annihilate. For a values larger than this there is only one fixed point. In between the values $a \simeq -4.8$ and $a \simeq +4.8$ there are two incomplete period-doubling cascades.

Exercises

J.2.1. Figure 2.1 displays intersections between the curve $y = f(a, b; x)$ and the line $y = x$, and we have seen how these intersections are related to the blue-sky bifurcation at $a \simeq -4.8$. Show that the blue-sky merger at $a \simeq +4.8$ can also be understood in terms of intersections between the curve $y = f(a, b; x)$ and the line $y = x$.

J.3 Complete Cascade Achieved

We have seen, from Figure 1.2, that for $b = 11.5$ the period-doubling cascades fail to run to completion. By contrast in Figure 3.1, for which $b = 11.7$, each Feigenbaum cascade runs to completion followed by a region of chaos. Then, it is fascinating to see, each cascade undoes itself by successive mergers as a is further increased until eventually there is again only the one stable fixed point.¹ Note also there that there are two visible windows of stability at $a \simeq -3.3$ and $a \simeq -3.05$. These windows contain stable period-twelve fixed points (as well as numerous unstable fixed points that do not appear because this is a Feigenbaum diagram).

¹In fact there are dynamical systems for which a large or even infinite number of period doubling cascades followed by inverse cascades occur.

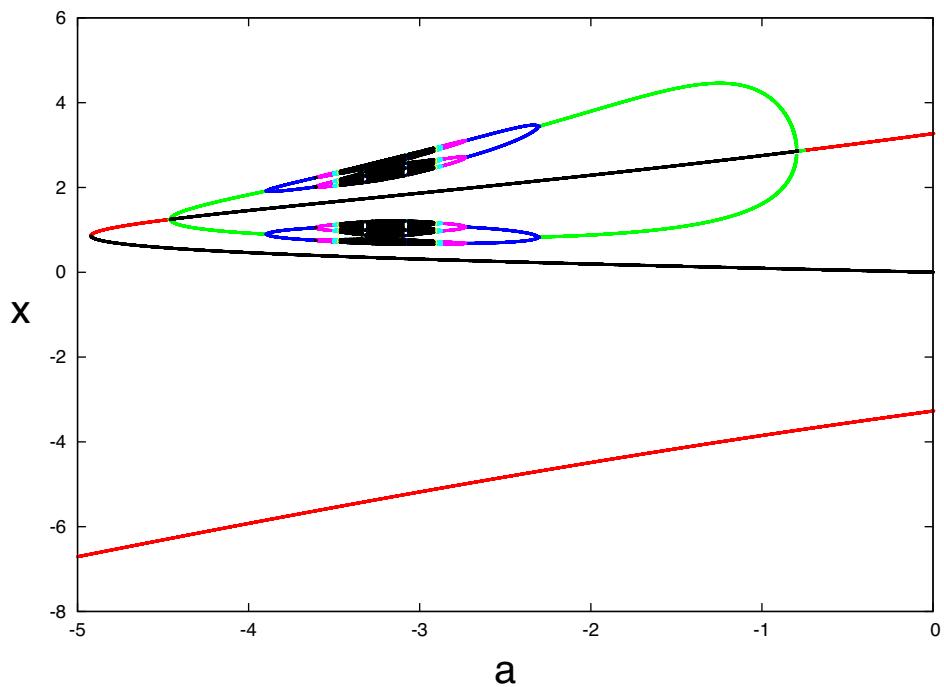


Figure J.3.1: A portion of the Feigenbaum diagram for the map (1.1) with $b = 11.7$. Also shown are the paths of all period-one fixed points, both stable and unstable. The full diagram is similar to that of Figure 1.2 except that both period-doubling cascades now run to completion. Specifically, for the upper cascade shown here, a blue-sky bifurcation again occurs and, as a is further increased, the stable fixed point begins a Feigenbaum period-doubling cascade that now runs to completion followed by a region of chaos. But then, as a is increased still further, the cascades undo themselves until there is again only a single stable fixed point. The behavior for the lower cascade is analogous.

Bibliography

- [1] M. Bier and T. Bountis, “Remerging Feigenbaum trees in dynamical systems”, *Phys. Lett.*, **104A**, 239-244 (1984).
- [2] S. Dawson, C. Grebogi, J. Yorke, I. Kan, and H. Koçak, “Antimonotonicity: inevitable reversals of period-doubling cascades”, *Phys. Lett. A*, **162**, 249-254 (1992).

Appendix K

Supplement to Chapter 17

K.1 Computation of On-Axis Gradients from Spinning Coil Data

A widely used method to measure the magnetic field in magnets for beam optics relies on spinning coils [*]. By using spinning coils one can achieve very accurate measurements of the angular Fourier components of the magnetic field. In this section, which is restricted to the case of straight elements, we show how it is possible using a short length (i.e. with a length shorter than the region of the magnet where the fields are z -dependent) rectangular spinning coil to recover the full z -dependent profiles of the fields and in particular the profiles of the on-axis gradients that are necessary to compute accurately both the linear and nonlinear parts of the transfer map for the magnet.

We consider the case of a rectangular coil rotating in such a way that one side of the coil is always positioned along the magnetic axis of the magnet. The idea is to make repeated measurements of angular field data (integrated over the coil area) by moving the coil along the magnet axis by small steps. The Fourier transforms of the experimental data for each angular harmonic are then calculated, multiplied by a suitable kernel, and then Fourier transformed back to obtain the desired on-axis gradients.

For the kind of coil we consider in this section the only relevant component of the magnetic field is B_ϕ because it is the only one generating a flux linked to the coil. It should be mentioned that tangential coils are also used for which the relevant component of the magnetic field is B_ρ . The treatment of that case would follow the same lines as for the kind of coils considered here.

The E.M.F. produced by a rectangular spinning coil (or set of coils, in a realistic setup), with barycenter positioned at z is given by

$$\mathcal{E}(z, t) = - \int_{z-\ell_c}^{z+\ell_c} dz' \int_0^R \frac{dB_\phi}{dt} d\rho, \quad (\text{K.1.1})$$

where $2\ell_c$ is the length of coil and R is its radius. The E.M.F. can be written in terms of a Fourier series in time,

$$\mathcal{E}(z, t) = \sum_{m=0} \mathcal{E}_{m,s}(z) \sin(m\omega t) + \mathcal{E}_{m,c}(z) \cos(m\omega t), \quad (\text{K.1.2})$$

where we assume $\mathcal{E}_{m,s}(z)$ and $\mathcal{E}_{m,c}(z)$ can be experimentally determined over a sufficient number of locations in z in the end and fringe regions where the field varies with z . The angular frequency of the spinning coil is ω .

By using (2.3) we can write B_ϕ as

$$B_\phi = \frac{1}{\rho} \frac{\partial \psi}{\partial \phi} = \sum_{m=1}^{\infty} \int_{-\infty}^{\infty} dk e^{ikz} m \frac{I_m(k\rho)}{\rho} [\hat{b}_m(k) \cos m\phi - \hat{a}_m(k) \sin m\phi]. \quad (\text{K.1.3})$$

By substituting (5.3) into (5.1) with $\phi = \omega t$ we get

$$\mathcal{E}(z, t) = \frac{\omega}{\sqrt{2\pi}} \sum_{m=1}^{\infty} \int_{-\infty}^{\infty} dk e^{ikz} \frac{2m^2 \sin k\ell_c}{k} [\hat{b}_m(k) \sin m\omega t + \hat{a}_m(k) \cos m\omega t]. \quad (\text{K.1.4})$$

Then, by comparing (5.2) and (5.4), we find the relations

$$\mathcal{E}_{m,c}(z) = \frac{m^2 \omega}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk e^{ikz} \mathcal{I}_m(kR) \frac{2 \sin k\ell_c}{k} \hat{b}_m(k), \quad (\text{K.1.5})$$

$$\mathcal{E}_{m,s}(z) = \frac{m^2 \omega}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk e^{ikz} \mathcal{I}_m(kR) \frac{2 \sin k\ell_c}{k} \hat{a}_m(k). \quad (\text{K.1.6})$$

Here we have defined the new function

$$\mathcal{I}_m(kR) = \int_0^R \frac{I_m(k\rho)}{\rho} d\rho = \int_0^{kR} \frac{I_m(x)}{x} dx. \quad (\text{K.1.7})$$

Finally use of (5.5) and (5.6) allows us to write the expression for the on-axis gradients,

$$C_{m,\alpha}^{[n]}(z) = \frac{i^n}{2^{m+1} m! m^2 \omega} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk e^{ikz} \frac{k^{m+n+1} \tilde{\mathcal{E}}_{m,\alpha}(k)}{\mathcal{I}_m(kR) \sin k\ell_c}, \quad (\text{K.1.8})$$

where the $\tilde{\mathcal{E}}_{m,\alpha}(k)$ are the Fourier transforms of the experimental data,

$$\tilde{\mathcal{E}}_{m,\alpha}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz e^{-ikz} \mathcal{E}_{m,\alpha}(z). \quad (\text{K.1.9})$$

Notice that because of the asymptotic form of the Bessel function I_m , as $k \rightarrow \infty$ the function $\mathcal{I}_m(kR)$ grows exponentially at infinity as e^{kR}/\sqrt{k} . Because the function $\mathcal{I}_m(kR)$ is in the denominator of the integrand in (5.8), there is again an effective cut-off in k .

The integral (5.7) for a general m cannot be carried out analytically, but it can easily be reduced to an infinite series either in the Bessel functions or, most conveniently, directly in kR :

$$\mathcal{I}_m(kR) = \frac{1}{kR} \frac{2}{m} \sum_{n=0}^{\infty} (-1)^n (m+2n+1) I_{m+2n+1}(kR), \quad (\text{K.1.10})$$

$$\mathcal{I}_m(kR) = \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+m)\ell!(\ell+m)!} \left(\frac{kR}{2}\right)^{m+2\ell}. \quad (\text{K.1.11})$$

In the particular case $m = 2$ we have

$$\mathcal{I}_2(kR) = \frac{I_1(kR)}{kR} - \frac{1}{2}. \quad (\text{K.1.12})$$

For numerical purposes use of (5.11) may be perfectly adequate (in particular if speed is not an issue). We will see later that we are often interested in values of kR that satisfy the condition $kR < 20$. For such kR values, one can obtain values for $\mathcal{I}_m(kR)$ that are accurate through 15 digits by retaining the first 30 terms in the series.

K.2 Computation of On-Axis Gradients from Coil Geometry and Current Data

Bibliography

- [1] M. Bassetti and C. Biscari, “Analytic Formulae for Magnetic Multipoles”, *Particle Accelerators*, **52**, 221-250 (1996).

Appendix L

Spline Routines

```
! The following are double precision versions of the subroutines
! "spline" and "splint" used for 1-D cubic spline interpolation, found in Numerical
! Recipes pp. 107-110. Instructions for use:
!
! 1) Call spline(x,y,n,yp1,ypn,y2).
!
!Here x={x_k} is an array of length n containing the x-values on which the function is given, and y i
!array of the same length containing the corresponding function values {f(x_k)}. Also, yp1 and ypn a
!the first derivatives of the function at the points x_1 and x_n, respectively. The routine returns
!array y2 of length n, which contains the second derivatives of the interpolating function at the■
!tabulated points x_n.
!
!The subroutine spline is called only once for a given data set, to set up the array y2.
!
! 2) For a given point x at which the interpolating function is desired, call
    splint(xa,ya,y2a,n,x,y).
!
!Here xa and ya are the arrays {x_n} and {f(x_n)} as above. The array y2 is the output from the■
!subroutine "spline" above. Again, n is the number of points in x. Finally, the double precision nu
!x is the value at which the interpolating function is to be evaluated. The resulting value f(x) is
!as the double precision number y.
!
! C. E. M. 5/27/08
```

```
SUBROUTINE splint(xa,ya,y2a,n,x,y)
INTEGER n
double precision x,y,xa(n),y2a(n),ya(n)
INTEGER k,khi,klo
double precision a,b,h
klo=1
khi=n
1  if (khi-klo.gt.1) then
      k=(khi+klo)/2
      if(xa(k).gt.x)then
          khi=k
      else
          klo=k
```

```

        endif
      goto 1
    endif
    h=xa(khi)-xa(klo)
    if (h.eq.0.d0) pause 'bad xa input in splint'
    a=(xa(khi)-x)/h
    b=(x-xa(klo))/h
    y=a*ya(klo)+b*ya(khi)+((a**3-a)*y2a(klo)+(b**3-b)*y2a(khi))*(h**2)/6.d0
    return
  END

SUBROUTINE spline(x,y,n,yp1,ypn,y2)
implicit none
INTEGER n,NMAX
double precision yp1,ypn,x(n),y(n),y2(n)
PARAMETER (NMAX=10001)
INTEGER i,k
double precision p,qn,sig,un,u(NMAX)
c   if (yp1.gt..99e30) then
c We set the natural bc with vanishing second derivative.
  if (yp1.gt..99e30) then
    y2(1)=0.d0
    u(1)=0.d0
  else
    y2(1)=-0.5d0
    u(1)=(3.d0/(x(2)-x(1)))*((y(2)-y(1))/(x(2)-x(1))-yp1)
  c   u(1)=yp1
  endif
  do 11 i=2,n-1
    sig=(x(i)-x(i-1))/(x(i+1)-x(i-1))
    p=sig*y2(i-1)+2.d0
    y2(i)=(sig-1.d0)/p
    u(i)=(6.d0*((y(i+1)-y(i))/(x(i+1)-x(i))-(y(i)-y(i-1))/(x(i)-x(i-1)))/(x(i+1)-x(i-1))-sig*u(i-1)/p
11  continue
  if (ypn.gt..99e30) then
  c We set the natural upper bc with second derivative = 0.
    qn=0.d0
    un=0.d0
  else
    qn=0.5d0
    un=(3.d0/(x(n)-x(n-1)))*(ypn-(y(n)-y(n-1))/(x(n)-x(n-1)))
  c   un=ypn
  endif
  y2(n)=(un-qn*u(n-1))/(qn*y2(n-1)+1.d0)
  do 12 k=n-1,1,-1
    y2(k)=y2(k)*y2(k+1)+u(k)
12  continue
  return
END

```

PROGRAM PERSPLINE

```

C
C =====
C      Periodic cubic spline interpolation.
C =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
DIMENSION x(801),y(801),y2(801),xa(801),ya(801)

pi=4.d0*atan(1.d0)
n=20
open(unit=26,file='output',status='new')
c      Define the data points to be used for interpolation.
      z=0.d0
dz=(2*pi)/(n-2)
do 10 i=1,n
  y(i)=cos(3*z)
  x(i)=z
  ya(i)=y(i)
  xa(i)=x(i)
  z=z+dz
10      continue
call pspline(x,y,n,y2)
c write(*,*) 'Values of M0, MN = ',y2(1),y2(2),y2(3),y2(4)
c write(*,*) 'x=',xa(1),xa(2),xa(3),'...',xa(n)
write(*,*) 'Spline calls completed.'
c      Compute interpolated values.
      z=0.0d0
nmax=801
dz=(2*pi)/(nmax-1)
do 20 j=1,nmax
  call splint(xa,ya,y2,n,z,C1)
  exact=cos(3*z)
  write(26,*) z,C1,exact
  z=z+dz
20      continue
      end

SUBROUTINE pspline(x,y,n,y2)
c      Takes as input vectors x(n), y(n) defining evaluation of
c      the periodic function at its sampling points.
c      Produces output vectors x - solution to A'x = r
c                      y2 - solution to A'z = u
c      Uses the tridiagonal algorithm for LU decomposition to solve
c      both systems simultaneously.
c      Outputs the vector y2 of second derivatives y'' at sampling points.
      INTEGER n,NMAX
      REAL*8 yp1,ypn,x(n),y(n),y2(n)
      PARAMETER (NMAX=500)

```

```

INTEGER i,k
REAL*8 p,qn,sig,un,u1(NMAX),u2(NMAX)
write(*,*) 'Inside spline'
c Set boundary conditions for lower end. Here u1 is the intermediate
c solution for A'x=r in the LU decomposition, and u2 is the
c intermediate solution of A'z=u in the LU decomposition.
y2(1)=-1.0d0
u1(1)=0.0d0
u2(1)=-1.0d0
do 11 i=2,n-1
sig=(x(i)-x(i-1))/(x(i+1)-x(i-1))
write(*,*) 'sig =',sig
p=sig*y2(i-1)+2.0d0
y2(i)=(sig-1.d0)/p
write(*,*) 'Beta, gamma = ',p,-1.d0*y2(i)
u1(i)=(6.d0*((y(i+1)-y(i))/(x(i+
& 1)-x(i))-(y(i)-y(i-1))/(x(i)-x(i-1)))/(x(i+1)-x(i-1))-sig*
& u1(i-1))/p
u2(i)=((sig-1.d0)*u2(i-1))/p
write(*,*) 'u1,u2 = ',u1(i),u2(i)
11 continue
c Set boundary conditions for upper end.
x(n)=-1.d0*u1(n-1)/(y2(n-1)+1.d0)
y2(n)=-1.d0*(1.d0+u2(n-1))/(y2(n-1)+1.d0)
write(*,*) 'xn, y2n = ',x(n),y2(n)
do 12 k=n-1,1,-1
x(k)=y2(k)*x(k+1)+u1(k)
y2(k)=y2(k)*y2(k+1)+u2(k)
write(*,*) 'x,y2 = ',x(k),y2(k)
12 continue
c Given the two solutions x(k) and y2(k), we use the Sherman-Morrison
c formula to construct the solution to the periodic spline system with
c its off-diagonal terms. Here 'fact' is the correction to the
c intermediate solution vector x due to the off-diagonal terms.
fact = (x(2)+x(n-1))/(1.d0+y2(2)+y2(n-1))
do 10 i=1,n
y2(i) = x(i) - fact*y2(i)
write(*,*) 'y2(i) = ',y2(i)
10 continue
return
END

SUBROUTINE splint(xa,ya,y2a,n,x,y)
implicit double precision(a-h,o-z)
INTEGER n
REAL*8 x,y,xa(n),y2a(n),ya(n)
INTEGER k,khi,klo
REAL*8 a,b,h
klo=1
khi=n
1 if (khi-klo.gt.1) then
k=(khi+klo)/2
if(xa(k).gt.x)then
khi=k

```

```
else
  klo=k
endif
goto 1
endif
h=xa(khi)-xa(klo)
if (h.eq.0.0d0) pause 'bad xa input in splint'
a=(xa(khi)-x)/h
b=(x-xa(klo))/h
y=a*ya(klo)+b*ya(khi)+((a**3-a)*y2a(klo) +
&           (b**3-b)*y2a(khi))*(h**2)/6.d0
return
END
```


Appendix M

Routines for Mathieu Separation Constants $a_n(q)$ and $b_n(q)$

```
SUBROUTINE CVA2(KD,M,Q,A)
C
C      =====
C      Purpose: Calculate a specific characteristic value of
C              Mathieu functions
C      Input :  m   --- Order of Mathieu functions
C              q   --- Parameter of Mathieu functions
C              KD  --- Case code
C                      KD=1 for cem(x,q)  ( m = 0,2,4,...)
C                      KD=2 for cem(x,q)  ( m = 1,3,5,...)
C                      KD=3 for sem(x,q)  ( m = 1,3,5,...)
C                      KD=4 for sem(x,q)  ( m = 2,4,6,...)
C      Output: A   --- Characteristic value
C      Routines called:
C          (1) REFINE for finding accurate characteristic
C              values using an iteration method
C          (2) CVO for finding initial characteristic
C              values using polynomial approximation
C          (3) CVQM for computing initial characteristic
C              values for q  3*m
C          (3) CVQL for computing initial characteristic
C              values for q  m*m
C      =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
IF (M.LE.12.OR.Q.LE.3.0*M.OR.Q.GT.M*M) THEN
    CALL CVO(KD,M,Q,A)
    IF (Q.NE.0.0D0) CALL REFINE(KD,M,Q,A,1)
ELSE
    NDIV=10
    DELTA=(M-3.0)*M/NDIV
    IF ((Q-3.0*M).LE.(M*M-Q)) THEN
5       NN=INT((Q-3.0*M)/DELTA)+1
        DELTA=(Q-3.0*M)/NN
        Q1=2.0*M
```

```

CALL CVQM(M,Q1,A1)
Q2=3.0*M
CALL CVQM(M,Q2,A2)
QQ=3.0*M
DO 10 I=1,NN
QQ=QQ+DELTA
A=(A1*Q2-A2*Q1+(A2-A1)*QQ)/(Q2-Q1)
IFLAG=1
IF (I.EQ.NN) IFLAG=-1
CALL REFINE(KD,M,QQ,A,IFLAG)
Q1=Q2
Q2=QQ
A1=A2
A2=A
10          CONTINUE
IF (IFLAG.EQ.-10) THEN
NDIV=NDIV*2
DELTA=(M-3.0)*M/NDIV
GO TO 5
ENDIF
ELSE
15          NN=INT((M*M-Q)/DELTA)+1
DELTA=(M*M-Q)/NN
Q1=M*(M-1.0)
CALL CVQL(KD,M,Q1,A1)
Q2=M*M
CALL CVQL(KD,M,Q2,A2)
QQ=M*M
DO 20 I=1,NN
QQ=QQ-DELTA
A=(A1*Q2-A2*Q1+(A2-A1)*QQ)/(Q2-Q1)
IFLAG=1
IF (I.EQ.NN) IFLAG=-1
CALL REFINE(KD,M,QQ,A,IFLAG)
Q1=Q2
Q2=QQ
A1=A2
A2=A
20          CONTINUE
IF (IFLAG.EQ.-10) THEN
NDIV=NDIV*2
DELTA=(M-3.0)*M/NDIV
GO TO 15
ENDIF
ENDIF
RETURN
END

SUBROUTINE REFINE(KD,M,Q,A,IFLAG)
C
C      =====
C      Purpose: calculate the accurate characteristic value

```

```

C           by the secant method
C   Input : m --- Order of Mathieu functions
C           q --- Parameter of Mathieu functions
C           A --- Initial characteristic value
C   Output: A --- Refineed characteristic value
C   Routine called: CVF for computing the value of F for
C                   characteristic equation
C =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
EPS=1.0D-14
MJ=10+M
CA=A
DELTA=0.0D0
X0=A
CALL CVF(KD,M,Q,X0,MJ,F0)
X1=1.002*A
CALL CVF(KD,M,Q,X1,MJ,F1)
5      DO 10 IT=1,100
      MJ=MJ+1
      X=X1-(X1-X0)/(1.0D0-F0/F1)
      CALL CVF(KD,M,Q,X,MJ,F)
      IF (ABS(1.0-X1/X).LT.EPS.OR.F.EQ.0.0) GO TO 15
      X0=X1
      F0=F1
      X1=X
10      F1=F
15      A=X
IF (DELTA.GT.0.05) THEN
  A=CA
  IF (IFLAG.LT.0) THEN
    IFLAG=-10
  ENDIF
  RETURN
ENDIF
IF (ABS((A-CA)/CA).GT.0.05) THEN
  X0=CA
  DELTA=DELTA+0.005D0
  CALL CVF(KD,M,Q,X0,MJ,F0)
  X1=(1.0D0+DELTA)*CA
  CALL CVF(KD,M,Q,X1,MJ,F1)
  GO TO 5
ENDIF
RETURN
END

```

```

SUBROUTINE CVF(KD,M,Q,A,MJ,F)
C
C =====
C   Purpose: Compute the value of F for characteristic
C             equation of Mathieu functions
C   Input : m --- Order of Mathieu functions
C           q --- Parameter of Mathieu functions

```

```

C           A --- Characteristic value
C           Output: F --- Value of F for characteristic equation
C   =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
B=A
IC=INT(M/2)
L=0
L0=0
J0=2
JF=IC
IF (KD.EQ.1) L0=2
IF (KD.EQ.1) J0=3
IF (KD.EQ.2.OR.KD.EQ.3) L=1
IF (KD.EQ.4) JF=IC-1
T1=0.0D0
DO 10 J=MJ,IC+1,-1
10      T1=-Q*Q/((2.0D0*J+L)**2-B+T1)
IF (M.LE.2) THEN
  T2=0.0D0
  IF (KD.EQ.1.AND.M.EQ.0) T1=T1+T1
  IF (KD.EQ.1.AND.M.EQ.2) T1=-2.0*Q*Q/(4.0-B+T1)-4.0
  IF (KD.EQ.2.AND.M.EQ.1) T1=T1+Q
  IF (KD.EQ.3.AND.M.EQ.1) T1=T1-Q
ELSE
  IF (KD.EQ.1) T0=4.0D0-B+2.0D0*Q*Q/B
  IF (KD.EQ.2) T0=1.0D0-B+Q
  IF (KD.EQ.3) T0=1.0D0-B-Q
  IF (KD.EQ.4) T0=4.0D0-B
  T2=-Q*Q/T0
  DO 15 J=J0,JF
15      T2=-Q*Q/((2.0D0*J-L-L0)**2-B+T2)
ENDIF
F=(2.0D0*IC+L)**2+T1+T2-B
RETURN
END

```

```

SUBROUTINE CV0(KD,M,Q,A0)
C
C   =====
C   Purpose: Compute the initial characteristic value of
C             Mathieu functions for m  12 or q  300 or
C             q  m*m
C   Input : m --- Order of Mathieu functions
C             q --- Parameter of Mathieu functions
C   Output: A0 --- Characteristic value
C   Routines called:
C             (1) CVQM for computing initial characteristic
C                 value for q  3*m
C             (2) CVQL for computing initial characteristic
C                 value for q  m*m
C   =====
C

```

```

IMPLICIT DOUBLE PRECISION (A-H,O-Z)
Q2=Q*Q
IF (M.EQ.0) THEN
  IF (Q.LE.1.0) THEN
    A0=((.0036392*Q2-.0125868)*Q2+.0546875)*Q2-.5)*Q2
  ELSE IF (Q.LE.10.0) THEN
    A0=((3.999267D-3*Q-9.638957D-2)*Q-.88297)*Q
    &           +.5542818
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.1) THEN
  IF (Q.LE.1.0.AND.KD.EQ.2) THEN
    A0=(((-6.51E-4*Q-.015625)*Q-.125)*Q+1.0)*Q+1.0
  ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
    A0=(((-6.51E-4*Q+.015625)*Q-.125)*Q-1.0)*Q+1.0
  ELSE IF (Q.LE.10.0.AND. KD.EQ.2) THEN
    A0=(((-4.94603D-4*Q+1.92917D-2)*Q-.3089229)
    &           *Q+1.33372)*Q+.811752
  ELSE IF (Q.LE.10.0.AND.KD.EQ.3) THEN
    A0=((1.971096D-3*Q-5.482465D-2)*Q-1.152218)
    &           *Q+1.10427
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.2) THEN
  IF (Q.LE.1.0.AND.KD.EQ.1) THEN
    A0=(((-.0036391*Q2+.0125888)*Q2-.0551939)*Q2
    &           +.416667)*Q2+4.0
  ELSE IF (Q.LE.1.0.AND.KD.EQ.4) THEN
    A0=(.0003617*Q2-.0833333)*Q2+4.0
  ELSE IF (Q.LE.15.AND.KD.EQ.1) THEN
    A0=((3.200972D-4*Q-8.667445D-3)*Q
    &           -1.829032D-4)*Q+.9919999)*Q+3.3290504
  ELSE IF (Q.LE.10.0.AND.KD.EQ.4) THEN
    A0=((2.38446D-3*Q-.08725329)*Q-4.732542D-3)
    &           *Q+4.00909
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.3) THEN
  IF (Q.LE.1.0.AND.KD.EQ.2) THEN
    A0=((6.348E-4*Q+.015625)*Q+.0625)*Q2+9.0
  ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
    A0=((6.348E-4*Q-.015625)*Q+.0625)*Q2+9.0
  ELSE IF (Q.LE.20.0.AND.KD.EQ.2) THEN
    A0=((3.035731D-4*Q-1.453021D-2)*Q
    &           +.19069602)*Q-.1039356)*Q+8.9449274
  ELSE IF (Q.LE.15.0.AND.KD.EQ.3) THEN
    A0=((9.369364D-5*Q-.03569325)*Q+.2689874)*Q
    &           +8.771735
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF

```

```

ELSE IF (M.EQ.4) THEN
  IF (Q.LE.1.0.AND.KD.EQ.1) THEN
    A0=(-2.1E-6*Q2+5.012E-4)*Q2+.0333333)*Q2+16.0
  ELSE IF (Q.LE.1.0.AND.KD.EQ.4) THEN
    A0=(3.7E-6*Q2-3.669E-4)*Q2+.0333333)*Q2+16.0
  ELSE IF (Q.LE.25.0.AND.KD.EQ.1) THEN
    A0=(((1.076676D-4*Q-7.9684875D-3)*Q
      &           +.17344854)*Q-.5924058)*Q+16.620847
  ELSE IF (Q.LE.20.0.AND.KD.EQ.4) THEN
    A0=(-7.08719D-4*Q+3.8216144D-3)*Q
      &           +.1907493)*Q+15.744
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.5) THEN
  IF (Q.LE.1.0.AND.KD.EQ.2) THEN
    A0=((6.8E-6*Q+1.42E-5)*Q2+.0208333)*Q2+25.0
  ELSE IF (Q.LE.1.0.AND.KD.EQ.3) THEN
    A0=(-6.8E-6*Q+1.42E-5)*Q2+.0208333)*Q2+25.0
  ELSE IF (Q.LE.35.0.AND.KD.EQ.2) THEN
    A0=(((2.238231D-5*Q-2.983416D-3)*Q
      &           +.10706975)*Q-.600205)*Q+25.93515
  ELSE IF (Q.LE.25.0.AND.KD.EQ.3) THEN
    A0=(-7.425364D-4*Q+2.18225D-2)*Q
      &           +4.16399D-2)*Q+24.897
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.6) THEN
  IF (Q.LE.1.0) THEN
    A0=.4D-6*Q2+.0142857)*Q2+36.0
  ELSE IF (Q.LE.40.0.AND.KD.EQ.1) THEN
    A0=(((1.66846D-5*Q+4.80263D-4)*Q
      &           +2.53998D-2)*Q-.181233)*Q+36.423
  ELSE IF (Q.LE.35.0.AND.KD.EQ.4) THEN
    A0=(-4.57146D-4*Q+2.16609D-2)*Q-2.349616D-2)*Q
      &           +35.99251
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.EQ.7) THEN
  IF (Q.LE.10.0) THEN
    CALL CVQM(M,Q,A0)
  ELSE IF (Q.LE.50.0.AND.KD.EQ.2) THEN
    A0=(((1.411114D-5*Q+9.730514D-4)*Q
      &           -3.097887D-3)*Q+3.533597D-2)*Q+49.0547
  ELSE IF (Q.LE.40.0.AND.KD.EQ.3) THEN
    A0=(-3.043872D-4*Q+2.05511D-2)*Q
      &           -9.16292D-2)*Q+49.19035
  ELSE
    CALL CVQL(KD,M,Q,A0)
  ENDIF
ELSE IF (M.GE.8) THEN
  IF (Q.LE.3.*M) THEN

```

```

    CALL CVQM(M,Q,A0)
    ELSE IF (Q.GT.M*M) THEN
        CALL CVQL(KD,M,Q,A0)
    ELSE
        IF (M.EQ.8.AND.KD.EQ.1) THEN
            A0=((8.634308D-6*Q-2.100289D-3)*Q+.169072)*Q
            & -4.64336)*Q+109.4211
        ELSE IF (M.EQ.8.AND.KD.EQ.4) THEN
            A0=(-6.7842D-5*Q+2.2057D-3)*Q+.48296)*Q+56.59
        ELSE IF (M.EQ.9.AND.KD.EQ.2) THEN
            A0=((2.906435D-6*Q-1.019893D-3)*Q+.1101965)*Q
            & -3.821851)*Q+127.6098
        ELSE IF (M.EQ.9.AND.KD.EQ.3) THEN
            A0=(-9.577289D-5*Q+.01043839)*Q+.06588934)*Q
            & +78.0198
        ELSE IF (M.EQ.10.AND.KD.EQ.1) THEN
            A0=((5.44927D-7*Q-3.926119D-4)*Q+.0612099)*Q
            & -2.600805)*Q+138.1923
        ELSE IF (M.EQ.10.AND.KD.EQ.4) THEN
            A0=(-7.660143D-5*Q+.01132506)*Q-.09746023)*Q
            & +99.29494
        ELSE IF (M.EQ.11.AND.KD.EQ.2) THEN
            A0=(((-5.67615D-7*Q+7.152722D-6)*Q+.01920291)*Q
            & -1.081583)*Q+140.88
        ELSE IF (M.EQ.11.AND.KD.EQ.3) THEN
            A0=(-6.310551D-5*Q+.0119247)*Q-.2681195)*Q
            & +123.667
        ELSE IF (M.EQ.12.AND.KD.EQ.1) THEN
            A0=((-2.38351D-7*Q-2.90139D-5)*Q+.02023088)*Q
            & -1.289)*Q+171.2723
        ELSE IF (M.EQ.12.AND.KD.EQ.4) THEN
            A0=((3.08902D-7*Q-1.577869D-4)*Q+.0247911)*Q
            & -1.05454)*Q+161.471
        ENDIF
    ENDIF
ENDIF
RETURN
END

```

```

SUBROUTINE CVQL(KD,M,Q,A0)
C
C =====
C Purpose: Compute the characteristic value of Mathieu
C           functions for q 3m
C Input : m --- Order of Mathieu functions
C           q --- Parameter of Mathieu functions
C Output: A0 --- Initial characteristic value
C =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
IF (KD.EQ.1.OR.KD.EQ.2) W=2.0D0*M+1.0D0
IF (KD.EQ.3.OR.KD.EQ.4) W=2.0D0*M-1.0D0
W2=W*W

```

```

W3=W*W2
W4=W2*W2
W6=W2*W4
D1=5.0+34.0/W2+9.0/W4
D2=(33.0+410.0/W2+405.0/W4)/W
D3=(63.0+1260.0/W2+2943.0/W4+486.0/W6)/W2
D4=(527.0+15617.0/W2+69001.0/W4+41607.0/W6)/W3
C1=128.0
P2=Q/W4
P1=DSQRT(P2)
CV1=-2.0*Q+2.0*W*DSQRT(Q)-(W2+1.0)/8.0
CV2=(W+3.0/W)+D1/(32.0*P1)+D2/(8.0*C1*P2)
CV2=CV2+D3/(64.0*C1*P1*P2)+D4/(16.0*C1*C1*P2*P2)
AO=CV1-CV2/(C1*P1)
RETURN
END

```

```

SUBROUTINE CVQM(M,Q,A0)
C
C      =====
C      Purpose: Compute the characteristic value of Mathieu
C              functions for q  m*m
C      Input :  m --- Order of Mathieu functions
C              q   --- Parameter of Mathieu functions
C      Output: A0 --- Initial characteristic value
C      =====
C
IMPLICIT DOUBLE PRECISION (A-H,O-Z)
HM1=.5*Q/(M*M-1.0)
HM3=.25*HM1**3/(M*M-4.0)
HM5=HM1*HM3*Q/((M*M-1.0)*(M*M-9.0))
A0=M*M+Q*(HM1+(5.0*M*M+7.0)*HM3
&      +(9.0*M**4+58.0*M*M+29.0)*HM5)
RETURN
END

FUNCTION fac(n)
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
INTEGER n,J
IF (n.EQ.0) fac = 1.d0
      f=1.d0
DO 11 J=1,n
      f = J*f
11      CONTINUE
      fac = f
RETURN
END

```

Appendix N

Mathieu-Bessel Connection Coefficients

As a consequence of the symmetry properties (13.9.40) through (13.9.43), the Fourier expansions (13.10.144) and (13.10.150) for the $\text{ce}_n(v, q)$ and $\text{se}_n(v, q)$ can be written in the form

$$\text{ce}_{2n}(v, q) = \sum_{m=0}^{\infty} A_{2m}^{2n}(q) \cos(2mv), \quad (\text{N.0.1})$$

$$\text{ce}_{2n+1}(v, q) = \sum_{m=0}^{\infty} A_{2m+1}^{2n+1}(q) \cos[(2m+1)v], \quad (\text{N.0.2})$$

$$\text{se}_{2n+1}(v, q) = \sum_{m=0}^{\infty} B_{2m+1}^{2n+1}(q) \sin[(2m+1)v], \quad (\text{N.0.3})$$

$$\text{se}_{2n+2}(v, q) = \sum_{m=0}^{\infty} B_{2m+2}^{2n+2}(q) \sin[(2m+2)v]. \quad (\text{N.0.4})$$

In all these relations $n = 0, 1, 2, 3, \dots$. Put another way, the symmetry properties require that the coefficients A_m^n and B_m^n vanish unless both m and n are even or both m and n are odd.

In this appendix we will see that the same symmetry properties hold for the Mathieu-Bessel connection coefficients α_m^n and β_m^n . That is, formulas (13.9.64) and (13.9.65) can be written in the corresponding form

$$\text{Ce}_{2n}(u, q) \text{ ce}_{2n}(v, q) = \sum_{m=0}^{\infty} \alpha_{2m}^{2n}(k) I_{2m}(k\rho) \cos(2m\phi), \quad (\text{N.0.5})$$

$$\text{Ce}_{2n+1}(u, q) \text{ ce}_{2n+1}(v, q) = \sum_{m=0}^{\infty} \alpha_{2m+1}^{2n+1}(k) I_{2m+1}(k\rho) \cos[(2m+1)\phi], \quad (\text{N.0.6})$$

$$\text{Se}_{2n+1}(u, q) \text{ se}_{2n+1}(v, q) = \sum_{m=0}^{\infty} \beta_{2m+1}^{2n+1}(k) I_{2m+1}(k\rho) \sin[(2m+1)\phi], \quad (\text{N.0.7})$$

$$\text{Se}_{2n+2}(u, q) \text{ se}_{2n+2}(v, q) = \sum_{m=0}^{\infty} \beta_{2m+2}^{2n+2}(k) I_{2m+2}(k\rho) \sin[(2m+2)\phi]. \quad (\text{N.0.8})$$

Moreover, there are the relations

$$\alpha_{2m}^{2n}(k) = g_c^{2n}(k) A_{2m}^{2n}(q), \quad (\text{N.0.9})$$

$$\alpha_{2m+1}^{2n+1}(k) = g_c^{2n+1}(k) A_{2m+1}^{2n+1}(q), \quad (\text{N.0.10})$$

$$\beta_{2m+1}^{2n+1}(k) = g_s^{2n+1}(k) B_{2m+1}^{2n+1}(q), \quad (\text{N.0.11})$$

$$\beta_{2m+2}^{2n+2}(k) = g_s^{2n+2}(k) B_{2m+2}^{2n+2}(q), \quad (\text{N.0.12})$$

where

$$g_c^{2n}(k) = [\text{ce}_{2n}(\pi/2, q) \text{ ce}_{2n}(0, q)]/A_0^{2n}(q), \quad (\text{N.0.13})$$

$$g_c^{2n+1}(k) = -2[\text{ce}'_{2n+1}(\pi/2, q) \text{ ce}_{2n+1}(0, q)]/[kf A_1^{2n+1}(q)], \quad (\text{N.0.14})$$

$$g_s^{2n+1}(k) = 2[\text{se}_{2n+1}(\pi/2, q) \text{ se}'_{2n+1}(0, q)]/[kf B_1^{2n+1}(q)], \quad (\text{N.0.15})$$

$$g_s^{2n+2}(k) = [\text{se}'_{2n+2}(\pi/2, q) \text{ se}'_{2n+2}(0, q)]/[q B_2^{2n+2}(q)]. \quad (\text{N.0.16})$$

Here a \prime denotes d/dv .

Appendix O

Quadratic Forms

O.1 Background

Let L be a real $m \times m$ matrix, let w be a real m -component vector, and let $(*, *)$ denote the usual real inner product. Define a quadratic form $Q(w)$ by the rule

$$Q(w) = (w, Lw). \quad (\text{O.1.1})$$

The matrix L can be uniquely decomposed into symmetric and antisymmetric parts S and A by writing

$$L = S + A \quad (\text{O.1.2})$$

with

$$S = (1/2)(L + L^T) \quad (\text{O.1.3})$$

and

$$A = (1/2)(L - L^T). \quad (\text{O.1.4})$$

Then, since only the symmetric part of L contributes to Q , we may equally well write

$$Q(w) = (w, Sw). \quad (\text{O.1.5})$$

According to standard matrix theory, any real $m \times m$ symmetric matrix S has m real eigenvalues and m associated real eigenvectors that can be arranged to form an orthonormal basis. (Note that no assumption needs to be made about the eigenvalues being distinct.) Call the eigenvalues σ_j and the associated orthonormal eigenvectors v_j . Then we may write S in the dyadic form

$$S = \sum_{j=1}^m \sigma_j |v_j\rangle (v_j|). \quad (\text{O.1.6})$$

With the aid of this representation for S we find that Q takes the form

$$Q(w) = (w, Sw) = \sum_{j=1}^m \sigma_j (w, v_j) (v_j, w) = \sum_{j=1}^m \sigma_j (w, v_j)^2. \quad (\text{O.1.7})$$

We see that Q will be positive definite if all $\sigma_j > 0$. Conversely, since the v_j are orthonormal, it is evident that all the σ_j will be positive if Q is positive definite. Similarly, Q will be negative definite if all $\sigma_j < 0$, and conversely. Finally, Q will be indefinite if not all σ_j have the same sign or some are zero.

O.2 Effect of Small Perturbations in the Definite Case

Now suppose, for example, that Q is positive definite and that all the eigenvalues σ_j of S are substantially different from 0. Next suppose that L is slightly perturbed so that S is also slightly perturbed. Thus, we may write that

$$S = S^0 + S^1 \quad (\text{O.2.1})$$

where S^0 is the initial S before perturbation and S^1 is a small symmetric matrix that describes the perturbation. Now the quadratic form Q becomes Q' with

$$Q'(w) = (w, [S^0 + S^1]w) = (w, S^0 w) + (w, S^1 w) = Q(w) + (w, S^1 w). \quad (\text{O.2.2})$$

It is easy to see from (1.7) that

$$Q(w) \geq \sigma_{\min} \sum_{j=1}^m (w, v_j)^2 = \sigma_{\min}(w, w) = \sigma_{\min} \|w\|^2 \quad (\text{O.2.3})$$

where σ_{\min} is the smallest eigenvalue of S . Also, we have the estimate

$$|(w, S^1 w)| \leq \|w\| \|S^1 w\| \leq \|w\| \|S^1\| \|w\| = \|S^1\| \|w\|^2. \quad (\text{O.2.4})$$

It follows that Q' will also be positive definite providing

$$\|S^1\| < \sigma_{\min}. \quad (\text{O.2.5})$$

We conclude that if Q is positive definite, it will remain positive definite under small perturbations of L . Similarly, if Q is negative definite, it will remain negative definite under small perturbations of L .

Even more can be said. The *rank* of Q is defined to be the number of nonzero eigenvectors of S , and the *signature* is defined to be the number of positive eigenvalues minus the number of negative eigenvalues. It can be shown that if under a continuous change in S the rank does not change (no eigenvalue passes through the value 0), then the signature also does not change. This result follows from the fact that the eigenvalues of S are continuous functions of the matrix elements of S .

Bibliography

- [1] F. Gantmacher, *The Theory of Matrices*, Vols. I and II, Chelsea (1959). See page 309 in Vol. I, which discusses quadratic forms.

Appendix P

Parameterization of the Coset Space $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$

P.1 Introduction

Suppose that $M \in GL(2n, \mathbb{R})$ has a symplectic polar decomposition,

$$M = QR \quad (\text{P.1.1})$$

where Q is J -symmetric and R is symplectic.¹ We know that such a decomposition is possible for M sufficiently near the symplectic group and is unique. We know that the ordinary (orthogonal) polar decomposition can be made globally. By contrast, from counter examples, we know that symplectic polar decomposition is not possible globally. We also see that, by construction, J -symmetric matrices Q are related to the cosets $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$. We want to find what restrictions must be imposed on M for a symplectic polar decomposition to be possible, and suspect that these restrictions are related to coset structure.

P.2 M Must Have Positive Determinant

From (1.1) we find

$$\det M = (\det Q)(\det R) = \det Q. \quad (\text{P.2.1})$$

According to Lemma 3.6 of Section 4.3 of *Lie Methods*, any J -symmetric matrix Q can be written in the form

$$Q = JA \quad (\text{P.2.2})$$

where A is real and antisymmetric. It follows that

$$\det Q = (\det J)(\det A) = \det A \geq 0. \quad (\text{P.2.3})$$

Here we have used the fact that a real antisymmetric matrix cannot have a negative determinant. It follows from (1.1), (2.1), and (2.3) that, if M is to be nonsingular and have a symplectic polar decomposition, it must have positive determinant,

$$\det M > 0. \quad (\text{P.2.4})$$

¹We adopt the terminology of Chapter 4 of *Lie Methods* . . .

Then, from (2.1), we also have

$$\det Q > 0. \quad (\text{P.2.5})$$

P.3 It is Sufficient to Consider $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$

Suppose N is any matrix in $GL(2n, \mathbb{R})$ and $\det N > 0$. Define an associated matrix M by the rule

$$M = (\det N)^{-1/(2n)} N. \quad (\text{P.3.1})$$

By construction, M will be in $SL(2n, \mathbb{R})$.

Next assume that M has the symplectic polar decomposition (1.1). Then (1.1) and (3.1) imply that

$$N = (\det N)^{1/(2n)} M = (\det N)^{1/(2n)} QR = Q'R \quad (\text{P.3.2})$$

where

$$Q' = (\det N)^{1/(2n)} Q. \quad (\text{P.3.3})$$

By the lemmas of Section 4.3 of *Lie Methods*, Q' will also be J -symmetric, and therefore N has a symplectic polar decomposition. Thus, it is sufficient to study whether any $M \in SL(2n, \mathbb{R})$ has a symplectic polar decomposition.

P.4 Some Symmetries

Consider the map Σ of $SL(2n, \mathbb{R})$ into itself defined by the rule

$$\Sigma(M) = J(M^T)^{-1} J^{-1}. \quad (\text{P.4.1})$$

We will now explore the properties of Σ .

Suppose M_1 and M_2 are any two $SL(2n, \mathbb{R})$ matrices. Then we find the relation

$$\begin{aligned} \Sigma(M_1 M_2) &= J[(M_1 M_2)^T]^{-1} J^{-1} = J[M_2^T M_1^T]^{-1} J^{-1} \\ &= J(M_1^T)^{-1} (M_2^T)^{-1} J^{-1} = J(M_1^T)^{-1} J^{-1} J(M_2^T)^{-1} J^{-1} \\ &= \Sigma(M_1) \Sigma(M_2). \end{aligned} \quad (\text{P.4.2})$$

Thus, Σ is a homomorphism.

Next we observe that

$$\Sigma(I) = I \quad (\text{P.4.3})$$

and

$$\Sigma(J) = J(J^T)^{-1} J^{-1} = J(-J)^{-1} J^{-1} = JJJ^{-1} = J. \quad (\text{P.4.4})$$

Similarly, we find

$$\Sigma(J^{-1}) = J^{-1}. \quad (\text{P.4.5})$$

Also, there is the property

$$\Sigma(M^{-1}) = J[(M^{-1})^T]^{-1} J^{-1} = JM^T J^{-1} = [J(M^T)^{-1} J^{-1}]^{-1} = [\Sigma(M)]^{-1}. \quad (\text{P.4.6})$$

[This result also follows from (4.2) and (4.3).] Consequently, Σ is an isomorphism.

We claim that Σ acts as the identity map on $Sp(2n, \mathbb{R})$. That is, all elements $R \in Sp(2n, \mathbb{R})$ are fixed points of Σ . Indeed, suppose that $R \in Sp(2n, \mathbb{R})$. Then we have the result

$$RJR^T = J, \quad (\text{P.4.7})$$

which is equivalent to the relation

$$J^{-1} = (RJR^T)^{-1} = (R^T)^{-1}J^{-1}R^{-1}, \quad (\text{P.4.8})$$

which in turn is equivalent to the relation

$$R = J(R^T)^{-1}J^{-1} = \Sigma(R). \quad (\text{P.4.9})$$

Note that (4.3) through (4.5) are special cases of (4.9).

Let us next find the action of Σ on any J -symmetric matrix Q . We find the result

$$\Sigma(Q) = J(Q^T)^{-1}J^{-1} = (JQ^TJ^{-1})^{-1} = Q^{-1}. \quad (\text{P.4.10})$$

Note that (2.5) guarantees that Q^{-1} exists.

Upon combining (4.9) and (4.10) we find that the effect of Σ on any matrix M having the factorization (1.1) is given by the relation

$$\Sigma(M) = \Sigma(QR) = \Sigma(Q)\Sigma(R) = Q^{-1}R. \quad (\text{P.4.11})$$

Finally, Σ is an involution. By calculating we find that

$$\begin{aligned} \Sigma^2(M) &= \Sigma[\Sigma(M)] = \Sigma[J(M^T)^{-1}J^{-1}] = \Sigma(J)\Sigma[(M^T)^{-1}]\Sigma(J^{-1}) \\ &= J\Sigma[(M^T)^{-1}]J^{-1} = JJ\{(M^T)^{-1}\}^{-1}J^{-1}J^{-1} \\ &= JJJ^{-1}J^{-1} = (-I)M(-I) = M. \end{aligned} \quad (\text{P.4.12})$$

We have found a symmetry for $SL(2n, \mathbb{R})$. We will now see that Σ produces an associated symmetry σ on the Lie algebra $sl(2n, \mathbb{R})$. Let B be any element in the Lie algebra $sl(2n, \mathbb{R})$. Let σ be the associated induced map in the Lie algebra defined by the relation

$$\Sigma[\exp(B)] = \exp[\sigma(B)]. \quad (\text{P.4.13})$$

By calculation we find

$$\begin{aligned} \Sigma[\exp(B)] &= J\{[\exp(B)]^T\}^{-1}J^{-1} = J\{\exp(B^T)\}^{-1}J^{-1} \\ &= J\exp(-B^T)J^{-1} = \exp(-JB^TJ^{-1}). \end{aligned} \quad (\text{P.4.14})$$

Upon comparing (4.13) and (4.14) in the vicinity of the identity, we conclude that

$$\sigma(B) = -JB^TJ^{-1}. \quad (\text{P.4.15})$$

Let us explore the properties of σ . Any element $B \in sl(2n, \mathbb{R})$ can be written uniquely in the form

$$B = JS + JA \quad (\text{P.4.16})$$

where S is symmetric, A is antisymmetric, and JA is traceless. The elements JS are a basis for $sp(2n, \mathbb{R})$, and together with the elements of the form JA form a basis for $sl(2n, \mathbb{R})$. Computation gives the results

$$\sigma(JS) = -J(JS)^T J^{-1} = -JS^T J^T J^{-1} = -JS(-J)J^{-1} = JS, \quad (\text{P.4.17})$$

$$\sigma(JA) = -J(JA)^T J^{-1} = -JA^T J^T J^{-1} = JA(-J)J^{-1} = -JA. \quad (\text{P.4.18})$$

We see from (4.17) that $sp(2n, \mathbb{R})$ is invariant under σ . That is, σ acts as the identity on $sp(2n, \mathbb{R})$. This is the local consequence of the global result that Σ acts as the identity on $Sp(2n, \mathbb{R})$. And, since σ is manifestly linear, we have

$$\sigma(B) = \sigma(JS + JA) = \sigma(JS) + \sigma(JA) = JS - JA. \quad (\text{P.4.19})$$

From (4.16) through (4.19) we find that

$$\sigma^2(B) = \sigma[\sigma(B)] = B. \quad (\text{P.4.20})$$

Thus, σ is an involution on $sl(2n, \mathbb{R})$.

P.5 Connection between Symmetries and Being J -Symmetric

A matrix Q is called J -symmetric if it satisfies the condition

$$JQ^T J^{-1} = Q, \quad (\text{P.5.1})$$

which is equivalent to the condition

$$\sigma(Q) = -Q. \quad (\text{P.5.2})$$

Suppose we represent Q in the form

$$Q = JX \quad (\text{P.5.3})$$

where the properties of X are yet to be determined. Then we find

$$\sigma(Q) = \sigma(JX) = -J(JX)^T J^{-1} = -JX^T J^T J^{-1} = JX^T. \quad (\text{P.5.4})$$

Thus, in this representation, the J -symmetric condition (5.2) yields the requirement

$$JX^T = -JX, \quad (\text{P.5.5})$$

from which it follows that

$$X^T = -X. \quad (\text{P.5.6})$$

That is,

$$Q = JA' \quad (\text{P.5.7})$$

where A' is antisymmetric.

Suppose we instead represent Q in the form

$$Q = \exp(JX). \quad (\text{P.5.8})$$

Then we find the relation

$$\begin{aligned} \sigma(Q) &= \sigma[\exp(JX)] = -J[\exp(JX)]^T J^{-1} = -J \exp[(JX)^T] J^{-1} \\ &= -J \exp(-X^T J) J^{-1} = -\exp(-JX^T). \end{aligned} \quad (\text{P.5.9})$$

In this context requiring (5.2) produces the relation

$$\exp(JX) = \exp(-JX^T), \quad (\text{P.5.10})$$

which again produces (5.5) and hence (5.6) and consequently

$$Q = \exp(JA) \quad (\text{P.5.11})$$

where A is antisymmetric.

As a side comment, we have discovered, upon comparing (5.7) and (5.11), the relation

$$JA' = \exp(JA), \quad (\text{P.5.12})$$

or,

$$A' = -J \exp(JA), \quad (\text{P.5.13})$$

which maps antisymmetric matrices A into antisymmetric matrices A' . Keeping the first few terms in the power series we find that

$$\begin{aligned} A' &= -J[I + JA + (JA)^2/2! + (JA)^3/3! + \dots] \\ &= -J + A - J(JA)^2/2! - J(JA)^3/3! + \dots \\ &= -J + A + AJA/2! + AJAJA/3! + \dots. \end{aligned} \quad (\text{P.5.14})$$

In this form we see that *any* Taylor series in JA would have the same mapping property.

Finally, let us apply Σ to Q as given by (5.11). We find the result

$$\Sigma(Q) = \Sigma[\exp(JA)] = \exp[\sigma(JA)] = \exp(-JA) = Q^{-1} \quad (\text{P.5.15})$$

as before.

P.6 Relation to Darboux Matrices

According to *Lie Methods* the matrix $N(M)$ is J -symmetric, and we seek a J -symmetric matrix Q such that

$$Q^2 = N(M). \quad (\text{P.6.1})$$

Use the result of Lemma 3.6 of *Lie Methods* to write the representations

$$N(M) = JA' \quad (\text{P.6.2})$$

and

$$Q = JA \quad (\text{P.6.3})$$

where A' and A are antisymmetric. With these representations (6.1) becomes

$$JAJA = JA', \quad (\text{P.6.4})$$

which yields the relation

$$A' = AJA. \quad (\text{P.6.5})$$

Since A is assumed to be antisymmetric, (6.5) can also be written in the form

$$-A' = AJA^T, \quad (\text{P.6.6})$$

which shows that A , if it exists, is a Darboux matrix connecting $-A'$ and J . Thus, the problem of finding Q is equivalent to showing that it is possible to find a Darboux matrix connecting $-A'$ and J that is also antisymmetric.

P.7 Some Observations on $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$

By its nature, $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ is a homogeneous space. That is, $SL(2n, \mathbb{R})$ when acting on this space can send any point into any other point. See Section 5.12 for a description of group action on cosets. Between the JS and the JA there are the relations (4.3.2) through (4.3.4). Consequently $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ is a reductive homogeneous space. Since σ is an involution on $sl(2n, \mathbb{R})$ that leaves $sp(2n, \mathbb{R})$ invariant, it follows that $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ is a symmetric space. Moreover, at least for the cases $n = 2$ and $n = 3$ and presumably for all n , the elements of the form JA with JA traceless transform irreducibly under the action of $sp(2n, \mathbb{R})$. Therefore $SL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ is an irreducible symmetric space. For example, in the case $n = 2$ the elements of the form JA with JA traceless carry the irreducible representation $\Gamma(0, 1)$ of $sp(4, \mathbb{R})$; and in the case $n = 3$ they carry the irreducible representation $\Gamma(0, 1, 0)$ of $sp(6, \mathbb{R})$. See the weight diagrams 27.5.4 and 27.8.4 in Chapter 27.

Digesting Goodman Notes

P.8 Action of σ on $sl(2n, \mathbb{R})$

Consider the action of σ on $sl(2n, \mathbb{R})$. First we see that, for a real symmetric matrix,

$$\sigma(S) = -JS^TJ^{-1} = -JSJ^{-1} = JSJ \quad (\text{P.8.1})$$

so that

$$[\sigma(S)]^T = [JSJ]^T = J^T S^T J^T = JSJ = \sigma(S). \quad (\text{P.8.2})$$

Therefore, σ maps the space of real symmetric matrices into itself.

Next suppose B and B' are two matrices in $s\ell(2n, \mathbb{R})$. Define their inner product to be

$$(B, B') = \text{tr}(BB'). \quad (\text{P.8.3})$$

Then we find that σ preserves the inner product,

$$\begin{aligned} (\sigma(B), \sigma(B')) &= \text{tr}[(-J)B^T J^{-1}(-J)(B')^T J^{-1}] = \text{tr}[JB^T (B')^T J^{-1}] = \text{tr}[B^T (B')^T] \\ &= \text{tr}[(B'B)^T] = \text{tr}[B'B] = \text{tr}[BB'] = (B, B'). \end{aligned} \quad (\text{P.8.4})$$

Suppose we write for any real symmetric matrix S the decomposition

$$S = S^a + S^c \quad (\text{P.8.5})$$

where S^a anticommutes with J and S^c commutes with J . See Section 3.8 of *Lie Methods*. Then we find that

$$\sigma(S^a) = -JS^aJ^{-1} = S^aJJ^{-1} = S^a, \quad (\text{P.8.6})$$

and

$$\sigma(S^c) = -JS^cJ^{-1} = -S^cJJ^{-1} = -S^c. \quad (\text{P.8.7})$$

We see that σ , which we already know is a linear operator that maps the space of real symmetric matrices into itself, has eigenvalues ± 1 . This is to be expected because σ is an involution.

As an application of this result, we find that

$$(S^a, S^c) = -(\sigma(S^a), \sigma(S^c)) = -(S^a, S^c) \quad (\text{P.8.8})$$

from which it follows that

$$(S^a, S^c) = 0 \quad (\text{P.8.9})$$

for any S^a, S^c pair.

P.9 Lie Triple System

Let S , S' , and S'' be real symmetric matrices. Then we have

$$\{S', S\} = A \quad (\text{P.9.1})$$

where A is an antisymmetric matrix. And,

$$\{S'', \{S', S\}\} = S''' \quad (\text{P.9.2})$$

where S''' is again a symmetric matrix. Thus symmetric matrices comprise a Lie triple system.

Let S^a , $S^{a'}$, and $S^{a''}$ be real symmetric matrices that anticommute with J . Then we have

$$\{S^{a'}, S^a\} = A^c \quad (\text{P.9.3})$$

where A^c is an antisymmetric matrix that commutes with J . And,

$$\{S^{a''}, \{S^{a'}, S^a\}\} = S^{a'''}$$
 (P.9.4)

where $S^{a'''}$ is again a symmetric matrix that anticommutes with J . Thus symmetric matrices that anticommute with J also comprise a Lie triple system.

Similarly, it can be verified that real symmetric matrices of the form S^c comprise a Lie triple system,

$$\{S^{c''}, \{S^{c'}, S^c\}\} = S^{c'''}$$
 (P.9.5)

where $S^{c'''}$ is again a symmetric matrix that commutes with J .

P.10 A Factorization Theorem (Theorem 1.1 of Goodman)

P.10.1 A Particular Mapping from Real Symmetric Matrices to Positive-Definite Matrices

Define for any S an associated matrix $P(S)$ by the rule

$$P(S) = \exp(S^a) \exp(S^c) \exp(S^a).$$
 (P.10.1)

Evidently, P is real, symmetric and nonsingular. It is also positive definite because we have

$$\begin{aligned} (v, Pv) &= (v, \exp(S^a) \exp(S^c) \exp(S^a)v) = (\exp(S^a)v, \exp(S^c) \exp(S^a)v) \\ &= ([\exp(S^a)v], \exp(S^c)[\exp(S^a)v]) > 0 \text{ if } v \neq 0 \end{aligned}$$
 (P.10.2)

because $\exp(S^a)$ is symmetric and invertible and $\exp(S^c)$ is positive definite.

We will eventually see that the map (10.1) is invertible. That is, given any real symmetric positive-definite matrix P , there are (unique) matrices S^a and S^c such that (10.1) holds. Thus, (10.1) provides a factorization of any real symmetric positive-definite matrix P .

P.10.2 The Map Is Real Analytic

Evidently, by the nature of the exponential function, $P(S)$ is a real analytic function of S^a and S^c . We claim that $P(S)$ is also an analytic function of S . Note that

$$S^a = (S - J^{-1}SJ)/2$$
 (P.10.3)

and

$$S^c = (S + J^{-1}SJ)/2$$
 (P.10.4)

Therefore S^a and S^c are analytic functions of S . Since the various exponential functions appearing in (10.1) are analytic functions of their arguments, it follows that $P(S)$ is an analytic function of S .

P.10.3 Trace and Determinant Properties

From (10.3) we find

$$\begin{aligned}\text{tr}(S^a) &= (1/2)[\text{tr}(S) - \text{tr}(J^{-1}SJ)] = (1/2)[\text{tr}(S) - \text{tr}(JJ^{-1}S)] \\ &= (1/2)[\text{tr}(S) - \text{tr}(S)] = 0.\end{aligned}\tag{P.10.5}$$

From (10.4) we find

$$\begin{aligned}\text{tr}(S^c) &= (1/2)[\text{tr}(S) + \text{tr}(J^{-1}SJ)] = (1/2)[\text{tr}(S) + \text{tr}(JJ^{-1}S)] \\ &= (1/2)[\text{tr}(S) + \text{tr}(S)] = \text{tr}(S).\end{aligned}\tag{P.10.6}$$

Now take the determinant of both sides of (10.1). So doing gives the result

$$\det[P(S)] = \exp[\text{tr}(S^a)] \exp[\text{tr}(S^c)] \exp[\text{tr}(S^a)] = \exp[\text{tr}(S)].\tag{P.10.7}$$

P.10.4 Study of the Inverse of the Map

Conversely, it is claimed that S is an analytic function of P . If true, then, by (10.3) and (10.4), S^a and S^c are also analytic functions of P . Proceed as follows: Since P is real, symmetric, and positive definite, there is a real symmetric matrix Z such that

$$P(S) = \exp(Z) = \exp(S^a) \exp(S^c) \exp(S^a),\tag{P.10.8}$$

and Z will be analytic in P and therefore in S . What we want to do is find S^a and S^c in terms of Z .

P.10.5 Formula for S^a in terms of Z

Begin by finding S^a in terms of Z . Apply Σ to both sides of (10.8) to find the result

$$\Sigma[P(S)] = \Sigma[\exp(Z)] = \Sigma[\exp(S^a)]\Sigma[\exp(S^c)]\Sigma[\exp(S^a)],\tag{P.10.9}$$

from which it follows that

$$\exp[\sigma(Z)] = \exp[\sigma(S^a)] \exp[\sigma(S^c)] \exp[\sigma(S^a)],\tag{P.10.10}$$

from which it follows that

$$\exp[\sigma(Z)] = \exp(S^a) \exp(-S^c) \exp(S^a).\tag{P.10.11}$$

Next take inverses of both sides of (10.8) to find the relation

$$\exp(-Z) = \exp(-S^a) \exp(-S^c) \exp(-S^a),\tag{P.10.12}$$

from which it follows that

$$\exp(-S^c) = \exp(S^a) \exp(-Z) \exp(S^a).\tag{P.10.13}$$

Use (10.13) in (10.11) to get the relation

$$\exp[\sigma(Z)] = \exp(2S^a) \exp(-Z) \exp(2S^a), \quad (\text{P.10.14})$$

which is a relation between Z and S^a .

Next show that, given Z , (10.14) has, in fact, a unique solution S^a . In particular, we want to verify the assertion

$$\exp(2S^a) = \exp(Z/2) \exp(T) \exp(Z/2) \quad (\text{P.10.15})$$

where

$$\exp(2T) = \exp(-Z/2) \exp[\sigma(Z)] \exp(-Z/2). \quad (\text{P.10.16})$$

Note that the right side of (10.16) is real, symmetric, and positive definite. Therefore T and consequently $\exp(T)$ are well defined (real analytic) functions of Z . Correspondingly, the right side of (10.15) is well defined, real, symmetric, and positive definite. Therefore S^a is well defined, and a real analytic function of Z . We also observe that (10.16) can be rewritten in the form

$$\exp[\sigma(Z)] = \exp(Z/2) \exp(2T) \exp(Z/2). \quad (\text{P.10.17})$$

To prove the assertion, take (10.15) and (10.16) to be the definition of S^a . Then, using (10.15), we find that

$$\begin{aligned} \exp(2S^a) \exp(-Z) \exp(2S^a) &= \\ \exp(Z/2) \exp(T) \exp(Z/2) \exp(-Z) \exp(Z/2) \exp(T) \exp(Z/2) &= \\ \exp(Z/2) \exp(2T) \exp(Z/2) &= \exp[\sigma(Z)]. \end{aligned} \quad (\text{P.10.18})$$

Here, in the last step, we have also used (10.17). Thus, we see that (10.14) is satisfied.

P.10.6 Uniqueness of Solution for S^a

What about uniqueness? Suppose \hat{S}^a also satisfies (10.14). That is, assume

$$\exp[\sigma(Z)] = \exp(2\hat{S}^a) \exp(-Z) \exp(2\hat{S}^a). \quad (\text{P.10.19})$$

Substitute (10.19) into (10.16) to get

$$\begin{aligned} \exp(2T) &= \exp(-Z/2) \exp[\sigma(Z)] \exp(-Z/2) \\ &= \exp(-Z/2) \{\exp(2\hat{S}^a) \exp(-Z) \exp(2\hat{S}^a)\} \exp(-Z/2) \\ &= \exp(-Z/2) \{\exp(2\hat{S}^a) \exp(-Z/2) \exp(-Z/2) \exp(2\hat{S}^a)\} \exp(-Z/2) \\ &= [\exp(-Z/2) \exp(2\hat{S}^a) \exp(-Z/2)]^2. \end{aligned} \quad (\text{P.10.20})$$

Therefore, by the uniqueness of the positive-definite square root of a positive-definite matrix, we have

$$\exp(T) = \exp(-Z/2) \exp(2\hat{S}^a) \exp(-Z/2), \quad (\text{P.10.21})$$

which can be rewritten in the form

$$\exp(2\hat{S}^a) = \exp(Z/2) \exp(T) \exp(Z/2). \quad (\text{P.10.22})$$

Now compare (10.15) and (10.22) to get

$$\exp(2S^a) = \exp(2\hat{S}^a), \quad (\text{P.10.23})$$

from which it follows that

$$S^a = \hat{S}^a. \quad (\text{P.10.24})$$

P.10.7 Verification of Expected Symmetry for S^a

Also, does the S^a just found satisfy (8.6)? Apply Σ to both sides of (10.14) to get the relation

$$\exp\{\sigma[\sigma(Z)]\} = \exp[2\sigma(S^a)] \exp[-\sigma(Z)] \exp[2\sigma(S^a)], \quad (\text{P.10.25})$$

form which it follows by (4.20) that

$$\exp(Z) = \exp[2\sigma(S^a)] \exp[-\sigma(Z)] \exp[2\sigma(S^a)]. \quad (\text{P.10.26})$$

Rewrite (10.26) in the form

$$\exp[-2\sigma(S^a)] \exp(Z) \exp[-2\sigma(S^a)] = \exp[-\sigma(Z)]. \quad (\text{P.10.27})$$

Now invert both sides of (10.27) to get

$$\exp[\sigma(Z)] = \exp[2\sigma(S^a)] \exp(-Z) \exp[2\sigma(S^a)]. \quad (\text{P.10.28})$$

Upon comparing (10.14) and (10.28) we see that S^a and $\sigma(S^a)$ obey the same equation. Therefore, from the uniqueness of the solution, we have

$$\sigma(S^a) = S^a, \quad (\text{P.10.29})$$

which is (8.6).

P.10.8 Formula for S^c in Terms of Z

The last thing to do is, given Z , find S^c . Look at (10.8). It can be rewritten in the form

$$\exp(S^c) = \exp(-S^a) \exp(Z) \exp(-S^a). \quad (\text{P.10.30})$$

And, since S^a is now known, we may regard (10.30) as a formula for S^c .

P.10.9 Verification of Expected Symmetry for S^c

However, it would be good to check that (8.7) holds. Apply Σ to both sides of (10.30) and manipulate to find

$$\begin{aligned}\exp[\sigma(S^c)] &= \exp[-\sigma(S^a)] \exp[\sigma(Z)] \exp[-\sigma(S^a)] \\ &= \exp(-S^a) \exp[\sigma(Z)] \exp(-S^a) \\ &= \exp(-S^a) \{\exp(2S^a) \exp(-Z) \exp(2S^a)\} \exp(-S^a) \\ &= \exp(S^a) \exp(-Z) \exp(S^a) \\ &= \exp(-S^c),\end{aligned}\tag{P.10.31}$$

from which it follows that

$$\sigma(S^c) = -S^c,\tag{P.10.32}$$

as required by (8.7). Here we used (10.14) and the relation

$$\exp(-S^c) = \exp(S^a) \exp(-Z) \exp(S^a)\tag{P.10.33}$$

which follows from (10.30).

P.10.10 Conclusion

In summary, we have learned that both S^a and S^c are real-analytic functions of Z .

P.10.11 Motivation for Mapping

Suppose P is a real, symmetric, and positive-definite matrix. Use it to define a matrix Q by the relation

$$Q = P^{-1/2} J P^{-1/2}.\tag{P.10.34}$$

(Note that Goodman defines Q by $Q = P^{1/2} J P^{-1/2}$, but presumably this is a misprint.) By calculation we find that

$$Q^T = -P^{-1/2} J P^{-1/2} = -Q.\tag{P.10.35}$$

Evidently Q is real, antisymmetric, and nonsingular. Then \hat{P} given by

$$\hat{P} = Q^T Q\tag{P.10.36}$$

will be real, symmetric, and positive definite. Goodman claims that

$$P = \exp(X) \exp(Y) \exp(X)\tag{P.10.37}$$

with

$$\exp(X) = (P^{1/2} \hat{P}^{1/2} P^{1/2})^{1/2}\tag{P.10.38}$$

and

$$\exp(Y) = \exp(-X) P \exp(-X).\tag{P.10.39}$$

Let us see if this is true. Evidently (10.37) and (10.39) are logically equivalent for any matrices P , X , and Y . So, perhaps we should examine the properties of X and Y .

Evidently X is real analytic in P . Squaring both sides of (10.38) gives

$$\exp(2X) = P^{1/2} \hat{P}^{1/2} P^{1/2}. \quad (\text{P.10.40})$$

Define Z by writing

$$\exp(Z) = P. \quad (\text{P.10.41})$$

Evidently Z is real analytic in P and

$$\exp(Z/2) = P^{1/2}. \quad (\text{P.10.42})$$

With this definition, (10.40) can be rewritten in the form

$$\exp(2X) = \exp(Z/2) \hat{P}^{1/2} \exp(Z/2). \quad (\text{P.10.43})$$

Next, define T by writing

$$\exp(T) = \hat{P}^{1/2}. \quad (\text{P.10.44})$$

Then we have the result

$$\exp(2T) = \hat{P}. \quad (\text{P.10.45})$$

Also, (10.43) can now be written in the form

$$\exp(2X) = \exp(Z/2) \exp(T) \exp(Z/2). \quad (\text{P.10.46})$$

Now work out an expression for \hat{P} . From (10.34) through (10.36) we find that

$$\begin{aligned} \hat{P} &= -P^{-1/2} J P^{-1/2} P^{-1/2} J P^{-1/2} = P^{-1/2} J P^{-1} J^{-1} P^{-1/2} \\ &= \exp(-Z/2) J \exp(-Z) J^{-1} \exp(-Z/2) = \exp(-Z/2) \exp(-J Z J^{-1}) \exp(-Z/2) \\ &= \exp(-Z/2) \exp[\sigma(Z)] \exp(-Z/2). \end{aligned} \quad (\text{P.10.47})$$

Thus, we get the result

$$\exp(2T) = \exp(-Z/2) \exp[\sigma(Z)] \exp(-Z/2). \quad (\text{P.10.48})$$

We see that (10.46) is the counterpart to (10.22), and (10.48) is the counterpart to (10.16). Therefore we have the relations

$$X = S^a \quad (\text{P.10.49})$$

and

$$Y = S^c. \quad (\text{P.10.50})$$

P.11 Theorem 1.2 of Goodman Due to Mostow

Consider the triplet $\{k \in O(2n, \mathbb{R}), S^c, S^a\}$. Use it to construct $g \in GL(2n, \mathbb{R})$ by the rule

$$g = k \exp(S^c) \exp(S^a). \quad (\text{P.11.1})$$

It is claimed that (11.1) provides an analytic isomorphism between the triplet and $GL(2n, \mathbb{R})$.

Let us pause to do a dimension count. The dimension of S^c plus the dimension of S^a is the dimension of all symmetric matrices S . And the dimension of $O(2n, \mathbb{R})$ is the dimension of all antisymmetric matrices A . Taken together, these dimensions add up to the dimension of all $2n \times 2n$ matrices, which is just the dimension of $GL(2n, \mathbb{R})$.

To continue, first we verify that (11.1) is injective. That is, different triplets must yield different elements in $GL(2n, \mathbb{R})$. For suppose that two triplets yield the same element $g \in GL(2n, \mathbb{R})$,

$$g = k_1 \exp(S_1^c) \exp(S_1^a) \quad (\text{P.11.2})$$

and

$$g = k_2 \exp(S_2^c) \exp(S_2^a). \quad (\text{P.11.3})$$

From (11.2) we find

$$g^T g = \exp(S_1^a) \exp(S_1^c) k_1^T k_1 \exp(S_1^c) \exp(S_1^a) = \exp(S_1^a) \exp(2S_1^c) \exp(S_1^a), \quad (\text{P.11.4})$$

and from (11.3) we find

$$g^T g = \exp(S_2^a) \exp(S_2^c) k_2^T k_2 \exp(S_2^c) \exp(S_2^a) = \exp(S_2^a) \exp(2S_2^c) \exp(S_2^a). \quad (\text{P.11.5})$$

Thus, we have the relation

$$\exp(S_1^a) \exp(2S_1^c) \exp(S_1^a) = \exp(S_2^a) \exp(2S_2^c) \exp(S_2^a). \quad (\text{P.11.6})$$

From Theorem 1.1 and (11.6) we conclude that

$$S_1^a = S_2^a \quad (\text{P.11.7})$$

and

$$S_1^c = S_2^c. \quad (\text{P.11.8})$$

Then, from (11.2) and (11.3), we see that

$$k_1 = k_2. \quad (\text{P.11.9})$$

Next, we verify that any $g \in GL(2n, \mathbb{R})$ can be written in the form (11.1). Given any $g \in GL(2n, \mathbb{R})$, define a real symmetric positive-definite matrix P by the rule

$$P = g^T g. \quad (\text{P.11.10})$$

Evidently P is real analytic in g . Therefore, by the factorization theorem, there are unique matrices S^a and S^c , that depend real-analytically on P and therefore real-analytically on g , such that

$$P = \exp(S^a) \exp(2S^c) \exp(S^a). \quad (\text{P.11.11})$$

Now define k by the rule

$$k = (g^T)^{-1} \exp(S^a) \exp(S^c). \quad (\text{P.11.12})$$

Then k is also real-analytic in g . Moreover, we find that

$$k^T = \exp(S^c) \exp(S^a) (g)^{-1}, \quad (\text{P.11.13})$$

from which it follows that

$$\begin{aligned}
 k^T k &= \exp(S^c) \exp(S^a) (g)^{-1} (g^T)^{-1} \exp(S^a) \exp(S^c) \\
 &= \exp(S^c) \exp(S^a) P^{-1} \exp(S^a) \exp(S^c) \\
 &= \exp(S^c) \exp(S^a) \exp(-S^a) \exp(-2S^c) \exp(-S^a) \exp(S^a) \exp(S^c) \\
 &= \exp(S^c) \exp(-2S^c) \exp(S^c) = I.
 \end{aligned} \tag{P.11.14}$$

Therefore $k \in O(2n, \mathbb{R})$.

Finally, suppose that $g \in SL(2n, \mathbb{R})$. In this case, take the determinant of both sides of (11.1) to get the result

$$1 = \det(g) = \det(k) \exp[\text{tr}(S^c)] \exp[\text{tr}(S^a)] = \det(k) \exp[\text{tr}(S^c)] \tag{P.11.15}$$

where we have used (10.5). We know that

$$\det(k) = \pm 1. \tag{P.11.16}$$

We see that, in order for (11.15) to be satisfied, we must have the relations

$$\det(k) = +1 \text{ so that } k \in SO(2n, \mathbb{R}) \tag{P.11.17}$$

and

$$\text{tr}(S^c) = 0. \tag{P.11.18}$$

We conclude the following: Consider the triplet $\{k \in SO(2n, \mathbb{R}), S^c \text{ with } \text{tr}(S^c) = 0, S^a\}$. Use it to construct $g \in SL(2n, \mathbb{R})$ by the rule

$$g = k \exp(S^c) \exp(S^a). \tag{P.11.19}$$

Then (11.19) provides an analytic isomorphism between the triplet and $SL(2n, \mathbb{R})$.

P.12 Goodman's Work on Symplectic Polar Decomposition

Consider the group $G = SL(2n, \mathbb{R})$, and its subgroups $H = Sp(2n, \mathbb{R})$ and $K = SO(2n, \mathbb{R})$. We want to study the coset space G/H .

P.12.1 Some More Symmetry Operations

To do so it is useful to introduce some additional symmetry operations on G . The operation Σ has already been defined by (4.1). We will define two more.

Introduce the operation Θ by the rule

$$\Theta(M) = (M^T)^{-1} \tag{P.12.1}$$

for any $M \in G$. Evidently this map preserves the condition $\det(M) = 1$, and therefore sends G unto itself. Also we find that

$$\Theta(I) = I, \tag{P.12.2}$$

$$\Theta(J) = J, \quad (\text{P.12.3})$$

$$\Theta(M_1 M_2) = [(M_1 M_2)^T]^{-1} = [(M_2^T M_1^T)]^{-1} = (M_1^T)^{-1} (M_2^T)^{-1} = \Theta(M_1) \Theta(M_2), \quad (\text{P.12.4})$$

$$\Theta[\Theta(M)] = \{[(M^T)^{-1}]^T\}^{-1} = M. \quad (\text{P.12.5})$$

Thus Θ , like Σ , is an isomorphism and an involution.

Next, we discover that Σ and Θ commute. From the definition of Σ we find that

$$\Sigma[\Theta(M)] = J\{\Theta(M)\}^{-1}J^{-1}. \quad (\text{P.12.6})$$

But, from (12.1),

$$\{\Theta(M)\}^{-1} = \{[(M^T)^{-1}]^T\}^{-1} = M. \quad (\text{P.12.7})$$

Therefore, we conclude that

$$\Sigma[\Theta(M)] = JM J^{-1}. \quad (\text{P.12.8})$$

Applying the symmetries in opposite order gives

$$\Theta[\Sigma(M)] = \{\Sigma(M)\}^{-1}. \quad (\text{P.12.9})$$

But, from (4.1), we have the relations

$$[\Sigma(M)]^T = [J(M^T)^{-1}J^{-1}]^T = JM^{-1}J^{-1} \quad (\text{P.12.10})$$

and

$$\{\Sigma(M)\}^{-1} = [JM^{-1}J^{-1}]^{-1} = JM J^{-1}. \quad (\text{P.12.11})$$

It follows that

$$\Theta[\Sigma(M)] = JM J^{-1}. \quad (\text{P.12.12})$$

We see that the right sides of (12.8) and (12.12) agree, and therefore

$$\Sigma[\Theta(M)] = \Theta[\Sigma(M)] \quad (\text{P.12.13})$$

or, in operator notation,

$$\Sigma\Theta = \Theta\Sigma. \quad (\text{P.12.14})$$

Next define the operation Υ as the product

$$\Upsilon = \Sigma\Theta = \Theta\Sigma. \quad (\text{P.12.15})$$

From (12.8) or (12.2) we find that

$$\Upsilon(M) = JM J^{-1}. \quad (\text{P.12.16})$$

We see that Υ is also an isomorphism. And, from (12.16), we see that

$$\Upsilon[\Upsilon(M)] = J[JMJ^{-1}]J^{-1} = M, \quad (\text{P.12.17})$$

and therefore Υ is an involution as is expected for the product of commuting involutions.

Let θ and τ be the associated induced maps in the Lie algebra defined by

$$\Theta[\exp(B)] = \exp[\theta(B)], \quad (\text{P.12.18})$$

$$\Upsilon[\exp(B)] = \exp[\tau(B)]. \quad (\text{P.12.19})$$

For (12.18) we find

$$\Theta[\exp(B)] = \{[\exp(B)]^T\}^{-1} = [\exp(B^T)]^{-1} = \exp(-B^T), \quad (\text{P.12.20})$$

and conclude that

$$\theta(B) = -B^T. \quad (\text{P.12.21})$$

For (12.19) we find

$$\Upsilon[\exp(B)] = J \exp(B) J^{-1} = \exp(JBJ^{-1}), \quad (\text{P.12.22})$$

and conclude that

$$\tau(B) = JBJ^{-1}. \quad (\text{P.12.23})$$

Let us check some expected relations: First, we find the results

$$\theta[\theta(B)] = \theta[-B^T] = -[-B^T]^T = B \quad (\text{P.12.24})$$

so that θ , like σ , is an involution, as expected. Second, we find the results

$$\sigma[\theta(B)] = \sigma[-B^T] = -J[-B^T]^T J^{-1} = JBJ^{-1} = \tau(B), \quad (\text{P.12.25})$$

$$\theta[\sigma(B)] = \theta[-JB^TJ^{-1}] = -[-JB^TJ^{-1}]^T = JBJ^{-1} = \tau(B). \quad (\text{P.12.26})$$

Thus, we conclude that

$$\sigma\theta = \theta\sigma = \tau. \quad (\text{P.12.27})$$

It follows that τ is also an involution, as is also obvious from (12.22). Moreover, we note the relations

$$\sigma\tau = \tau\sigma = \theta, \quad (\text{P.12.28})$$

$$\theta\tau = \tau\theta = \sigma. \quad (\text{P.12.29})$$

Finally we should check the effects of θ and τ on scalar products. First we see that, for a real symmetric matrix,

$$\theta(S) = -S^T = -S \quad (\text{P.12.30})$$

so that

$$[\theta(S)]^T = -S^T = -S = \theta(S). \quad (\text{P.12.31})$$

Therefore, θ maps the space of real symmetric matrices into itself.

Next suppose B and B' are two matrices in $sl(2n, \mathbb{R})$. As before, define their inner product to be

$$(B, B') = \text{tr}(BB'). \quad (\text{P.12.32})$$

Then we find that θ preserves the inner product,

$$\begin{aligned} (\theta(B), \theta(B')) &= \text{tr}[(-B^T)(-B')^T] = \text{tr}[B^T(B')^T] = \text{tr}[(B'B)^T] \\ &= \text{tr}[B'B] = \text{tr}[BB'] = (B, B'). \end{aligned} \quad (\text{P.12.33})$$

Using (12.25), because σ and θ preserve the inner product, we see that τ also preserves the inner product,

$$(\tau(B), \tau(B')) = (\sigma[\theta(B)], \sigma[\theta(B')]) = ([\theta(B)], [\theta(B')]) = (B, B'). \quad (\text{P.12.34})$$

P.12.2 Fixed-Point Subgroups Associated with Symmetry Operations

In Section 4 we found that the fixed points of Σ comprise the subgroup $Sp(2n, \mathbb{R})$. Now we will find that the fixed points of Θ and Υ also yield subgroups of $GL(2n, \mathbb{R})$.

Suppose g is a fixed point of Θ . Then we find that

$$\Theta(g) = g \Leftrightarrow (g^T)^{-1} = g \Leftrightarrow g^T g = I. \quad (\text{P.12.35})$$

Thus, the fixed points of Θ comprise $SO(2n, \mathbb{R})$. [Here we have already assumed $g \in SL(2n, \mathbb{R})$ so that we know that $\det g = 1$. Otherwise the fixed points of Θ comprise $O(2n, \mathbb{R})$.]

Suppose g is a fixed point of Υ . Then we find that

$$\Upsilon(g) = g \Leftrightarrow JgJ^{-1} = g \Leftrightarrow Jg = gJ. \quad (\text{P.12.36})$$

Thus, the fixed points of Υ comprise the matrices in $GL(2n, \mathbb{R})$ that commute with J . They also obviously form a group. But what is this group?

Write g in the block form

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad (\text{P.12.37})$$

where the matrices a, b, c , and d are real and $n \times n$. Then we find the results

$$Jg = \begin{pmatrix} c & d \\ -a & -b \end{pmatrix}, \quad (\text{P.12.38})$$

and

$$gJ = \begin{pmatrix} -b & a \\ -d & c \end{pmatrix}. \quad (\text{P.12.39})$$

Therefore, requiring that g commute with J yields the restrictions

$$c = -b \quad (\text{P.12.40})$$

and

$$d = a. \quad (\text{P.12.41})$$

Thus, g is of the form

$$g = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}. \quad (\text{P.12.42})$$

Next define matrices A and B by the rules

$$A = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \quad (\text{P.12.43})$$

and

$$B = \begin{pmatrix} b & 0 \\ 0 & b \end{pmatrix}. \quad (\text{P.12.44})$$

Then both A and B commute with J , and we also have the relation

$$JB = \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}. \quad (\text{P.12.45})$$

Therefore, we may also write

$$g = A + JB. \quad (\text{P.12.46})$$

Suppose g_1 and g_2 are two matrices that commute with J and we use the representation (12.46) to write

$$g_k = A_k + JB_k \quad (\text{P.12.47})$$

Then, recalling that the A_k and B_k commute with J and that $J^2 = -I$, we find the product relation

$$g_1 g_2 = (A_1 A_2 - B_1 B_2) + J(A_1 B_2 + B_1 A_2). \quad (\text{P.12.48})$$

We see that, in (12.47) and (12.48), the matrix J plays a role analogous to the imaginary number i .

This analogy can be made explicit using the machinery of Section 3.9 of *Lie Methods*. Suppose m is an arbitrary $n \times n$ matrix with possibly complex entries. Evidently it can be written in the form

$$m = a + ib \quad (\text{P.12.49})$$

where a and b are real $n \times n$ matrices. Let us multiply two such matrices together. So doing gives the result

$$m_1 m_2 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + b_1 a_2). \quad (\text{P.12.50})$$

Note the resemblance between the pairs (12.46), (12.49) and (12.48), (12.50).

To pursue the analogy further, let W be the unitary and (complex) symplectic matrix

$$W = \frac{1}{\sqrt{2}} \begin{pmatrix} I & iI \\ iI & I \end{pmatrix}. \quad (\text{P.12.51})$$

Here each block in W is $n \times n$. Now, as in Section 3.9 of *Lie Methods*, define an associated $2n \times 2n$ matrix $g(m)$ by the rule

$$g(m) = M(m) = W \begin{pmatrix} m & 0 \\ 0 & \bar{m} \end{pmatrix} W^{-1}. \quad (\text{P.12.52})$$

Then it is easily verified that there are the relations

$$g(I) = I, \quad (\text{P.12.53})$$

$$g(m_1 m_2) = g(m_1)g(m_2), \quad (\text{P.12.54})$$

$$g(m^{-1}) = g^{-1}(m). \quad (\text{P.12.55})$$

Also, if (12.52) is multiplied out explicitly, we find the result

$$g(m) = \begin{pmatrix} \operatorname{Re}(m) & \operatorname{Im}(m) \\ -\operatorname{Im}(m) & \operatorname{Re}(m) \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}. \quad (\text{P.12.56})$$

It follows that $g(m)$ is real for any m .

Matrices of the form (12.49) constitute the group $SL(n, \mathbb{C})$ provided we add the condition

$$\det(m) = 1. \quad (\text{P.12.57})$$

Now take the determinant of both sides of (12.52). Doing so gives the result

$$\begin{aligned} \det(g) &= [\det(W)][\det(m)][\det(\bar{m})][\det(W^{-1})] \\ &= [\det(m)][\det(\bar{m})] = |\det(m)|^2 \geq 0. \end{aligned} \quad (\text{P.12.58})$$

If (12.57) holds, then from (12.58) we also have the condition

$$\det(g) = 1. \quad (\text{P.12.59})$$

From (12.49) through (12.59) we conclude that the set of matrices $g \in SL(2n, \mathbb{R})$ that also commute with J constitutes a group that is isomorphic to $SL(n, \mathbb{C})$. More precisely, the set of matrices $g \in SL(2n, \mathbb{R})$ that also commute with J constitutes a group that is the representation $SL(n, \mathbb{C}) \oplus \overline{SL(n, \mathbb{C})}$ of $SL(n, \mathbb{C})$. If we relax the determinant condition, we conclude that the set of matrices $g \in GL(2n, \mathbb{R}, +)$ that also commute with J constitutes a group that is the representation $GL(n, \mathbb{C}) \oplus \overline{GL(n, \mathbb{C})}$ of $GL(n, \mathbb{C})$.

To summarize, let G^Σ be the fixed-point group associated with the symmetry Σ . Then we have the result

$$G^\Sigma = Sp(2n, \mathbb{R}) = H. \quad (\text{P.12.60})$$

Similarly, we have

$$G^\Theta = SO(2n, \mathbb{R}) = K, \quad (\text{P.12.61})$$

and

$$G^\Upsilon \cong SL(n, \mathbb{C}). \quad (\text{P.12.62})$$

Also, from Section 3.9 of *Lie Methods*, we know that

$$K \cap H = SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \cong U(n) \oplus \overline{U(n)}. \quad (\text{P.12.63})$$

Finally, suppose

$$m = \exp(i\phi)I, \quad (\text{P.12.64})$$

which corresponds to

$$a = \cos(\phi)I \quad (\text{P.12.65})$$

and

$$b = \sin(\phi)I. \quad (\text{P.12.66})$$

Then we find the result

$$g(m) = \begin{pmatrix} \cos(\phi)I & \sin(\phi)I \\ -\sin(\phi)I & \cos(\phi)I \end{pmatrix} = I \cos(\phi) + J \sin(\phi) = \exp(\phi J). \quad (\text{P.12.67})$$

P.13 Decomposition of Lie Algebras

Denote the Lie algebras of $G = SL(2n, \mathbb{R})$, $H = Sp(2n, \mathbb{R})$, and $K = SO(2n, \mathbb{R})$ by the symbols \mathfrak{g} , \mathfrak{h} , and \mathfrak{k} , respectively. The effects of σ , θ , and τ on $\mathfrak{g} = sl(2n, \mathbb{R})$ have already been determined in (4.15), (12.21), and (12.23), respectively. To recapitulate, we find the results

$$\sigma(B) = -JB^TJ^{-1} = B \text{ for } B \in \mathfrak{h} \quad (\text{P.13.1})$$

and

$$\theta(B) = -B^T = B \text{ for } B \in \mathfrak{k}. \quad (\text{P.13.2})$$

That is, \mathfrak{h} is the +1 eigenspace of σ in \mathfrak{g} , and \mathfrak{k} (the antisymmetric matrices) is the +1 eigenspace of θ in \mathfrak{g} .

Define the subspace \mathfrak{p} by the requirement

$$\mathfrak{p} = \{B \in \mathfrak{g} \mid \theta(B) = -B\}. \quad (\text{P.13.3})$$

That is, \mathfrak{p} consists of the symmetric traceless matrices, and is the -1 eigenspace of θ in \mathfrak{g} . Then we have the direct sum decomposition (± 1 eigenspaces of θ)

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}, \quad (\text{P.13.4})$$

which is just the familiar statement that any matrix can be uniquely decomposed into antisymmetric and symmetric parts. These parts are also mutually orthogonal relative to the trace form. Indeed, we have

$$(A, S) = \text{tr}(AS) = \text{tr}[(AS)^T] = \text{tr}(S^TA^T) = \text{tr}(-SA) = \text{tr}(-AS) = -(A, S) \quad (\text{P.13.5})$$

and therefore

$$(A, S) = 0. \quad (\text{P.13.6})$$

Here A and S are antisymmetric and symmetric matrices, respectively.

Likewise, define the subspace \mathfrak{q} by the requirement

$$\mathfrak{q} = \{B \in \mathfrak{g} \mid \sigma(B) = -B\}. \quad (\text{P.13.7})$$

Then we have the direct sum decomposition (± 1 eigenspaces of σ)

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q}. \quad (\text{P.13.8})$$

We should check that these eigenspaces are also mutually orthogonal relative to the trace form. The elements of \mathfrak{h} are matrices of the form JS where S is symmetric. From (4.18) we know that the elements of \mathfrak{q} are matrices of the form JA where A is antisymmetric. For their inner product we find that

$$\begin{aligned} (JS, JA) &= \text{tr}(JSJA) = \text{tr}[(JSJA)^T] = \text{tr}[A^TJ^TS^TJ^T] \\ &= \text{tr}[-AJSJ] = \text{tr}[-JSJA] = -(JS, JA), \end{aligned} \quad (\text{P.13.9})$$

from which we conclude that

$$(JS, JA) = 0. \quad (\text{P.13.10})$$

Finally, we observe that the relation (13.8) is just the assertion that every traceless matrix can be written as the sum $(JS + JA)$ where A is chosen to make JA traceless. (Note that JS is automatically traceless for any S .)

Next we can refine the decompositions (13.4) and (13.8) using both θ and σ . For example, the decomposition (13.4) used θ to decompose \mathfrak{g} into $\mathfrak{k} + \mathfrak{p}$. We can now further decompose \mathfrak{k} and \mathfrak{p} using σ . Alternatively, the decomposition (13.8) used σ to decompose \mathfrak{g} into $\mathfrak{h} + \mathfrak{q}$. We can now further decompose \mathfrak{h} and \mathfrak{q} using θ . These refinements are possible because θ and σ commute. The eigenspaces of θ are invariant under the action of σ , and the eigenspaces of σ are invariant under the action of θ .

Let s and s' be *sign* variables that take on the values ± 1 . Define subspaces $\mathfrak{g}_{ss'}$ by the requirements

$$\sigma(B) = sB \quad (\text{P.13.11})$$

and

$$\theta(B) = s'B \quad (\text{P.13.12})$$

for any $B \in \mathfrak{g}_{ss'}$. Then we have the direct sum decomposition

$$\mathfrak{g} = \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{1,-1} \oplus \mathfrak{g}_{-1,1} \oplus \mathfrak{g}_{-1,-1}. \quad (\text{P.13.13})$$

Also, we have the result

$$\tau(B) = ss'B \quad (\text{P.13.14})$$

for any $B \in \mathfrak{g}_{ss'}$.

Let us examine the contents of each subspace $\mathfrak{g}_{ss'}$. Begin with $\mathfrak{g}_{1,1}$. From (4.17) we see that it must be of the form JS . From (12.23) and (13.14) we see that it must be of the form JS^c . Thus we have

$$B \in \mathfrak{g}_{1,1} \Leftrightarrow B = JS^c. \quad (\text{P.13.15})$$

Similarly, we find that elements in $\mathfrak{g}_{1,-1}$ must be of the form JS^a ,

$$B \in \mathfrak{g}_{1,-1} \Leftrightarrow B = JS^a. \quad (\text{P.13.16})$$

Together $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{1,-1}$ span $\mathfrak{h} = sp(2n, \mathbb{R})$,

$$\mathfrak{h} = \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{1,-1}. \quad (\text{P.13.17})$$

Next consider $\mathfrak{g}_{-1,1}$. By (4.18) it must be of the form JA . By (13.14) it must be of the form JA^a where A^a is an antisymmetric matrix that anticommutes with J ,

$$B \in \mathfrak{g}_{-1,1} \Leftrightarrow B = JA^a. \quad (\text{P.13.18})$$

Finally, any $B \in \mathfrak{g}_{-1,-1}$ must be of the form JA^c where A^c is an antisymmetric matrix that commutes with J ,

$$B \in \mathfrak{g}_{-1,-1} \Leftrightarrow B = JA^c. \quad (\text{P.13.19})$$

In summary, the claim is that any matrix $B \in sl(2n, \mathbb{R})$ can be uniquely be decomposed as the sum

$$B = JS^c + JS^a + JA^a + JA^c, \quad (\text{P.13.20})$$

which is, in fact, almost obvious upon inspection.

These elements are also mutually orthogonal. We find from (8.9) the result

$$(JS^c, JS^a) = \text{tr}(JS^c JS^a) = \text{tr}(S^c J JS^a) = -\text{tr}(S^c S^a) = -(S^c, S^a) = 0. \quad (\text{P.13.21})$$

From (13.10) we find the results

$$(JS^c, JA^a) = (JS^c, JA^c) = (JS^a, JA^a) = (JS^a, JA^c) = 0. \quad (\text{P.13.22})$$

Lastly, we find

$$(JA^a, JA^c) = \text{tr}(JA^a JA^c) = -\text{tr}(A^a JJA^c) = \text{tr}(A^a A^c) = (A^a, A^c). \quad (\text{P.13.23})$$

But we also find

$$(JA^a, JA^c) = \text{tr}(JA^a JA^c) = \text{tr}(JA^a A^c J) = \text{tr}(JJA^a A^c) = -(A^a, A^c), \quad (\text{P.13.24})$$

from which we conclude that

$$(JA^a, JA^c) = 0. \quad (\text{P.13.25})$$

These orthogonality proofs just provided are brute force. A more elegant proof, in the style of (8.8) and (8.9), can be given based on (13.1) and (13.2) and the fact that σ and θ preserve the inner product.

Together $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{-1,1}$ span $\mathfrak{k} = so(2n, \mathbb{R})$,

$$\mathfrak{k} = \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{-1,1}. \quad (\text{P.13.26})$$

Inspection of (3.15) and (3.18) shows that

$$B \in \mathfrak{g}_{1,1} \oplus \mathfrak{g}_{-1,1} \Leftrightarrow B = JS^c + JA^a. \quad (\text{P.13.27})$$

Matrices of the form ($B = JS^c + JA^a$) are evidently antisymmetric. It is also easily verified that matrices of the form ($JS^a + JA^c$) are symmetric. Since all the matrices appearing in (13.20), when taken together, span $gl(2n, \mathbb{R})$, it follows that matrices of the form ($B = JS^c + JA^a$) span the full space of antisymmetric matrices. See also (13.28) and (13.30) below.

It remains to be seen what matrices in (13.20) are traceless. We already know that JS^c and JS^a are traceless, and JA^a is also traceless because it is antisymmetric. The remaining candidate is JA^c . It contains the possibility $JJ = -I$, which is not traceless.

Finally, using the notation of intersecting sets, we may write

$$\mathfrak{g}_{1,1} = \mathfrak{h} \cap \mathfrak{k} = \text{matrices of the form } JS^c \cong A^c, \quad (\text{P.13.28})$$

$$\mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p} = \text{matrices of the form } JS^a \cong S^a, \quad (\text{P.13.29})$$

$$\mathfrak{g}_{-1,1} = \mathfrak{q} \cap \mathfrak{k} = \text{matrices of the form } JA^a \cong A^a, \quad (\text{P.13.30})$$

$$\mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p} = \text{matrices of the form } JA^c \cong S^c. \quad (\text{P.13.31})$$

Here we have included the fact that various categories of matrices are are isomorphic. For example, matrices of the form JS^c are evidently antisymmetric, and commute with J . Therefore they are of the form A^c . We also note, from the discussion of the previous paragraph, that all the $\mathfrak{g}_{s,s'}$ are traceless with the possible exception of $\mathfrak{g}_{-1,-1}$.

By using these isomorphisms in (13.20), we find that any matrix $B \in sl(2n, \mathbb{R})$ can also be uniquely be decomposed as the sum

$$B = A^c + S^a + A^a + S^c, \quad (\text{P.13.32})$$

which is also obvious upon inspection.

We have already noted in (13.17) that $\mathfrak{g}_{1,1}$ and $\mathfrak{g}_{1,-1}$ together span $\mathfrak{h} = sp(2n, \mathbb{R})$. They, in fact, provide the Cartan decomposition of $sp(2n, \mathbb{R})$.

Finally, Goodman makes the claims

$$\mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p} = \{B \in \mathcal{E} \mid \text{tr}B = 0\}, \quad (\text{P.13.33})$$

$$\mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p} = \{B \in \mathcal{F} \mid \text{tr}B = 0\}. \quad (\text{P.13.34})$$

But, \mathcal{E} is defined by the requirements

$$\mathcal{E} = \{B \mid B^T = B \text{ and } \sigma(B) = B\}, \quad (\text{P.13.35})$$

which is equivalent to the requirements

$$\mathcal{E} = \{B \mid \theta(B) = -B \text{ and } \sigma(B) = B\}. \quad (\text{P.13.36})$$

In our notation, these requirements simply state that

$$\mathcal{E} = \mathfrak{g}_{1,-1}. \quad (\text{P.13.37})$$

We already know that matrices in $\mathfrak{g}_{1,-1}$ are traceless, and so (13.33) is verified. Moreover, \mathcal{F} is defined by the requirements

$$\mathcal{F} = \{B \mid B^T = B \text{ and } \sigma(B) = -B\}, \quad (\text{P.13.38})$$

which is equivalent to the requirements

$$\mathcal{F} = \{B \mid \theta(B) = -B \text{ and } \sigma(B) = -B\}. \quad (\text{P.13.39})$$

In our notation, these requirements simply state that

$$\mathcal{F} = \mathfrak{g}_{-1,-1}. \quad (\text{P.13.40})$$

We know from (13.19) that in this case there is a matrix that has trace, namely JJ , and therefore in this case (13.34) is also verified provided the JJ case is excluded.

P.14 Preparation for Lemma 2.1 of Goodman

Let S_d consist of all the real *diagonal* symmetric matrices s_d of the form

$$s_d = \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} \quad (\text{P.14.1})$$

where

$$d = \text{diag}[d_1, \dots, d_n] \text{ and } \text{tr}(d) = 0. \quad (\text{P.14.2})$$

(Goodman uses the symbol A for what we call S_d . However, we have already used A to denote antisymmetric matrices.) Then, from Section 12.2, we have the relation

$$\tau(s_d) = s_d. \quad (\text{P.14.3})$$

And, from (12.21), we see that

$$\theta(s_d) = -s_d. \quad (\text{P.14.4})$$

Therefore, from (12.29), we conclude that

$$\sigma(s_d) = -s_d, \quad (\text{P.14.5})$$

and consequently

$$S_d \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}. \quad (\text{P.14.6})$$

Let S_d^+ be the open subset of S_d consisting of the matrices s_d with

$$d_1 > d_2 > \dots > d_n. \quad (\text{P.14.7})$$

Let S_d^{+c} be the closure of S_d^+ . Then we also have the relations

$$S_d^+ \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}, \quad (\text{P.14.8})$$

and

$$S_d^{+c} \subset \mathfrak{g}_{-1,-1} = \mathfrak{q} \cap \mathfrak{p}. \quad (\text{P.14.9})$$

Also, let D^+ be the set of matrices d of the form (14.2) with (14.7) satisfied, and let D^{+c} denote its closure.

P.15 Lemma 2.1 of Goodman

Suppose x is of the form $x = \exp(B)$ where $B \in \mathfrak{g}_{-1,-1}$. Then we know there is a matrix of the form JA^c such that

$$x = \exp(JA^c). \quad (\text{P.15.1})$$

We may also require that JA^c be traceless. By inspection, JA^c is symmetric and commutes with J . Therefore, consistent with (13.31), there is a real traceless symmetric matrix S^c such that

$$JA^c = S^c, \quad (\text{P.15.2})$$

and we have the relation

$$x = \exp(S^c). \quad (\text{P.15.3})$$

It follows from (15.3) that x is real, symmetric, and positive definite.

Let us see what can be said about S^c . Since it commutes with J , by the work of Section 12.2, it must be of the form

$$S^c = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}. \quad (\text{P.15.4})$$

Since S^c is traceless, α must be traceless. Since S^c is symmetric, the matrices α and β must have the properties

$$\alpha^T = \alpha, \quad (\text{P.15.5})$$

$$\beta^T = -\beta. \quad (\text{P.15.6})$$

Note that, by (15.6), the matrix β is traceless. Define the matrix m by the rule

$$m = \alpha + i\beta. \quad (\text{P.15.7})$$

Then, we have the relation

$$S^c = S^c(m) = M(m) = W \begin{pmatrix} m & 0 \\ 0 & \overline{m} \end{pmatrix} W^{-1}. \quad (\text{P.15.8})$$

We also observe that

$$m^\dagger = \alpha^T - i\beta^T = \alpha + i\beta = m \quad (\text{P.15.9})$$

so that m is Hermitian. Since both α and β are traceless, m is also traceless.

Since m is Hermitian and traceless, there is a unitary matrix v and a matrix $d \in D^{+c}$ such that

$$m = vdv^{-1}. \quad (\text{P.15.10})$$

Since d is real, we also have the relation

$$\overline{m} = \overline{v}d\overline{v^{-1}}. \quad (\text{P.15.11})$$

It follows that S^c has the representation

$$S^c(m) = W \begin{pmatrix} v & 0 \\ 0 & \overline{v} \end{pmatrix} \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} \begin{pmatrix} v^{-1} & 0 \\ 0 & \overline{v^{-1}} \end{pmatrix} W^{-1}. \quad (\text{P.15.12})$$

Insert factors of W and W^{-1} into (15.12) to rewrite it in the form

$$S^c(m) = W \begin{pmatrix} v & 0 \\ 0 & \overline{v} \end{pmatrix} W^{-1} W \begin{pmatrix} d & 0 \\ 0 & d \end{pmatrix} W^{-1} W \begin{pmatrix} v^{-1} & 0 \\ 0 & \overline{v^{-1}} \end{pmatrix} W^{-1}. \quad (\text{P.15.13})$$

Since d is real, we see that (15.13) can be rewritten in the form

$$S^c(m) = M(v)M(d)M(v^{-1}). \quad (\text{P.15.14})$$

We also know, since d is real, that we have

$$M(d) = s_d, \quad (\text{P.15.15})$$

with $s_d \in S_d^{+c}$, so that there is also the relation

$$S^c = M(v)s_dM(v^{-1}). \quad (\text{P.15.16})$$

Define the matrix O by the rule

$$O = M(v). \quad (\text{P.15.17})$$

Since v is unitary, O will be real, symplectic, and orthogonal. Thus we have the result

$$S^c = Os_dO^{-1}. \quad (\text{P.15.18})$$

Finally, exponentiate both sides of (15.18). So doing gives the result

$$x = O \exp(s_d)O^{-1} \quad (\text{P.15.19})$$

with

$$O \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \cong U(n) \oplus \overline{U(n)}. \quad (\text{P.15.20})$$

Note that since O is orthogonal, (15.19) can also be written in the form

$$x = O \exp(s_d)O^T, \quad (\text{P.15.21})$$

from which we again see that x is real, symmetric, and positive definite.

Goodman claims that s_d depends analytically on x when the eigenvalues of x , which will be the quantities $\exp(d_j)$, are distinct. We will worry about proving this later. We know from (15.3) that S^c is real analytic in x , and from (15.4) through (15.7) we see that m is analytic in S^c . What remains to be shown is that the eigenvalues of m , with m Hermitian and traceless, are analytic in m under the assumption that the eigenvalues are distinct.

P.16 Preparation for Theorem 2.1 of Goodman

Let u be any $n \times n$ unitary matrix, and consider $M(u)$. From Section 3.9 of *Lie methods* we know that

$$M(u) \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \cong U(n) \oplus \overline{U(n)}, \quad (\text{P.16.1})$$

and given any $M' \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})$ there is a unique $u \in U(n)$ such that

$$M(u) = M'. \quad (\text{P.16.2})$$

Let $L \subset SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})$ be the subgroup of matrices g such that

$$gs_dg^{-1} = s_d \text{ for all } s_d \in S_d. \quad (\text{P.16.3})$$

By definition for any such s_d there is a corresponding d given by (14.1) and (14.2), and we have the relation

$$s_d = M(d). \quad (\text{P.16.4})$$

Also, given any $g \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R})$, there is a unique $u \in U(n)$ such that

$$M(u) = g. \quad (\text{P.16.5})$$

With these definitions, the relation (16.3) becomes

$$M(u)M(d)[M(u)]^{-1} = M(d) \text{ for all } d. \quad (\text{P.16.6})$$

By the isomorphic property of M this relation is equivalent to the requirement

$$udu^{-1} = d \text{ for all } d. \quad (\text{P.16.7})$$

All such u must be diagonal unitary matrices, and therefore have the explicit form

$$u = \text{diag}[\exp(i\phi_1), \dots, \exp(i\phi_n)]. \quad (\text{P.16.8})$$

Thus, we have the isomorphism

$$L \cong T^n. \quad (\text{P.16.9})$$

Moreover, since $SO(2n, \mathbb{R})$ has rank n , L is a maximal torus in $SO(2n, \mathbb{R})$. Goodman says that this means that $SO(2n, \mathbb{R})/L$ is the flag manifold for $SO(2n, \mathbb{R})$.

By the definition of L we see from (16.3) that

$$\ell s_d \ell^{-1} = s_d \text{ for all } \ell \in L \text{ and all } s_d \in S_d. \quad (\text{P.16.10})$$

It follows that

$$\ell \exp(s_d) \ell^{-1} = \exp(\ell s_d \ell^{-1}) = \exp(s_d). \quad (\text{P.16.11})$$

Introduce the notation

$$\hat{s}_d = \exp(s_d). \quad (\text{P.16.12})$$

Then, for future use, we have the relation

$$\ell \hat{s}_d = \hat{s}_d \ell \text{ for all } \ell \in L \text{ and all } \hat{s}_d \in \exp(S_d). \quad (\text{P.16.13})$$

Also, we see directly that

$$\ell \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \quad (\text{P.16.14})$$

since ℓ is of the form $M(u)$ with u given by (16.8).

For insight, let us work out the explicit matrix form of ℓ . We have the relation

$$\ell = M(u) = \begin{pmatrix} \text{Re}(u) & \text{Im}(u) \\ -\text{Im}(u) & \text{Re}(u) \end{pmatrix} = \begin{pmatrix} C & S \\ -S & C \end{pmatrix}. \quad (\text{P.16.15})$$

Here C and S are $n \times n$ diagonal matrices given by the relations

$$C = \begin{pmatrix} \cos(\phi_1) & & & \\ & \cos(\phi_2) & & \\ & & \ddots & \\ & & & \cos(\phi_n) \end{pmatrix}, \quad (\text{P.16.16})$$

$$S = \begin{pmatrix} \sin(\phi_1) & & & \\ & \sin(\phi_2) & & \\ & & \ddots & \\ & & & \sin(\phi_n) \end{pmatrix}. \quad (\text{P.16.17})$$

P.17 Theorem 2.1 of Goodman

Recall that $G = SL(2n, \mathbb{R})$, $H = Sp(2n, \mathbb{R})$, and $K = SO(2n, \mathbb{R})$. Goodman claims that any $g \in G$ has the decomposition

$$g = k\hat{s}_d h \quad (\text{P.17.1})$$

where $k \in K$, $h \in H$, and $\hat{s}_d \in \exp(S_d^{+c})$.

The argument goes as follows. Recall that, according to Theorem 1.2, every $g \in SL(2n, \mathbb{R})$ has the factorization

$$g = k \exp(S^c) \exp(S^a) \quad (\text{P.17.2})$$

with $k \in SO(2n, \mathbb{R})$ and $\text{tr}(S^c) = 0$. Also, we know from (15.3) and (15.20) that there is the representation

$$\exp(S^c) = O \exp(s_d) O^T, \quad (\text{P.17.3})$$

so that we also have the factorization

$$g = k O \exp(s_d) O^T \exp(S^a). \quad (\text{P.17.4})$$

Rewrite this relation in the form

$$g = [kO][\exp(s_d)][O^T \exp(S^a)]. \quad (\text{P.17.5})$$

Note from (13.29) that $S^a \in \mathfrak{g}_{1,-1} = \mathfrak{h} \cap \mathfrak{p}$ and $S^a \cong JS^a$. Therefore,

$$\exp(S^a) \in Sp(2n, \mathbb{R}). \quad (\text{P.17.6})$$

Also O is in both $SO(2n, \mathbb{R})$ and $Sp(2n, \mathbb{R})$. Consequently there are the following group relations,

$$kO \in SO(2n, \mathbb{R}), \quad (\text{P.17.7})$$

$$O^T \exp(S^a) \in Sp(2n, \mathbb{R}). \quad (\text{P.17.8})$$

Define group elements k' , \hat{s}_d , and h by the rules

$$k' = kO, \quad (\text{P.17.9})$$

$$\hat{s}_d = \exp(s_d), \quad (\text{P.17.10})$$

$$h = O^T \exp(S^a). \quad (\text{P.17.11})$$

Then, we may write

$$g = k' \hat{s}_d h, \quad (\text{P.17.12})$$

which is a factorization of the form (17.1).

Consider the map

$$(K, S_d) \mapsto G \quad (\text{P.17.13})$$

given by the rule

$$g(k, s_d) = k[\exp(s_d)]. \quad (\text{P.17.14})$$

Then, we find that

$$g(k\ell, s_d) = k\ell[\exp(s_d)] = k[\exp(s_d)]\ell \quad (\text{P.17.15})$$

where we have used (16.12). Also we know that

$$\ell \in SO(2n, \mathbb{R}) \cap Sp(2n, \mathbb{R}) \quad (\text{P.17.16})$$

and therefore

$$\ell \in Sp(2n, \mathbb{R}). \quad (\text{P.17.17})$$

Thus, we may write (17.15) in the form

$$g(k\ell, s_d) = k\ell[\exp(s_d)] = k[\exp(s_d)]\ell = g(k, s_d)h \quad (\text{P.17.18})$$

with

$$h = \ell \text{ and } h \in Sp(2n, \mathbb{R}). \quad (\text{P.17.19})$$

We see that $g(k\ell, s_d)$ and $g(k, s_d)$ are in the same coset in the coset space G/H ,

$$g(k\ell, s_d) \sim g(k, s_d) \pmod{Sp(2n, \mathbb{R})}. \quad (\text{P.17.20})$$

Also, we know that $k\ell$ and k are in the same coset in the coset space K/L ,

$$k\ell \sim k \pmod{L}. \quad (\text{P.17.21})$$

Therefore (17.14) provides a map

$$(K/L, S_d) \mapsto G/H. \quad (\text{P.17.22})$$

But, by (17.1), we know that every element of G/H can be obtained in this way. Thus, we conjecture that there is a correspondence of the form

$$[K/L] \times S_d^{+c} \leftrightarrow G/H. \quad (\text{P.17.23})$$

Let us check dimensions. We want to check the relation

$$\dim K - \dim L + \dim S_d^{+c} = \dim G - \dim H. \quad (\text{P.17.24})$$

We have the counts

$$\dim K = n(2n - 1), \quad (\text{P.17.25})$$

$$\dim L = n, \quad (\text{P.17.26})$$

$$\dim S_d^{+c} = n - 1, \quad (\text{P.17.27})$$

$$\dim H = n(2n + 1), \quad (\text{P.17.28})$$

$$\dim G = (2n)^2 - 1. \quad (\text{P.17.29})$$

Therefore we have to check the relation

$$n(2n - 1) - n + (n - 1) = [(2n)^2 - 1] - n(2n + 1)? \quad (\text{P.17.30})$$

A little algebra shows that both sides of (17.30) simplify to the expression $(2n^2 - n - 1)$ so that the relation does indeed hold.

We still have to check uniqueness. Suppose (k_1, s_{d1}) and (k_2, s_{d2}) are sent to g_1 and g_2 under (17.14),

$$k_1[\exp(s_{d1})] = g_1, \quad (\text{P.17.31})$$

$$k_2[\exp(s_{d2})] = g_2. \quad (\text{P.17.32})$$

Also suppose that

$$g_2 \sim g_1 \bmod H. \quad (\text{P.17.33})$$

Then, we want to show that

$$k_2 \sim k_1 \bmod L \quad (\text{P.17.34})$$

and

$$s_{d2} = s_{d1}. \quad (\text{P.17.35})$$

Suppose (17.33) holds. Then there is an $h \in H$ such that

$$g_2 = g_1 h, \quad (\text{P.17.36})$$

and therefore, from (17.31) and (17.32), there is the relation

$$k_2[\exp(s_{d2})] = k_1[\exp(s_{d1})]h. \quad (\text{P.17.37})$$

To be continued.

Application to Dragt's Symplectic Polar Decomposition

P.18 Search for Counter Examples

In view of the results of the previous section, we will also get elements g' in all possible cosets $GL(2n, \mathbb{R})/Sp(2n, \mathbb{R})$ by the rule

$$g'(k, s'_d) = ks'_d \quad (\text{P.18.1})$$

where s'_d is of the form (14.1) but d is no longer required to be traceless. Comparison of this rule with (17.14) shows that we may get some overlap by this procedure, but (18.1) appears easier to work with.

The search for counter examples can be begun with the case $k = I$ for which

$$g'(I, s'_d) = Is'_d = s'_d, \quad (\text{P.18.2})$$

and we have found counter examples in the 4×4 matrix context. They demonstrate that symplectic polar decomposition of a matrix M is not possible globally even with the restriction $\det(M) > 0$. See Section 4.3.5 of *Lie Methods*.

We can next examine the case $s'_d = I$ for which the g' are matrices of the form

$$g'(k, I) = k. \quad (\text{P.18.3})$$

When k is in the vicinity of the identity, g' will be near the identity, and therefore have a symplectic polar decomposition. But what happens when we consider all $k \in SO(2n, \mathbb{R})$? Exercise 4.3.19 of *Lie Methods* studies a particular one-parameter subgroup of $SO(4, \mathbb{R})$ and shows that for all such elements symplectic polar decomposition is always possible, but the ray $\lambda^2 N(M)$ need not always intersect the unit ball around I . Exercise 4.3.22 shows that all elements of $SO(4, \mathbb{R})$ have symplectic polar decompositions.

Let H be the subgroup consisting of all elements $g \in GL(2n, \mathbb{R}, +)$ such that

$$\{g, J\} = 0. \quad (\text{P.18.4})$$

We know that H is isomorphic to $GL(n, \mathbb{C})$. Exercise 4.3.21 shows that all elements of H have symplectic polar decompositions.

We should now think about more general elements of $GL(4, \mathbb{R})$. For example, Exercise 4.3.20 describes a one parameter closed path of elements that includes elements that do and do not have symplectic polar decompositions. It would be interesting to see where the decomposition first fails.

Also, suppose symplectic polar decomposition is possible for some matrix M . What can be said about matrices in the neighborhood of M ? Similarly, suppose symplectic polar decomposition is impossible for some matrix M . What can be said then about matrices in the neighborhood of M ? What is the nature of the transition from having to not having a symplectic polar decomposition?

Bibliography

- [1] The essential contents of this Appendix, including all major insights, were provided in kind correspondence from Professor Roe Goodman.
- [2] G. Heckman and H. Schlichtkrull, “Harmonic Analysis and Special Functions on Symmetric Spaces”, *Perspectives in Mathematics* **16**, Academic Press (1994).
- [3] G. Mostow, “Some New Decomposition Theorems for Semisimple Groups”, *Memoirs of the Amer. Math. Soc.* **14**, 31-54 (1955).
- [4] G. Hochschild, *The Structure of Lie Groups*, Holden-Day (1965).

Appendix Q

Improving Convergence of Fourier Representation

Q.1 Introduction

Suppose $f(u)$ is a function defined on the interval $u \in [0, 2\pi]$, and suppose f is continuous and has a continuous first derivative. What can be said about a Fourier representation of f over this interval? We observe that, by definition, a Fourier series produces a periodic function, and straight-forward application of Fourier's theorem to f produces a function with period 2π . But, f may not have a periodic extension unless some kind of singularity (say a discontinuity in f or one of its derivatives) is introduced at the points $u = 0, \pm 2\pi, \pm 4\pi, \dots$. For example, if $f(0) \neq f(2\pi)$, the periodic extension of f cannot be continuous. The net effect of this discontinuity is that in this case the Fourier coefficients of f can fall off no faster than $(1/n)$ for large n . In Section 14.5 we saw that this situation can be improved somewhat by doubling the domain of definition for f and imposing an *evenness* condition on its extension. The net result is a modified Fourier representation over the domain $[-2\pi, 2\pi]$ whose coefficients fall off like $(1/n)^2$. The purpose of this appendix is to describe and apply a further trick that makes it possible to obtain a Fourier-like representation for which the coefficients fall off still faster.

Begin by writing f in the form

$$f(u) = c + [d/(2\pi)]u + g(u) \quad (\text{Q.1.1})$$

where

$$c = f(0) \quad \text{and} \quad d = f(2\pi) - f(0). \quad (\text{Q.1.2})$$

Then the function $g(u)$ is also defined for $u \in [0, 2\pi]$, is continuous, and has a continuous first derivative. Moreover, it has the property

$$g(0) = g(2\pi) = 0. \quad (\text{Q.1.3})$$

Extend g to the interval $[-2\pi, 0]$ by requiring that g be odd,

$$g(u) = -g(-u) \quad \text{for } u \in [-2\pi, 0]. \quad (\text{Q.1.4})$$

Then, because of (1.3) and (1.4), the extended g is continuous at $u = 0$. Moreover, we find that

$$g(-2\pi) = -g(2\pi) = 0. \quad (\text{Q.1.5})$$

Finally, from (1.4) we find that

$$g'(u) = g'(-u) \quad \text{for } u \in [-2\pi, 0], \quad (\text{Q.1.6})$$

from which it follows that the extended g' is continuous at $u = 0$. Thus, g has now been defined for $u \in [-2\pi, 2\pi]$, and is continuous and has a continuous first derivative in the open interval $u \in (-2\pi, 2\pi)$.

Further extend g to the full interval $u \in (-\infty, \infty)$ by requiring that g be periodic with period 4π ,

$$g(u + 4\pi) = g(u). \quad (\text{Q.1.7})$$

We will now see that this extension results in a g that is also continuous and has a continuous first derivative at the points $u = \pm 2\pi$ and their 4π periodic extensions. Thus the net result is that, by these extensions, g has been defined everywhere and is continuous and has a continuous first derivative everywhere. Let us check first the case $u = 2\pi$. From the definitions so far we have the relations

$$g(2\pi + \epsilon) = g(-2\pi + \epsilon) = -g(2\pi - \epsilon). \quad (\text{Q.1.8})$$

In view of (1.3) and (1.8), continuity at $u = 2\pi$ has been established. Also, using periodicity and (1.6), we have the relations

$$g'(2\pi + \epsilon) = g'(-2\pi + \epsilon) = g'(2\pi - \epsilon), \quad (\text{Q.1.9})$$

from which it follows that g has a continuous first derivative at $u = 2\pi$. Similarly, we find that

$$g(-2\pi - \epsilon) = g(2\pi - \epsilon) = -g(-2\pi + \epsilon) \quad (\text{Q.1.10})$$

and

$$g'(-2\pi - \epsilon) = g'(2\pi - \epsilon) = g'(-2\pi + \epsilon), \quad (\text{Q.1.11})$$

from which it follows that g is continuous and has a continuous first derivative at $u = -2\pi$.

We are now ready to invoke the results of Fourier. Since g is 4π periodic, it has an expansion over the interval $u \in [-2\pi, 2\pi]$ of the form

$$g(u) = \sum_{n=0}^{\infty} a_n \cos(nu/2) + \sum_{n=1}^{\infty} b_n \sin(nu/2). \quad (\text{Q.1.12})$$

Since g is odd, all the a_n must vanish, and we are left with

$$g(u) = \sum_{n=1}^{\infty} b_n \sin(nu/2). \quad (\text{Q.1.13})$$

The coefficients b_n are given by the integrals

$$b_n = [1/(2\pi)] \int_{-2\pi}^{2\pi} du g(u) \sin(nu/2) = (1/\pi) \int_0^{2\pi} du g(u) \sin(nu/2). \quad (\text{Q.1.14})$$

Note that the integrals on the far right side of (1.14) depend only on the knowledge of g , and hence f , in the original interval $u \in [0, 2\pi]$. Finally, since g will generally have a discontinuous second derivative at the points $u = 0, \pm 2\pi, \pm 4\pi, \dots$, the coefficients b_n will generally fall off according to the rule

$$b_n \sim (1/n)^3 \quad \text{as } n \rightarrow \infty. \quad (\text{Q.1.15})$$

The net result of our efforts is that f has the representation

$$f(u) = c + [d/(2\pi)]u + \sum_{n=1}^{\infty} b_n \sin(nu/2) \quad (\text{Q.1.16})$$

with the coefficients b_n obeying (1.15).

Consider the function $h(v)$ defined by

$$h(v) = f(v + \pi). \quad (\text{Q.1.17})$$

It is defined on the interval $v \in [-\pi, \pi]$. According to (1.16) it has the expansion

$$\begin{aligned} h(v) &= c + [d/(2\pi)](v + \pi) + \sum_{n=1}^{\infty} b_n \sin[n(v + \pi)/2] \\ &= [h(\pi) + h(-\pi)]/2 + \{[h(\pi) - h(-\pi)]/(2\pi)\}v + \sum_{n=1}^{\infty} b_n \sin[n(v + \pi)/2] \end{aligned} \quad (\text{Q.1.18})$$

with

Q.2 Application

Bibliography

[1] ***

Appendix R

Abstract Lie Group Theory

The purpose of this appendix is to show that the Jacobi identity in a Lie algebra is related to the assumed associativity of group multiplication in the corresponding Lie group. When a Lie group is realized in terms of matrices, the associative condition for group multiplication is automatically satisfied because matrix multiplication is associative. Correspondingly, the Jacobi identity is readily verified for Lie algebras realized in terms of matrices with the Lie product taken to be the matrix commutator. Treating the case of an abstract Lie group requires somewhat more effort.

Bibliography

- [1] Eugene Lerman, *Notes on Lie Groups* (2012).

<https://faculty.math.illinois.edu/~lerman/519/s12/427notes.pdf>

Observe that the equation in Corollary 6.7a should read

$$\exp[(t_1 + t_2)X] = [\exp(t_1X)][\exp(t_2X)].$$

Appendix S

Mathematica Realization of TPSA and Taylor Map Computation

S.1 Background

The forward integration method (Section 10.12.4) for computing Taylor maps can be implemented by a code employing the tools of *automatic differentiation* (AD) described by Neidinger [1].¹ In this approach arrays of Taylor coefficients of various functions are referred to as AD variables or *pyramids* since, as will be seen, they have a hyper-pyramidal structure. Generally the first entry in the array will be the value of the function about some expansion point, and the remaining entries will be the higher-order Taylor coefficients about the expansion point and truncated beyond some specified order. Such truncated Taylor expansions are also commonly called *jets*. Recall Section 7.5.

In our application elements in these arrays will be addressed and manipulated with the aid of scalar indices and associated look-up and look-back tables generated at run time. We have also replaced the original APL implementation of Neidinger with a code written in the language of *Mathematica* (Version 6, or 7) [2,3]. Where necessary, for those unfamiliar with the details of *Mathematica*, we will explain the consequences of various *Mathematica* commands. Recall that we wish to integrate equations of the form

$$\dot{z}_a = f_a(\mathbf{z}, t), \quad a = 1, m \quad (\text{S.1.1})$$

and their associated complete variational equations. The inputs to the code are the right sides (RS) of (1.1). Other input parameters are the number of variables m , the desired order of the Taylor map p , and the *initial* conditions $(z_a^d)^i$ for the design solution.

Various AD tools for describing and manipulating pyramids are outlined in Section S.2. There we show how pyramid operations are encoded in the case of polynomial RS, as needed, for example, for the Duffing equation. For brevity, we omit the cases of rational, fractional power, and transcendental RS. These cases can also be handled using various methods based on functional identities and known Taylor coefficients, or the differential equations that such

¹Some authors refer to AD as *truncated power series algebra* (TPSA) since AD algorithms arise from manipulating multivariable truncated power series. Other authors refer to AD as *Differential Algebra* (DA). There is a substantial literature on this subject. See the Web site <http://www.autodiff.org/>.

functions obey along with certain recursion relations [1]. In Section S.3, based on the work of Section S.2, we in effect obtain and integrate numerically the complete variational equations (10.12.36) in pyramid form, i.e. valid for any map order and any number of variables. Section S.4 treats the specific case of the Duffing equation. A final Section S.5 describes in more detail the relation between integrating equations for pyramids and the complete variational equations.

S.2 AD Tools

This section describes how arithmetic expressions representing $f_a(\mathbf{z}, t)$, the right sides of (1.1) where \mathbf{z} denotes the dependent variables, are replaced with expressions for arrays (pyramids) of Taylor coefficients. These pyramids in turn constitute the input to our code. Such an ad-hoc replacement, according to the problem at hand, as opposed to operator overloading where the kind of operation depends on the type of its argument, is also the approach taken in [1,4,5].

Let u, v, w be general arithmetic expressions, i.e. scalar-valued functions of \mathbf{z} . They contain various arithmetic operations such as addition/subtraction (\pm), multiplication ($*$), and raising to a power (\wedge). (They may also entail the computation of various transcendental functions such as the sine function, etc. However, as stated earlier, for simplicity we will omit these cases.) The arguments of these operations may be a constant, a single variable or multiple variables z_a , or even some other expression. The idea of AD is to redefine the arithmetic operations in such a way (see Definition 1), that all functions u, v, w can be consistently replaced with the arrays of coefficients of their Taylor expansions. For example, by redefining the usual product of numbers ($*$) and introducing the pyramid operation PROD, $u * v$ is replaced with PROD[U, V].

We use upper typewriter font for pyramids (U, V, \dots) and for operations on pyramids (PROD, POW, \dots). Everywhere, equalities written in typewriter fonts have equivalent *Mathematica* expressions. That is, they have associated realizations in *Mathematica* and directly correspond to various operations and commands in *Mathematica*. In effect, our code operates entirely on pyramids. However, as we will see, any pyramid expression contains, as its first entry, its usual arithmetic counterpart.

We begin with a description of our method of monomial labeling. In brief, we list all monomials in a polynomial in some sequence, and label them by where they occur in the list. Next follow Definition 1 and the recipes for encoding operations on pyramids. Subsequently, by using Definition 2, which simply states the rule by which an arithmetic expression is replaced with its pyramid counterpart, we show how a general expression can be encoded by using only the pyramids of a constant and those of the various variables involved.

S.2.1 Labeling Scheme

A monomial $G_j(\mathbf{z})$ in m variables is of the form

$$G_j(\mathbf{z}) = (z_1)^{j_1} (z_2)^{j_2} \cdots (z_m)^{j_m}. \quad (\text{S.2.1})$$

Here we have introduced an exponent vector \mathbf{j} by the rule

$$\mathbf{j} = (j_1, j_2, \dots, j_m). \quad (\text{S.2.2})$$

Evidently \mathbf{j} is an m -tuple of non-negative integers. The degree of $G_{\mathbf{j}}(\mathbf{z})$, denoted by $|\mathbf{j}|$, is given by the sum of exponents,

$$|\mathbf{j}| = j_1 + j_2 + \dots + j_m. \quad (\text{S.2.3})$$

The set of all exponents for monomials in m variables with degree less than or equal to p will be denoted by Γ_m^p ,

$$\Gamma_m^p = \{\mathbf{j} \mid |\mathbf{j}| \leq p\}. \quad (\text{S.2.4})$$

According to Section 32.1, this set has $L(m, p)$ entries with $L(m, p)$ given by a binomial coefficient,

$$L(m, p) = S_0(m, p) = \binom{p+m}{p}. \quad (\text{S.2.5})$$

With this notation, a Taylor series expansion (about the origin) of a scalar-valued function u of m variables $\mathbf{z} = (z_1, z_2, \dots, z_m)$, truncated beyond terms of degree p , can be written in the form

$$u(\mathbf{z}) = \sum_{\mathbf{j} \in \Gamma_m^p} \mathbf{U}(\mathbf{j}) G_{\mathbf{j}}(\mathbf{z}). \quad (\text{S.2.6})$$

Assuming that m and p are fixed input variables, we will often simply write Γ and L . Here, for now, \mathbf{U} simply denotes an array of numerical coefficients. When employed in code that has symbolic manipulation capabilities, each $\mathbf{U}(\mathbf{j})$ may also be a symbolic quantity.

To proceed, what is needed is some way of listing monomials systematically. With such a list, as described in Subsections 32.3.3 and 32.3.4, we may assign a label r to each monomial based on where it appears in the list. We will use a variant of *modified glex sequencing*, the only change being that we will begin the list with the monomial of degree 0. For example, Table 2.1 shows a list of monomials in three variables. As one goes down the list, first the monomial of degree $D = 0$ appears, then the monomials of degree $D = 1$, etc. Within each group of monomials of fixed degree the individual monomials appear in descending lex order. Note that Table 2.1 is similar to Table 32.2.4 except that it begins with the monomial of degree 0. Other possible listings include ascending true glex order in which monomials appear in ascending lex order within each group of degree D , and lex order for the whole monomial list as in [1].

Table S.2.1: A labeling scheme for monomials in three variables.

r	j_1	j_2	j_3	D
1	0	0	0	0
2	1	0	0	1
3	0	1	0	1
4	0	0	1	1
5	2	0	0	2
6	1	1	0	2
7	1	0	1	2
8	0	2	0	2
9	0	1	1	2
10	0	0	2	2
11	3	0	0	3
12	2	1	0	3
13	2	0	1	3
14	1	2	0	3
15	1	1	1	3
16	1	0	2	3
17	0	3	0	3
18	0	2	1	3
19	0	1	2	3
20	0	0	3	3
.
.
.
28	1	2	1	4
.
.
.

With the aid of the scalar index r the relation (2.6) can be rewritten in the form

$$u(\mathbf{z}) = \sum_{r=1}^{L(m,p)} \mathbf{U}(r) G_r(\mathbf{z}), \quad (\text{S.2.7})$$

because (by construction and with fixed m) for each positive integer r there is a unique exponent $\mathbf{j}(r)$, and for each \mathbf{j} there is a unique r . Here \mathbf{U} may be viewed as a vector with entries $\mathbf{U}(\mathbf{r})$, and $G_r(\mathbf{z})$ denotes $G_{\mathbf{j}(r)}(\mathbf{z})$.

Consider, in an m -dimensional space, the points defined by the heads of the vectors $\mathbf{j} \in \Gamma_m^p$. See (2.4). Figure 2.1 displays them in the case $m = 3$ and $p = 4$. Evidently they form a grid that lies on the surface and interior of what can be viewed as an m -dimensional *pyramid* in m -dimensional space. At each grid point there is an associated coefficient $\mathbf{U}(\mathbf{r})$.

Because of its association with this pyramidal structure, we will refer to the entire set of coefficients in (2.6) or (2.7) as the *pyramid U* of $u(\mathbf{z})$.

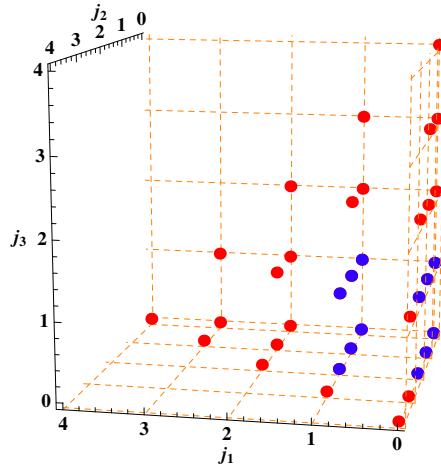


Figure S.2.1: A grid of points representing the set Γ_3^4 . For future reference a subset of Γ_3^4 , called a *box*, is shown in blue.

S.2.2 Implementation of Labeling Scheme

We have seen that use of modified glex sequencing, for any specified number of variables m , provides a labeling rule such that for each positive integer r there is a unique exponent $\mathbf{j}(r)$, and for each \mathbf{j} there is a unique r . That is, there is a invertible function $r(\mathbf{j})$ that provides a 1-to-1 correspondence between the positive integers and the exponent vectors \mathbf{j} . To proceed further, it would be useful to have this function and its inverse in more explicit form.

From the work of Subsection 32.2.6, we already know a formula for $r(\mathbf{j})$ based on the Giorgilli formula (32.2.15),

$$r(\mathbf{j}) = r(j_1, \dots, j_m) = 1 + i(j_1, \dots, j_m). \quad (\text{S.2.8})$$

Below is simple *Mathematica* code that implements this formula (which we call *Gfor*) in the case of three variables, and evaluates it for selected exponents \mathbf{j} . Observe that these

evaluations agree with results in Table 2.1.

```

Gfor[j1_, j2_, j3_] := (
  s1 = j3; s2 = 1 + j3 + j2; s3 = 2 + j3 + j2 + j1;
  t1 = Binomial[s1, 1]; t2 = Binomial[s2, 2]; t3 = Binomial[s3, 3];
  r = 1 + t1 + t2 + t3; r
)
Gfor[0, 0, 0]
Gfor[1, 0, 0]
Gfor[2, 0, 1]
Gfor[1, 2, 1]
1
2
13
28

```

(S.2.9)

For the inverse relation we have found it convenient to introduce a rectangular matrix associated with the set Γ_m^p . By abuse of notation, it will also be called Γ . It has $L(m, p)$ rows and m columns with entries

$$\Gamma_{r,a} = j_a(r). \quad (\text{S.2.10})$$

For example, looking at Table 2.1, we see (when $m = 3$) that $\Gamma_{1,1} = 0$ and $\Gamma_{17,2} = 3$. Indeed, if the first and last columns of Table 2.1 are removed, what remains (when $m = 3$) is the matrix $\Gamma_{r,a}$. In the language of Subsection 32.2.9, Γ is a *look up table* that, given r , produces the associated j . In our *Mathematica* implementation Γ is the matrix **GAMMA** with elements **GAMMA[[r, a]]**.

The matrix **GAMMA** is constructed using the *Mathematica* code illustrated below,

```

Needs["Combinatorica`"];
m = 3; p = 4;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA]
r = 17; a = 2;
GAMMA[[r]]
GAMMA[[r, a]]
35
{0, 3, 0}
3

```

(S.2.11)

It employs the *Mathematica* commands **Compositions**, **Reverse**, and **Join**.

We will now describe the ingredients of this code and illustrate the function of each:

- The command `Needs["Combinatorica`"]`; loads a combinatorial package.
- The command `Compositions[i, m]` produces, as a list of arrays (a rectangular array), all *compositions* (under addition) of the integer i into m integer parts. Furthermore, the compositions appear in *ascending* lex order. For example, the command `Compositions[0, 3]` produces the single row

$$\begin{array}{ccc} 0 & 0 & 0 \end{array} \quad (\text{S.2.12})$$

As a second example, the command `Compositions[1, 3]` produces the rectangular array

$$\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array} \quad (\text{S.2.13})$$

As a third example, the command `Compositions[2, 3]` produces the rectangular array

$$\begin{array}{ccc} 0 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 0 & 0 \end{array} \quad (\text{S.2.14})$$

- The command `Reverse` acts on the list of arrays, and reverses the order of the list while leaving the arrays intact. For example, the nested sequence of commands `Reverse[Compositions[1, 3]]` produces the rectangular array

$$\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \quad (\text{S.2.15})$$

As a second example, the nested sequence of commands `Reverse[Compositions[2, 3]]` produces the rectangular array

$$\begin{array}{ccc} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{array} \quad (\text{S.2.16})$$

Now the compositions appear in *descending* lex order.

- Look, for example, at Table 2.1. We see that the exponents j_a for the $r = 1$ entry are those appearing in (2.12). Next, exponents for the $r = 2$ through $r = 4$ entries are those appearing in (2.15). Following them, the exponents for the $r = 5$ through $r = 10$ entries, are those appearing in (2.16), etc. Evidently, to produce the exponent list of Table 2.1, what we must do is successively *join* various lists. That is what the *Mathematica* command `Join` accomplishes.

We are now ready to describe how **GAMMA** is constructed:

- The second line in (2.11) sets the values of m and p . They are assigned the values $m = 3$ and $p = 4$ for this example, which will construct **GAMMA** for the case of Table 2.1. The third line in (2.11) initially sets **GAMMA** to a row of m zeroes. The fourth line is a `Do` loop that successively redefines **GAMMA** by generating and joining to it successive descending lex order compositions. The net result is the exponent list of Table 2.1.
- The quantity $L = L(m, p)$ is obtained by applying the *Mathematica* command `Length` to the the rectangular array **GAMMA**.
- The last 6 lines of (2.11) illustrate that L is computed properly and that the command `GAMMA[[r, a]]` accesses the array **GAMMA** in the desired fashion. Specifically, in this example, we find from (2.5) that $L(3, 4) = 35$ in agreement with the *Mathematica* output for L . Moreover, `GAMMA[[17]]` produces the exponent array $\{0, 3, 0\}$, in agreement with the $r = 17$ entry in Table 2.1, and `GAMMA[[17, 2]]` produces $\Gamma_{17,2} = 3$, as expected.

S.2.3 Pyramid Operations: General Procedure

Here we *derive* the pyramid operations in terms of \mathbf{j} -vectors by using the ordering previously described, and provide scripts to *encode* them in the r -representation (2.7).

Definition 1. Suppose that $w(\mathbf{z})$ arises from carrying out various *arithmetic operations* on $u(\mathbf{z})$ and $v(\mathbf{z})$, and the associated pyramids **U** and **V** are known. The corresponding pyramid operation on **U** and **V** is so defined that it yields the pyramid **W** of $w(\mathbf{z})$.

Here we assume that u, v, w are polynomials such as (2.6).

S.2.4 Pyramid Operations: Scalar Multiplication and Addition

We begin with the operations of scalar multiplication and addition, which are easy to define and implement. If

$$w(\mathbf{z}) = c u(\mathbf{z}), \quad (\text{S.2.17})$$

then

$$\mathbf{W}(r) = c \mathbf{U}(r), \quad (\text{S.2.18})$$

and we write

$$\mathbf{W} = c \mathbf{U}. \quad (\text{S.2.19})$$

If

$$w(\mathbf{z}) = u(\mathbf{z}) + v(\mathbf{z}), \quad (\text{S.2.20})$$

then

$$W(r) = U(r) + V(r), \quad (\text{S.2.21})$$

and we write

$$W = U + V. \quad (\text{S.2.22})$$

In both cases all operations are performed coordinate-wise (as for vectors).

Implementation of scalar multiplication and vector addition is easy in *Mathematica* because, as the example below illustrates, it has built in vector routines. There we define two vectors, multiply them by scalars, and add the resulting vectors.

```
Unprotect[V];
U = {1, 2, 3};
V = {4, 5, 6};
W = .1U + .2V
{.9, 1.2, 1.5} \quad (\text{S.2.23})
```

Since *V* is a “protected” symbol in the *Mathematica* language, and, for purposes of illustration, we wish to use it as an ordinary vector variable, it must first be unprotected as in line 1 above. The last line shows that the *Mathematica* output is indeed the desired result.

S.2.5 Pyramid Operations: Background for Polynomial Multiplication

The operation of polynomial multiplication is more involved. Now we have the relation

$$w(\mathbf{z}) = u(\mathbf{z}) * v(\mathbf{z}), \quad (\text{S.2.24})$$

and we want to encode

$$W = \text{PROD}[U, V]. \quad (\text{S.2.25})$$

Shown below is *Mathematica* code that implements this operation,

$$\text{PROD}[U_, V_] := \text{Table}[U[[B[[k]]]] \cdot V[[\text{Brev}[[k]]]], \{k, 1, L, 1\}]; \quad (\text{S.2.26})$$

Our next task is to describe and explain the ingredients in (2.26).

Let us write $u(\mathbf{z})$ in the form (2.6), but with a change of dummy indices, so that it has the representation

$$u(\mathbf{z}) = \sum_{\mathbf{i} \in \Gamma_m^p} U(\mathbf{i}) G_{\mathbf{i}}(\mathbf{z}). \quad (\text{S.2.27})$$

Similarly, write $v(\mathbf{z})$ in the form

$$v(\mathbf{z}) = \sum_{\mathbf{j} \in \Gamma_m^p} V(\mathbf{j}) G_{\mathbf{j}}(\mathbf{z}). \quad (\text{S.2.28})$$

Then there is the result

$$u(\mathbf{z}) * v(\mathbf{z}) = \sum_{\mathbf{i} \in \Gamma_m^p} \sum_{\mathbf{j} \in \Gamma_m^p} U(\mathbf{i}) V(\mathbf{j}) G_{\mathbf{i}}(\mathbf{z}) * G_{\mathbf{j}}(\mathbf{z}). \quad (\text{S.2.29})$$

From (2.1) we observe that

$$\begin{aligned} G_{\mathbf{i}}(\mathbf{z}) * G_{\mathbf{j}}(\mathbf{z}) &= (z_1)^{i_1}(z_2)^{i_2} \cdots (z_m)^{i_m} * (z_1)^{j_1}(z_2)^{j_2} \cdots (z_m)^{j_m} \\ &= (z_1)^{i_1+j_1}(z_2)^{i_2+j_2} \cdots (z_m)^{i_m+j_m} = G_{\mathbf{i}+\mathbf{j}}(\mathbf{z}). \end{aligned} \quad (\text{S.2.30})$$

Therefore, we may also write

$$u(\mathbf{z}) * v(\mathbf{z}) = \sum_{\mathbf{i} \in \Gamma_m^p} \sum_{\mathbf{j} \in \Gamma_m^p} U(\mathbf{i})V(\mathbf{j})G_{\mathbf{i}+\mathbf{j}}(\mathbf{z}). \quad (\text{S.2.31})$$

Now we see that there are two complications. First, there may be terms on the right side of (2.31) whose degree is higher than p and therefore need not be computed. Second, there are generally many terms on the right side of (2.31) that contribute to a given monomial term in $w(\mathbf{z}) = u(\mathbf{z}) * v(\mathbf{z})$. Suppose we write

$$w(\mathbf{z}) = \sum_{\mathbf{k}} W(\mathbf{k}) G_{\mathbf{k}}(\mathbf{z}). \quad (\text{S.2.32})$$

Upon comparing (2.31) and (2.32) we conclude that there is the multidimensional *Cauchy* product rule

$$W(\mathbf{k}) = \sum_{\mathbf{i}+\mathbf{j}=\mathbf{k}} U(\mathbf{i})V(\mathbf{j}) = \sum_{\mathbf{j} \leq \mathbf{k}} U(\mathbf{k}-\mathbf{j})V(\mathbf{j}). \quad (\text{S.2.33})$$

Here, by $\mathbf{j} \leq \mathbf{k}$, we mean that the sum ranges over all \mathbf{j} such that $j_a \leq k_a$ for all $a \in [1, m]$. That is,

$$\mathbf{j} \leq \mathbf{k} \Leftrightarrow j_a \leq k_a \text{ for all } a \in [1, m]. \quad (\text{S.2.34})$$

Evidently, to implement the relation (2.33) in terms of r labels, we need to describe the exponent relation $\mathbf{j} \leq \mathbf{k}$ in terms of r labels. Suppose \mathbf{k} is some exponent vector with label $r(\mathbf{k})$ as, for example, in Table 2.1. Introduce the notation

$$k = r(\mathbf{k}). \quad (\text{S.2.35})$$

This notation may be somewhat confusing because k is not the norm of the vector \mathbf{k} , but rather the label associated with \mathbf{k} . However, this notation is very convenient. Now, given a label k , we can find \mathbf{k} . Indeed, from (2.10), we have the result

$$k_a = \Gamma_{k,a}. \quad (\text{S.2.36})$$

Having found \mathbf{k} , we define a set of exponents B_k by the rule

$$B_k = \{\mathbf{j} | \mathbf{j} \leq \mathbf{k}\}. \quad (\text{S.2.37})$$

This set of exponents is called the k^{th} box. Note that the heads of the vectors \mathbf{j} that satisfy (2.37) for some fixed vector \mathbf{k} do indeed lie within some hyper-rectangular volume (box). For example (when $m = 3$), suppose $k = 28$. Then we see from Table 2.1 that $\mathbf{k} = (1, 2, 1)$. Table 2.2 lists, in modified glex order, all the vectors in B_{28} , i.e. all vectors \mathbf{j} such that

Table S.2.2: The vectors in $B_{28} = \{\mathbf{j} | \mathbf{j} \leq (1, 2, 1)\}$.

r	j_1	j_2	j_3	D
1	0	0	0	0
2	1	0	0	1
3	0	1	0	1
4	0	0	1	1
6	1	1	0	2
7	1	0	1	2
8	0	2	0	2
9	0	1	1	2
14	1	2	0	3
15	1	1	1	3
18	0	2	1	3
28	1	2	1	4

$\mathbf{j} \leq (1, 2, 1)$. These are the vectors whose heads are shown in blue in Figure 2.1. Finally, with this notation, we can rewrite (2.33) in the form

$$W(\mathbf{k}) = \sum_{\mathbf{j} \in B_k} U(\mathbf{k} - \mathbf{j}) V(\mathbf{j}). \quad (\text{S.2.38})$$

What can be said about the vectors $(\mathbf{k} - \mathbf{j})$ as \mathbf{j} ranges over B_ℓ ? Table 2.3 lists, for example, the vectors $\mathbf{j} \in B_{28}$ and the associated vectors \mathbf{i} with $\mathbf{i} = (\mathbf{k} - \mathbf{j})$. Also listed are the labels $r(\mathbf{j})$ and $r(\mathbf{i})$. Compare columns 2,3,4, which specify the $\mathbf{j} \in B_{28}$, with columns 5,6,7, which specify the associated \mathbf{i} vectors. We see that every vector that appears in the \mathbf{j} list also occurs somewhere in the \mathbf{i} list, and vice versa. This is to be expected because the operation of multiplication is commutative: we can also write (2.33) in the form

$$W(\mathbf{k}) = \sum_{\mathbf{j} \in B_k} U(\mathbf{j}) V(\mathbf{k} - \mathbf{j}). \quad (\text{S.2.39})$$

We also observe the more remarkable feature that the two lists are *reverses* of each other: running down the \mathbf{j} list gives the same vectors as running up the \mathbf{i} list, and vice versa. This feature is a consequence of our ordering procedure.

As indicated earlier, what we really want is a version of (2.33) that involves labels instead of exponent vectors. Looking at Table 2.3, we see that this is easily done. We may equally well think of B_k as containing a collection of labels $r(\mathbf{j})$, and we may introduce a *reversed* array $Brev_k$ of *complementary* labels $r^c(\mathbf{j})$ where

$$r^c(\mathbf{j}) = r(\mathbf{i}). \quad (\text{S.2.40})$$

That is, for example, B_{28} would consist of the first column of Table 2.3 and $Brev_{28}$ would consist of the last column of Table 2.3. Finally, we have already introduced k as being the

label associated with \mathbf{k} . With these understandings in mind, we may rewrite (2.33) in the label form

$$W(k) = \sum_{r \in B_k} U(r^c)V(r) = \sum_{r \in B_k} U(r)V(r^c). \quad (\text{S.2.41})$$

This is the rule $W = \text{PROD}[U, V]$ for multiplying pyramids. In the language of Section 32.7, B_k and $Brev_k$, when taken together, provide a *look back table* that, given k , look back to find all monomial pairs with labels r, r^c which produce, when multiplied, the monomial with label k .

Table S.2.3: The vectors \mathbf{j} and $\mathbf{i} = (\mathbf{k} - \mathbf{j})$ for $\mathbf{j} \in B_{28}$ and $k_a = \Gamma_{28,a}$.

$r(\mathbf{j})$	j_1	j_2	j_3	i_1	i_2	i_3	$r(\mathbf{i})$
1	0	0	0	1	2	1	28
2	1	0	0	0	2	1	18
3	0	1	0	1	1	1	15
4	0	0	1	1	2	0	14
6	1	1	0	0	1	1	9
7	1	0	1	0	2	0	8
8	0	2	0	1	0	1	7
9	0	1	1	1	1	0	6
14	1	2	0	0	0	1	4
15	1	1	1	0	1	0	3
18	0	2	1	1	0	0	2
28	1	2	1	0	0	0	1

S.2.6 Pyramid Operations: Implementation of Multiplication

The code shown below in (2.42) illustrates how B_k and $Brev_k$ are constructed using *Mathematica*.

```

JSK[list_, K_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, K]]] & /@ list, True] // Flatten;
B = Table[JSK[GAMMA, GAMMA[[k]]], {k, 1, L}];
Brev = Reverse /@ B;                                     (S.2.42)

```

As before, some explanation is required. The main tasks are to implement the $\mathbf{j} \leq \mathbf{k}$ operation (2.34) and then to employ this implementation. We will begin by implementing the $\mathbf{j} \leq \mathbf{k}$ operation. Several steps are required, and each of them is described briefly below:

- When *Mathematica* is presented with a statement of the form $j \leq k$, with j and k being integers, it replies with the answer True or the answer False. (Here $j \leq k$

denotes $j \leq k$.) Two sample *Mathematica* runs are shown below:

$$\begin{aligned} 3 &\leq 4 \\ \text{True} & \end{aligned} \tag{S.2.43}$$

$$\begin{aligned} 5 &\leq 4 \\ \text{False} & \end{aligned} \tag{S.2.44}$$

- A *Mathematica* function can be constructed that does the same thing. It takes the form

$$\#1 \leq \#2 \& [j, k] \tag{S.2.45}$$

Here the symbols **#1** and **#2** set up two *slots* and the symbol **&** means the operation to its left is to be regarded as a function and is to be applied to the arguments to its right by inserting the arguments into the slots. Below is a *Mathematica* run illustrating this feature.

$$\begin{aligned} j &= 3; k = 4; \\ \#1 &\leq \#2 \& [j, k] \\ \text{True} & \end{aligned} \tag{S.2.46}$$

Observe that the output of this run agrees with that of (2.43).

- The same operation can be performed on pairs of *arrays* (rather than pairs of numbers) in such a way that corresponding entries from each array are compared, with the output then being an array of True and False answers. This is done using the *Mathematica* command **Thread**. Below is a *Mathematica* run illustrating this feature.

$$\begin{aligned} j &= \{1, 2, 3\}; k = \{4, 5, 1\}; \\ \text{Thread}[\#1 &\leq \#2 \& [j, k]] \\ \{\text{True}, \text{True}, \text{False}\} & \end{aligned} \tag{S.2.47}$$

Note that the first two answers in the output array are True because the statements $1 \leq 4$ and $2 \leq 5$ are true. The last answer in the output array is False because the statement $3 \leq 1$ is false.

- Suppose, now, that we are given two arrays ***j*** and ***k*** and we want to determine if $\mathbf{j} \leq \mathbf{k}$ in the sense of (2.34). This can be done by *applying* the logical **And** operation (using the *Mathematica* command **Apply**) to the True/False output array described above. Below is a *Mathematica* run illustrating this feature.

$$\begin{aligned} j &= \{1, 2, 3\}; k = \{4, 5, 1\}; \\ \text{Apply}[\text{And}, \text{Thread}[\#1 &\leq \#2 \& [j, k]]] \\ \text{False} & \end{aligned} \tag{S.2.48}$$

Note that the output answer is False because at least one of the entries in the output array in (2.47) is False. The output answer would be True if, and only if, all entries in the output array in (2.47) were True.

- Now that the $\mathbf{j} \leq \mathbf{k}$ operation has been defined for two exponent arrays, we would like to construct a related operator/function, to be called JSK. (Here the letter S stands for *smaller than or equal to*.) It will depend on the exponent array \mathbf{k} , and its task will be to search a list of exponent arrays to find those \mathbf{j} within it that satisfy $\mathbf{j} \leq \mathbf{k}$. The first step in this direction is to slightly modify the function appearing in (2.48). Below is a *Mathematica* run that specifies this modified function and illustrates that it has the same effect.

```
j = {1, 2, 3}; k = {4, 5, 1};
Apply[And, Thread[#1 <= #2 & [#, k]]] & [j]
False
```

(S.2.49)

Comparison of the functions in (2.48) and (2.49) reveals that what has been done is to replace the argument j in (2.48) by a slot $\#$, then follow the function by the character $\&$, and finally add the symbols $[j]$. What this modification does is to redefine the function in such a way that it acts on what follows the second $\&$.

- The next step is to extend the function appearing in (2.49) so that it acts on a list of exponent arrays. To do this, we replace the symbols $[j]$ by the symbols $/@$ list. The symbols $/@$ indicate that what stands to their left is to act on what stands to their right, and what stands to their right is a list of exponent arrays. The result of this action will be a list of True/False results with one result for each exponent array in the list. Below is a *Mathematica* run that illustrates how the further modified function acts on lists.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list
{True, False, True}
```

(S.2.50)

Observe that the output answer list is $\{\text{True}, \text{False}, \text{True}\}$ because $\{3, 4, 1\} \leq \{4, 5, 1\}$ is True, $\{1, 2, 3\} \leq \{4, 5, 1\}$ is False, and $\{1, 2, 1\} \leq \{4, 5, 1\}$ is True.

- What we would really like to know is where the True items are in the list, because that will tell us where the \mathbf{j} that satisfy $\mathbf{j} \leq \mathbf{k}$ reside. This can be accomplished by use of the *Mathematica* command *Position* in conjunction with the result True. Below is a *Mathematica* run that illustrates how this works.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]
{{1}, {3}}
```

(S.2.51)

Note that the output is an array of positions in the list for which $j \leq k$. There is, however, still one defect. Namely, the output array is an array of single-element subarrays, and we would like it to be simply an array of location numbers. This defect can be remedied by appending the *Mathematica* command `Flatten`, preceded by `//`, to the instruction string in (2.51). The *Mathematica* run below illustrates this modification.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten
{1, 3}                                         (S.2.52)
```

Now the output is a simple array containing the positions in the list for which $j \leq k$.

- The last step is to employ the ingredients in (2.52) to define the operator `JSK[list, k]`. The *Mathematica* run below illustrates how this can be done.

```
k = {4, 5, 1};
ja = {3, 4, 1}; jb = {1, 2, 3}; jc = {1, 2, 1};
list = {ja, jb, jc};
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
JSK[list, k]
{1, 3}                                         (S.2.53)
```

Lines 4 and 5 above define the operator `JSK[list, k]`, line 6 invokes it, and line 7 displays its output, which agrees with the output of (2.52).

- With the operator `JSK[list, k]` in hand, we are prepared to construct tables B and $Brev$ that will contain the B_k and the $Brev_k$. The *Mathematica* run below illustrates how this can be done.

```
B = Table[JSK[GAMMA, GAMMA[[k]]], {k, 1, L, 1}];
Brev = Reverse /@ B;
B[[8]]
Brev[[8]]
B[[28]]
Brev[[28]]
{1, 3, 8}
{8, 3, 1}
{1, 2, 3, 4, 6, 7, 8, 9, 14, 15, 18, 28}
{28, 18, 15, 14, 9, 8, 7, 6, 4, 3, 2, 1}          (S.2.54)
```

The first line employs the *Mathematica* command **Table** in combination with an implied Do loop to produce a two-dimensional array **B**. Values of k in the range $[1, L]$ are selected sequentially. For each k value the associated exponent array $\mathbf{k}(k) = \text{GAMMA}[[\mathbf{k}]]$ is obtained. The operator **JSK** then searches the full **GAMMA** array to find the list of r values associated with the $\mathbf{j} \leq \mathbf{k}$. All these r values are listed in a row. Thus, the array **B** consists of list of L rows, of varying width. The rows are labeled by $k \in [1, L]$, and in each row are the r values associated with the $\mathbf{j} \leq \mathbf{k}$. In the second line the *Mathematica* command **Reverse** is applied to **B** to produce a second array called **Brev**. Its rows are the reverse of those in **B**. For example, as the *Mathematica* run illustrates, **B[[8]]**, which is the 8th row of **B**, contains the list {1, 3, 8}, and **Brev[[8]]** contains the list {8, 3, 1}. Inspection of the $r = 8$ monomial in Table 2.1, that with exponents {0, 2, 0}, shows that it has the monomials with exponents {0,0,0}, {0,1,0}, and {0,2,0} as factors. And further inspection of Table 2.1 shows that the exponents of these factors have the r values {1, 3, 8}. Similarly **B[[28]]**, which is the 28th row of **B**, contains the same entries that appear in the first column of Table 2.3. And **Brev[[28]]**, which is the 28th row of **Brev**, contains the same entries that appear in the last column of Table 2.3.

Finally, we need to explain how the arrays **B** and **Brev** can be employed to carry out polynomial multiplication. This can be done using the *Mathematica* dot product command:

- The exhibit below shows a simple *Mathematica* run that illustrates the use of the dot product command.

```

Unprotect[V];
U = {.1, .2, .3, .4, .5, .6, .7, .8};
V = {1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8};
U.V
u = {1, 3, 5};
v = {6, 4, 2};
U[[u]]
V[[v]]
U[[u]].V[[v]]
5.64
{.1, .3, .5}
{1.6, 1.4, 1.2}
1.18
(S.2.55)

```

As before, **V** must be unprotected. See line 1. The rest of the first part this run (lines 2 through 4) defines two vectors **U** and **V** and then computes their dot product. Note that if we multiply the entries in **U** and **V** pairwise and add, we get the result

$$.1 \times 1.1 + .2 \times 1.2 + \cdots + .8 \times 1.8 = 5.64,$$

which agrees with the *Mathematica* result for $\mathbf{U} \cdot \mathbf{V}$. See line 10.

The second part of this *Mathematica* run, lines 5 through 9, illustrates a powerful feature of the *Mathematica* language. Suppose, as illustrated, we define two arrays \mathbf{u} and \mathbf{v} of integers, and use these arrays as *arguments* for the vectors by writing $\mathbf{U}[[\mathbf{u}]]$ and $\mathbf{V}[[\mathbf{v}]]$. Then *Mathematica* uses the integers in the two arrays \mathbf{u} and \mathbf{v} as labels to select the corresponding entries in \mathbf{U} and \mathbf{V} , and from these entries it makes new corresponding vectors. In this example, the 1st, 3rd, and 5th entries in \mathbf{U} are .1, .3, and .5. And the 6th, 4th, and 2nd entries in \mathbf{V} are 1.6, 1.4, and 1.2. Consequently, we find that

$$\mathbf{U}[[\mathbf{u}]] = \{.1, .3, .5\},$$

$$\mathbf{V}[[\mathbf{v}]] = \{1.6, 1.4, 1.2\},$$

in agreement with lines 11 and 12 of the *Mathematica* results. Correspondingly, we expect that $\mathbf{U}[[\mathbf{u}]] \cdot \mathbf{V}[[\mathbf{v}]]$ will have the value

$$\mathbf{U}[[\mathbf{u}]] \cdot \mathbf{V}[[\mathbf{v}]] = .1 \times 1.6 + .3 \times 1.4 + .5 \times 1.2 = 1.18,$$

in agreement with the last line of the *Mathematica* output.

- Now suppose, as an example, that we set $k = 8$ and use $\mathbf{B}[[\mathbf{k}]]$ and $\mathbf{Brev}[[\mathbf{k}]]$ in place of the arrays \mathbf{u} and \mathbf{v} . The *Mathematica* fragment below shows what happens when this is done.

```

k = 8;
B[[k]]
Brev[[k]]
U[[B[[k]]]]
V[[Brev[[k]]]]
U[[B[[k]]]] . V[[Brev[[k]]]]
{1,3,8}
{8,3,1}
{.1,.3,.8}
{1.8,1.3,1.1}
1.45
(S.2.56)

```

From (2.54) we see that $\mathbf{B}[[8]] = \{1, 3, 8\}$ and $\mathbf{Brev}[[8]] = \{8, 3, 1\}$ in agreement with lines 7 and 8 of the *Mathematica* output above. Also, the 1st, 3rd, and 8th entries in \mathbf{U} are .1, .3, and .8. And the 8th, 3rd, and 1st entries in \mathbf{V} are 1.8, 1.3, and 1.1. Therefore we expect the results

$$\mathbf{U}[[\mathbf{B}[[\mathbf{k}]]]] = \{.1, .3, .8\},$$

$$\mathbf{V}[[\mathbf{Brev}[[\mathbf{k}]]]] = \{1.8, 1.3, 1.1\},$$

$$\mathbf{U}[[\mathbf{B}[[\mathbf{k}]]]] \cdot \mathbf{V}[[\mathbf{Brev}[[\mathbf{k}]]]] = .1 \times 1.8 + .3 \times 1.3 + .8 \times 1.1 = 1.45,$$

in agreement with the last three lines of (2.56).

- Finally, suppose we carry out the operation $U[[B[[k]]]] \cdot V[[\text{Brev}[[k]]]]$ for all $k \in [1, L]$ and put the results together in a Table with entries labeled by k . According to (2.41), the result will be the pyramid for the product of the two polynomials whose individual pyramids are U and V . The *Mathematica* fragment (2.26), which is displayed again below, shows how this can be done to define a *product* function, called PROD, that acts on general pyramids U and V , using the command Table with an implied Do loop over k .

```
PROD[U_, V_] := Table[U[[B[[k]]]] \cdot V[[\text{Brev}[[k]]]], {k, 1, L, 1}];
```

Let us verify that this whole multiplication procedure works for a simple example. For the sake of brevity, we will consider the case of $m = 2$ variables and work through terms of degree $p = 3$. In this case pyramids have the modest length $L(2, 3) = 10$. Table 2.4 provides a labeling scheme for monomials in two variables using our standard modified glex sequencing.

Table S.2.4: A labeling scheme for monomials in two variables.

r	j_1	j_2
1	0	0
2	1	0
3	0	1
4	2	0
5	1	1
6	0	2
7	3	0
8	2	1
9	1	2
10	0	3
.	.	.
.	.	.

Suppose, for example, that u and v are the functions

$$u(\mathbf{z}) = 1 + 2z_1 + 3z_2 + 4z_1z_2 \quad (\text{S.2.57})$$

and

$$v(\mathbf{z}) = 5 + 6z_1 + 7z_2^2. \quad (\text{S.2.58})$$

From Table 2.4 we find that the corresponding pyramids U and V are

$$U = \{1, 2, 3, 0, 4, 0, 0, 0, 0, 0\} \quad (\text{S.2.59})$$

and

$$V = \{5, 6, 0, 0, 0, 7, 0, 0, 0, 0\}. \quad (\text{S.2.60})$$

Polynomial multiplication gives the result

$$\begin{aligned}
 w(\mathbf{z}) &= u(\mathbf{z}) * v(\mathbf{z}) \\
 &= 5 + 16z_1 + 15z_2 + 12z_1^2 + 38z_1z_2 + 7z_2^2 + 24z_1^2z_2 + 14z_1z_2^2 + 21z_2^3 + 28z_1z_2^3.
 \end{aligned} \tag{S.2.61}$$

Correspondingly, through terms of degree 3, the pyramid $W = \text{PROD}[U, V]$ is given by

$$W = \{5, 16, 15, 12, 38, 7, 0, 24, 14, 21\}. \tag{S.2.62}$$

Below is an execution of a *Mathematica* program illustrating the use of the product function for the polynomials u and v given by (2.57) and (2.58).

```

Clear["Global`*"];
Needs["Combinatorica`"];
m = 2; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA]
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
U = {1, 2, 3, 0, 4, 0, 0, 0, 0, 0};
V = {5, 6, 0, 0, 0, 7, 0, 0, 0, 0};
10
PROD[U, V]
{5, 16, 15, 12, 38, 7, 0, 24, 14, 21} \tag{S.2.63}

```

The first 11 lines of the code set up the necessary arrays and define the product function in pyramid form. The next two lines specify the pyramids U and V given in (2.59) and (2.60). The third line from the bottom, which results from the command in line 6, illustrates that indeed $L(2, 3) = 10$. The final two lines show that use of the product function when applied to the pyramids U and V does indeed product the pyramid W given by (2.62).

S.2.7 Pyramid Operations: Implementation of Powers

With operation of multiplication in hand, it is easy to implement the operation of raising a pyramid to a power. The code shown below in (2.64) demonstrates how this can be done.

```

POWER[U_, 0] := C1;
POWER[U_, 1] := U;
POWER[U_, 2] := PROD[U, U];
POWER[U_, 3] := PROD[U, POWER[U, 2]];
...

```

(S.2.64)

Here $C1$ is the pyramid for the Taylor series having *one* as its *constant* term and all other terms zero,

$$C1 = \{1, 0, 0, 0, \dots\}. \quad (\text{S.2.65})$$

It can be set up by the *Mathematica* code

$$C1 = \text{Table}[\text{KroneckerDelta}[k, 1], \{k, 1, L, 1\}]; \quad (\text{S.2.66})$$

which employs the Table command, the Kronecker delta function, and an implied Do loop over k . This code should be executed before executing (2.64), but after the value of L has been established.

S.2.8 Replacement Rule and Automatic Differentiation

Definition 2. The transformation $A(z) \rightsquigarrow A$ means replacement of every real variable z_a in the *arithmetic expression* $A(z)$ with an associated pyramid, and of every operation on real variables in $A(z)$ with the associated operation on pyramids.

Automatic differentiation is based on the following corollary: if $A(z) \rightsquigarrow A$, then A is the pyramid of $A(z)$.

For simplicity, we will begin our discussion of the replacement rule with examples involving only a single variable z . In this case monomial labeling, the relation between labels and exponents, is given directly by the simple rules

$$r(j) = 1 + j \text{ and } j(r) = r - 1. \quad (\text{S.2.67})$$

See Table 2.5.

As a first example, consider the expression

$$A = 2 + 3(z * z). \quad (\text{S.2.68})$$

We have agreed to consider the case $m = 1$. Suppose we also set $p = 2$, in which case $L = 3$. In ascending glex order, see Table 2.5, the pyramid for A is then

$$2 + 3z^2 \rightsquigarrow A = (2, 0, 3). \quad (\text{S.2.69})$$

Now imagine that A was not such a simple polynomial, but some complicated expression. Then the pyramid A could be generated by computing derivatives of A at $z = 0$ and dividing

Table S.2.5: A labeling scheme for monomials in one variable.

r	j
1	0
2	1
3	2
4	3
.	.
.	.

them by the appropriate factorials. Automatic differentiation offers another way to find \mathbf{A} . Assume that all operations in the arithmetic expression A have been encoded according to *Definition 1*. For our example, these are $+$ and PROD. Let $\mathbf{C1}$ and \mathbf{Z} be the pyramids associated with 1 and z ,

$$1 \rightsquigarrow \mathbf{C1} = (1, 0, 0), \quad (\text{S.2.70})$$

$$z \rightsquigarrow \mathbf{Z} = (0, 1, 0). \quad (\text{S.2.71})$$

The quantity $2 + 3z^2$ results from performing various arithmetic operations on 1 and z . *Definition 1* says that the pyramid of $2 + 3z^2$ is identical to the pyramid obtained by performing the same operations on the pyramids $\mathbf{C1}$ and \mathbf{Z} . That is, suppose we replace 1 and z with their associated pyramids $\mathbf{C1}$ and \mathbf{Z} , and also replace $*$ with PROD. Then, upon evaluating PROD, multiplying by the appropriate scalar coefficients, and summing, the result will be the same pyramid \mathbf{A} ,

$$2 \mathbf{C1} + 3 \text{PROD}[\mathbf{Z}, \mathbf{Z}] = \mathbf{A}. \quad (\text{S.2.72})$$

In this way, by knowing only the basic pyramids $\mathbf{C1}$ and \mathbf{Z} (prepared beforehand), one can compute the pyramid of an arbitrary $A(z)$. Finally, in contrast to numerical differentiation, all numerical operations involved are accurate to machine precision. *Mathematica* code that implements (2.72) will be presented shortly in (2.73).

Frequently, if $A(z)$ is some complicated expression, the replacement rule will result in a long chain of nested pyramid operations. At every step in the chain the present pyramid, the pyramid resulting from the previous step, will be combined with some other pyramid to produce a new pyramid. Each such operation has two arguments (the present pyramid and some other pyramid), and *Definition 1* applies to each step in the chain. Upon evaluating all pyramid operations, the final result will be the pyramid of $A(z)$.

By using the replacement operation the above procedure can be represented as:

$$1 \rightsquigarrow \mathbf{C1}, \quad z \rightsquigarrow \mathbf{Z}, \quad A \rightsquigarrow \mathbf{A}.$$

The following general recipe then applies: In order to derive the pyramid associated with some arithmetic expression, apply the \rightsquigarrow rule to all its variables, or parts, and replace all operations with operations on pyramids. Here “apply the \rightsquigarrow rule” to something means replace that something with the associated pyramid. And the term “parts” means subexpressions. *Definition 1* guarantees that the result will be the same pyramid \mathbf{A} no matter how

we split the arithmetic expression A into subexpressions. It is only necessary to recognize, in case of using subexpressions, that one pyramid expression should be viewed as a function of another.

For illustration, suppose we regard the A given by (2.68) to be the composition of two functions, $F(z) = 2 + 3z$ and $G(z) = z^2$, so that $A(z) = F(G(z))$. Instead of associating a constant and a single variable with their respective pyramids, let us now associate whole subexpressions. In addition, let us label the pyramid expressions on the right of \rightsquigarrow with some names, F and G :

$$2 + 3z \rightsquigarrow 2 \text{ C1} + 3 \text{ Z} = F[\text{Z}]$$

$$z^2 \rightsquigarrow \text{PROD}[\text{Z}, \text{Z}] = G[\text{Z}]$$

$$A(z) \rightsquigarrow F[G[\text{Z}]] = A.$$

We have indicated the explicit dependence on Z . It is important to note that $F[\text{Z}]$ is a pyramid *expression* prior to executing any the pyramid operations, i.e it is not yet a pyramid, but is simply the result of formal replacements that follow the association rule.

Mathematica code for the simple example (2.72) is shown below,

$$\begin{aligned} \text{C1} &= \{1, 0, 0\}; \\ \text{Z} &= \{0, 1, 0\}; \\ 2 \text{ C1} + 3 \text{ PROD}[\text{Z}, \text{Z}] & \\ \{2, 0, 3\} & \end{aligned} \tag{S.2.73}$$

Note that the result (2.73) agrees with (2.69). This example does not use any nested expressions. We will now illustrate how the same results can be obtained using nested expressions.

We begin by displaying a simple *Mathematica* program/execution, that employs ordinary variables, and uses *Mathematica*'s intrinsic abilities to handle nested expressions. The program/execution is

$$\begin{aligned} f[z_]:=2+3z; \\ g[z_]:=z^2; \\ f[g[z]] \\ 2+3z^2 \end{aligned} \tag{S.2.74}$$

With *Mathematica* the underscore in $z_$ indicates that z is a dummy variable name, and the symbols $:=$ indicate that f is defined with a delayed assignment. That is what is done in line one above. The same is done in line two for g . Line three requests evaluation of the nested function $f(g(z))$, and the result of this evaluation is displayed in line four. Note that the result agrees with (2.68).

With this background, we are ready to examine a program with analogous nested pyramid operations. The same comments apply regarding the use of underscores and delayed

assignments. The program is

$$\begin{aligned}
 \text{C1} &= \{1, 0, 0\}; \\
 \text{Z} &= \{0, 1, 0\}; \\
 \text{F[Z]} &:= 2 \text{ C1} + 3 \text{ Z}; \\
 \text{G[Z]} &:= \text{PROD}[\text{Z}, \text{Z}]; \\
 \text{F[G[Z]]} & \\
 \{2, 0, 3\} &
 \end{aligned} \tag{S.2.75}$$

Note that line (2.75) agrees with line (2.73), and is consistent with line (2.69).

S.2.9 Taylor Rule

We close this section with an important consequence of the replacement rule and nested operations, which we call the *Taylor* rule. We begin by considering functions of a single variable. Suppose the function $G(x)$ has the special form

$$G(x) = z^d + x \tag{S.2.76}$$

where z^d is some constant. Let F be some other function. Consider the composite (nested) function A defined by

$$A(x) = F(G(x)) = F(z^d + x). \tag{S.2.77}$$

Then, assuming the necessary analyticity and by the chain rule, A evidently has a Taylor expansion in x about the origin of the form

$$\begin{aligned}
 A &= A(0) + A'(0)x + (1/2)A''(0)x^2 + \dots \\
 &= F(z^d) + F'(z^d)x + (1/2)F''(z^d)x^2 + \dots
 \end{aligned} \tag{S.2.78}$$

We conclude that if we know the Taylor expansion of A about the origin, then we also know the Taylor expansion of F about z^d , and vice versa. Suppose, for example, that

$$F(z) = 1 + 2z + 3z^2 \tag{S.2.79}$$

and

$$z^d = 4. \tag{S.2.80}$$

Then there is the result

$$A(x) = F(G(x)) = F(z^d + x) = 1 + 2(4 + x) + 3(4 + x)^2 = 57 + 26x + 3x^2. \tag{S.2.81}$$

We now show that this same result can be obtained using pyramids. The *Mathematica* fragment below illustrates how this can be done.

$$\begin{aligned}
 \text{C1} &= \{1, 0, 0\}; \\
 \text{X} &= \{0, 1, 0\}; \\
 \text{zd} &= 4; \\
 \text{F[Z]} &:= 1 \text{ C1} + 2 \text{ Z} + 3 \text{ PROD}[\text{Z}, \text{Z}]; \\
 \text{G[X]} &:= \text{zd} \text{ C1} + \text{X}; \\
 \text{F[G[X]]} & \\
 \{57, 26, 3\} &
 \end{aligned} \tag{S.2.82}$$

Note that (2.82) agrees with (2.81). See also Table 2.5.

Let us also illustrate the Taylor rule in the two-variable case. Let $F(z_1, z_2)$ be some function of two variables. Introduce the functions $G(x_1)$ and $H(x_1)$ having the special forms

$$G(x_1) = z_1^d + x_1, \quad (\text{S.2.83})$$

$$H(x_2) = z_2^d + x_2, \quad (\text{S.2.84})$$

where z_1^d and z_2^d are some constants. Consider the function A defined by

$$A(x_1, x_2) = F(G(x_1), H(x_2)) = F(z_1^d + x_1, z_2^d + x_2). \quad (\text{S.2.85})$$

Then, again assuming the necessary analyticity and by the chain rule, A evidently has a Taylor expansion in x_1 and x_2 about the origin $(0, 0)$ of the form

$$\begin{aligned} A &= A(0, 0) + [\partial_1 A(0, 0)]x_1 + [\partial_2 A(0, 0)]x_2 \\ &\quad + (1/2)[(\partial_1)^2 A(0, 0)]x_1^2 + [\partial_1 \partial_2 A(0, 0)]x_1 x_2 + (1/2)[(\partial_2)^2 A(0, 0)]x_2^2 + \dots \\ &= F(z_1^d, z_2^d) + [\partial_1 F(z_1^d, z_2^d)]x_1 + [\partial_2 F(z_1^d, z_2^d)]x_2 \\ &\quad + (1/2)[(\partial_1)^2 F(z_1^d, z_2^d)]x_1^2 + [\partial_1 \partial_2 F(z_1^d, z_2^d)]x_1 x_2 + (1/2)[(\partial_2)^2 F(z_1^d, z_2^d)]x_2^2 + \dots \end{aligned} \quad (\text{S.2.86})$$

where

$$\partial_1 = \partial/\partial x_1, \quad \partial_2 = \partial/\partial x_2 \quad (\text{S.2.87})$$

when acting on A , and

$$\partial_1 = \partial/\partial z_1, \quad \partial_2 = \partial/\partial z_2 \quad (\text{S.2.88})$$

when acting on F . We conclude that if we know the Taylor expansion of A about the origin $(0, 0)$, then we also know the Taylor expansion of F about (z_1^d, z_2^d) , and vice versa.

As a concrete example, suppose that

$$F(z_1, z_2) = 1 + 2z_1 + 3z_2 + 4z_1^2 + 5z_1 z_2 + 6z_2^2 \quad (\text{S.2.89})$$

and

$$z_1^d = 7, \quad z_2^d = 8. \quad (\text{S.2.90})$$

Then, hand calculation shows that $F(G(x_1), H(x_2))$ takes the form

$$\begin{aligned} F(z_1^d + x_1, z_2^d + x_2) &= F(G(x_1), H(x_2)) \\ &= 899 + 98x_1 + 4x_1^2 + 134x_2 + 5x_1 x_2 + 6x_2^2. \end{aligned} \quad (\text{S.2.91})$$

Below is a *Mathematica* execution that finds the same result,

```

F[z1_,z2_]:=1+2 z1+3 z2+4 z12+5 z1 z2+6 z22
G[x1_]:=zd1+x1;
H[x2_]:=zd2+x2;
zd1=7;
zd2=8;
A=F[G[x1],H[x2]]
Expand[A]
1+2 (7+x1)+4 (7+x1)2+3 (8+x2)+5 (7+x1) (8+x2)+6 (8+x2)2
899+98 x1+4 x12+134 x2+5 x1 x2+6 x22
(S.2.92)

```

The calculation above dealt with the case of a function of two ordinary variables. We now illustrate, for the same example, that there is an analogous result for pyramids. Following the replacement rule, we should make the substitutions

$$z_1^d + x_1 \rightsquigarrow zd1 C1 + X1, \quad (\text{S.2.93})$$

$$z_2^d + x_2 \rightsquigarrow zd2 C1 + X2, \quad (\text{S.2.94})$$

$$\begin{aligned}
& 1+2 z_1+3 z_2+4 z_1^2+5 z_1 z_2+6 z_2^2 \rightsquigarrow \\
& C1+2 Z1+3 Z2+4 \text{PROD}[Z1, Z1]+5 \text{PROD}[Z1, Z2]+6 \text{PROD}[Z2, Z2].
\end{aligned} \quad (\text{S.2.95})$$

The *Mathematica* fragment below, executed for the case $m = 2$ and $p = 2$, in which case $L = 6$, illustrates how the analogous result is obtained using pyramids,

```

C1={1,0,0,0,0,0};
X1={0,1,0,0,0,0};
X2={0,0,1,0,0,0};
F[Z1_,Z2_]:=C1+2 Z1+3 Z2+4 PROD[Z1,Z1]+5 PROD[Z1,Z2]
+6 PROD[Z2,Z2];
G[X1_]:=z01 C1 + X1;
H[X2_]:=z02 C1 + X2;
zd1=7;
zd2=8;
F[G[X1],H[X2]]
{899,98,134,4,5,6}
(S.2.96)

```

Note that, when use is made of Table 2.4, the last line of (2.96) agrees with (2.91) and the last line of (2.92).

S.3 Numerical Integration and Replacement Rule

S.3.1 Numerical Integration

Consider the set of differential equations (1.1). As described in Chapter 2, a standard procedure for their numerical integration from an initial time $t^i = t^0$ to some final time t^f is to divide the time axis into a large number of steps N , each of small duration h , thereby introducing successive times t^n defined by the relation

$$t^n = t^0 + nh \quad \text{with } n = 0, 1, \dots, N. \quad (\text{S.3.1})$$

By construction, there will also be the relation

$$Nh = t^f - t^0. \quad (\text{S.3.2})$$

The goal is to compute the vectors \mathbf{z}^n , where

$$\mathbf{z}^n = \mathbf{z}(t^n), \quad (\text{S.3.3})$$

starting from the vector \mathbf{z}^0 . The vector \mathbf{z}^0 is assumed given as a set of definite numbers, i.e. the initial conditions at t^0 .

If we assume for the solution piece-wise analyticity in t , or at least sufficient differentiability in t (which will be the case if the f_a are piece-wise analytic or at least have sufficient differentiability in t), we may convert the set of differential equations (1.1) into a set of recursion relations for the \mathbf{z}^n in such a way that the \mathbf{z}^n obtained by solving the recursion relations differ from the true \mathbf{z}^n by only small truncation errors of order h^m . (Here m is *not* the number of variables, but rather some fixed integer describing the accuracy of the integration method.) One such procedure, a fourth-order *Runge Kutta* (RK4) method, is the set of marching/recursion rules

$$\mathbf{z}^{n+1} = \mathbf{z}^n + \frac{1}{6}(\mathbf{a} + 2\mathbf{b} + 2\mathbf{c} + \mathbf{d}) \quad (\text{S.3.4})$$

where, at each step,

$$\mathbf{a} = h\mathbf{f}(\mathbf{z}^n, t^n), \quad (\text{S.3.5})$$

$$\mathbf{b} = h\mathbf{f}\left(\mathbf{z}^n + \frac{1}{2}\mathbf{a}, t^n + \frac{1}{2}h\right),$$

$$\mathbf{c} = h\mathbf{f}\left(\mathbf{z}^n + \frac{1}{2}\mathbf{b}, t^n + \frac{1}{2}h\right),$$

$$\mathbf{d} = h\mathbf{f}(\mathbf{z}^n + \mathbf{c}, t^n + h).$$

Thanks to the genius of Runge and Kutta, the relations (3.4) and (3.5) have been constructed in such a way that the method is locally (at each step) correct through order h^4 , and makes local truncation errors of order h^5 . Recall Section 2.3.2

In the case of a single variable, and therefore a single differential equation, the relations (3.4) and (3.5) may be encoded in the *Mathematica* form shown below. Here **Zvar** is the dependent variable, **t** is the time, **Zt** is a temporary variable, **tt** is a temporary time, and

`ns` is the number of integration steps. The program employs a Do loop over `i` so that the operations (3.4) and (3.5) are carried out `ns` times.

```
RK4 := (
  t0 = t;
  Do[
    Aa = h F[Zvar, t];
    Zt = Zvar + (1/2)Aa;
    tt = t + h/2;
    Bb = h F[Zt, tt];
    Zt = Zvar + (1/2)Bb;
    Cc = h F[Zt, tt];
    Zt = Zvar + Cc;
    tt = t + h;
    Dd = h F[Zt, tt];
    Zvar = Zvar + (1/6)(Aa + 2 Bb + 2 Cc + Dd);
    t = t0 + i h;
    {i, 1, ns, 1}
  ]
)
```

(S.3.6)

S.3.2 Replacement Rule, Single Equation/Variable Case

We now make what, for our purposes, is a fundamental observation: The operations that occur in the Runge Kutta recursion rules (3.4) and (3.5) and realized in the code above can be extended to pyramids by application of the replacement rule. In particular, the dependent variable z can be replaced by a pyramid, and the various operations involved in the recursion rules can be replaced by pyramid operations. Indeed if we look at the code above, apart from the evaluation of F , we see that the quantities $Zvar$, Zt , Aa , Bb , Cc , and Dd can be viewed, if we wish, as pyramids since the only operations involved are scalar multiplication and addition. The only requirement for a pyramidal interpretation of the `RK4` *Mathematica* code is that the right side of the differential equation, $F[*, *]$, be defined for pyramids. Finally, we remark that the features that make it possible to interpret the `RK4` *Mathematica* code either in terms of ordinary variables or pyramidal variables will hold for *Mathematica* realizations of many other familiar numerical integration methods including other forms of Runge Kutta, predictor-corrector methods, and extrapolation methods.

To make these ideas concrete, and to understand their implications, let us begin with a simple example. Suppose, in the single variable case, that the right side of the differential equation has the simple form

$$f(z, t) = -2tz^2. \quad (\text{S.3.7})$$

The differential equation with this right side can be integrated analytically to yield the solution

$$z(t) = z^0/[1 + z^0(t - t^0)^2]. \quad (\text{S.3.8})$$

In particular, for the case $t^0 = 0$, $z^0 = 1$, and $t = 1$, there is the result

$$z(1) = z^0/[1 + z^0] = 1/2. \quad (\text{S.3.9})$$

Let us also integrate the differential equation with the right side (3.7) numerically. Shown below is the result of running the associated *Mathematica* Runge Kutta code for this case.

```
Clear["Global`*"];
F[Z_, t_] := -2 t Z^2;
h = .1;
ns = 10;
t = 0;
Zvar = 1.;
RK4;
t
Zvar
1.
0.500001
(S.3.10)
```

Note that the last line of (3.10) agrees with (3.9) save for a “1” in the last entry. As expected, and as experimentation shows, this small difference, due to accumulated truncation error, becomes even smaller if h is decreased (and correspondingly, ns is increased).

Suppose we expand the solution (3.9) about the design initial condition $z^{d0} = 1$ by replacing z^0 by $z^{d0} + x$ and expanding the result in a Taylor series in x about the point $x=0$. Below is a *Mathematica* run that performs this task.

```
zd0 = 1;
Series[(zd0 + x)/(1 + zd0 + x), {x, 0, 5}]

$$\frac{1}{2} + \frac{x}{4} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{x^4}{32} + \frac{x^5}{64} + O[x]^6
(S.3.11)$$

```

We will now see that the same Taylor series can be obtained by the operation of numerical integration applied to pyramids. The *Mathematica* code below shows, for our example

differential equation, the application of numerical integration to pyramids.

```

Clear["Global`*"];
Needs["Combinatorica`"];
m = 1; p = 5;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
F[Z_, t_] := -2 t PROD[Z, Z];
h = .01;
ns = 100;
t = 0;
zd0 = 1;
C1 = {1, 0, 0, 0, 0, 0};
X = {0, 1, 0, 0, 0, 0};
Zvar = zd0 C1 + X;
RK4;
t
Zvar
1.
{0.5, 0.25, -0.125, 0.0625, -0.03125, 0.015625}                                (S.3.12)

```

The first 11 lines of the code set up what should be by now the familiar procedure for labeling and multiplying pyramids. In particular, $m = 1$ because we are dealing with a single variable, and $p = 5$ since we wish to work through fifth order. The line

$$F[Z, t] := -2 t PROD[Z, Z] \quad (S.3.13)$$

defines $F[*, *]$ for the case of pyramids, and is the result of applying the replacement rule to the right side of f as given by (3.7),

$$-2 t z^2 \rightsquigarrow -2 t PROD[Z, Z]. \quad (S.3.14)$$

Lines 13 through 15 play the same role as lines 3 through 5 in (3.10) except that, in order to improve numerical accuracy, the step size h has been decreased and correspondingly the number of steps ns has been increased. Lines 16 through 19 now initialize $Zvar$ as a pyramid with a constant part $zd0$ and first-order monomial part with coefficient 1,

$$Zvar = zd0 C1 + X. \quad (S.3.15)$$

These lines are the pyramid equivalent of line 6 in (3.10). Finally lines 20 through 22 are the same as lines 7 through 9 in (3.10). In particular, the line RK4 in (3.10) and the line RK4 in (3.12) refer to exactly the *same* code, namely that in (3.6).

Let us now compare the outputs of (3.10) and (3.12). Comparing the penultimate lines in each we see that the final time $t = 1$ is the same in each case. Comparing the last lines shows that the output $Zvar$ for (3.12) is a pyramid whose first entry agrees with the last line of (3.10). Finally, all the entries in the pyramid output agree with the Taylor coefficients in the expansion (3.11). We see, in the case of numerical integration (of a single differential equation), that replacing the dependent variable by a pyramid, with the initial value of the pyramid given by (3.15), produces a Taylor expansion of the final condition in terms of the initial condition.

What accounts for this near miraculous result? It's the Taylor rule described described in Subsection 2.9. We have already learned that to expand some function $F(z)$ about some point z^d we must evaluate $F(z^d + x)$. See (2.77). We know that the final $Zvar$, call it $Zvar^{fin}$, is an analytic function of the initial $Zvar$, call it $Zvar^{in}$, so that we may write

$$Zvar^{fin} = Zvar^{fin}(Zvar^{in}) = g(Zvar^{in}) \quad (\text{S.3.16})$$

where g is the function that results from following the trajectory from $t = t^{in}$ to $t = t^{fin}$. Therefore, by the Taylor rule, to expand $Zvar^{fin}$ about $Zvar^{in} = z^{d0}$, we must evaluate $Zvar^{fin}(z^{d0} + x)$. That, with the aid of pyramids, is what the code (3.12) accomplishes.

S.3.3 Multi Equation/Variable Case

Because of *Mathematica*'s built-in provisions for handling arrays, the work of the previous section can easily be extended to the case of several differential equations. Consider, as an example, the two-variable case for which \mathbf{f} has the form

$$\begin{aligned} f_1(\mathbf{z}, t) &= -z_1^2, \\ f_2(\mathbf{z}, t) &= +2z_1z_2. \end{aligned} \quad (\text{S.3.17})$$

The differential equations associated with this \mathbf{f} can be solved in closed form to yield, with the understanding that $t^0 = 0$, the solution

$$\begin{aligned} z_1(t) &= z_1^0/(1 + tz_1^0), \\ z_2(t) &= z_2^0(1 + tz_1^0)^2. \end{aligned} \quad (\text{S.3.18})$$

For the final time $t = 1$ we find the result

$$\begin{aligned} z_1(1) &= z_1^0/(1 + z_1^0), \\ z_2(1) &= z_2^0(1 + z_1^0)^2. \end{aligned} \quad (\text{S.3.19})$$

Let us expand the solution (3.19) about the design initial conditions

$$\begin{aligned} z_1^{d0} &= 1, \\ z_2^{d0} &= 2, \end{aligned} \quad (\text{S.3.20})$$

by writing

$$\begin{aligned} z_1^0 &= z_1^{d0} + x_1 = 1 + x_1, \\ z_2^0 &= z_2^{d0} + x_2 = 2 + x_2. \end{aligned} \quad (\text{S.3.21})$$

Doing so gives the results

$$\begin{aligned} z_1(1) &= (1 + x_1)/(2 + x_1) = (2 + x_1 - 1)/(2 + x_1) = 1 - 1/(2 + x_1) = \\ &= 1 - (1/2)(1 + x_1/2)^{-1} = 1 - (1/2)[1 - x_1/2 + (x_1/2)^2 - (x_1/2)^3 + \dots] \\ &= (1/2) + (1/4)x_1 - (1/8)x_1^2 + (1/16)x_1^3 + \dots, \end{aligned} \quad (\text{S.3.22})$$

$$\begin{aligned} z_2(1) &= (2 + x_2)(2 + x_1)^2 \\ &= 8 + 8x_1 + 4x_2 + 2x_1^2 + 4x_1x_2 + x_1^2x_2. \end{aligned} \quad (\text{S.3.23})$$

We will now explore how this same result can be obtained using the replacement rule applied to the operation of numerical integration. As before, we will label individual monomials by an integer r . Recall that Table 2.5 shows our standard modified glex sequencing applied to the case of two variables.

The *Mathematica* code below shows, for our two-variable example differential equation, the application of numerical integration to pyramids. Before describing the code in some detail, we take note of the bottom two lines. When interpreted with the aid of Table 2.4, we see that the penultimate line of (3.24) agrees with (3.22), and the last line of (3.24) nearly agrees with (3.23). The only discrepancy is that for the monomial with label $r = 7$ in the last line of (3.24). In the *Mathematica* output it has the value -1.16563×10^{-7} while, according to (3.23), the true value should be zero. This small discrepancy arises from the truncation error inherent in the RK4 algorithm, and becomes smaller as the step size h is decreased (and ns is correspondingly increased), or if some more accurate integration algorithm is used. We conclude that, with the use of pyramids, it is also possible in the two-variable case to obtain Taylor expansions of the final conditions in terms of the initial conditions. Indeed, what is involved is again the Taylor rule applied, in this instance, to the case of two variables.

```

Clear["Global`*"];
Needs["Combinatorica`"];
m = 2; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] :=
Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
F[Z_, t_] := {-PROD[Z[[1]], Z[[1]]], 2. PROD[Z[[1]], Z[[2]]]};
h = .01;
ns = 100;
t = 0;
zd0 = {1., 2.};
C1 = Table[KroneckerDelta[k, 1], {k, 1, L, 1}];
X[1] = Table[KroneckerDelta[k, 2], {k, 1, L, 1}];
X[2] = Table[KroneckerDelta[k, 3], {k, 1, L, 1}];
Zvar = {zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2]};
RK4;
t
Zvar
1.
{{0.5, 0.25, 0., -0.125, 0., 0., 0.0625, 0., 0., 0.},  

{8., 8., 4., 2., 4., 0., -1.16563 × 10-7, 1., 0., 0.}} (S.3.24)

```

Let us compare the structures of the routines for the single variable case and multi (two) variable case as illustrated in (3.12) and (3.24). The first difference occurs at line 3 where the number of variables m and the maximum degree p are specified. In (3.24) m is set to 2 because we wish to treat the case of two variables, and p is set to 3 simply to limit the lengths of the output arrays. The next difference occurs in line 12 where the right side F of the differential equation is specified. The major feature of the definition of F in (3.24) is that it is specified as two pyramids because the right side of the definition has the structure $\{*, *\}$ where each item $*$ is an instruction for computing a pyramid. In particular, the two pyramids are those for the two components of f as given by (3.17) and use of the replacement rule,

$$-z_1^2 \rightsquigarrow -\text{PROD}[Z[[1]], Z[[1]]], \quad (\text{S.3.25})$$

$$2z_1 z_2 \rightsquigarrow 2. \text{PROD}[Z[[1]], Z[[2]]]. \quad (\text{S.3.26})$$

The next differences occur in lines 16 through 20 of (3.24). In line 16, since specification of the initial conditions now requires two numbers, see (3.20), `zd0` is specified as a two-component array. In lines 17 and 18 of (3.12) the pyramids `C1` and `X` are set up explicitly for the case $p = 5$. By contrast, in lines 17 through 19 of (3.24), the pyramids `C1`, `X[1]`, and `X[2]` are set up for general p with the aid of the `Table` command and the Kronecker delta function. Recall (2.66) and observe from Tables 2.1, 2.4, and 2.5 that, no matter what the values of m and p , the constant monomial has the label $r = 1$ and the monomial x_1 has the label $r = 2$. Moreover, as long as $m \geq 2$ and no matter what the value of p , the x_2 monomial has the label $r = 3$. Finally, compare line 19 in (3.12) with line 20 in (3.24), both of which define the initial `Zvar`. We see that the difference is that in (3.12) `Zvar` is defined as a single pyramid while in (3.24) it is defined as a pair of pyramids of the form $\{*, *\}$. Most remarkably, all other corresponding lines in (3.12) and (3.24) are the same. In particular, the *same* RK4 code, namely that given by (3.6), is used in the scalar case (3.10), the single pyramid case (3.12), and the two-pyramid case (3.24). This multi-use is possible because of the convenient way in which *Mathematica* handles arrays.

We conclude that the pattern for the multivariable case is now clear. Only the following items need to be specified in an m dependent way:

- The value of m .
- The entries in `F` with entries entered as an array $\{*, *, \dots\}$ of m pyramids.
- The design initial condition array `zd0`.
- The pyramids for `C1` and `X[1]` through `X[m]`.
- The entries for the initial `Zvar` specified as an array $\{zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2], \dots, zd0[[m]] C1 + X[m]\}$ of m pyramids.

S.4 Duffing Equation Application

Let us now apply the methods just developed to the case of the Duffing equation with parameter dependence as described by the relations (10.12.133) through (10.12.138). *Mathematica* code for this purpose is shown below. By looking at the final lines that result from executing this code, we see that the final output is an array of the form $\{\{*\}, \{*\}, \{*\}\}$. That is, the final output is an array of three pyramids. This is what we expect, because now we are dealing with three variables. See line 3 of the code, which sets $m = 3$. Also, for convenience of viewing, results are calculated and displayed only through third order as a consequence of setting $p = 3$.

```

Clear["Global`*"];
Needs["Combinatorica`"];
m = 3; p = 3;
GAMMA = Compositions[0, m];
Do[GAMMA = Join[GAMMA, Reverse[Compositions[d, m]]], {d, 1, p, 1}];
L = Length[GAMMA];
JSK[list_, k_] := Position[Apply[And, Thread[#1 <= #2 & [#, k]]] & /@ list, True]//Flatten;
B = Table[JSK[GAMMA, GAMMA[[r]]], {r, 1, L, 1}];
Brev = Reverse/@ B;
PROD[U_, V_] := Table[U[[B[[k]]]].V[[Brev[[k]]]], {k, 1, L, 1}];
POWER[U_, 2] := PROD[U, U];
POWER[U_, 3] := PROD[U, POWER[U, 2]];
C0 = Table[0, {k, 1, L, 1}];
F[Z_, t_] := {Z[[2]],
-2. beta PROD[Z[[3]], Z[[2]]] - PROD[POWER[Z[[3]], 2], Z[[1]]]-
POWER[Z[[1]], 3] - eps Sin[t] POWER[Z[[3]], 3],
C0};
ns = 100;
t = 0;
h = (2Pi)/ns;
beta = .1; eps = 1.5;
zd0 = {.3, .4, .5};
C1 = Table[KroneckerDelta[k, 1], {k, 1, L, 1}];
X[1] = Table[KroneckerDelta[k, 2], {k, 1, L, 1}];
X[2] = Table[KroneckerDelta[k, 3], {k, 1, L, 1}];
X[3] = Table[KroneckerDelta[k, 4], {k, 1, L, 1}];
Zvar = {zd0[[1]] C1 + X[1], zd0[[2]] C1 + X[2], zd0[[3]] C1 + X[3]};
RK4;
t
Zvar

```

$$\begin{aligned}
& 2\pi \\
& \{ \{-0.0493158, 0.973942, -0.110494, 5.51271, 3.54684, 3.46678, \\
& \quad 11.2762, 2.36463, 1.0985, 23.3332, -1.03541, -3.23761, -12.8064, \\
& \quad 4.03421, -23.4342, -17.8967, 1.96148, 5.07403, -36.9009, 25.1379\}, \\
& \{0.439713, 1.05904, 0.427613, 3.3177, 0.0872459, 0.635397, -3.02822, \\
& \quad 1.77416, -4.10115, 3.16981, -2.43002, -5.33643, -7.77038, -6.08476, \\
& \quad -0.541465, -21.1672, -1.4091, -9.54326, 14.6334, -39.2312\}, \\
& \{0.5, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\} \}
\end{aligned} \tag{S.4.1}$$

The first unusual fragments in the code are lines 12 and 13, which define functions that implement the calculation of second and third powers of pyramids. Recall Subsection 2.7. The first new fragment is line 14, which defines the pyramid $C0$ with the aid of the Table command and an implied Do loop. As a result of executing this code, $C0$ is an array of L zeroes. The next three lines, lines 15 through 18, define F , which specifies the right sides of equations (10.12.133) through (10.12.135). See (10.12.136) through (10.12.138). The right side of F is of the form $\{*, *, *\}$, an array of three pyramids. By looking at (10.12.136) and recalling the replacement rule, we see that the first pyramid should be $Z[[2]]$,

$$z_2 \rightsquigarrow Z[[2]]. \tag{S.4.2}$$

The second pyramid on the right side of F is more complicated. It arises by applying the replacement rule to the right side of (10.12.137) to obtain the associated pyramid,

$$\begin{aligned}
& -2\beta z_3 z_2 - z_3^2 z_1 - z_1^3 - \epsilon z_3^3 \sin t \rightsquigarrow \\
& -2. \text{beta PROD}[Z[[3]], Z[[2]]] - \text{PROD}[\text{POWER}[Z[[3]], 2], Z[[1]]] - \\
& \text{POWER}[Z[[1]], 3] - \text{eps Sin}[t] \text{POWER}[Z[[3]], 3].
\end{aligned} \tag{S.4.3}$$

The third pyramid on the right side of F is simplicity itself. From (10.12.138) we see that this pyramid should be the result of applying the replacement rule to the number 0. Hence, this pyramid is $C0$,

$$0 \rightsquigarrow C0 = \{0, 0, \dots, 0\}. \tag{S.4.4}$$

The remaining lines of the code require little comment. Line 20 sets the initial time to 0, and line 21 defines h in such a way that the final value of t will be 2π . Line 22 establishes the parameter values $\beta = .1$ and $\epsilon = 1.5$, which are those for Figure 1.4.9. Line 23 specifies that the design initial condition is

$$z_1(0) = z_1^{d0} = .3, \quad z_2(0) = z_2^{d0} = .4, \quad z_3(0) = z_3^{d0} = .5 = \sigma, \tag{S.4.5}$$

and consequently

$$\omega = 1/\sigma = 2. \tag{S.4.6}$$

See (10.12.104). Also, it follows from (10.12.103) and (10.12.106) that

$$q(0) = \omega Q(0) = \omega z_1(0) = (2)(.3) = .6, \tag{S.4.7}$$

$$q'(0) = \omega^2 \dot{Q}(0) = \omega^2 z_2(0) = (2^2)(.4) = 1.6. \quad (\text{S.4.8})$$

Next, lines 24 through 28 specify that the expansion is to be carried out about the initial conditions (7.124). Finally, line 29 invokes the `RK4` code given by (3.6). That is, as before, *no* modifications are required in the integration code.

A few more comments about the output are appropriate. Line 32 shows that the final time t is indeed 2π , as desired. The remaining output lines display the three pyramids that specify the final value of `Zvar`. From the first entry in each pyramid we see that

$$z_1(2\pi) = -0.0493158, \quad (\text{S.4.9})$$

$$z_2(2\pi) = 0.439713, \quad (\text{S.4.10})$$

$$z_3(2\pi) = .5, \quad (\text{S.4.11})$$

when there are no deviations in the initial conditions. The remaining entries in the pyramids are the coefficients in the Taylor series that describe the changes in the final conditions that occur when changes are made in the initial conditions (including the parameter σ). We are, of course, particularly interested in the first two pyramids. The third pyramid has entries only in the first place and the fourth place, and these entries are the same as those in the third pyramid pyramid for `Zvar` at the start of the integration, namely those in `zd0[3] C1 + X[3]`. The fact that the third pyramid in `Zvar` remains constant is the expected consequence of (10.12.138).

At this point we should also describe how the \mathcal{M}_8 employed in Section 22.12 was actually computed. It could have been computed by setting $p = 8$ in (4.1) and specifying a small step size h and a great number of steps ns to insure good accuracy. Of course, when $p = 8$, the pyramids are large. Therefore, one does not usually print them out, but rather writes them to files or sends them directly to other programs for further use.

However, rather than using `RK4` in (4.1), we replaced it with an adaptive 4-5th order Runge-Kutta-Fehlberg routine that dynamically adjusts the time step h during the course of integration to achieve a specified local accuracy, and we required that the error at each step be no larger than 10^{-12} . (Recall Subsection 2.1.1.) Like the `RK4` routine, the Runge-Kutta-Fehlberg routine, when implemented in *Mathematica*, has the property that it can integrate any number of equations both in scalar variable and pyramid form without any changes in the code.²

S.5 Relation to the Complete Variational Equations

At this point it may not be obvious to the reader that the use of pyramids in integration routines to obtain Taylor expansions is the same as integrating the complete variational equations. We now show that the integration of pyramid equations is equivalent to the forward integration of the complete variational equations. For simplicity, we will examine the single variable case with no parameter dependence. The reader who has mastered this case should be able to generalize the results obtained to the general case.

²A *Mathematica* version of this code is available from Dobrin Kaltchev (kaltchev@triumf.ca) upon request.

In the single variable case with no parameter dependence (1.1) becomes

$$\dot{z} = f(z, t). \quad (\text{S.5.1})$$

Let $z^d(t)$ be some design solution and introduce a deviation variable ζ by writing

$$z = z^d + \zeta. \quad (\text{S.5.2})$$

Then the equation of motion (5.1) takes the form

$$\dot{z}^d + \dot{\zeta} = f(z^d + \zeta, t). \quad (\text{S.5.3})$$

Also, the relations (10.12.14) and (10.12.15) take the form

$$f(z^d + \zeta, t) = f(z^d, t) + g(z^d, t, \zeta) \quad (\text{S.5.4})$$

where g has an expansion of the form

$$g(z^d, t, \zeta) = \sum_{j=1}^{\infty} g^j(t) \zeta^j. \quad (\text{S.5.5})$$

Finally, (10.12.16) and (10.12.17) become

$$\dot{z}^d = f(z^d, t), \quad (\text{S.5.6})$$

$$\dot{\zeta} = g(z^d, t, \zeta) = \sum_{j=1}^{\infty} g^j(t) \zeta^j, \quad (\text{S.5.7})$$

and (10.12.18) becomes

$$\zeta = \sum_{j=1}^{\infty} h^j(t) (\zeta_i)^j. \quad (\text{S.5.8})$$

Insertion of (5.8) into both sides of (5.7) and equating like powers of ζ_i now yields the set of differential equations

$$\dot{h}^{j''}(t) = \sum_{j=1}^{\infty} g^j(t) U_j^{j''}(h^s) \text{ with } j, j'' \geq 1 \quad (\text{S.5.9})$$

where the (universal) functions $U_j^{j''}(h^s)$ are given by the relations

$$\left(\sum_{j'=1}^{\infty} h^{j'}(\zeta_i)^{j'} \right)^j = \sum_{j''=1}^{\infty} U_j^{j''}(h^s)(\zeta_i)^{j''}. \quad (\text{S.5.10})$$

The equations (5.6) and (5.9) are to be integrated from $t = t^{\text{in}} = t^0$ to $t = t^{\text{fin}}$ with the initial conditions

$$z^d(t^0) = z^{d0}, \quad (\text{S.5.11})$$

$$h^1(t^0) = 1, \quad (\text{S.5.12})$$

$$h^{j''}(t^0) = 0 \text{ for } j'' > 1. \quad (\text{S.5.13})$$

Let us now consider the numerical integration of pyramids. Upon some reflection, we see that the numerical integration of pyramids is equivalent to finding the numerical solution to a differential equation with pyramid arguments. For example, in the single-variable case, let $\mathbf{Zvar}(t)$ be the pyramid appearing in the integration process. Then, its integration is equivalent to solving numerically the pyramid differential equation

$$(d/dt)\mathbf{Zvar}(t) = \mathbf{F}(\mathbf{Zvar}, t). \quad (\text{S.5.14})$$

We now work out the consequences of this observation. By the inverse of the replacement rule, we may associate a Taylor series with the pyramid $\mathbf{Zvar}(t)$ by writing

$$\mathbf{Zvar}(t) \rightsquigarrow c_0(t) + \sum_{j \geq 1} c_j(t)x^j. \quad (\text{S.5.15})$$

By (5.15) it is intended that the entries in the pyramid $\mathbf{Zvar}(t)$ be used to construct a corresponding Taylor series with variable x . In view of (3.15), there are the initial conditions

$$c_0(t_0) = z^d(t_0), \quad (\text{S.5.16})$$

$$c_1(t_0) = 1, \quad (\text{S.5.17})$$

$$c_j(t_0) = 0 \text{ for } j > 1. \quad (\text{S.5.18})$$

We next seek the differential equations that determine the time evolution of the $c_j(t)$. Under the inverse replacement rule, there is also the correspondence

$$(d/dt)\mathbf{Zvar}(t) \rightsquigarrow \dot{c}_0(t) + \sum_{j \geq 1} \dot{c}_j(t)x^j. \quad (\text{S.5.19})$$

We have found a representation for the left side of (5.14). We need to do the same for the right side. That is, we need the Taylor series associated with the pyramid $\mathbf{F}(\mathbf{Zvar}, t)$. By the inverse replacement rule, it will be given by the relation

$$\mathbf{F}(\mathbf{Zvar}, t) \rightsquigarrow f\left(\sum_{j \geq 0} c_j(t)x^j, t\right). \quad (\text{S.5.20})$$

Here it is understood that the right side of (5.20) is to be expanded in a Taylor series about $x = 0$. From (5.4), (5.5), and (5.10) we have the relations

$$\begin{aligned} f\left(\sum_{j \geq 0} c_j(t)x^j, t\right) &= f(c_0(t)) + g(c_0(t), t, \sum_{j \geq 1} c_j(t)x^j) \\ &= f(c_0(t)) + \sum_{k \geq 1} g^k(t) \left(\sum_{j \geq 1} c_j(t)x^j\right)^k \\ &= f(c_0(t)) + \sum_{k \geq 1} g^k(t) \sum_{j \geq 1} U_k^j(c_\ell)x^j. \end{aligned} \quad (\text{S.5.21})$$

Therefore, there is the inverse replacement rule

$$F(Zvar, t) \rightsquigarrow f(c_0(t)) + \sum_{k \geq 1} g^k(t) \sum_{j \geq 1} U_k^j(c_\ell) x^j. \quad (\text{S.5.22})$$

Upon comparing like powers of x in (5.19) and (5.22), we see that the pyramid differential equation (5.14) is equivalent to the set of differential equations

$$\dot{c}_0(t) = f(c_0(t)), \quad (\text{S.5.23})$$

$$\dot{c}_j(t) = \sum_{k \geq 1} g^k(t) U_k^j(c_\ell). \quad (\text{S.5.24})$$

Finally, compare the initial conditions (5.11) through (5.13) with the initial conditions (5.16) through (5.18), and compare the differential equations (5.6) and (5.9) with the differential equations (5.23) and (5.24). We conclude that that there must be the relations

$$c_0(t) = z^d(t), \quad (\text{S.5.25})$$

$$c_j(t) = h^j(t) \text{ for } j \geq 1. \quad (\text{S.5.26})$$

We have verified, in the single variable case, that the use of pyramids in integration routines is equivalent to the solution of the complete variational equations using forward integration. As stated earlier, verification of the analogous m -variable result is left to the reader.

We also observe the wonderful convenience that, when pyramid operations are implemented and employed, it is not necessary to explicitly work out the forcing terms $g_a^r(t)$ of Subsection 10.12.1 and the universal functions $U_r^{r''}(h_n^s)$ of Subsection 10.12.3, nor is it necessary to explicitly set up the complete variational equations (10.12.36). All these complications are handled implicitly and automatically by the pyramid routines.

S.6 Acknowledgment

Dobrin Kaltchev made major contributions to the work of this appendix.

Exercises

S.6.1. Verify, in the general m variable case, that the use of pyramids in integration routines is equivalent to the solution of the complete variational equations using forward integration.

Bibliography

- [1] R. Neidinger, “Computing Multivariable Taylor Series to Arbitrary Order”, *Proc. of Intern. Conf. on Applied programming languages*, San Antonio, pp. 134-144 (1995).
- [2] R. Neidinger, “Introduction to Automatic Differentiation and MATLAB Object-Oriented Programming”, *SIAM Review* **52**, 545-563 (2010).
- [3] Wolfram Research, Inc., *Mathematica*, Version 7.0, Champaign, IL (2008).
- [4] Dobrin Kaltchev (*TRIUMF, 4004 Wesbrook Mall, Vancouver, B.C., Canada V6T 2A3*) designed and wrote all the *Mathematica* code for this appendix, and he and Alex Dragt coauthored the text. D. Kaltchev wishes to thank his colleagues from TRIUMF and CERN, especially Richard Abram Baartman, for their interest and support.
- [5] D. Kalman and R. Lindell, “A recursive approach to multivariate automatic differentiation”, *Optimization Methods and Software*, Volume 6, Issue 3, pp. 161-192 (1995).
- [6] M. Berz, “Differential algebraic description of beam dynamics to very high orders”, *Particle Accelerators* 24, p. 109 (1989).
- [7] U. Naumann, *The Art of Differentiating Computer Programs: An Introduction to Algorithmic Differentiation*, SIAM (2012).
- [8] Alex Haro, “Automatic Differentiation Tools in Computational Dynamical Systems”. See the Web site <http://www.maia.ub.es/~alex/ad/adhds.pdf>
- [9] A. Haro, M. Canadell, J-L. Figueras, A. Luque, and J-M. Mondelo, *The Parameterization Method for Invariant Manifolds: From Rigorous Results to Effective Computations*, Applied Mathematical Sciences Volume 195, Springer (2016).

Appendix T

Quadrature and Cubature Formulas

T.1 Quadrature Formulas

T.1.1 Introduction

Suppose we wish to integrate some function $f(x)$ over some (finite) interval. Without loss of generality, by suitable translation and scaling, we may take this interval to be $[0, 1]$. A *quadrature formula* is a set of k *sampling points* x_i in the interval $[0, 1]$ and *weights* w_i such that

$$\int_0^1 dx f(x) \simeq \sum_{i=1}^k w_i f(x_i). \quad (\text{T.1.1})$$

The challenge is to select the sampling points and weights in such a way that the approximation (1.1) is optimal and to define what is meant by *optimal*. From the *Weierstrass* approximation theorem we know that the monomials are dense on any bounded domain. Also, according to Taylor's theorem, monomials are the building blocks for analytic functions. Therefore, for our purposes, we will define optimal to mean that the relation (1.1) is to hold exactly for polynomials in x of as high a degree as possible. That is, for a given set of sampling points, we select the w_i in such a way that

$$\sum_{i=1}^k w_i (x_i)^\ell = \int_0^1 dx x^\ell = 1/(\ell + 1) \quad (\text{T.1.2})$$

for $\ell = 0, 1, 2, \dots$ up to as large an ℓ value (for a given k) as possible.

At this point some discussion is required. Suppose we reason as follows: Let m be an integer with $m \geq 0$. Assume $f(x)$ is a polynomial of maximum degree m . It will then have

$$k = m + 1 \quad (\text{T.1.3})$$

coefficients in its Taylor series representation, and these coefficients can be found by sampling the value of f at k different points x_i on the interval $[0, 1]$. Put another way, suppose we are given k values v_1, v_2, \dots, v_k . Then $f(x)$ is the unique polynomial of degree m whose graph passes through the points $\{x_i, v_i\}$. Now, with the coefficients known, the Taylor series can be integrated to determine the left side of (1.1). Thus, with a knowledge of f at k sampling

sampling points, it is in principle possible to integrate exactly polynomials of degree $\ell \leq m$ with m given by $m = k - 1$.

We next observe that once the k sampling points have been determined, the k weights w_i are uniquely determined. Let $L_i(x)$, called the *Lagrange* polynomial, be the degree m polynomial that takes on the value 1 at the sampling point x_i and has the value 0 at all the other sampling points,

$$L_i(x_j) = \delta_{ij}. \quad (\text{T.1.4})$$

It is given by the construction

$$L_i(x) = \left[\prod_{j \neq i} (x - x_j) \right] / \left[\prod_{j \neq i} (x_i - x_j) \right]. \quad (\text{T.1.5})$$

Evidently there are k such polynomials. Also, as a result of (1.4), it follows that $f(x)$ given by

$$f(x) = \sum_{i=1}^k v_i L_i(x) \quad (\text{T.1.6})$$

has the property

$$f(x_j) = v_j. \quad (\text{T.1.7})$$

Moreover, we see that

$$\int_0^1 dx f(x) = \sum_i v_i \int_0^1 dx L_i(x). \quad (\text{T.1.8})$$

Therefore, in view of (1.7) and the desire (1.1), we make the definition

$$w_i = \int_0^1 dx L_i(x) \quad (\text{T.1.9})$$

to achieve the result

$$\int_0^1 dx f(x) = \sum_i w_i f(x_i). \quad (\text{T.1.10})$$

We conclude that given k sampling points, there are k weights w_i , uniquely determined by (1.9), such that (1.2) holds for $\ell \leq m$ with $m = k - 1$.

Given the k sampling points x_i , and the associated weights w_i , can it happen that (1.2) also holds for some ℓ values with $\ell > m$, i.e. $\ell > k - 1$? Let ℓ_{\max} be the largest integer for which (1.2) holds. More precisely, we require that (1.2) holds for $\ell \leq \ell_{\max}$, but not for $\ell = \ell_{\max} + 1$. We will see that exactly how large ℓ_{\max} can be depends on how the sampling points are chosen. In particular, we will learn that there is a unique optimum sampling procedure (called Legendre Gauss) for which ℓ_{\max} has the optimum value

$$\ell_{\max} = 2k - 1. \quad (\text{T.1.11})$$

Upon combining (1.3) and (1.11) we see that for any sampling procedure there is the range relation

$$k - 1 \leq \ell_{\max} \leq 2k - 1. \quad (\text{T.1.12})$$

Put another way, if ℓ_{\max} is specified, we will see that there can be sampling procedures such that (1.2) holds for all $\ell \leq \ell_{\max}$ with $k < \ell_{\max} + 1$. Indeed, if ℓ_{\max} is odd, k can be as small as k_{\min} with

$$k_{\min} = (\ell_{\max} + 1)/2. \quad (\text{T.1.13})$$

Since we know that $\ell_{\max} + 1$ sampling points are necessary to determine the Taylor coefficients of a polynomial of degree ℓ_{\max} , how can it be that we can exactly integrate over the interval $[0, 1]$ such polynomials using just k sampling points with $k < \ell_{\max} + 1$? The answer is that the *only* thing we want to know is the value of the integral over the *fixed* interval $[0, 1]$, and *not* the values of the individual Taylor coefficients. What happens is that the value of the integral depends only on the value of a certain combination of the Taylor coefficients, and the value of this combination can be found by sampling the function at fewer than $\ell_{\max} + 1$ points providing these points are judiciously chosen. See Exercise 1.2.

T.1.2 Newton Cotes

One sampling option is to space the x_i evenly with $x_1 = 0$ and $x_k = 1$,

$$x_i = (i - 1)/(k - 1). \quad (\text{T.1.14})$$

Doing so gives the family of *Newton-Cotes* quadrature formulas.¹ For example, for the case $k = 3$, the sampling points are

$$(x_1, x_2, x_3) = (0, 1/2, 1), \quad (\text{T.1.15})$$

the associated weights are found, using (1.9), to be

$$(w_1, w_2, w_3) = (1/6, 4/6, 1/6), \quad (\text{T.1.16})$$

thereby yielding the celebrated *Simpson's rule* 1-4-1 formula

$$\int_0^1 dx f(x) \simeq (1/6)f(0) + (4/6)f(1/2) + (1/6)f(1). \quad (\text{T.1.17})$$

In this case (1.2) holds for $\ell = 0, 1, 2$, and 3; and errors first begin to appear for $\ell \geq 4$. For $\ell = 4$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^3 w_i(x_i)^4 = 1/(4+1) + 4!/[90(2^5)] = 1/5 + 4!/2880. \quad (\text{T.1.18})$$

Correspondingly, assuming that f is sufficiently differentiable, use of Newton Cotes gives (for the case $k = 3$) the approximation

$$\int_0^1 dx f(x) = (1/6)f(0) + (4/6)f(1/2) + (1/6)f(1) - (1/2880)f^{(4)}(\xi) \quad (\text{T.1.19})$$

¹Newton's student Roger Cotes urged and inspired Newton to write a second and enlarged edition of his *Principia*, and wrote the preface to this edition. He died of a violent fever at the early age of 33. At Cotes' death Newton remarked, "If he had lived we would have known something." It is interesting to note that several years before Euler wrote his famous formula $\exp(i\theta) = \cos \theta + i \sin \theta$, Cotes wrote the inverse relation $\log(\cos \theta + i \sin \theta) = i\theta$.

where $\xi \in [0, 1]$.

As a second example, for the case $k = 4$ the sampling points are

$$(x_1, x_2, x_3, x_4) = (0, 1/3, 2/3, 1), \quad (\text{T.1.20})$$

the associated weights are found to be

$$(w_1, w_2, w_3, w_4) = (1/8, 3/8, 3/8, 1/8), \quad (\text{T.1.21})$$

and we have the equally celebrated Simpson's 3/8 rule

$$\begin{aligned} \int_0^1 dx f(x) &= (1/8)f(0) + (3/8)f(1/3) + (3/8)f(2/3) + (1/8)f(1) \\ &\quad - (1/6480)f^{(4)}(\xi) \end{aligned} \quad (\text{T.1.22})$$

where again $\xi \in [0, 1]$.

Table 1.1 lists ℓ_{\max} as a function of k for Newton Cotes.² Evidently, when k is odd, there is no increase in order when using instead the next even value of k . Doing so does result in a decrease in the coefficient in the error term, but this decrease is fairly modest. For this reason, odd values of k are frequently preferred. Note that for even k the entries in the Table take the floor value $\ell_{\max} = k - 1$ guaranteed by (1.12), and beat this value for odd k .

Table T.1.1: Maximum Order ℓ_{\max} for k Newton-Cotes Sampling Points.

k	1	2	3	4	5	6	7	8	9	10
ℓ_{\max}	1	1	3	3	5	5	7	7	9	9

T.1.3 Legendre Gauss

Another appealing sampling option is not to space the x_i evenly, but rather to select them in such a way that (for a fixed k) the number of successive ℓ values for which (1.2) holds is maximized. This choice produces the family of *Legendre-Gauss* quadrature formulas. Legendre-Gauss quadrature is most naturally described in terms of the interval $[-1, 1]$. It can be shown that the *Legendre* polynomial $P_k(x)$ has k distinct zeroes on the interval $[-1, 1]$, and it is these zeroes that are used in k -sampling point Legendre-Gauss quadrature.

Three examples of Legendre Gauss are given below. For ease of comparison with Newton-Cotes quadrature, we have transformed Legendre-Gauss results to the interval $[0, 1]$.

For $k = 3$, the (transformed) Legendre-Gauss sampling points are given by

$$(x_1, x_2, x_3) = (1/2 - \sqrt{15}/10, 1/2, 1/2 + \sqrt{15}/10). \quad (\text{T.1.23})$$

²Strictly speaking, for the Newton-Cotes we have been describing, the $k = 1$ table entry is meaningless. However, it does apply in the case of *open* Newton Cotes. See Exercise 1.1.

The corresponding weights, calculated using (1.9), are given by

$$(w_1, w_2, w_3) = (5/18, 8/18, 5/18). \quad (\text{T.1.24})$$

Correspondingly, there is the formula

$$\int_0^1 dx f(x) \simeq (5/18)f(1/2 - \sqrt{15}/10) + (8/18)f(1/2) + (5/18)f(1/2 + \sqrt{15}/10). \quad (\text{T.1.25})$$

In this case (1.2) holds for $\ell = 0$ through 5, but not $\ell = 6$. Therefore $\ell_{\max} = 5$. And for $\ell = 6$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^3 w_i(x_i)^6 = 1/7 - 6!/2016000. \quad (\text{T.1.26})$$

Correspondingly, again assuming that f is sufficiently differentiable, use of Legendre Gauss gives (for the case $k = 3$) the approximation

$$\begin{aligned} \int_0^1 dx f(x) &= (5/18)f(1/2 - \sqrt{15}/10) + (8/18)f(1/2) + (5/18)f(1/2 + \sqrt{15}/10) \\ &\quad + (1/2016000)f^{(6)}(\xi). \end{aligned} \quad (\text{T.1.27})$$

As another example, consider the case $k = 2$. Then there is the formula

$$\int_0^1 dx f(x) \simeq (1/2)f(1/2 - \sqrt{3}/6) + (1/2)f(1/2 + \sqrt{3}/6). \quad (\text{T.1.28})$$

It corresponds to sampling points and weights given by the rules

$$(x_1, x_2) = (1/2 - \sqrt{3}/6, 1/2 + \sqrt{3}/6), \quad (\text{T.1.29})$$

$$(w_1, w_2) = (1/2, 1/2). \quad (\text{T.1.30})$$

In this case (1.2) holds for $\ell = 0$ through 3, but not for $\ell = 4$. That is, $\ell_{\max} = 3$. For $\ell = 4$ the sum on the left side of (1.2) has the value

$$\sum_{i=1}^2 w_i(x_i)^4 = 1/5 - 4!/4320. \quad (\text{T.1.31})$$

Correspondingly, use of Legendre Gauss gives (for the case $k = 2$) the approximation

$$\int_0^1 dx f(x) = (1/2)f(1/2 - \sqrt{3}/6) + (1/2)f(1/2 + \sqrt{3}/6) + (1/4320)f^{(4)}(\xi). \quad (\text{T.1.32})$$

Compare (1.32), Legendre Gauss for $k = 2$, with (1.19), Newton Cotes for $k = 3$. They are of the same order, but Legendre Gauss has a somewhat smaller error coefficient. Thus, Legendre Gauss for $k = 2$ not only requires one less evaluation of f than Newton Cotes for $k = 3$, it is also slightly more accurate.

Finally, consider the case $k = 1$. It has sampling point and weight given by

$$x_1 = 1/2, \quad (\text{T.1.33})$$

$$w_1 = 1, \quad (\text{T.1.34})$$

and yields the midpoint rule

$$\int_0^1 dx f(x) \simeq f(1/2). \quad (\text{T.1.35})$$

In this case (1.2) holds for $\ell = 0$ and 1 , but not for $\ell = 2$. Therefore $\ell_{\max} = 1$. And for $\ell = 2$ the sum on the left side of (1.2) has the value

$$w_1(x_1)^2 = 1/3 - 2!/24. \quad (\text{T.1.36})$$

Correspondingly, use of Legendre Gauss gives (for the case $k = 1$) the approximation

$$\int_0^1 dx f(x) = f(1/2) + (1/24)f^{(2)}(\xi). \quad (\text{T.1.37})$$

It can be shown that for Legendre Gauss the relation (1.11) holds. Moreover, for a given k value, there is no sampling procedure with larger ℓ_{\max} than that of Legendre Gauss. All other sampling-point choices yield a smaller value of ℓ_{\max} . Comparison of Table 1.1 and (1.11) shows that, with increasing k , Legendre Gauss rapidly becomes far more efficient than Newton Cotes.

T.1.4 Clenshaw Curtis

Like Legendre Gauss, *Clenshaw-Curtis* quadrature is most naturally described on the interval $[-1, 1]$. It uses the *Chebyshev* points as sampling points. For $k = 1$ the Chebyshev point is 0 .³ For $k = 2$ the Chebyshev points are ∓ 1 , and for $k = 3$ the Chebyshev points are $\{-1, 0, 1\}$. For $k \geq 2$ select k equally spaced angles θ_i with $\theta_1 = -\pi$ and $\theta_k = \pi$,

$$\theta_i = -\pi + 2\pi(i - 1)/(k - 1). \quad (\text{T.1.38})$$

For $k \geq 2$ the Chebyshev points are given by the rule

$$x_i = \cos(\theta_i). \quad (\text{T.1.39})$$

Let $T_{k-1}(x)$ be the Chebyshev polynomial of degree $k - 1$. It can be shown, for $k \geq 2$, that T_{k-1} has k extrema on the interval $[-1, 1]$ (all of which are ∓ 1). Moreover, for $k \geq 2$, these extrema are the k Chebyshev sampling points defined by (1.38) and (1.39).

Evidently for $k \leq 3$ the Chebyshev points, when transformed to the interval $[0, 1]$, are the same as the sampling points for Newton Cotes, and therefore Table 1 provides the value of ℓ_{\max} in these cases. It can be shown that in fact Table 1 provides the correct value of

³Some authors omit this point since it is really a Chebyshev point of the second kind.

ℓ_{\max} for Clenshaw Curtis in all cases. That is, Table 1 holds for both Newton Cotes and Clenshaw Curtis.⁴

Although the order of Clenshaw-Curtis quadrature is the same as that for Newton Cotes, and therefore much less than that of Legendre Gauss, its use has two advantages. First, if the value of $k - 1$ is doubled, thereby doubling the order, essentially half of the sampling points are the same as before. Therefore, in the same spirit as embedded Runge Kutta, it is relatively easy to devise adaptive integration schemes. A second advantage, as described in the next subsection, has to do with convergence.

T.1.5 Convergence

What happens in the limit $k \rightarrow \infty$? For the interval $[-1, 1]$ it can be shown that Newton-Cotes quadrature converges to the correct result providing $f(x)$ is *analytic* in a disk centered about $x = 0$ and having radius slightly larger than 1; and is divergent if f fails to be analytic in the open unit disk centered about $x = 0$. (The case where f has singularities on the boundary of this open unit disk requires a more refined analysis.)⁵ When the interval is $[0, 1]$, this result for convergence translates to the requirement that f be analytic in a disk centered about $x = 1/2$ and having radius slightly larger than $1/2$.

By contrast, and working in the interval $[-1, 1]$, Legendre-Gauss quadrature is guaranteed to converge under the much less restrictive condition that $f(x)$ simply be *sufficiently smooth* for $x \in [-1, 1]$. An adequate condition for sufficiently smooth, which is generally realized in practice, is that $f(x)$ be *Lipschitz* continuous for $x \in [-1, 1]$. Correspondingly, when transformed to the interval $[0, 1]$, Legendre-Gauss quadrature is guaranteed to converge if $f(x)$ is sufficiently smooth for $x \in [0, 1]$.

The reason for this difference is that Taylor series on the interval $[-1, 1]$ converge in disks about $x = 0$ whereas Legendre polynomial expansions in the interval $[-1, 1]$ converge in ellipses (sometimes called *Bernstein* ellipses) with foci at $x = \mp 1$.

Although for a given k Clenshaw-Curtis quadrature has relatively low order (the same as Newton Cotes) compared to Legendre Gauss, its convergence properties are similar to those for Legendre Gauss. It too, when transformed to the interval $[0, 1]$, is guaranteed to converge under the much less restrictive condition that $f(x)$ simply be sufficiently smooth for $x \in [0, 1]$.

What can be said about Legendre-Gauss and Clenshaw-Curtis convergence in the case that f is analytic? Again it is most convenient to employ the interval $[-1, 1]$ with the understanding that the results obtained for this interval can be easily be transformed to the interval $[0, 1]$. Let ρ be a real number with $\rho > 1$. Consider the points z in the complex plane given by the rule

$$z = \rho \exp(i\phi) \tag{T.1.40}$$

⁴From (1.38) and (1.39) it follows that the Chebyshev points are symmetrically distributed about $x = 0$. Correspondingly the weights for the points $\mp x_i$ are the same. Therefore, by symmetry, Clenshaw Curtis is exact (yields the value 0) for all odd-degree monomials.

⁵Runge considered the function $f(x) = 1/(1 + 25x^2)$, now called the Runge function, on the interval $x \in [-1, 1]$ and showed that Newton-Cotes quadrature diverges for this example. Note that Runge's f is analytic on this interval, but has poles at the points $x = \pm i/5$, and these poles lie inside the unit disk centered about the origin.

with $\phi \in [0, 2\pi)$. Evidently they lie on a circle about the origin with radius ρ . Next consider points w related to points z by the rule

$$w = (1/2)(z + z^{-1}). \quad (\text{T.1.41})$$

It can be verified that the image of the circle (1.40) under the transformation (1.41) is an ellipse in the complex plane (a Bernstein ellipse E_ρ) with foci ∓ 1 ,

$$\text{semi-major axis} = (1/2)(\rho + 1/\rho), \quad (\text{T.1.42})$$

and

$$\text{semi-minor axis} = (1/2)(\rho - 1/\rho). \quad (\text{T.1.43})$$

Suppose f is analytic on $[-1, 1]$ and that it can be analytically continued (see Figure 32.4.9) into the interior of some E_ρ without encountering a singularity, and also suppose that there are no singularities on the boundary (E_ρ itself). Then it can be shown that for large k the error in Legendre-Gauss quadrature must go to zero at least as fast as

$$\text{Legendre-Gauss quadrature error} \sim \exp[-2k \log(\rho)].$$

And for Clenshaw-Curtis quadrature the error must go to zero at least as fast as

$$\text{Clenshaw-Curtis quadrature error} \sim \exp[-k \log(\rho)].$$

Thus, in both cases, the error goes to zero *exponentially* with increasing k . And the larger the value of ρ can be without there being singularities inside or on E_ρ , the more rapid the exponential decrease.

T.1.6 Quadrature on a Circle/One-Sphere

Suppose we wish to integrate some function $f(\theta)$ over the interval $[0, 2\pi]$. If we view θ as being an angular coordinate, then we may view the integral in question as being an integral over a circle. And if we view a circle as being a one-dimensional *sphere* S^1 , then we may say that the integral in question is an integral over S^1 . With this background in mind, a quadrature formula for S^1 is a set of k sampling points θ_j in the interval $[0, 2\pi]$ and weights w_j such that

$$\int_0^{2\pi} d\theta f(\theta) \simeq \sum_{j=1}^k w_j f(\theta_j). \quad (\text{T.1.44})$$

As usual, the challenge is to select the sampling points and weights in such a way that the approximation (1.44) is optimal and to define what is meant by *optimal*.

It seems reasonable to treat all points in S^1 in the same way, and therefore we take the weights to be equal and the angles θ_j to be equally spaced,

$$\theta_j = (j - 1)(2\pi)/k. \quad (\text{T.1.45})$$

By *optimal* we will require that (1.44) be exact for all $f(\theta) = \exp(in\theta)$ with $|n| = 0 \cdots n_{\max}$ and n_{\max} as large as possible. For these $f(\theta)$ the left side of (1.44) becomes

$$\int_0^{2\pi} d\theta f(\theta) = \int_0^{2\pi} d\theta \exp(in\theta) = 2\pi\delta_{n0}. \quad (\text{T.1.46})$$

The right side of (1.44) becomes

$$\sum_{j=1}^k w_j f(\theta_j) = w \sum_{j=1}^k \exp(in\theta_j) = wk \text{ when } n = 0, \quad (\text{T.1.47})$$

and

$$\begin{aligned} \sum_{j=1}^k w_j f(\theta_j) &= w \sum_{j=1}^k \exp(in\theta_j) = \\ &w(1 + \exp[in(1/k)(2\pi)] + \exp[in(2/k)(2\pi)] + \cdots + \exp\{in[(k-1)/k](2\pi)\}) = \\ &w(1 - \exp[in(2\pi)])/(1 - \exp[in(1/k)(2\pi)]) = 0 \text{ for } 0 < |n| < k. \end{aligned} \quad (\text{T.1.48})$$

Upon viewing (1.46) through (1.48) we see that the proposed quadrature formula is optimal provided

$$w = 2\pi/k \text{ and } |n| < k \text{ so that } n_{\max} = k - 1. \quad (\text{T.1.49})$$

So the quadrature formula (1.44) becomes

$$\int_0^{2\pi} d\theta f(\theta) \simeq (2\pi/k) \sum_{j=1}^k f(\theta_j) \quad (\text{T.1.50})$$

with the θ_j given (1.45) and exactness for the functions $\exp(in\theta)$ with $|n| \leq n_{\max} = k - 1$.

Consider the functions $g_{\ell m}(\theta)$ defined by

$$g_{\ell m}(\theta) = \cos^\ell \theta \sin^m \theta. \quad (\text{T.1.51})$$

Recall that

$$\cos \theta = (1/2)[\exp(i\theta) + \exp(-i\theta)] \text{ and } \sin \theta = [1/(2i)][\exp(i\theta) - \exp(-i\theta)]. \quad (\text{T.1.52})$$

It follows from (1.52) that $g_{\ell m}(\theta)$ can contain factors of $\exp(in\theta)$ with $|n|$ no greater than $(\ell + m)$. Consequently, there are the *exact* quadrature results

$$\int_0^{2\pi} d\theta g_{\ell m}(\theta) = (2\pi/k) \sum_{j=1}^k g_{\ell m}(\theta_j) \text{ provided } k > \ell + m. \quad (\text{T.1.53})$$

How is this discussion related to the subject of discrete Fourier transforms? Error analysis and analyticity.

Exercises

T.1.1. The sampling option (1.14) involves use of the endpoints 0 and 1, and for this reason is more precisely called *closed* Newton Cotes. It is also possible to employ a sampling procedure, called *open* Newton Cotes, for which the x_i are still equally spaced but the

endpoints are not used. Consider, for example, the case $k = 3$ and the two equally spaced sampling procedures

$$(x_1, x_2, x_3) = (0, 1/2, 1) \text{ closed Newton Cotes}, \quad (\text{T.1.54})$$

and

$$(x_1, x_2, x_3) = (1/4, 1/2, 3/4) \text{ open Newton Cotes}. \quad (\text{T.1.55})$$

For $k = 3$ open Newton Cotes the weights are

$$(w_1, w_2, w_3) = (2/3, -1/3, 2/3). \quad (\text{T.1.56})$$

Show that, just as for the case of closed Newton Cotes, (1.2) holds for $k = 3$ open Newton Cotes when $\ell = 0, 1, 2, 3$. Show that, for $k = 3$ open Newton Cotes,

$$\sum_{i=1}^3 w_i(x_i)^4 = 1/(4+1) + (14)(4!)/[45(4^5)] = 1/5 + (4!)(7/23040). \quad (\text{T.1.57})$$

Correspondingly, assuming that f is sufficiently differentiable, use of $k = 3$ open Newton Cotes gives the approximation

$$\int_0^1 dx f(x) = (2/3)f(1/4) - (1/3)f(1/2) + (2/3)f(3/4) + (7/23040)f^{(4)}(\xi) \quad (\text{T.1.58})$$

where $\xi \in [0, 1]$.

For $k = 4$ open Newton Cotes the sampling points are

$$(x_1, x_2, x_3, x_4) = (1/5, 2/5, 3/5, 4/5), \quad (\text{T.1.59})$$

and the weights are

$$(w_1, w_2, w_3, w_4) = (11/24, 1/24, 1/24, 11/24). \quad (\text{T.1.60})$$

Correspondingly, assuming that f is sufficiently differentiable, use of $k = 4$ open Newton Cotes gives the approximation

$$\begin{aligned} \int_0^1 dx f(x) &= (11/24)f(1/5) + (1/24)f(2/5) + (1/24)f(3/5) + (11/24)f(4/5) \\ &\quad + (19/90000)f^{(4)}(\xi) \end{aligned} \quad (\text{T.1.61})$$

where again $\xi \in [0, 1]$.

It can be shown that, for a given value of k , both closed and open Newton Cotes have the same ℓ_{\max} . Thus Table 1.1 holds for both open and closed Newton Cotes. Comparison of (1.19) and (1.48) shows that for $k = 3$ the error coefficient for open Newton Cotes is slightly smaller than that for closed Newton Cotes. This case is an anomaly. For $k \geq 4$ closed Newton Cotes has a smaller error coefficient than open Newton Cotes. Finally, we remark that $k = 1$ Legendre Gauss may also be viewed as being $k = 1$ open Newton Cotes.

T.1.2. Consider quadrature on the interval $[-1, 1]$. Let $f(x)$ be the third-order polynomial

$$f(x) = a + bx + cx^2 + dx^3. \quad (\text{T.1.62})$$

Verify that

$$\int_{-1}^1 dx f(x) = 2a + (2/3)c. \quad (\text{T.1.63})$$

Observe that the right side of (1.53) is a particular linear combination of the Taylor coefficients for f , and the value of the integral on the left side of (1.53) depends only on the value of this particular combination.

The $k = 2$ Legendre-Gauss sampling points and weights on the interval $[-1, 1]$ are

$$(x_1, x_2) = (-1/\sqrt{3}, 1/\sqrt{3}), \quad (\text{T.1.64})$$

$$(w_1, w_2) = (1, 1). \quad (\text{T.1.65})$$

Verify that

$$\sum_{i=1}^2 w_i f(x_i) = 2a + (2/3)c, \quad (\text{T.1.66})$$

and therefore $k = 2$ Legendre-Gauss quadrature is exact for all polynomials of degree 3 or less. Verify that Legendre Gauss is the unique $k = 2$ quadrature rule with this property. Also verify that $k = 2$ Legendre-Gauss quadrature fails for x^4 .

T.1.3. There is another way of viewing quadrature formulas that is more akin to the numerical integration of differential equations. Consider the single-variable differential equation

$$dy/dt = g(t) \quad (\text{T.1.67})$$

with the initial condition

$$y(0) = 0. \quad (\text{T.1.68})$$

Note that g does not depend on y , and therefore (1.57) has the immediate solution

$$y(t) = \int_0^t dt' g(t'). \quad (\text{T.1.69})$$

Suppose, in the spirit of a local stepping formula, we wish to compute $y(h)$ through some order in h . From (1.59) we find

$$y(h) = \int_0^h dt g(t). \quad (\text{T.1.70})$$

Bring the right side of (1.60) to the \int_0^1 standard form by making the change of variable and definition

$$t = xh, \quad (\text{T.1.71})$$

$$f(x) = g(xh). \quad (\text{T.1.72})$$

Show that (1.60) then becomes

$$y(h) = h \int_0^1 dx f(x). \quad (\text{T.1.73})$$

Suppose the right side of (1.63) is evaluated using Newton Cotes with k odd. Show that there is the result

$$y(h) = h \sum_{i=1}^k w_i f(x_i) + c_{k+1} h f^{k+1}(\xi) \quad (\text{T.1.74})$$

where c_{k+1} is a coefficient that can be read off from formulas such as (1.19) and (1.48). Finally, show that

$$f^{k+1}(\xi) = h^{k+1} g^{k+1}(\tau), \quad (\text{T.1.75})$$

where $\tau \in [0, h]$, so that (1.64) becomes

$$y(h) = h \sum_{i=1}^k w_i g(x_i h) + c_{k+1} h^{k+2} g^{k+1}(\tau), \quad k \text{ odd.} \quad (\text{T.1.76})$$

We see that, for k function evaluations with k odd, Newton Cotes provides a stepping formula with local error of order h^{k+2} , and therefore local accuracy through terms of order h^{k+1} .

Suppose the right side of (1.63) is evaluated using Newton Cotes with k even. Show that then there is the result

$$y(h) = h \sum_{i=1}^k w_i f(x_i) + c_k h f^k(\xi) \quad (\text{T.1.77})$$

where c_k is a coefficient that can be read off from formulas such as (1.22) and (1.51). Finally, show that

$$f^k(\xi) = h^k g^k(\tau), \quad (\text{T.1.78})$$

where $\tau \in [0, h]$, so that (1.67) becomes

$$y(h) = h \sum_{i=1}^k w_i g(x_i h) + c_k h^{k+1} g^k(\tau), \quad k \text{ even.} \quad (\text{T.1.79})$$

We see that, for k function evaluations with k even, Newton Cotes provides a stepping formula with local error of order h^{k+1} , and therefore local accuracy through terms of order h^k .

Suppose, instead, that the right side of (1.63) is evaluated using Legendre Gauss. Then there is the result

$$y(h) = h \sum_{i=1}^k w_i f(x_i) + c_{2k} h f^{2k}(\xi) \quad (\text{T.1.80})$$

where c_{2k} is a coefficient that can be read off from formulas such as (1.27), (1.32), and (1.37). Now show that

$$f^{2k}(\xi) = h^{2k} g^{2k}(\tau), \quad (\text{T.1.81})$$

where $\tau \in [0, h]$, so that (1.70) becomes

$$y(h) = h \sum_{i=1}^k w_i g(x_i h) + c_{2k} h^{2k+1} g^{2k}(\tau). \quad (\text{T.1.82})$$

We see that, for k function evaluations, Legendre Gauss provides a stepping formula with local error of order h^{2k+1} , and therefore local accuracy through terms of order h^{2k} .

T.1.4. Verify that E_ρ given by (1.40) and (1.41) is indeed an ellipse with the advertised properties.

T.2 Cubature Formulas

T.2.1 Introduction

Cubature formulas extend the concept of quadrature formulas to the case of domains D having dimension greater than one. All such formulas are called cubature formulas, no matter what the dimension of D , as long as this dimension is greater than one.

Let D be some domain having dimension m . Label points within D by m -dimensional vectors

$$x = (x_1, x_2, \dots, x_m), \quad (\text{T.2.1})$$

and let $f(x)$ be a function defined on D . Then a cubature formula is a set of k sampling points in D , now call them x^i , and weights w^i such that

$$\int_D d^m x f(x) \simeq \sum_{i=1}^k w^i f(x^i). \quad (\text{T.2.2})$$

Again the challenge is to select the sampling points and weights in such a way that the approximation (2.2) is optimal and to define what is meant by optimal. Results are known for some standard domains. For our purposes, we are interested in the cases where D is either a square, rectangle, or S^2 (the surface of a 2-sphere).

T.2.2 Cubature on a Square

The unit m -cube, denoted by C^m , is defined to be the domain

$$-1 \leq x_c \leq 1 \quad c = 1, \dots, m. \quad (\text{T.2.3})$$

A variety of cubature formulas are known for the C^m for all m . We will be particularly interested in the case C^2 , which we call the unit square.

Let P_{j_1, j_2} denote the monomial

$$P_{j_1, j_2}(x) = x_1^{j_1} x_2^{j_2}. \quad (\text{T.2.4})$$

It has degree

$$\hat{d} = j_1 + j_2. \quad (\text{T.2.5})$$

From the work of Exercise 7.10.2 we know that the number of such monomials of degree 0 through d is given by

$$S_0(2, d) = (2 + d)!/(2!d!). \quad (\text{T.2.6})$$

See (7.10.17). For convenience, the values of $S_0(2, d)$ are tabulated below for the first few values of d .

Table T.2.1: $S_0(2, d)$ as a Function of d .

d	0	1	2	3	4	5	6	7
$S_0(2, d)$	1	3	6	10	15	21	28	36

Let D be the domain C^2 . Again, because of the results of Weierstrass and Taylor, as our definition of optimal we will seek k sampling points x^i and weights w^i such that

$$\int_D d^2x P_{j_1, j_2}(x) = \sum_{i=1}^k w^i P_{j_1, j_2}(x^i) \quad (\text{T.2.7})$$

for all monomials of degree less than or equal to d . We note that the left side of (2.7) is easily evaluated to give the result

$$\int_D d^2x P_{j_1, j_2}(x) = 4/[(j_1 + 1)(j_2 + 1)] \quad j_1, j_2 \text{ both even}, \quad (\text{T.2.8})$$

$$= 0 \quad \text{otherwise.} \quad (\text{T.2.9})$$

Observe that, for the 2-dimensional case, the specification of each sampling point x^i requires 2 values. Consequently, the specification of k sampling points and k weights involves the specification of $2k + k = 3k$ values. On the other hand, there are $S_0(2, d)$ conditions of the form (2.7) that need to be met. Therefore naive counting suggests that (2.7) can be satisfied providing

$$3k \geq S_0(2, d). \quad (\text{T.2.10})$$

For example, for the case $d = 5$, we see from Table 2.1 that (2.10) yields the requirement

$$3k \geq 21, \text{ or } k \geq 7. \quad (\text{T.2.11})$$

We note that a formula that works for $d = 5$ can be constructed by converting the integral over the unit square into two iterated integrals, and these integrals (since they must be exact for single-variable monomials of degrees 0 through 5) can each be approximated using $[-1, 1]$ variants of the 3-point Legendre-Gauss formula (1.25). Doing so produces what is called a *product* cubature formula, and evidently has $k = 3 \times 3 = 9$.

Remarkably, for the unit square and $d = 5$, there are two cubature formulas for which $k = 7$. Stroud, in his book on the approximate calculation of multiple integrals, refers to them as $*C^2:5\text{-}1$ and $*C^2:5\text{-}2$. In these formulas several sampling points have the same weight. The formulas are described below. Each description consists of weights and the sampling points having these weights. Also shown, in Figures 2.1 and 2.2, are the locations of the sampling points in the x_1, x_2 plane. Note that all sampling points lie *within* the unit square, a feature that is essential for our purposes; and all have positive weights, a feature that is numerically desirable.

The Unit Square Cubature Formula $*C^2:5-1$

weights	points
8/7	(0, 0)
20/36	($\pm r, \pm s$)
20/63	(0, $\pm t$)

$r =$	$\sqrt{3/5} \simeq .775$	(T.2.12)
$s =$	$\sqrt{1/3} \simeq .577$	
$t =$	$\sqrt{14/15} \simeq .966$	

The Unit Square Cubature Formula $*C^2:5-2$

weights	points
8/7	(0, 0)
100/168	$\pm (r, r)$
20/48	$\pm (s, -t)$
20/48	$\pm (t, -s)$

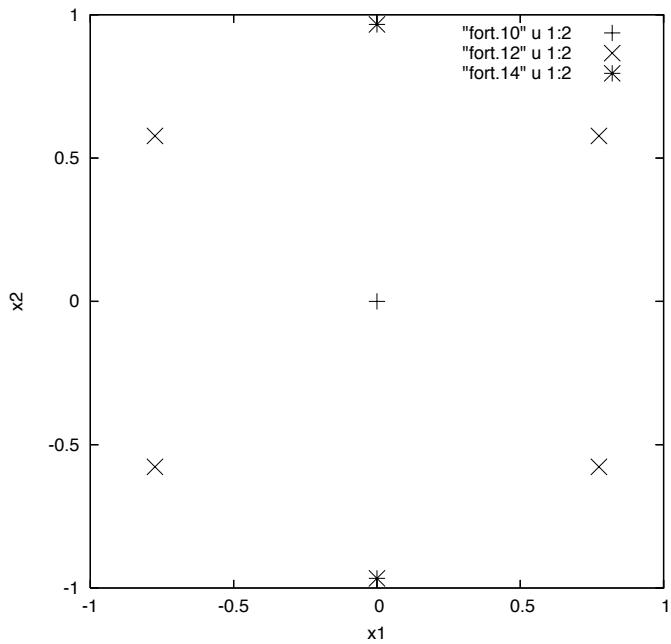
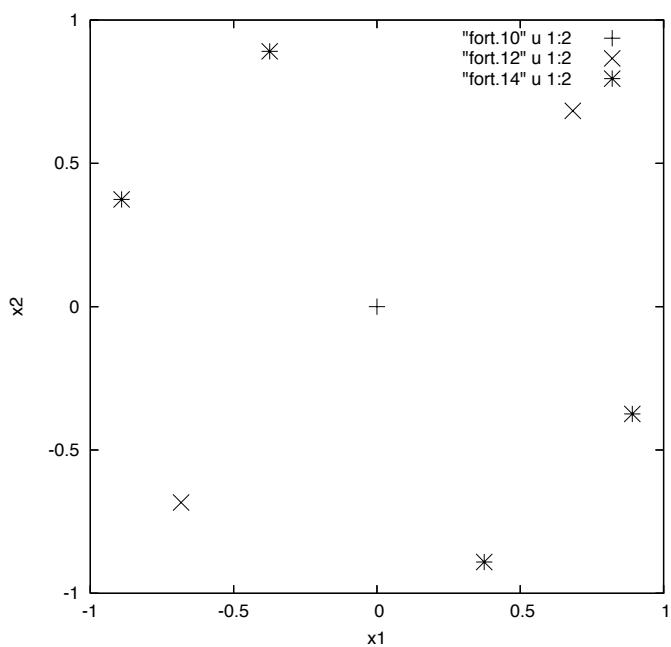
$r =$	$\sqrt{7/15} \simeq .683$	(T.2.13)
$s =$	$[(7 + \sqrt{24})/15]^{1/2} \simeq .891$	
$t =$	$[(7 - \sqrt{24})/15]^{1/2} \simeq .374$	

Rough Error Analysis

Analysis of the errors involved with the use of cubature formulas is a complicated subject. However, one simple thing that can be done is to see how well (2.7) works for monomials having degree $d + 1$ when the method has been constructed to work for monomials having degrees less than or equal to d . For example, for the two unit-square $d = 5$ and $k = 7$ cubature formulas we have been discussing, we can examine how well they work for monomials of degree $5 + 1 = 6$.

Observe, for these two unit-square cubature formulas, that the $(0, 0)$ sampling point does not contribute to the result for any first or higher degree monomial. With regard to the other sampling points, the sampling points for each weight are symmetrically arranged in such a way that their net contribution for any monomial $x_1^{j_1} x_2^{j_2}$ is identically zero if either j_1 or j_2 is odd, but not both are odd. Therefore these cubature formulas automatically satisfy (2.7) and (2.9) exactly for this subset of monomials,

$$\sum_{i=1}^5 w^i P_{j_1, j_2}(x^i) = \int_D d^2 x P_{j_1, j_2}(x) = 0 \quad \text{if either } j_1 \text{ or } j_2 \text{ is odd, but not both are odd.} \quad (\text{T.2.14})$$

Figure T.2.1: Sampling points for $*C^2:5-1$ Figure T.2.2: Sampling points for $*C^2:5-2$

Note that, by this symmetry, (2.7) is automatically satisfied for *all* odd degree polynomials.

With regard to monomials of degree 6, we need to examine the performance of the methods (2.12) and (2.13) separately. Comparison of Figures 2.1 and 2.2 shows that the sampling points for the method (2.12) are more symmetrically arranged. In particular, for this method we have the stronger result that for all monomials

$$\sum_{i=1}^5 w^i P_{j_1, j_2}(x^i) = 0 \quad \text{if either } j_1 \text{ or } j_2 \text{ is odd, or both are odd.} \quad (\text{T.2.15})$$

Therefore this cubature formula automatically satisfies (2.9) exactly for all monomials of any degree if either j_1 or j_2 is odd, or both are odd. What remains to be examined for monomials of degree 6 are the cases x_1^6 , $x_1^4 x_2^2$, $x_1^2 x_2^4$, and x_2^6 , for which, according to (2.8), the exact results are $4/7$, $4/15$, $4/15$, and $4/7$, respectively. Numerical calculations show that, for $*C^2:5-1$, there are the results

$$\sum_{i=1}^5 w^i P_{6,0}(x^i) = 4/7 - .0914 \dots = (4/7)(1 - .16 \dots), \quad (\text{T.2.16})$$

$$\sum_{i=1}^5 w^i P_{4,2}(x^i) = 4/15 + 0, \quad (\text{T.2.17})$$

$$\sum_{i=1}^5 w^i P_{2,4}(x^i) = 4/15 - .1185 \dots = (4/15)(1 - .44 \dots), \quad (\text{T.2.18})$$

$$\sum_{i=1}^5 w^i P_{0,6}(x^i) = 4/7 + .0271 \dots = (4/7)(1 + .05 \dots). \quad (\text{T.2.19})$$

Surprisingly, (2.17) shows that $*C^2:5-1$ is exact for $x_1^4 x_2^2$! We conclude, with regard to monomials of degrees 0 through 7, that $*C^2:5-1$ is exact for all such monomials save for the monomials x_1^6 , $x_1^2 x_2^4$, and x_2^6 where it makes errors of approximately 16%, 44%, and 5%, respectively.

For $*C^2:5-2$, the sampling points are less symmetrically arranged so that a larger number of degree 6 monomials need to be considered separately. Numerical calculations show that there are the results

$$\sum_{i=1}^5 w^i P_{6,0}(x^i) = 4/7 - .032 \dots = (4/7)(1 - .06 \dots), \quad (\text{T.2.20})$$

$$\sum_{i=1}^5 w^i P_{5,1}(x^i) = 0 - .059 \dots, \quad (\text{T.2.21})$$

$$\sum_{i=1}^5 w^i P_{4,2}(x^i) = 4/15 - .059 \dots = (4/15)(1 - .22 \dots), \quad (\text{T.2.22})$$

$$\sum_{i=1}^5 w^i P_{3,3}(x^i) = 0 + .059 \dots, \quad (\text{T.2.23})$$

$$\sum_{i=1}^5 w^i P_{2,4}(x^i) = 4/15 - .059 \dots = (4/15)(1 - .22 \dots), \quad (\text{T.2.24})$$

$$\sum_{i=1}^5 w^i P_{1,5}(x^i) = 0 - .059 \dots, \quad (\text{T.2.25})$$

$$\sum_{i=1}^5 w^i P_{0,6}(x^i) = 4/7 - .032 \dots = (4/7)(1 - .06 \dots). \quad (\text{T.2.26})$$

We conclude, with regard to monomials of degrees 0 through 7, that $*C^2:5-2$ is exact for all such monomials save for all the monomials of degree 6 where it makes the errors indicated above.

Finally, comparison of (2.16) through (2.19) for $*C^2:5-1$ with (2.20) through (2.26) for $*C^2:5-2$ shows that, save for the case $x_1^2 x_2^4$, $*C^2:5-1$ generally makes smaller errors than $*C^2:5-2$, and therefore might be slightly preferred.

T.2.3 Cubature on a Rectangle

Let $R(a, b)$ be a rectangle with sides $2a$ and $2b$ parameterized by coordinates ξ_1, ξ_2 such that

$$\xi_1 \in [-a, a], \quad \xi_2 \in [-b, b]. \quad (\text{T.2.27})$$

Let $g(\xi)$ be some function defined on $R(a, b)$ and suppose we wish to evaluate the integral

$$\int_{R(a,b)} d^2\xi g(\xi). \quad (\text{T.2.28})$$

Our goal is to find k sampling points ξ^i and k weights \bar{w}^i such that

$$\int_{R(a,b)} d^2\xi g(\xi) \simeq \sum_{i=1}^k \bar{w}^i g(\xi^i). \quad (\text{T.2.29})$$

Introduce new variables x_1, x_2 by the rules

$$\xi_1 = ax_1, \quad \xi_2 = bx_1 \quad (\text{T.2.30})$$

so that (2.30) maps C^2 into $R(a, b)$. See (2.3). Then there is the relation

$$\int_{R(a,b)} d^2\xi g(\xi) = ab \int_{C^2} d^2x f(x) \quad (\text{T.2.31})$$

where

$$f(x) = g(ax_1, bx_2). \quad (\text{T.2.32})$$

From (2.2) we have the approximation

$$\int_{C^2} d^2x f(x) \simeq \sum_{i=1}^k w^i f(x^i). \quad (\text{T.2.33})$$

It follows that

$$\int_{R(a,b)} d^2\xi g(\xi) \simeq ab \sum_{i=1}^k w^i f(x^i) = ab \sum_{i=1}^k w^i g(ax_1^i, bx_2^i). \quad (\text{T.2.34})$$

We conclude that (2.29) holds providing we make the definitions

$$\bar{w}^i = abw^i, \quad (\text{T.2.35})$$

$$\xi_1^i = ax_1^i, \quad \xi_2^i = bx_2^i. \quad (\text{T.2.36})$$

Exercises

T.2.1. The purpose of this exercise is to study how the accuracy of a cubature formula depends on the size of the integration domain and the analytic properties of the function being integrated. For simplicity, we will consider the case (2.29), cubature on a rectangle.

Suppose that g has a Taylor expansion of the form

$$g(\xi) = \sum_{j_1, j_2} g_{j_1, j_2} \xi_1^{j_1} \xi_2^{j_2}. \quad (\text{T.2.37})$$

Rearrange this expansion into one in homogenous polynomials by writing

$$g(\xi) = \sum_{\ell} P_{\ell}(\xi) \quad (\text{T.2.38})$$

where

$$P_{\ell}(\xi) = \sum_{j_1 + j_2 = \ell} g_{j_1, j_2} \xi_1^{j_1} \xi_2^{j_2}. \quad (\text{T.2.39})$$

Since we will be employing cubature formulas that are exact for polynomials through degree d , let us rewrite (2.38) in the form

$$g(\xi) = \sum_{\ell=0}^d P_{\ell}(\xi) + \Delta(\xi) \quad (\text{T.2.40})$$

where

$$\Delta(\xi) = \sum_{\ell=d+1}^{\infty} P_{\ell}(\xi) = \sum_{\ell=d+1}^{\infty} \sum_{j_1 + j_2 = \ell} g_{j_1, j_2} \xi_1^{j_1} \xi_2^{j_2}. \quad (\text{T.2.41})$$

In a moment, we will treat Δ as an error term. First, assuming suitable analyticity for g , use the Cauchy bound (33.2.18) on Taylor coefficients to show that, for $\xi \in R(a, b)$, Δ has the bound

$$|\Delta(\xi)| \leq \sum_{\ell=d+1}^{\infty} \sum_{j_1 + j_2 = \ell} |g_{j_1, j_2}| |\xi_1^{j_1}| |\xi_2^{j_2}| \leq L \quad (\text{T.2.42})$$

where

$$L = K \sum_{\ell=d+1}^{\infty} \sum_{j_1 + j_2 = \ell} (a/R'_1)^{j_1} (b/R'_2)^{j_2}. \quad (\text{T.2.43})$$

Verify that the series for L converges, when $a < R'_1$ and $b < R'_2$, by showing that there is the relation

$$\begin{aligned} \sum_{\ell=d+1}^{\infty} \sum_{j_1+j_2=\ell} (a/R'_1)^{j_1} (b/R'_2)^{j_2} &\leq \sum_{\ell=0}^{\infty} \sum_{j_1+j_2=\ell} (a/R'_1)^{j_1} (b/R'_2)^{j_2} = \sum_{j_1,j_2} (a/R'_1)^{j_1} (b/R'_2)^{j_2} \\ &= [1/(1-a/R'_1)][1/(1-b/R'_2)]. \end{aligned} \quad (\text{T.2.44})$$

Therefore, L is well defined.

Next, from (2.40), verify the relations

$$\int_{R(a,b)} d^2\xi g(\xi) = \int_{R(a,b)} d^2\xi \sum_{\ell=0}^d P_\ell(\xi) + \int_{R(a,b)} d^2\xi \Delta(\xi), \quad (\text{T.2.45})$$

$$\sum_{i=1}^k \bar{w}^i g(\xi^i) = \sum_{i=1}^k \bar{w}^i \sum_{\ell=0}^d P_\ell(\xi^i) + \sum_{i=1}^k \bar{w}^i \Delta(\xi^i). \quad (\text{T.2.46})$$

Show that, by the construction of the sampling points ξ^i and the weights \bar{w}^i , there must be the relation

$$\int_{R(a,b)} d^2\xi \sum_{\ell=0}^d P_\ell(\xi) = \sum_{i=1}^k \bar{w}^i \sum_{\ell=0}^d P_\ell(\xi^i). \quad (\text{T.2.47})$$

Then subtract (2.46) from (2.45) to show that

$$\int_{R(a,b)} d^2\xi g(\xi) - \sum_{i=1}^k \bar{w}^i g(\xi^i) = \int_{R(a,b)} d^2\xi \Delta(\xi) - \sum_{i=1}^k \bar{w}^i \Delta(\xi^i). \quad (\text{T.2.48})$$

Consequently, verify that there is the inequality

$$| \int_{R(a,b)} d^2\xi g(\xi) - \sum_{i=1}^k \bar{w}^i g(\xi^i) | \leq | \int_{R(a,b)} d^2\xi \Delta(\xi) | + | \sum_{i=1}^k \bar{w}^i \Delta(\xi^i) |. \quad (\text{T.2.49})$$

To continue, verify the bounds

$$| \int_{R(a,b)} d^2\xi \Delta(\xi) | \leq 4abL, \quad (\text{T.2.50})$$

$$| \sum_{i=1}^k \bar{w}^i \Delta(\xi^i) | \leq 4abL. \quad (\text{T.2.51})$$

Thus, show that there is the final inequality

$$| \int_{R(a,b)} d^2\xi g(\xi) - \sum_{i=1}^k \bar{w}^i g(\xi^i) | \leq 8abL. \quad (\text{T.2.52})$$

Your challenge is to determine how the error term L depends on the dimensions of the rectangle $R(a, b)$. To this end suppose a and b are scaled by a common factor of σ ,

$$a(\sigma) = \sigma a_1, \quad b(\sigma) = \sigma b_1 \quad (\text{T.2.53})$$

so that $R(a, b)$ is what we may call the original rectangle $R(a_1, b_1)$ when $\sigma = 1$, and $R(a, b)$ becomes ever smaller as $\sigma \rightarrow 0$.

Based on (2.43), define $L(\sigma)$ by writing

$$L(\sigma) = K \sum_{\ell=d+1}^{\infty} \sum_{j_1+j_2=\ell} (\sigma a_1/R'_1)^{j_1} (\sigma b_1/R'_2)^{j_2}. \quad (\text{T.2.54})$$

Show that

$$\begin{aligned} L(\sigma) &= \sigma^{d+1} K \sum_{\ell=d+1}^{\infty} \sigma^{\ell-d-1} \sum_{j_1+j_2=\ell} (a_1/R'_1)^{j_1} (b_1/R'_2)^{j_2} \\ &= \sigma^{d+1} K' + O(\sigma^{d+2}) \end{aligned} \quad (\text{T.2.55})$$

where

$$K' = K \sum_{j_1+j_2=d+1} (a_1/R'_1)^{j_1} (b_1/R'_2)^{j_2}. \quad (\text{T.2.56})$$

Insert (2.55) into (2.52) to obtain the inequality

$$| \int_{R(a,b)} d^2\xi g(\xi) - \sum_{i=1}^k \bar{w}^i g(\xi^i) | \leq 8\sigma^2 a_1 b_1 L(\sigma) = \sigma^{d+3} 8a_1 b_1 K' + O(\sigma^{d+4}). \quad (\text{T.2.57})$$

We expect that each of the two terms on the left side of (2.57) scales as σ^2 . You have shown that their difference, the error, scales as σ^{d+3} . Therefore, the *relative* error scales as σ^{d+1} .

T.2.2. Review Exercise 2.1. The purpose of this exercise is to find a more refined error bound for the specific case of the cubature formula ${}^*C^2:5-1$ by exploiting the relations (2.16) through (2.19). Define two linear functionals I and Q by the rules

$$I[g] = \int_{R(a,b)} d^2\xi g(\xi), \quad (\text{T.2.58})$$

$$Q[g] = \sum_{i=1}^5 \bar{w}^i g(\xi^i). \quad (\text{T.2.59})$$

Show, when ${}^*C^2:5-1$ is used in (2.35) and (2.36), that

$$Q[P_{j_1,j_2}] = I[P_{j_1,j_2}] \quad (\text{T.2.60})$$

for all P_{j_1,j_2} having degree less than 6, and also for all P_{j_1,j_2} having degree 7.

Monomials of degree 6 need to be treated separately. With reference to (2.16) through (2.19), define constants $\lambda_{6,0}$ etc. by the rules

$$\lambda_{6,0} = .0914 \dots, \quad (\text{T.2.61})$$

$$\lambda_{2,4} = .1185 \dots, \quad (\text{T.2.62})$$

$$\lambda_{0,6} = .0271 \dots. \quad (\text{T.2.63})$$

Show that, when $*C^2:5-1$ is used in (2.35) and (2.36), there are the results

$$Q[P_{6,0}] = I[P_{6,0}] - a^7 b \lambda_{6,0}, \quad (\text{T.2.64})$$

$$Q[P_{2,4}] = I[P_{2,4}] - a^3 b^5 \lambda_{2,4}, \quad (\text{T.2.65})$$

$$Q[P_{0,6}] = I[P_{0,6}] + ab^7 \lambda_{0,6}, \quad (\text{T.2.66})$$

and that (2.60) holds for all the remaining monomials of degree 6.

Using the machinery just developed show that, when $*C^2:5-1$ is used in (2.35) and (2.36), there is the result

$$\int_{R(a,b)} d^2\xi g(\xi) = \sum_{i=1}^5 \bar{w}^i g(\xi^i) + ab[\lambda_{6,0}g_{6,0}a^6 + \lambda_{2,4}g_{2,4}a^2b^4 - \lambda_{0,6}g_{0,6}b^6] + O(\sigma^{10}). \quad (\text{T.2.67})$$

Note that

$$g_{6,0} = (1/6!)g^{[6,0]}(0,0), \quad g_{2,4} = (1/2!)(1/4!)g^{[2,4]}(0,0), \quad g_{0,6} = (1/6!)g^{[0,6]}(0,0). \quad (\text{T.2.68})$$

Show that (2.67) can also be written in the form

$$\int_{R(a,b)} d^2\xi g(\xi) = \sum_{i=1}^5 \bar{w}^i g(\xi^i) + \sigma^8 a_1 b_1 [\lambda_{6,0}g_{6,0}a_1^6 + \lambda_{2,4}g_{2,4}a_1^2b_1^4 - \lambda_{0,6}g_{0,6}b_1^6] + O(\sigma^{10}). \quad (\text{T.2.69})$$

Therefore, for the case where $*C^2:5-1$ is used in (2.35) and (2.36), you have found an exact result for the $O(\sigma^{d+3})$ term in (2.57) and have shown that the $O(\sigma^{d+4})$ term, in fact, vanishes.

Examination of (2.67) and (2.68) indicates that there is still some freedom left in the choice of orientation of the sampling point scheme and the aspect ratio of the rectangle $R(a, b)$. In some cases it may be possible to minimize the leading error term in (2.67) by exploiting this freedom. Suppose that something is known about the relative sizes of the terms $g_{6,0}$, $g_{4,2}$, $g_{2,4}$, and $g_{0,6}$. If, for example, $g_{4,2}$ is much smaller than $g_{2,4}$, then it might be advantageous to rotate the sampling points shown in Figure 2.1 by 90° so that the cubature formula error no longer involves $g_{2,4}$ but instead involves $g_{4,2}$. With regard to the aspect ratio of $R(a, b)$, in the case that (2.67) is employed and assuming that $g_{6,0}$ and $g_{0,6}$ are comparable, it might be advantageous to use rectangles for which $a < b$ since $\lambda_{6,0} > \lambda_{0,6}$. Obviously, when contemplating this strategy, attention should also be paid to the size of the $\lambda_{2,4}g_{2,4}a^2b^4$ error term.

T.2.4 Cubature on the Two-Sphere

Bibliography

Quadrature Formulas

- [1] P. Davis and P. Rabinowitz, *Methods of Numerical Integration*, Second Edition, Dover (2007).
- [2] F.B. Hildebrand, *Introduction to Numerical Analysis*, Second Edition, Dover (1987).
- [3] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover (1972). Also available on the Web by Googling “abramowitz and stegun 1972”. We note that the original June 1964 edition of this venerable reference has an error with regard to Gaussian quadrature. There formula 25.4.30 for R_n has an erroneous factor of 2^{2n+1} . This error has been corrected in the 1972 edition.
- [4] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). See also the Web site <http://dlmf.nist.gov/>.
- [5] L.N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM (2013).
- [6] L.N. Trefethen, *Six Myths of Polynomial Interpolation and Quadrature*, http://people.maths.ox.ac.uk/trefethen/publication/PDF/2011_139.pdf.
- [7] B. Beers and A.J. Dragt, “New Theorems about Spherical Harmonic Expansions and $SU(2)$ ”, *J. Math. Phys.* **11**, 2313 (1970).

Cubature Formulas

- [8] A.H. Stroud, *Approximate calculation of multiple integrals*, Prentice-Hall (1971).
- [9] H. Engels, *Numerical Quadrature and Cubature*, Academic Press (1980).
- [10] R. Cools, see, for example, the Web sites
<http://nines.cs.kuleuven.be/research/ecf/Teksten/accepted.pdf>
<http://nines.cs.kuleuven.be/ecf/>

Appendix U

Rotational Classification and Properties of Polynomials and Analytic/Polynomial Vector Fields

U.1 Introduction

Suppose $\mathbf{A}(\mathbf{r})$ is an analytic vector field in three dimensions. That is, suppose there are three component functions $A_x(\mathbf{r})$, $A_y(\mathbf{r})$, and $A_z(\mathbf{r})$, all of which are analytic in some common domain in the variables x , y , and z . Without loss of generality, we may take this domain to be centered on the origin. Doing so brings us to the notationally easier problem of studying polynomial vector fields, vector fields whose components are polynomials in the variables x , y , and z . In this appendix we will study how to use $SO(3)$, the rotation group in 3 dimensions, as a tool for labeling/classifying both such polynomials and such polynomial vector fields. We will also study some of their properties.

U.2 Polynomials and Spherical Polynomials

U.2.1 Polynomials

The first step in studying multi-variable polynomials is to decompose them into homogeneous polynomials. According to (7.3.40) the number of monomials of degree n in $d = 3$ variables is given by $N(n, 3)$. And, according to (7.10.17), the total number of monomials in $d = 3$ variables having degrees 0 through n is given by $S_0(n, 3)$. Table 2.1 below shows values of $N(n, 3)$ and $S_0(n, 3)$ for various values of n . Also shown are the quantities $3N(n, 3)$ and $3S_0(n, 3)$. The quantity $3N(n, 3)$ is the number of parameters required to specify 3 homogeneous polynomials of degree n in $d = 3$ variables. And the quantity $3S_0(n, 3)$ is the number of parameters required to specify a $d = 3$ dimensional vector field in $d = 3$ variables through terms of degree n . Thus, for example, to specify a 3-dimensional vector field through terms of degree $n = 4$ requires $3S_0(4, 3) = 105$ parameters. Finally, the table displays $S_B(n)$, the number of parameters required to specify a source-free magnetic field through terms of degree n . See (2.7). Thus, for example, to specify a source-free magnetic

field through terms of degree $n = 4$ requires $S_B(4) = 35$ parameters. Note that the requirement that the field be source free, namely divergence and curl free, reduces the parameter count substantially even for modest values of n .

Table U.2.1: $N(n, 3)$, $S_0(n, 3)$, $3N(n, 3)$, $3S_0(n, 3)$, and $S_B(n)$ as functions of n .

n	$N(n, 3)$	$S_0(n, 3)$	$3N(n, 3)$	$3S_0(n, 3)$	$S_B(n)$
0	1	1	3	3	3
1	3	4	9	12	8
2	6	10	18	30	15
3	10	20	30	60	24
4	15	35	45	105	35
5	21	56	63	168	48
6	28	84	84	252	63
7	36	120	108	360	80
8	45	165	135	495	99

U.2.2 Spherical Polar Coordinates and Harmonic Polynomials

Introduce spherical polar coordinates in the usual way as in Section 15.2.2:

$$r^2 = x^2 + y^2 + z^2, \quad (\text{U.2.1})$$

$$x = r \sin(\theta) \cos(\phi), \quad (\text{U.2.2})$$

$$y = r \sin(\theta) \sin(\phi), \quad (\text{U.2.3})$$

$$z = r \cos \theta. \quad (\text{U.2.4})$$

Let $Y_\ell^m(\theta, \phi)$ denote the usual *spherical harmonics*,

$$\begin{aligned} Y_\ell^m(\theta, \phi) &= \{[(2\ell+1)(\ell-m)!]/[4\pi(\ell+m)!]\}^{1/2} P_\ell^m(\cos \theta) \exp(im\phi) \\ &\text{with } -\ell \leq m \leq \ell. \end{aligned} \quad (\text{U.2.5})$$

Here the P_ℓ^m are the usual associated Legendre functions. Consider the functions

$$H_\ell^m(\mathbf{r}) = r^\ell Y_\ell^m(\theta, \phi). \quad (\text{U.2.6})$$

They are homogeneous polynomials of degree ℓ in the variables x , y , and z . They are also harmonic functions, and are variously called *harmonic polynomials* or *solid harmonics*. For a given ℓ there are $2\ell+1$ such polynomials as m ranges over $-\ell \leq m \leq \ell$.

At this point we are prepared to compute $S_B(n)$. Evidently a source-free magnetic field homogeneous of degree ℓ results from the gradient of a harmonic polynomial of degree $\ell+1$, and there are $2(\ell+1)+1 = 2\ell+3$ such harmonic polynomials. Therefore we have the result

$$S_B(n) = \sum_{\ell=0}^n (2\ell+3) = 3 \sum_{\ell=0}^n 1 + 2 \sum_{\ell=0}^n \ell = 3(n+1) + 2(n/2)(n+1) = (n+1)(n+3). \quad (\text{U.2.7})$$

U.2.3 Examples of Harmonic Polynomials and Missing Homogeneous Polynomials

Continuing with the discussion of harmonic polynomials we find, for example, the results

$$H_0^0(\mathbf{r}) = 1/\sqrt{4\pi}; \quad (\text{U.2.8})$$

$$\begin{aligned} H_1^1(\mathbf{r}) &= \sqrt{3/(4\pi)}(-1/\sqrt{2})(x + iy) = -\sqrt{3/(8\pi)}(x + iy), \\ H_1^0(\mathbf{r}) &= \sqrt{3/(4\pi)}z, \\ H_1^{-1}(\mathbf{r}) &= \sqrt{3/(4\pi)}(1/\sqrt{2})(x - iy) = \sqrt{3/(8\pi)}(x - iy); \end{aligned} \quad (\text{U.2.9})$$

$$\begin{aligned} H_2^2(\mathbf{r}) &= \sqrt{15/(32\pi)}(x + iy)^2, \\ H_2^1(\mathbf{r}) &= -\sqrt{15/(8\pi)}(x + iy)z, \\ H_2^0(\mathbf{r}) &= \sqrt{5/(16\pi)}(2z^2 - x^2 - y^2), \\ H_2^{-1}(\mathbf{r}) &= \sqrt{15/(8\pi)}(x - iy)z, \\ H_2^{-2}(\mathbf{r}) &= \sqrt{15/(32\pi)}(x - iy)^2. \end{aligned} \quad (\text{U.2.10})$$

Note that there is one function for the case $\ell = 0$, three functions for the case $\ell = 1$, and five functions in the case $\ell = 2$. Comparison of this function count with the first three lines of Table 2.1 shows that all the homogeneous monomials of degrees 0 and 1 have been accounted for, but one homogenous polynomial of degree 2 is missing. This missing polynomial is evidently proportional to $r^2 = x^2 + y^2 + z^2$, and may be taken to be $r^2 H_0^0(\mathbf{r})$.¹

What about the case $n = 3$? There are the polynomials $H_3^m(\mathbf{r})$, and there are $2\ell + 1 = 7$ such polynomials when $\ell = 3$. But from Table 2.1 we see that there should be, in total, ten polynomials when $n = 3$. What are the remaining three? The remaining three third-degree polynomials can be taken to be the polynomials $r^2 H_1^m(\mathbf{r})$.

U.2.4 Spherical Polynomials

The general picture should now be clear. Form the functions $S_{n\ell}^m(\mathbf{r})$, which we will call *spherical polynomials*, by the rule

$$S_{n\ell}^m(\mathbf{r}) = r^n Y_\ell^m(\theta, \phi) \text{ with } \ell = n, n-2, \dots, 0 \text{ for } n \text{ even}, \quad (\text{U.2.11})$$

$$S_{n\ell}^m(\mathbf{r}) = r^n Y_\ell^m(\theta, \phi) \text{ with } \ell = n, n-2, \dots, 1 \text{ for } n \text{ odd}. \quad (\text{U.2.12})$$

So doing produces polynomials of degree n , and these polynomials form a basis for the set of all polynomials of degree n . But still more can be said. The functions Y_ℓ^m have well-defined transformation properties under rotations, and r is unchanged by rotations. It follows that the $S_{n\ell}^m(\mathbf{r})$ have the same transformation properties as the Y_ℓ^m . Finally, in view of (2.6), (2.11), and (2.12), spherical polynomials of the special form $S_{nn}^m(\mathbf{r})$ are harmonic polynomials.

¹Note that quantities of the form $r^{2k} = (x^2 + y^2 + z^2)^k$ are *polynomials* in the variables x , y , and z .

Listed below, for possible future use, are the spherical polynomials for the cases $n = 0$, $n = 1$, and $n = 2$:

$$S_{00}^0(\mathbf{r}) = 1/\sqrt{4\pi}; \quad (\text{U.2.13})$$

$$\begin{aligned} S_{11}^1(\mathbf{r}) &= -\sqrt{3/(8\pi)}(x + iy), \\ S_{11}^0(\mathbf{r}) &= \sqrt{3/(4\pi)}z, \\ S_{11}^{-1}(\mathbf{r}) &= \sqrt{3/(8\pi)}(x - iy); \end{aligned} \quad (\text{U.2.14})$$

$$S_{20}^0(\mathbf{r}) = 1/\sqrt{4\pi}(x^2 + y^2 + z^2); \quad (\text{U.2.15})$$

$$\begin{aligned} S_{22}^2(\mathbf{r}) &= \sqrt{15/(32\pi)}(x + iy)^2, \\ S_{22}^1(\mathbf{r}) &= -\sqrt{15/(8\pi)}(x + iy)z, \\ S_{22}^0(\mathbf{r}) &= \sqrt{5/(16\pi)}(2z^2 - x^2 - y^2), \\ S_{22}^{-1}(\mathbf{r}) &= \sqrt{15/(8\pi)}(x - iy)z, \\ S_{22}^{-2}(\mathbf{r}) &= \sqrt{15/(32\pi)}(x - iy)^2. \end{aligned} \quad (\text{U.2.16})$$

U.3 Analytic/Polynomial Vector Fields and Spherical Polynomial Vector Fields

With the $S_{n\ell}^m(\mathbf{r})$ in hand, we next turn to the problem of classifying/labeling all vector fields whose components are polynomials in x , y , and z . Our construction will be analogous to that for vector spherical harmonics.

U.3.1 Vector Spherical Harmonics

Define a spherical basis $\mathbf{e}_{\pm 1}$, \mathbf{e}_0 by the rule

$$\begin{aligned} \mathbf{e}_{+1} &= -(1/\sqrt{2})(\mathbf{e}_x + i\mathbf{e}_y), \\ \mathbf{e}_0 &= \mathbf{e}_z, \\ \mathbf{e}_{-1} &= (1/\sqrt{2})(\mathbf{e}_x - i\mathbf{e}_y). \end{aligned} \quad (\text{U.3.1})$$

Note the resemblance between (3.1) and (2.9) and (2.14).² Indeed, suppose we define three functions $r_m(\mathbf{r})$ by the rules

$$\begin{aligned} r_{+1}(\mathbf{r}) &= \mathbf{r} \cdot \mathbf{e}_{+1} = -(1/\sqrt{2})(x + iy), \\ r_0(\mathbf{r}) &= \mathbf{r} \cdot \mathbf{e}_0 = z, \\ r_{-1}(\mathbf{r}) &= \mathbf{r} \cdot \mathbf{e}_{-1} = (1/\sqrt{2})(x - iy). \end{aligned} \quad (\text{U.3.2})$$

²Note also that the basis (3.1) differs slightly from that of Exercise 3.7.22, which was selected to make the $so(3)$ structure constants real.

Then (2.13) can be rewritten in the form

$$S_{11}^m(\mathbf{r}) = \sqrt{3/(4\pi)} r_m(\mathbf{r}) = \sqrt{3/(4\pi)} \mathbf{r} \cdot \mathbf{e}_m. \quad (\text{U.3.3})$$

Moreover, we have the relation

$$\begin{aligned} r_{-1}(\mathbf{r})\mathbf{e}_{+1} + r_{+1}(\mathbf{r})\mathbf{e}_{-1} &= -(1/2)[(x - iy)(\mathbf{e}_x + i\mathbf{e}_y) + (x + iy)(\mathbf{e}_x - i\mathbf{e}_y)] \\ &= -x\mathbf{e}_x - y\mathbf{e}_y. \end{aligned} \quad (\text{U.3.4})$$

It follows that there is the relation

$$-r_{-1}(\mathbf{r})\mathbf{e}_{+1} + r_0(\mathbf{r})\mathbf{e}_0 - r_{+1}(\mathbf{r})\mathbf{e}_{-1} = x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z = \mathbf{r}. \quad (\text{U.3.5})$$

But, we have digressed. The *vector spherical harmonics* $\mathbf{Y}_{\ell J}^M(\theta, \phi)$ are defined by the rules

$$\mathbf{Y}_{\ell J}^M(\theta, \phi) = \sum_{m_1, m_2} C_{m_1 m_2 M}^{\ell 1 J} Y_\ell^{m_1}(\theta, \phi) \mathbf{e}_{m_2}. \quad (\text{U.3.6})$$

Here the $C_{m_1 m_2 M}^{\ell 1 J}$ denote the *Clebsch-Gordan* coefficients that couple the angular momenta ℓ and 1 to produce angular momentum J .³ In particular, there are *range* rules:

$$\text{when } \ell = 0, \text{ then } J = 1; \quad (\text{U.3.7})$$

$$\text{when } \ell > 0, \text{ then } J \text{ can have the values } J = \ell - 1, \ell, \ell + 1. \quad (\text{U.3.8})$$

The specific Clebsch-Gordan coefficients needed for our purposes are given by the relations

$$C_{M-1,1,M}^{\ell,1,\ell+1} = \sqrt{(\ell + M)(\ell + M + 1)/[(2\ell + 1)(2\ell + 2)]}, \quad (\text{U.3.9})$$

$$C_{M,0,M}^{\ell,1,\ell+1} = \sqrt{(\ell - M + 1)(\ell + M + 1)/[(2\ell + 1)(\ell + 1)]}, \quad (\text{U.3.10})$$

$$C_{M+1,-1,M}^{\ell,1,\ell+1} = \sqrt{(\ell - M)(\ell - M + 1)/[(2\ell + 1)(2\ell + 2)]}; \quad (\text{U.3.11})$$

$$C_{M-1,1,M}^{\ell,1,\ell} = -\sqrt{(\ell + M)(\ell - M + 1)/[2\ell(\ell + 1)]}, \quad (\text{U.3.12})$$

$$C_{M,0,M}^{\ell,1,\ell} = M/\sqrt{\ell(\ell + 1)}, \quad (\text{U.3.13})$$

$$C_{M+1,-1,M}^{\ell,1,\ell} = \sqrt{(\ell - M)(\ell + M + 1)/[2\ell(\ell + 1)]}; \quad (\text{U.3.14})$$

$$C_{M-1,1,M}^{\ell,1,\ell-1} = \sqrt{(\ell - M)(\ell - M + 1)/[2\ell(2\ell + 1)]}, \quad (\text{U.3.15})$$

$$C_{M,0,M}^{\ell,1,\ell-1} = -\sqrt{(\ell - M)(\ell + M)/[\ell(2\ell + 1)]}, \quad (\text{U.3.16})$$

$$C_{M+1,-1,M}^{\ell,1,\ell-1} = \sqrt{(\ell + M + 1)(\ell + M)/[2\ell(2\ell + 1)]}. \quad (\text{U.3.17})$$

³Note that we write the subscripts on the vector spherical harmonics and elsewhere in the order ℓJ , the same order in which they appear in the Clebsch-Gordan coefficients. Thus, the last lower index and the upper index are *paired* and obey the rule $-J \leq M \leq J$. Many authors employ the opposite order, namely $J\ell$. We also remark that the Clebsch-Gordan coefficients are also sometimes called *Wigner* or vector-addition coefficients. Finally, we use the more compact notation $C_{m_1 m_2 M}^{\ell 1 J}$ for what some authors write as $C(\ell 1 J; m_1 m_2 M)$.

U.3.2 Spherical Polynomial Vector Fields

In a corresponding manner, we define *spherical polynomial vector fields* $\mathbf{S}_{n\ell J}^M(\mathbf{r})$ by the rule

$$\begin{aligned}\mathbf{S}_{n\ell J}^M(\mathbf{r}) &= \sum_{m_1, m_2} C_{m_1 m_2 M}^{\ell 1 J} S_{n\ell}^{m_1}(\mathbf{r}) \mathbf{e}_{m_2} \\ &= r^n \sum_{m_1, m_2} C_{m_1 m_2 M}^{\ell 1 J} Y_\ell^{m_1}(\theta, \phi) \mathbf{e}_{m_2} = r^n \mathbf{Y}_{\ell J}^M(\theta, \phi).\end{aligned}\quad (\text{U.3.18})$$

Note that by construction the components of $\mathbf{S}_{n\ell J}^M(\mathbf{r})$ are polynomial (analytic) functions of the variables x , y , and z .

For convenience, Table 3.1 lists the allowed values of the triplets $n\ell J$ in accord with the relations (2.10), (2.11), (3.7), and (3.8). And, of course, M lies in the range $-J \leq M \leq J$.

Table U.3.1: Allowed values of $n\ell J$

n	ℓ	J
0	0	1
1	1	0
1	1	1
1	1	2
2	0	1
2	2	1
2	2	2
2	2	3
3	1	0
3	1	1
3	1	2
3	3	2
3	3	3
3	3	4
.	.	.
.	.	.

U.3.3 Examples of and Counting Spherical Polynomial Vector Fields

Let us work out a first few examples. The simplest are those for $n = 0$. In this case we must have $\ell = 0$, see (2.10), and $J = 1$, see (3.7). For the \mathbf{S}_{001}^M we find from (3.18) the results

$$\mathbf{S}_{001}^M(\mathbf{r}) = C_{0MM}^{011} S_{00}^0(\mathbf{r}) \mathbf{e}_M. \quad (\text{U.3.19})$$

From (3.9) through (3.11) we see that all the C_{0MM}^{011} have value 1, and $S_{00}^0(\mathbf{r})$ is given by (2.12). Therefore, there is the final result

$$\mathbf{S}_{001}^M(\mathbf{r}) = (1/\sqrt{4\pi}) \mathbf{e}_M. \quad (\text{U.3.20})$$

Note that M can take the three values $-1, 0, +1$ corresponding to the fact that a constant vector field has 3 constant components.

The next simplest case is $n = 1$. Then we must have $\ell = 1$, see (2.11), and there are the possibilities $J = 0, 1, 2$. See (3.8). For the case $J = 0$ we find from (3.18) the result

$$\mathbf{S}_{110}^0(\mathbf{r}) = C_{-1,1,0}^{110} S_{11}^{-1}(\mathbf{r}) \mathbf{e}_{+1} + C_{000}^{110} S_{11}^0(\mathbf{r}) \mathbf{e}_0 + C_{1,-1,0}^{110} S_{11}^1(\mathbf{r}) \mathbf{e}_{-1}. \quad (\text{U.3.21})$$

From (3.15) through (3.17) we see that the required Clebsch-Gordan coefficients have the values

$$C_{-1,1,0}^{110} = \sqrt{1/3}, \quad (\text{U.3.22})$$

$$C_{000}^{110} = -\sqrt{1/3}, \quad (\text{U.3.23})$$

$$C_{1,-1,0}^{110} = \sqrt{1/3}. \quad (\text{U.3.24})$$

Also, the required $S_{11}^m(\mathbf{r})$ can be written in the form (3.3). Therefore, (3.21) can be rewritten in the final form

$$\mathbf{S}_{110}^0(\mathbf{r}) = [\sqrt{1/3}][\sqrt{3/(4\pi)}][r_{-1}(\mathbf{r}) \mathbf{e}_{+1} - r_0(\mathbf{r}) \mathbf{e}_0 + r_{+1}(\mathbf{r}) \mathbf{e}_{-1}] = -\sqrt{1/(4\pi)} \mathbf{r}. \quad (\text{U.3.25})$$

Here we have used (3.5).

For the case $J = 1$ we find from (3.18) and (3.3) the results

$$\mathbf{S}_{111}^M(\mathbf{r}) = \sum_{m_1, m_2} C_{m_1 m_2 M}^{111} S_{11}^{m_1}(\mathbf{r}) \mathbf{e}_{m_2} = \sqrt{3/(4\pi)} \sum_{m_1, m_2} C_{m_1 m_2 M}^{111} r_{m_1}(\mathbf{r}) \mathbf{e}_{m_2}. \quad (\text{U.3.26})$$

It follows that

$$\mathbf{S}_{111}^1(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{011}^{111} r_0(\mathbf{r}) \mathbf{e}_{+1} + C_{101}^{111} r_{+1}(\mathbf{r}) \mathbf{e}_0], \quad (\text{U.3.27})$$

$$\mathbf{S}_{111}^0(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{-1,1,0}^{111} r_{-1}(\mathbf{r}) \mathbf{e}_{+1} + C_{000}^{111} r_0(\mathbf{r}) \mathbf{e}_0 + C_{1,-1,0}^{111} r_{+1}(\mathbf{r}) \mathbf{e}_{-1}], \quad (\text{U.3.28})$$

$$\mathbf{S}_{111}^{-1}(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{-1,0,-1}^{111} r_{-1}(\mathbf{r}) \mathbf{e}_0 + C_{0,-1,-1}^{111} r_0(\mathbf{r}) \mathbf{e}_{-1}]. \quad (\text{U.3.29})$$

In this case the relevant Clebsch-Gordan coefficients are given by the following relations:

$$C_{011}^{111} = -1/\sqrt{2}, \quad (\text{U.3.30})$$

$$C_{101}^{111} = 1/\sqrt{2}, \quad (\text{U.3.31})$$

see (3.12) and (3.13);

$$C_{-1,1,0}^{111} = -1/\sqrt{2}, \quad (\text{U.3.32})$$

$$C_{000}^{111} = 0, \quad (\text{U.3.33})$$

$$C_{1,-1,0}^{111} = 1/\sqrt{2}, \quad (\text{U.3.34})$$

see (3.12) through (3.14);

$$C_{-1,0,-1}^{111} = -1/\sqrt{2}, \quad (\text{U.3.35})$$

$$C_{0,-1,-1}^{111} = 1/\sqrt{2}, \quad (\text{U.3.36})$$

see (3.13) and (3.14). Consequently, the relations (3.27) through (3.29) can be rewritten in the form

$$\mathbf{S}_{111}^1(\mathbf{r}) = [\sqrt{3/(8\pi)}][-r_0(\mathbf{r})\mathbf{e}_{+1} + r_{+1}(\mathbf{r})\mathbf{e}_0], \quad (\text{U.3.37})$$

$$\mathbf{S}_{111}^0(\mathbf{r}) = [\sqrt{3/(8\pi)}][-r_{-1}(\mathbf{r})\mathbf{e}_{+1} + r_{+1}(\mathbf{r})\mathbf{e}_{-1}], \quad (\text{U.3.38})$$

$$\mathbf{S}_{111}^{-1}(\mathbf{r}) = [\sqrt{3/(8\pi)}][-r_{-1}(\mathbf{r})\mathbf{e}_0 + r_0(\mathbf{r})\mathbf{e}_{-1}]. \quad (\text{U.3.39})$$

After a little struggle, we recognize that the expressions appearing on the right sides of (3.37) through (3.39) can be written more compactly in terms of the vector cross product. Doing so, we find the neat result

$$\mathbf{S}_{111}^M(\mathbf{r}) = -i[\sqrt{3/(8\pi)}][\mathbf{r} \times \mathbf{e}_M]. \quad (\text{U.3.40})$$

In retrospect, the cross-product result we have found should not be too surprising since we have, in effect, been combining two spin 1 objects to produce another spin 1 object, and that is just what the vector cross product operation does.

We also note, in view of (3.20), that there is the relation

$$\mathbf{S}_{111}^M(\mathbf{r}) = -i[\sqrt{3/2}][\mathbf{r} \times \mathbf{S}_{001}^M(\mathbf{r})]. \quad (\text{U.3.41})$$

Finally, consider the case $J = 2$. Now, from (3.18) and (3.3), we find that

$$\mathbf{S}_{112}^M(\mathbf{r}) = \sum_{m_1, m_2} C_{m_1 m_2 M}^{112} S_{11}^{m_1}(\mathbf{r}) \mathbf{e}_{m_2} = \sqrt{3/(4\pi)} \sum_{m_1, m_2} C_{m_1 m_2 M}^{112} r_{m_1}(\mathbf{r}) \mathbf{e}_{m_2}. \quad (\text{U.3.42})$$

It follows that

$$\mathbf{S}_{112}^2(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{112}^{112} r_1(\mathbf{r}) \mathbf{e}_1], \quad (\text{U.3.43})$$

$$\mathbf{S}_{112}^1(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{011}^{112} r_0(\mathbf{r}) \mathbf{e}_1 + C_{101}^{112} r_1(\mathbf{r}) \mathbf{e}_0], \quad (\text{U.3.44})$$

$$\mathbf{S}_{112}^0(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{-1,1,0}^{112} r_{-1}(\mathbf{r}) \mathbf{e}_1 + C_{000}^{112} r_0(\mathbf{r}) \mathbf{e}_0 + C_{1,-1,0}^{112} r_1(\mathbf{r}) \mathbf{e}_{-1}], \quad (\text{U.3.45})$$

$$\mathbf{S}_{112}^{-1}(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{-1,0,-1}^{112} r_{-1}(\mathbf{r}) \mathbf{e}_0 + C_{0,-1,-1}^{112} r_0(\mathbf{r}) \mathbf{e}_{-1}], \quad (\text{U.3.46})$$

$$\mathbf{S}_{112}^{-2}(\mathbf{r}) = [\sqrt{3/(4\pi)}][C_{-1,-1,-2}^{112} r_{-1}(\mathbf{r}) \mathbf{e}_{-1}]. \quad (\text{U.3.47})$$

To complete this calculation we need the Clebsch-Gordan coefficient values listed below:

$$C_{112}^{112} = 1, \quad (\text{U.3.48})$$

see (3.9);

$$C_{011}^{112} = 1/\sqrt{2}, \quad (\text{U.3.49})$$

$$C_{101}^{112} = 1/\sqrt{2}, \quad (\text{U.3.50})$$

see (3.9) and (3.10);

$$C_{-1,1,0}^{112} = 1/\sqrt{6}, \quad (\text{U.3.51})$$

$$C_{000}^{112} = \sqrt{2/3} = 2/\sqrt{6}, \quad (\text{U.3.52})$$

$$C_{1,-1,0}^{112} = 1/\sqrt{6}, \quad (\text{U.3.53})$$

see (3.9) through (3.11);

$$C_{-1,0,-1}^{112} = 1/\sqrt{2}, \quad (\text{U.3.54})$$

$$C_{0,-1,-1}^{112} = 1/\sqrt{2}, \quad (\text{U.3.55})$$

see (3.10) and (3.11);

$$C_{-1,-1,-2}^{112} = 1, \quad (\text{U.3.56})$$

see (3.11)).

Putting everything together gives the final results

$$\begin{aligned} \mathbf{S}_{112}^2(\mathbf{r}) &= [\sqrt{3/(4\pi)}][C_{112}^{112}r_1(\mathbf{r})\mathbf{e}_1] = [\sqrt{3/(4\pi)}]r_1(\mathbf{r})\mathbf{e}_1 \\ &= [\sqrt{3/(16\pi)}](x + iy)(\mathbf{e}_x + i\mathbf{e}_y), \end{aligned} \quad (\text{U.3.57})$$

$$\begin{aligned} \mathbf{S}_{112}^1(\mathbf{r}) &= [\sqrt{3/(4\pi)}][C_{011}^{112}r_0(\mathbf{r})\mathbf{e}_1 + C_{101}^{112}r_1(\mathbf{r})\mathbf{e}_0] \\ &= [\sqrt{3/(8\pi)}][r_0(\mathbf{r})\mathbf{e}_1 + r_1(\mathbf{r})\mathbf{e}_0] \\ &= -[\sqrt{3/(16\pi)}][z(\mathbf{e}_x + i\mathbf{e}_y) + (x + iy)\mathbf{e}_z], \end{aligned} \quad (\text{U.3.58})$$

$$\begin{aligned} \mathbf{S}_{112}^0(\mathbf{r}) &= [\sqrt{3/(4\pi)}][C_{-1,1,0}^{112}r_{-1}(\mathbf{r})\mathbf{e}_1 + C_{000}^{112}r_0(\mathbf{r})\mathbf{e}_0 + C_{1,-1,0}^{112}r_1(\mathbf{r})\mathbf{e}_{-1}] \\ &= [\sqrt{1/(8\pi)}][r_{-1}(\mathbf{r})\mathbf{e}_1 + 2r_0(\mathbf{r})\mathbf{e}_0 + r_1(\mathbf{r})\mathbf{e}_{-1}] \\ &= [\sqrt{1/(8\pi)}][(-x\mathbf{e}_x - y\mathbf{e}_y + 2z\mathbf{e}_z)], \end{aligned} \quad (\text{U.3.59})$$

$$\begin{aligned} \mathbf{S}_{112}^{-1}(\mathbf{r}) &= [\sqrt{3/(4\pi)}][C_{-1,0,-1}^{112}r_{-1}(\mathbf{r})\mathbf{e}_0 + C_{0,-1,-1}^{112}r_0(\mathbf{r})\mathbf{e}_{-1}] \\ &= [\sqrt{3/(8\pi)}][r_{-1}(\mathbf{r})\mathbf{e}_0 + r_0(\mathbf{r})\mathbf{e}_{-1}] \\ &= [\sqrt{3/(16\pi)}][(x - iy)\mathbf{e}_z + z(\mathbf{e}_x - i\mathbf{e}_y)], \end{aligned} \quad (\text{U.3.60})$$

$$\begin{aligned} \mathbf{S}_{112}^{-2}(\mathbf{r}) &= [\sqrt{3/(4\pi)}][C_{-1,-1,-2}^{112}r_{-1}(\mathbf{r})\mathbf{e}_{-1}] = [\sqrt{3/(4\pi)}]r_{-1}(\mathbf{r})\mathbf{e}_{-1} \\ &= [\sqrt{3/(16\pi)}](x - iy)(\mathbf{e}_x - i\mathbf{e}_y). \end{aligned} \quad (\text{U.3.61})$$

Let us do a count of the $n = 1$ spherical polynomial vector fields we have found. Observe that, when considering all $n = 1$ cases, there are $1 + 3 + 5 = 9$ possibilities, which is to be expected: When $n = 1$, $N(1, 3) = 3$. See Table 2.1. Moreover there are three components to be specified, and therefore there are $3N(1, 3) = 9$ parameters to be specified.

At this point the dubious reader may wonder at our counting calculations because complex numbers require two real numbers for their specification, and we appear to be working over the complex field. Should, therefore, all our counts be doubled? The answer is *no* because in our case there are implicit built-in constraints. Let $*$ denote complex conjugation. Then, for the vectors \mathbf{e}_m , there are the conjugation relations

$$(\mathbf{e}_m)^* = (-1)^m \mathbf{e}_{-m}. \quad (\text{U.3.62})$$

And for the functions Y_ℓ^m there are the conjugation relations

$$[Y_\ell^m(\theta, \phi)]^* = (-1)^m Y_l^{-m}(\theta, \phi). \quad (\text{U.3.63})$$

Finally, since the Clebsch-Gordan coefficients are real and satisfy the relation

$$C_{m_1 m_2 m_3}^{j_1 j_2 j_3} = (-1)^{j_1 + j_2 - j_3} C_{-m_1, -m_2, -m_3}^{j_1 j_2 j_3}, \quad (\text{U.3.64})$$

it follows from (3.18) and (3.62) through (3.64) that the $S_{n\ell J}^M(\mathbf{r})$ satisfy the conjugation relations

$$[S_{n\ell J}^M(\mathbf{r})]^* = (-1)^{\ell+J-M+1} S_{n\ell J}^{-M}(\mathbf{r}). \quad (\text{U.3.65})$$

U.4 Independence/Orthogonality/Integral Properties of Polynomials and Polynomial Vector Fields

U.4.1 Polynomial Results

The spherical polynomials $S_{n\ell}^m(\mathbf{r})$ and $S_{n'\ell'}^{m'}(\mathbf{r})$ are evidently linearly independent if their degrees n and n' differ. What can be said if $n = n'$? The spherical harmonics have the integral properties

$$\int d\Omega [Y_\ell^m(\theta, \phi)]^* Y_{\ell'}^{m'}(\theta, \phi) = \delta_{\ell\ell'} \delta_{mm'}. \quad (\text{U.4.1})$$

Here

$$\int d\Omega = \int_0^\pi \sin(\theta) d\theta \int_0^{2\pi} d\phi. \quad (\text{U.4.2})$$

From (2.9), (2.10), and (4.1) we conclude that the various $S_{n\ell}^m(\mathbf{r})$ are all linearly independent.

We know that the $S_{n\ell}^m(\mathbf{r})$ form a basis for the polynomials. It is convenient to introduce an inner product $\langle \cdot, \cdot \rangle$ among them. Since they are linearly independent if their degrees differ, we will *define* the inner product of two polynomials of different degrees to be zero. Moreover, we will go further to *stipulate* that the $S_{n\ell}^m(\mathbf{r})$ form an *orthonormal* basis,

$$\langle S_{n\ell}^m(\mathbf{r}), S_{n'\ell'}^{m'}(\mathbf{r}) \rangle = \delta_{nn'} \delta_{\ell\ell'} \delta_{mm'}. \quad (\text{U.4.3})$$

Note that once an inner product has been defined for a set of basis vectors, it is also defined, by linearity, for all vectors.

Let $A(\mathbf{r})$ and $\hat{A}(\mathbf{r})$ be two polynomials of at most degree n_{\max} . Since the $S_{n\ell}^m(\mathbf{r})$ form a basis, there will be expansions of the form

$$A(\mathbf{r}) = \sum_{n=0}^{n_{\max}} \sum_{m\ell} a_{nm\ell} S_{n\ell}^m(\mathbf{r}), \quad (\text{U.4.4})$$

$$\hat{A}(\mathbf{r}) = \sum_{n'=0}^{n_{\max}} \sum_{m'\ell'} \hat{a}_{n'm'\ell'} S_{n'\ell'}^{m'}(\mathbf{r}). \quad (\text{U.4.5})$$

Using these expansions, for the inner product $\langle A(\mathbf{r}), \hat{A}(\mathbf{r}) \rangle$ there will be the relation

$$\begin{aligned}\langle A(\mathbf{r}), \hat{A}(\mathbf{r}) \rangle &= \sum_{n=0}^{n_{\max}} \sum_{m\ell} \sum_{n'=0}^{n_{\max}} \sum_{m'\ell'} a_{nm\ell}^* \hat{a}_{n'm'\ell'} \langle S_{n\ell}^m(\mathbf{r}), S_{n'\ell'}^{m'} \rangle \\ &= \sum_{n=0}^{n_{\max}} \sum_{m\ell} \sum_{n'=0}^{n_{\max}} \sum_{m'\ell'} a_{nm\ell}^* \hat{a}_{n'm'\ell'} \delta_{nn'} \delta_{\ell\ell'} \delta_{mm'} \\ &= \sum_{n=0}^{n_{\max}} \sum_{m\ell} a_{nm\ell}^* \hat{a}_{nm\ell},\end{aligned}\tag{U.4.6}$$

as expected. And for the norm of A there will be the result

$$\|A(\mathbf{r})\|^2 = \langle A(\mathbf{r}), A(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \sum_{m\ell} |a_{nm\ell}|^2,\tag{U.4.7}$$

as also expected.

Given $A(\mathbf{r})$ and $\hat{A}(\mathbf{r})$ as polynomials, the coefficients a and \hat{a} in (4.4) and (4.5) can be determined by matching coefficients of monomials, and then inner products and norms can be found using (4.6) and (4.7). But is there a way of computing inner products and norms without first finding the coefficients a and \hat{a} ? There is, by exploiting the integral relations (4.1).

Suppose, for example, that $A(\mathbf{r})$ is decomposed into homogeneous polynomials of degree n ,

$$A(\mathbf{r}) = \sum_{n=0}^{n_{\max}} A^n(\mathbf{r}).\tag{U.4.8}$$

This is easily done since $A(\mathbf{r})$ is usually presented as a sum of monomials, and the degree of a monomial is easily ascertained. In terms of components a and the $S_{n\ell}^m(\mathbf{r})$, $A^n(\mathbf{r})$ is given by the relation

$$A^n(\mathbf{r}) = \sum_{m\ell} a_{nm\ell} S_{n\ell}^m(\mathbf{r}).\tag{U.4.9}$$

Let us use decompositions of the form (4.8) to compute the inner product $\langle A(\mathbf{r}), \hat{A}(\mathbf{r}) \rangle$. So doing gives the result

$$\langle A(\mathbf{r}), \hat{A}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \sum_{n'=0}^{n_{\max}} \langle A^n(\mathbf{r}), \hat{A}^{n'}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \langle A^n(\mathbf{r}), \hat{A}^n(\mathbf{r}) \rangle.\tag{U.4.10}$$

Here we have used the fact that polynomials of different degrees are defined to be orthogonal,

$$\langle A^n(\mathbf{r}), \hat{A}^{n'}(\mathbf{r}) \rangle = 0 \text{ for } n \neq n'.\tag{U.4.11}$$

Next use representations of the form (4.9) to write

$$\begin{aligned}\langle A^n(\mathbf{r}), \hat{A}^n(\mathbf{r}) \rangle &= \sum_{m\ell} \sum_{m'\ell'} a_{nm\ell}^* \hat{a}_{n'm'\ell'} \langle S_{n\ell}^m(\mathbf{r}), S_{n'\ell'}^{m'}(\mathbf{r}) \rangle \\ &= \sum_{m\ell} \sum_{m'\ell'} a_{nm\ell}^* \hat{a}_{n'm'\ell'} \delta_{nn'} \delta_{\ell\ell'} \delta_{mm'} \\ &= \sum_{m\ell} a_{nm\ell}^* \hat{a}_{nm\ell}.\end{aligned}\tag{U.4.12}$$

[Note that, if we employ (4.12) in the far right side of (4.10), we recover (4.6), as expected.]

Now consider the quantity $(1/r^{2n}) \int d\Omega [A^n(\mathbf{r})]^* \hat{A}^n(\mathbf{r})$. Let us evaluate it using representations of the form (4.9). So doing gives the result

$$\begin{aligned}
 (1/r^{2n}) \int d\Omega [A^n(\mathbf{r})]^* \hat{A}^n(\mathbf{r}) &= \\
 (1/r^{2n}) \int d\Omega \left[\sum_{m\ell} a_{nm\ell} S_{n\ell}^m(\mathbf{r}) \right]^* \sum_{m'\ell'} \hat{a}_{nm'\ell'} S_{n\ell'}^{m'}(\mathbf{r}) &= \\
 \int d\Omega \left[\sum_{m\ell} a_{nm\ell} Y_\ell^m(\mathbf{r}) \right]^* \sum_{m'\ell'} \hat{a}_{nm'\ell'} Y_{\ell'}^{m'}(\mathbf{r}) &= \\
 \sum_{m\ell} \sum_{m'\ell'} a_{nm\ell}^* \hat{a}_{nm'\ell'} \delta_{\ell\ell'} \delta_{mm'} &= \sum_{m\ell} a_{nm\ell}^* \hat{a}_{nm\ell} = \langle A^n(\mathbf{r}), \hat{A}^n(\mathbf{r}) \rangle. \quad (\text{U.4.13})
 \end{aligned}$$

Here we have used (2.11), (2.12), and (4.1).

Comparison of the far left and far right sides of (4.13) shows that we have been able to convert the inner product of two homogeneous polynomials of the same degree into an integral involving the polynomials. Finally, combine (4.10) and (4.13) to achieve the grand result

$$\langle A(\mathbf{r}), \hat{A}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} (1/r^{2n}) \int d\Omega [A^n(\mathbf{r})]^* \hat{A}^n(\mathbf{r}). \quad (\text{U.4.14})$$

We have been able to compute inner products of polynomials in terms of integrals involving homogeneous parts of the polynomials. And, using (4.14), we find as a special case the result

$$\|A(\mathbf{r})\|^2 = \langle A(\mathbf{r}), A(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} (1/r^{2n}) \int d\Omega [A^n(\mathbf{r})]^* A^n(\mathbf{r}). \quad (\text{U.4.15})$$

We have been able to compute polynomial norms in terms of integrals involving the homogeneous parts of the polynomial.

U.4.2 Polynomial Vector Field Results

The previous subsection showed how to compute inner products and norms of polynomials in terms of certain integrals. The purpose of this subsection is to provide analogous results for polynomial vector fields. Our presentation will parallel the case of polynomials.

The spherical polynomial vector fields $\mathbf{S}_{n\ell J}^M(\mathbf{r})$ and $\mathbf{S}_{n'\ell' J'}^{M'}(\mathbf{r})$ are evidently linearly independent if their degrees n and n' differ. What can be said if $n = n'$? It can be shown that the vector spherical harmonics have the integral properties

$$\int d\Omega [\mathbf{Y}_{\ell J}^M(\theta, \phi)]^* \cdot \mathbf{Y}_{\ell' J'}^{M'}(\theta, \phi) = \delta_{\ell\ell'} \delta_{JJ'} \delta_{MM'}. \quad (\text{U.4.16})$$

It follows from (3.18) and (4.16) that the various $\mathbf{S}_{n\ell J}^M(\mathbf{r})$ are all linearly independent. Indeed, they form a basis for the space of polynomial vector fields

It is useful to introduce an inner product $\langle \cdot, \cdot \rangle$ among them.⁴ Since they are linearly independent if their degrees differ, we will *define* the inner product of two polynomial vector fields of different degrees to be zero. We will also make the further definitions

$$\langle \mathbf{S}_{n\ell J}^M(\mathbf{r}), \mathbf{S}_{n'\ell'J'}^{M'}(\mathbf{r}) \rangle = \delta_{nn'}\delta_{\ell\ell'}\delta_{JJ'}\delta_{MM'} \quad (\text{U.4.17})$$

which, as we will see, enables us to exploit the relation (4.16). We again note that once an inner product has been defined for a set of basis vectors, it is also defined, by linearity, for all vectors. We will now work out some of the consequences of the definitions we have just made.

Let $\mathbf{A}(\mathbf{r})$ and $\hat{\mathbf{A}}(\mathbf{r})$ be two polynomial vector fields of at most degree n_{\max} . Since the $\mathbf{S}_{n\ell J}^M(\mathbf{r})$ form a basis for polynomial vector fields, there will be expansions of the form

$$\mathbf{A}(\mathbf{r}) = \sum_{n=0}^{n_{\max}} \sum_{\ell JM} a_{n\ell JM} \mathbf{S}_{n\ell J}^M(\mathbf{r}), \quad (\text{U.4.18})$$

$$\hat{\mathbf{A}}(\mathbf{r}) = \sum_{n'=0}^{n_{\max}} \sum_{\ell' J' M'} \hat{a}_{n'\ell' J' M'} \mathbf{S}_{n'\ell' J'}^{M'}(\mathbf{r}). \quad (\text{U.4.19})$$

Using these expansions, for the inner product $\langle \mathbf{A}(\mathbf{r}), \hat{\mathbf{A}}(\mathbf{r}) \rangle$ there will be the relation

$$\begin{aligned} \langle \mathbf{A}(\mathbf{r}), \hat{\mathbf{A}}(\mathbf{r}) \rangle &= \sum_{n=0}^{n_{\max}} \sum_{\ell JM} \sum_{n'=0}^{n_{\max}} \sum_{\ell' J' M'} a_{n\ell JM}^* \hat{a}_{n'\ell' J' M'} \langle \mathbf{S}_{n\ell J}^M(\mathbf{r}), \mathbf{S}_{n'\ell' J'}^{M'}(\mathbf{r}) \rangle \\ &= \sum_{n=0}^{n_{\max}} \sum_{\ell JM} \sum_{n'=0}^{n_{\max}} \sum_{\ell' J' M'} a_{n\ell JM}^* \hat{a}_{n'\ell' J' M'} \delta_{nn'} \delta_{\ell\ell'} \delta_{JJ'} \delta_{MM'} \\ &= \sum_{n=0}^{n_{\max}} \sum_{\ell JM} a_{n\ell JM}^* \hat{a}_{n\ell JM}, \end{aligned} \quad (\text{U.4.20})$$

as expected. And for the norm of $\mathbf{A}(\mathbf{r})$ there will be the result

$$\|\mathbf{A}(\mathbf{r})\|^2 = \langle \mathbf{A}(\mathbf{r}), \mathbf{A}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \sum_{\ell JM} |a_{n\ell JM}|^2, \quad (\text{U.4.21})$$

as also expected. Note that the norm $\|\mathbf{A}(\mathbf{r})\|$ vanishes iff $\mathbf{A}(\mathbf{r}) = 0$,

$$\|\mathbf{A}(\mathbf{r})\| = 0 \Leftrightarrow \mathbf{A}(\mathbf{r}) = 0. \quad (\text{U.4.22})$$

Given $A(\mathbf{r})$ and $\hat{A}(\mathbf{r})$ as polynomial vector fields, the coefficients a and \hat{a} in (4.18) and (4.19) can be determined by matching coefficients of like terms, these terms being unit vectors $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ multiplied by monomials in x, y, z . Once the a and \hat{a} have been determined, inner products and norms can then be found using (4.20) and (4.21). But is there a way of

⁴Although they are not the same, we have used the angular bracket notation $\langle \cdot, \cdot \rangle$ to denote the inner product in both the vector space of polynomials and the vector space of polynomial vector fields. What is meant when $\langle \cdot, \cdot \rangle$ is employed should be clear from the context.

computing inner products and norms without first finding the coefficients a and \hat{a} ? There is, by exploiting the integral relations (4.16).

Suppose, for example, that $\mathbf{A}(\mathbf{r})$ is decomposed into homogeneous polynomials of degree n ,

$$\mathbf{A}(\mathbf{r}) = \sum_{n=0}^{n_{\max}} \mathbf{A}^n(\mathbf{r}). \quad (\text{U.4.23})$$

This is easily done since $\mathbf{A}(\mathbf{r})$ is usually presented as a sum of monomials multiplying unit vectors, and the degree of a monomial is easily ascertained. In terms of components a and the $S_{n\ell J}^M(\mathbf{r})$, $\mathbf{A}^n(\mathbf{r})$ is given by the relation

$$\mathbf{A}^n(\mathbf{r}) = \sum_{\ell JM} a_{n\ell JM} S_{n\ell J}^M(\mathbf{r}). \quad (\text{U.4.24})$$

Let us use decompositions of the form (4.23) to compute the inner product $\langle \mathbf{A}(\mathbf{r}), \hat{\mathbf{A}}(\mathbf{r}) \rangle$. So doing gives the result

$$\langle \mathbf{A}(\mathbf{r}), \hat{\mathbf{A}}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \sum_{n'=0}^{n_{\max}} \langle \mathbf{A}^n(\mathbf{r}), \hat{\mathbf{A}}^{n'}(\mathbf{r}) \rangle = \sum_{n=0}^{n_{\max}} \langle \mathbf{A}^n(\mathbf{r}), \hat{\mathbf{A}}^n(\mathbf{r}) \rangle. \quad (\text{U.4.25})$$

Here we have used the fact that polynomials of different degrees are defined to be orthogonal,

$$\langle \mathbf{A}^n(\mathbf{r}), \hat{\mathbf{A}}^{n'}(\mathbf{r}) \rangle = 0 \text{ for } n \neq n'. \quad (\text{U.4.26})$$

Next use representations of the form (4.24) to write

$$\begin{aligned} \langle \mathbf{A}^n(\mathbf{r}), \hat{\mathbf{A}}^n(\mathbf{r}) \rangle &= \sum_{\ell JM} \sum_{\ell' J' M'} a_{n\ell JM}^* \hat{a}_{n\ell' J' M'} \langle S_{n\ell J}^M(\mathbf{r}), S_{n\ell' J'}^{M'}(\mathbf{r}) \rangle \\ &= \sum_{\ell JM} \sum_{\ell' J' M'} a_{n\ell JM}^* \hat{a}_{n\ell' J' M'} \delta_{\ell\ell'} \delta_{JJ'} \delta_{MM'} \\ &= \sum_{\ell JM} a_{n\ell JM}^* \hat{a}_{n\ell JM}. \end{aligned} \quad (\text{U.4.27})$$

[Note that, if we employ (4.27) in the far right side of (4.25), we recover (4.20), as expected.]

Now consider the quantity $(1/r^{2n}) \int d\Omega [\mathbf{A}^n(\mathbf{r})]^* \cdot \hat{\mathbf{A}}^n(\mathbf{r})$. Let us evaluate it using representations of the form (4.24). So doing gives the result

$$\begin{aligned} (1/r^{2n}) \int d\Omega [\mathbf{A}^n(\mathbf{r})]^* \cdot \hat{\mathbf{A}}^n(\mathbf{r}) &= \\ (1/r^{2n}) \int d\Omega \left[\sum_{\ell JM} a_{n\ell JM} S_{n\ell J}^M(\mathbf{r}) \right]^* \cdot \sum_{\ell' J' M'} \hat{a}_{n\ell' J' M'} S_{n\ell' J'}^{M'}(\mathbf{r}) &= \\ \int d\Omega \left[\sum_{\ell JM} a_{n\ell JM} \mathbf{Y}_{\ell J}^M(\mathbf{r}) \right]^* \cdot \sum_{\ell' J' M'} \hat{a}_{n\ell' J' M'} \mathbf{Y}_{\ell' J'}^{M'}(\mathbf{r}) &= \\ \sum_{\ell JM} \sum_{\ell' J' M'} a_{n\ell JM}^* \hat{a}_{n\ell' J' M'} \delta_{\ell\ell'} \delta_{JJ'} \delta_{MM'} &= \sum_{\ell JM} a_{n\ell JM}^* \hat{a}_{n\ell JM} = \langle \mathbf{A}^n(\mathbf{r}), \hat{\mathbf{A}}^n(\mathbf{r}) \rangle. \end{aligned} \quad (\text{U.4.28})$$

Here we have used (3.18), (4.16), and (4.27).

Comparison of the far left and far right sides of (4.28) shows that we have been able to convert the inner product of two homogeneous polynomial vector fields of the same degree into an integral involving the polynomial vector fields. Finally, combine (4.25) and (4.28) to achieve the grand result

$$< \mathbf{A}(\mathbf{r}), \hat{\mathbf{A}}(\mathbf{r}) > = \sum_{n=0}^{n_{\max}} (1/r^{2n}) \int d\Omega [\mathbf{A}^n(\mathbf{r})]^* \cdot \hat{\mathbf{A}}^n(\mathbf{r}). \quad (\text{U.4.29})$$

We have been able to compute inner products of polynomial vector fields in terms of integrals involving homogeneous parts of the polynomial vector fields. And, using (4.29), we find as a special case the result

$$\|\mathbf{A}(\mathbf{r})\|^2 = < \mathbf{A}(\mathbf{r}), \mathbf{A}(\mathbf{r}) > = \sum_{n=0}^{n_{\max}} (1/r^{2n}) \int d\Omega [\mathbf{A}^n(\mathbf{r})]^* \cdot \mathbf{A}^n(\mathbf{r}). \quad (\text{U.4.30})$$

We have been able to compute polynomial vector field norms in terms of integrals involving the homogeneous parts of the polynomial vector fields.

U.5 Differential Properties of Spherical Polynomials and Spherical Polynomial Vector Fields

The purpose of this section is to list various effects of the differential operator ∇ when acting on spherical polynomials and spherical polynomial vector fields.

U.5.1 Gradient Action on Spherical Polynomials

We begin with the action of ∇ on spherical polynomials. Suppose $f(r)$ is any function of r , and suppose $\ell \geq 1$. Then it can be shown that

$$\begin{aligned} \nabla[f(r)Y_\ell^m(\theta, \phi)] &= \sqrt{\ell/(2\ell+1)}\{f'(r) + [(\ell+1)/r]f(r)\}\mathbf{Y}_{\ell-1,\ell}^m(\theta, \phi) \\ &\quad - \sqrt{(\ell+1)/(2\ell+1)}\{f'(r) - (\ell/r)f(r)\}\mathbf{Y}_{\ell+1,\ell}^m(\theta, \phi). \end{aligned} \quad (\text{U.5.1})$$

For the case of the spherical polynomial functions $S_{n\ell}^m(\mathbf{r})$ we have

$$f(r) = r^n. \quad (\text{U.5.2})$$

See (2.10) and (2.11). It follows (again supposing $\ell \geq 1$) that

$$\begin{aligned} \nabla S_{n\ell}^m(\mathbf{r}) &= \sqrt{\ell/(2\ell+1)}(n+\ell+1)\mathbf{S}_{n-1,\ell-1,\ell}^m(\mathbf{r}) \\ &\quad - \sqrt{(\ell+1)/(2\ell+1)}(n-\ell)\mathbf{S}_{n-1,\ell+1,\ell}^m(\mathbf{r}). \end{aligned} \quad (\text{U.5.3})$$

What about the special case $\ell = 0$? Then n must be even. So we write

$$n = 2k. \quad (\text{U.5.4})$$

Also we must have $m = 0$. If $\ell = 0$, we might imagine evaluating (5.3) with the first term omitted since it contains $\sqrt{\ell}$. Doing so gives the result

$$\nabla S_{2k,0}^0(\mathbf{r}) = -2k S_{2k-1,1,0}^0(\mathbf{r}). \quad (\text{U.5.5})$$

This result is, in fact, correct, and can be verified directly. See Exercise 6.11.

We close this subsection by observing that a special case of (5.3) is the relation

$$\nabla S_{nn}^m(\mathbf{r}) = \sqrt{n(2n+1)} S_{n-1,n-1,n}^m(\mathbf{r}). \quad (\text{U.5.6})$$

U.5.2 Divergence Action on Spherical Polynomial Vector Fields

We continue with the case of spherical polynomial vector fields. Suppose again that $f(r)$ is any function of r . Then (assuming $\ell \geq 1$) it can be shown that

$$\nabla \cdot [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] = \sqrt{J/(2J+1)} \{f'(r) - [(J-1)/r]f(r)\} Y_J^M(\theta, \phi) \text{ when } J = \ell+1, \quad (\text{U.5.7})$$

$$\nabla \cdot [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] = 0 \text{ when } J = \ell, \quad (\text{U.5.8})$$

$$\nabla \cdot [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] = -\sqrt{(J+1)/(2J+1)} \{f'(r) + [(J+2)/r]f(r)\} Y_J^M(\theta, \phi) \text{ when } J = \ell-1. \quad (\text{U.5.9})$$

For the case of the spherical polynomial vector fields $\mathbf{S}_{n,\ell,J}^M(\mathbf{r})$ the relation (5.2) again holds. It follows (again assuming $\ell \geq 1$) that

$$\nabla \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = \sqrt{J/(2J+1)} (n-J+1) S_{n-1,J}^M(\mathbf{r}) \text{ when } J = \ell+1, \quad (\text{U.5.10})$$

$$\nabla \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = 0 \text{ when } J = \ell, \quad (\text{U.5.11})$$

$$\nabla \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = -\sqrt{(J+1)/(2J+1)} (n+J+2) S_{n-1,J}^M(\mathbf{r}) \text{ when } J = \ell-1. \quad (\text{U.5.12})$$

What about the special case $\ell = 0$? Then we must have $J = 1$. Moreover n must be even so that (5.4) again holds. Evidently when $\ell = 0$ and $J = 1$ the conditions relating J and ℓ in (5.11) and (5.12) do not hold. However, the condition in (5.10) does hold and so we might speculate that (5.10) should be evaluated with $J = 1$ to give the result

$$\nabla \cdot \mathbf{S}_{2k,0,1}^M(\mathbf{r}) = (\sqrt{1/3}) 2k S_{2k-1,1}^M(\mathbf{r}). \quad (\text{U.5.13})$$

This speculation is correct, and can be proved directly. See Exercise 6.13.

U.5.3 Curl Action on Spherical Polynomial Vector Fields

It can also be shown (assuming $\ell \geq 1$) that

$$\begin{aligned} \nabla \times [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] &= i\sqrt{(J+1)/(2J+1)} \{f'(r) - [(J-1)/r]f(r)\} \mathbf{Y}_{J,J}^M(\theta, \phi) \\ &\quad \text{when } J = \ell+1, \end{aligned} \quad (\text{U.5.14})$$

$$\begin{aligned} \nabla \times [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] &= i\sqrt{(J+1)/(2J+1)} \{f'(r) + [(J+1)/r]f(r)\} \mathbf{Y}_{J-1,J}^M(\theta, \phi) \\ &\quad + i\sqrt{J/(2J+1)} \{f'(r) - (J/r)f(r)\} \mathbf{Y}_{J+1,J}^M(\theta, \phi) \\ &\quad \text{when } J = \ell, \end{aligned} \quad (\text{U.5.15})$$

$$\nabla \times [f(r) \mathbf{Y}_{\ell,J}^M(\theta, \phi)] = i\sqrt{J/(2J+1)} \{ f'(r) + [(J+2)/r] f(r) \} \mathbf{Y}_{J,J}^M(\theta, \phi)$$

when $J = \ell - 1$. (U.5.16)

For the case of the spherical polynomial vector fields $\mathbf{S}_{n,\ell,J}^M(\mathbf{r})$ the relation (5.2) remains true. It follows (again assuming $\ell \geq 1$) that there are the relations:

$$\begin{aligned} \nabla \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{(J+1)/(2J+1)}(n-J+1) \mathbf{S}_{n-1,J,J}^M(\mathbf{r}) \\ &\text{when } J = \ell + 1. \text{ Equivalently, we have} \\ \nabla \times \mathbf{S}_{n,\ell,\ell+1}^M(\mathbf{r}) &= i\sqrt{(\ell+2)/(2\ell+3)}(n-\ell) \mathbf{S}_{n-1,\ell+1,\ell+1}^M(\mathbf{r}), \end{aligned} \quad (\text{U.5.17})$$

$$\begin{aligned} \nabla \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{(J+1)/(2J+1)}(n+J+1) \mathbf{S}_{n-1,J-1,J}^M(\mathbf{r}) \\ &\quad + i\sqrt{J/(2J+1)}(n-J) \mathbf{S}_{n-1,J+1,J}^M(\mathbf{r}) \\ &\text{when } J = \ell. \text{ Equivalently, we have} \\ \nabla \times \mathbf{S}_{n,\ell,\ell}^M(\mathbf{r}) &= i\sqrt{(\ell+1)/(2\ell+1)}(n+\ell+1) \mathbf{S}_{n-1,\ell-1,\ell}^M(\mathbf{r}) \\ &\quad + i\sqrt{\ell/(2\ell+1)}(n-\ell) \mathbf{S}_{n-1,\ell+1,\ell}^M(\mathbf{r}), \end{aligned} \quad (\text{U.5.18})$$

$$\begin{aligned} \nabla \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{J/(2J+1)}(n+J+2) \mathbf{S}_{n-1,J,J}^M(\mathbf{r}) \\ &\text{when } J = \ell - 1. \text{ Equivalently, we have} \\ \nabla \times \mathbf{S}_{n,\ell,\ell-1}^M(\mathbf{r}) &= i\sqrt{(\ell-1)/(2\ell-1)}(n+\ell+1) \mathbf{S}_{n-1,\ell-1,\ell-1}^M(\mathbf{r}). \end{aligned} \quad (\text{U.5.19})$$

Again we must consider the special case $\ell = 0$, in which case $J = 1$ and (5.4) holds. The conditions relating J and ℓ associated with (5.18) and (5.19) do not hold in this case, but the one associated with (5.17) does hold. We therefore speculate that (5.17) should be employed in the case $\ell = 0$ and $J = 1$ to give the result

$$\nabla \times \mathbf{S}_{2k,0,1}^M(\mathbf{r}) = i(\sqrt{2/3})(2k) \mathbf{S}_{2k-1,1,1}^M(\mathbf{r}). \quad (\text{U.5.20})$$

This speculation is also correct, and can be proved directly. See Exercise 6.17.

Note that in all cases there is the pleasant fact that the $\nabla \times$ operator *preserves* the top index and the last bottom index, the M and J indices, on $\mathbf{S}_{n,\ell,J}^M$. It can be shown that there are *total* angular momentum operators \mathcal{J}_1 , \mathcal{J}_2 , and \mathcal{J}_3 , and this preservation is a consequence of the fact that the operator $\nabla \times$ commutes with the total angular momentum operators.

We close this subsection by observing that a special case of (5.18) is the relation

$$\nabla \times \mathbf{S}_{n,n,n}^M(\mathbf{r}) = i\sqrt{(n+1)(2n+1)} \mathbf{S}_{n-1,n-1,n}^M(\mathbf{r}). \quad (\text{U.5.21})$$

U.6 Multiplicative Properties of Spherical Polynomials and Spherical Polynomial Vector Fields

This section deals with the effects of multiplication by \mathbf{r} .

U.6.1 Ordinary Multiplication

We begin with the case of spherical polynomials and consider ordinary multiplication. Suppose $\ell \geq 1$. Then it can be shown that

$$\mathbf{r}Y_\ell^m(\theta, \phi) = \sqrt{\ell/(2\ell+1)} \mathbf{r}\mathbf{Y}_{\ell-1,\ell}^m(\theta, \phi) - \sqrt{(\ell+1)/(2\ell+1)} \mathbf{r}\mathbf{Y}_{\ell+1,\ell}^m(\theta, \phi). \quad (\text{U.6.1})$$

In view of (3.18), it follows (again supposing $\ell \geq 1$) that

$$\mathbf{r}S_{n\ell}^m(\mathbf{r}) = \sqrt{\ell/(2\ell+1)} \mathbf{S}_{n+1,\ell-1,\ell}^m(\mathbf{r}) - \sqrt{(\ell+1)/(2\ell+1)} \mathbf{S}_{n+1,\ell+1,\ell}^m(\mathbf{r}). \quad (\text{U.6.2})$$

What about the special case $\ell = 0$? Then n must be even. So we write

$$n = 2k. \quad (\text{U.6.3})$$

Also we must have $m = 0$. If $\ell = 0$, we might imagine evaluating (6.1) with the first term omitted since it contains $\sqrt{\ell}$. Doing so gives the result

$$\mathbf{r}S_{2k,0}^0(\mathbf{r}) = -\mathbf{S}_{2k+1,1,0}^0(\mathbf{r}). \quad (\text{U.6.4})$$

This result is, in fact, correct, and can be verified directly. See Exercise 6.23.

U.6.2 Dot Product Multiplication

We continue with the case of spherical polynomial vector fields, and consider the case of dot product multiplication. Assume that $\ell \geq 1$. Then it can be shown that

$$\mathbf{r} \cdot \mathbf{Y}_{\ell,J}^M(\theta, \phi) = \sqrt{J/(2J+1)} \mathbf{r}Y_J^M(\theta, \phi) \text{ when } J = \ell + 1, \quad (\text{U.6.5})$$

$$\mathbf{r} \cdot \mathbf{Y}_{\ell,J}^M(\theta, \phi) = 0 \text{ when } J = \ell, \quad (\text{U.6.6})$$

$$\mathbf{r} \cdot \mathbf{Y}_{\ell,J}^M(\theta, \phi) = -\sqrt{(J+1)/(2J+1)} \mathbf{r}Y_J^M(\theta, \phi) \text{ when } J = \ell - 1. \quad (\text{U.6.7})$$

It follows from (3.18), again assuming $\ell \geq 1$, that

$$\mathbf{r} \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = \sqrt{J/(2J+1)} \mathbf{S}_{n+1,J}^M(\mathbf{r}) \text{ when } J = \ell + 1, \quad (\text{U.6.8})$$

$$\mathbf{r} \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = 0 \text{ when } J = \ell, \quad (\text{U.6.9})$$

$$\mathbf{r} \cdot \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) = -\sqrt{(J+1)/(2J+1)} \mathbf{S}_{n+1,J}^M(\mathbf{r}) \text{ when } J = \ell - 1. \quad (\text{U.6.10})$$

What about the special case $\ell = 0$? Then we must have $J = 1$. Moreover n must be even so that (5.4) again holds. Evidently when $\ell = 0$ and $J = 1$ the conditions relating J and ℓ in (6.9) and (6.10) do not hold. However, the condition in (6.8) does hold and so we might speculate that (6.8) should be evaluated with $J = 1$ to give the result

$$\mathbf{r} \cdot \mathbf{S}_{2k,0,1}^M(\mathbf{r}) = \sqrt{1/3} \mathbf{S}_{2k+1,1}^M(\mathbf{r}). \quad (\text{U.6.11})$$

This speculation is correct, and can be proved directly. See Exercise 6.25.

U.6.3 Cross Product Multiplication

Lastly, we consider the case of cross product multiplication of spherical polynomial vector fields. It can also be shown (assuming $\ell \geq 1$) that

$$\begin{aligned} \mathbf{r} \times \mathbf{Y}_{\ell,J}^M(\theta, \phi) &= i\sqrt{(J+1)/(2J+1)} r \mathbf{Y}_{J,J}^M(\theta, \phi) \\ &\quad \text{when } J = \ell + 1, \end{aligned} \quad (\text{U.6.12})$$

$$\begin{aligned} \mathbf{r} \times \mathbf{Y}_{\ell,J}^M(\theta, \phi) &= i\sqrt{(J+1)/(2J+1)} r \mathbf{Y}_{J-1,J}^M(\theta, \phi) \\ &\quad + i\sqrt{J/(2J+1)} r \mathbf{Y}_{J+1,J}^M(\theta, \phi) \\ &\quad \text{when } J = \ell, \end{aligned} \quad (\text{U.6.13})$$

$$\begin{aligned} \mathbf{r} \times [\mathbf{Y}_{\ell,J}^M(\theta, \phi)] &= i\sqrt{J/(2J+1)} r \mathbf{Y}_{J,J}^M(\theta, \phi) \\ &\quad \text{when } J = \ell - 1. \end{aligned} \quad (\text{U.6.14})$$

It follows from (3.18), again assuming $\ell \geq 1$, that there are the following results:

$$\begin{aligned} \mathbf{r} \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{(J+1)/(2J+1)} \mathbf{S}_{n+1,J,J}^M(\mathbf{r}) \\ &\quad \text{when } J = \ell + 1. \text{ Equivalently, we have} \\ \mathbf{r} \times \mathbf{S}_{n,\ell,\ell+1}^M(\mathbf{r}) &= i\sqrt{(\ell+2)/(2\ell+3)} \mathbf{S}_{n+1,\ell+1,\ell+1}^M(\mathbf{r}), \end{aligned} \quad (\text{U.6.15})$$

$$\begin{aligned} \mathbf{r} \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{(J+1)/(2J+1)} \mathbf{S}_{n+1,J-1,J}^M(\mathbf{r}) \\ &\quad + i\sqrt{J/(2J+1)} \mathbf{S}_{n+1,J+1,J}^M(\mathbf{r}) \\ &\quad \text{when } J = \ell. \text{ Equivalently, we have} \\ \mathbf{r} \times \mathbf{S}_{n,\ell,\ell}^M(\mathbf{r}) &= i\sqrt{(\ell+1)/(2\ell+1)} \mathbf{S}_{n+1,\ell-1,\ell}^M(\mathbf{r}) \\ &\quad + i\sqrt{\ell/(2\ell+1)} \mathbf{S}_{n+1,\ell+1,\ell}^M(\mathbf{r}), \end{aligned} \quad (\text{U.6.16})$$

$$\begin{aligned} \mathbf{r} \times \mathbf{S}_{n,\ell,J}^M(\mathbf{r}) &= i\sqrt{J/(2J+1)} \mathbf{S}_{n+1,J,J}^M(\mathbf{r}) \\ &\quad \text{when } J = \ell - 1. \text{ Equivalently, we have} \\ \mathbf{r} \times \mathbf{S}_{n,\ell,\ell-1}^M(\mathbf{r}) &= i\sqrt{(\ell-1)/(2\ell-1)} \mathbf{S}_{n+1,\ell-1,\ell-1}^M(\mathbf{r}). \end{aligned} \quad (\text{U.6.17})$$

Again we must consider the special case $\ell = 0$, in which case $J = 1$ and (5.4) holds. The conditions relating J and ℓ associated with (6.16) and (6.17) do not hold in this case, but the one associated with (6.15) does hold. We therefore speculate that (6.15) should be employed in the case $\ell = 0$ and $J = 1$ to give the result

$$\mathbf{r} \times \mathbf{S}_{2k,0,1}^M(\mathbf{r}) = i(\sqrt{2/3}) \mathbf{S}_{2k+1,1,1}^M(\mathbf{r}). \quad (\text{U.6.18})$$

This speculation is also correct, and can be proved directly. See Exercise 6.27.

Note that in all cases there is also the pleasant fact that the $\mathbf{r} \times$ operator also preserves the top index and the last bottom index, the M and J indices, on $\mathbf{S}_{n,\ell,J}^M$.

We close this subsection by making two useful observations. The first observation is that a special case of (6.15) yields the relation

$$\mathbf{S}_{nnn}^M(\mathbf{r}) = [-i\sqrt{(2n+1)/(n+1)}][\mathbf{r} \times \mathbf{S}_{n-1,n-1,n}^M(\mathbf{r})]. \quad (\text{U.6.19})$$

To verify this claim, evaluate (6.15) for the case

$$n = n' - 1 \quad (\text{U.6.20})$$

and

$$\ell = n' - 1; \quad (\text{U.6.21})$$

from which it follows that

$$\ell + 1 = n' \quad (\text{U.6.22})$$

and

$$(\ell + 2)/(2\ell + 3) = (n' + 1)/(2n' + 1). \quad (\text{U.6.23})$$

So doing gives the result

$$\mathbf{r} \times \mathbf{S}_{n'-1,n'-1,n'}^M(\mathbf{r}) = i\sqrt{(n'+1)/(2n'+1)}\mathbf{S}_{n'n'n'}^M(\mathbf{r}), \quad (\text{U.6.24})$$

from which (6.19) follows. Note that (3.41) is special case of (6.19).

The second observation is that combining (5.6) and (6.19) gives the relation

$$\begin{aligned} \mathbf{S}_{nnn}^M(\mathbf{r}) &= [-i\sqrt{(2n+1)/(n+1)}][\mathbf{r} \times \mathbf{S}_{n-1,n-1,n}^M(\mathbf{r})] \\ &= [-i/\sqrt{n(n+1)}][\mathbf{r} \times \nabla S_{nn}^M(\mathbf{r})]. \end{aligned} \quad (\text{U.6.25})$$

Note that if we define an *orbital* angular momentum operator \mathbf{L} by the rule

$$\mathbf{L} = \mathbf{r} \times \nabla, \quad (\text{U.6.26})$$

then (6.25) can be written in the form

$$\mathbf{S}_{nnn}^M(\mathbf{r}) = [-i/\sqrt{n(n+1)}]\mathbf{L}S_{nn}^M(\mathbf{r}). \quad (\text{U.6.27})$$

Exercises

U.6.1. Cognizant of the relation between ℓ and n given by (2.10) and (2.11) and the rule $-\ell \leq m \leq \ell$, count how many $S_{n\ell}^m$ there are for a given n . Show the result agrees with $N(n, 3)$.

U.6.2. Verify (3.3) through (3.5).

U.6.3. Verify (3.19) and (3.20).

U.6.4. Verify (3.21) through (3.25).

U.6.5. Verify (3.26) through (3.41).

U.6.6. Verify (3.42) through (3.61).

U.6.7. Verify (3.65) given (3.18), (3.62), (3.63), and (3.65). Verify (3.65) directly for the cases \mathbf{S}_{001}^M , \mathbf{S}_{110}^0 , \mathbf{S}_{111}^M , and \mathbf{S}_{112}^M worked out explicitly in Subsection 3.3.

U.6.8. Recall the relation between n , ℓ , and J given by (2.10), (2.11), (3.3), and (3.4). See Table 3.1. Recall also the rule $-J \leq M \leq J$. Count how many $\mathbf{S}_{n\ell J}^M$ there are for a given n . Show the result agrees with $3N(n, 3)$.

U.6.9. Show that

$$\int d\Omega S_{n\ell}^m = \sqrt{4\pi} r^n \delta_{\ell 0} \delta_{m0}. \quad (\text{U.6.28})$$

Recall Exercise 16.1.1.

U.6.10. Given (5.1) and (5.2), derive (5.3). Verify (5.6).

U.6.11. Show from the definition (2.10) that

$$S_{2k,0}^0(\mathbf{r}) = (1/\sqrt{4\pi})(x^2 + y^2 + z^2)^k. \quad (\text{U.6.29})$$

Show from the definition (3.18) and the result (3.25) that

$$\mathbf{S}_{2k-1,1,0}^0(\mathbf{r}) = (-1/\sqrt{4\pi})(x^2 + y^2 + z^2)^{k-1}\mathbf{r} = (-1/\sqrt{4\pi}) r^{2k-2}\mathbf{r}. \quad (\text{U.6.30})$$

Verify (5.5) by direct computation.

U.6.12. Given (5.2) and (5.8) through (5.9), derive (5.10) through (5.12).

U.6.13. Show from the definition (2.11) that

$$\mathbf{S}_{2k-1,1}^M(\mathbf{r}) = r^{2k-2} S_{11}^M(\mathbf{r}). \quad (\text{U.6.31})$$

Show from the definition (3.18) and the relation (3.20) that

$$\mathbf{S}_{2k,0,1}^M(\mathbf{r}) = r^{2k} \mathbf{S}_{001}^M(\mathbf{r}) = (1/\sqrt{4\pi}) r^{2k} \mathbf{e}_M. \quad (\text{U.6.32})$$

Verify (5.13) by direct computation. Hint: Use (3.3).

U.6.14. Verify that $S_{nn}^m(\mathbf{r})$ is a harmonic polynomial. Verify, using the rules (5.3) and (5.5) and (5.9) through (5.12), that

$$\nabla \cdot [\nabla S_{nn}^m(\mathbf{r})] = 0, \quad (\text{U.6.33})$$

as expected.

U.6.15. Show that

$$\nabla^2 S_{n\ell}^m = [n(n+1) - \ell(\ell+1)] S_{n-2,\ell}^m. \quad (\text{U.6.34})$$

U.6.16. Given (5.2) and (5.14) through (5.16), derive (5.17) through (5.19). Verify (5.21).

U.6.17. Show from the definition (3.18) and the relation (3.40) that

$$\mathbf{S}_{2k-1,1,1}^M(\mathbf{r}) = -i\sqrt{3/(4\pi)} r^{2k-2} \mathbf{r} \times \mathbf{e}_M. \quad (\text{U.6.35})$$

Review Exercise 6.13. Using the results (6.32) and (6.35) for $\mathbf{S}_{2k,0,1}^M$ and $\mathbf{S}_{2k-1,1,1}^M$, verify (5.20) by direct computation.

U.6.18. Using the rules (5.3) and (5.5) and (5.17) through (5.20), verify that

$$\nabla \times [\nabla S_{n\ell}^m(\mathbf{r})] = 0, \quad (\text{U.6.36})$$

as expected.

U.6.19. Using the rules (5.10) through (5.13) and (5.17) through (5.20), verify that

$$\nabla \cdot [\nabla \times \mathbf{S}_{n\ell J}^M(\mathbf{r})] = 0, \quad (\text{U.6.37})$$

as expected.

U.6.20. Verify the relations

$$\nabla \times \mathbf{S}_{110}^0(\mathbf{r}) = 0, \quad (\text{U.6.38})$$

$$\nabla \times \mathbf{S}_{111}^M(\mathbf{r}) = i\sqrt{6} \mathbf{S}_{001}^M(\mathbf{r}), \quad (\text{U.6.39})$$

$$\nabla \times \mathbf{S}_{112}^M(\mathbf{r}) = 0. \quad (\text{U.6.40})$$

U.6.21. Show that

$$\nabla \times \mathbf{S}_{201}^M(\mathbf{r}) = i\sqrt{8/3} \mathbf{S}_{111}^M(\mathbf{r}), \quad (\text{U.6.41})$$

$$\nabla \times \mathbf{S}_{223}^M(\mathbf{r}) = 0, \quad (\text{U.6.42})$$

$$\nabla \times \mathbf{S}_{222}^M(\mathbf{r}) = i\sqrt{15} \mathbf{S}_{112}^M(\mathbf{r}), \quad (\text{U.6.43})$$

$$\nabla \times \mathbf{S}_{221}^M(\mathbf{r}) = i\sqrt{25/3} \mathbf{S}_{111}^M(\mathbf{r}). \quad (\text{U.6.44})$$

Show that

$$\nabla \times \nabla \times \mathbf{S}_{201}^M(\mathbf{r}) = i\sqrt{8/3} \nabla \times \mathbf{S}_{111}^M(\mathbf{r}) = -4 \mathbf{S}_{001}^M(\mathbf{r}), \quad (\text{U.6.45})$$

$$\nabla \times \nabla \times \mathbf{S}_{223}^M(\mathbf{r}) = 0, \quad (\text{U.6.46})$$

$$\nabla \times \nabla \times \mathbf{S}_{222}^M(\mathbf{r}) = i\sqrt{15} \nabla \times \mathbf{S}_{112}^M(\mathbf{r}) = 0 \quad (\text{U.6.47})$$

$$\nabla \times \nabla \times \mathbf{S}_{221}^M(\mathbf{r}) = i\sqrt{25/3} \nabla \times \mathbf{S}_{111}^M(\mathbf{r}) = -\sqrt{50} \mathbf{S}_{001}^M(\mathbf{r}). \quad (\text{U.6.48})$$

U.6.22. Given (6.1), derive (6.2).

U.6.23. Verify (6.4) using (2.10) and (3.25).

U.6.24. Given (6.5) through (6.7), derive (6.8) through (6.10).

U.6.25. Review Exercise 6.13. Verify (6.11) using (6.32), (3.3), and (2.11).

U.6.26. Given (6.12) through (6.14), derive (6.15) through (6.17).

U.6.27. Review Exercise 6.13. Verify (6.18) using (6.32), (3.40), (3.18), and (2.11).

U.6.28. Verify the steps that connect (6.14) to (6.19).

U.6.29. Verify that combining (5.6) and (6.19) yields (6.25).

Bibliography

Group Theory of Angular Momentum

- [1] M. Rose, *Elementary Theory of Angular Momentum*, John Wiley & Sons (1957).
- [2] A. Edmonds, *Angular Momentum in Quantum Mechanics*, Princeton University Press (1957).
- [3] E. Condon and G. Shortley, *The Theory of Atomic Spectra*, Cambridge University Press (1935). A corrected 1999 version is available in paperback.
- [4] E.P. Wigner, *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra*, Academic Press (1959).
- [5] H. Weyl, *The Theory of Groups and Quantum Mechanics*, Dover (1950).

Harmonic Functions and Vector Spherical Harmonics

- [6] E. Stein and G. Weiss, *Introduction to Fourier analysis on Euclidean Spaces*, Princeton University Press (1971).
- [7] M. Rose, *Multipole Fields*, John Wiley & Sons (1955).
- [8] J. Mathews, *Tensor Spherical Harmonics*, California Institute of Technology (1981).
- [9] J. Blatt and V. Weisskopf, *Theoretical Nuclear Physics*, Appendix B, John Wiley (1958) and Dover (2010).
- [10] E. Hill, “Theory of Vector Spherical Harmonics”, *Am. J. Phys.* **22**, 211 (1954).
- [11] J. D. Jackson, *Classical Electrodynamics*, John Wiley (1999).
- [12] Google Vector Spherical Harmonics. See, for example, the University of Texas (at Austin) Professor Austin Gleeson Web site
<http://www.ph.utexas.edu/~gleeson/ElectricityMagnetismAppendixE.pdf>

Appendix V

PROT without and in the Presence of a Magnetic Field

V.1 The Case of No Magnetic Field

The material to be covered in this section is standard fare with results known through at least third order, but not yet completely documented.

V.2 The Constant Magnetic Field Case

V.2.1 Preliminaries

Recall from Exercise 1.6.2 that the Hamiltonian for charged-particle motion in an electromagnetic field, when employing cylindrical coordinates with the angle ϕ as the independent variable, is given by the relation

$$K = -\rho[(p_t + q\psi)^2/c^2 - m^2c^2 - (p_\rho - qA_\rho)^2 - (p_y - qA_y)^2]^{1/2} - q\rho A_\phi. \quad (\text{V.2.1})$$

Assume that $\psi = 0$ in accord with the desire that there be no electric field. Also stipulate that \mathbf{A} have the components

$$A_\rho = 0, \quad (\text{V.2.2})$$

$$A_y = 0, \quad (\text{V.2.3})$$

$$A_\phi = -(\rho/2)B. \quad (\text{V.2.4})$$

According to Exercise 1.5.8 this choice for \mathbf{A} results in a constant magnetic field

$$\mathbf{B} = B\mathbf{e}_y. \quad (\text{V.2.5})$$

With these provisos it follows that K takes the form

$$K = -\rho[(p_t/c)^2 - m^2c^2 - p_\rho^2 - p_y^2]^{1/2} + q(\rho^2/2)B. \quad (\text{V.2.6})$$

Note that (1.5.49) can be written in the vector/matrix form

$$\begin{pmatrix} A_\phi \\ A_\rho \end{pmatrix} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} A_z \\ A_x \end{pmatrix}, \quad (\text{V.2.7})$$

from which it follows immediately that there is the inverse relation

$$\begin{pmatrix} A_z \\ A_x \end{pmatrix} = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} A_\phi \\ A_\rho \end{pmatrix}. \quad (\text{V.2.8})$$

Upon combining (1.2) through (1.4), (1.8), and (1.5.33), we see that there is the relation

$$\mathbf{A} = (B/2)(-x\mathbf{e}_z + z\mathbf{e}_x). \quad (\text{V.2.9})$$

Evidently, \mathbf{A} is in the Poincaré-Coulomb gauge.

V.2.2 Dimensionless Variables and Limiting Hamiltonian

Change to new scaled variables by writing

$$\rho = \rho_0 \ell + \xi \ell, \quad (\text{V.2.10})$$

$$p_\rho = P_\xi p_0, \quad (\text{V.2.11})$$

$$y = Y \ell, \quad (\text{V.2.12})$$

$$p_y = P_y p_0, \quad (\text{V.2.13})$$

$$t = \tau \ell / c, \quad (\text{V.2.14})$$

$$p_t = p_t^0 + P_\tau p_0 c. \quad (\text{V.2.15})$$

Here ℓ is a scale length and p_0 is the design momentum. [Note that the variable Y in (2.12) is not to be confused with the Y associated with the \mathbf{R} of Section 15.9.] Correspondingly, there are the Poisson bracket relations

$$[\xi, P_\xi] = (p_0 \ell)^{-1} [\rho, p_\rho], \quad (\text{V.2.16})$$

$$[Y, P_y] = (p_0 \ell)^{-1} [y, p_y], \quad (\text{V.2.17})$$

$$[\tau, P_\tau] = (p_0 \ell)^{-1} [t, p_t]. \quad (\text{V.2.18})$$

Let \tilde{K} be the new Hamiltonian for these new variables. It is given by the relation

$$\tilde{K} = \lambda \ell \{ -(\rho_0 + \xi)[(p_t^0 + P_\tau p_0 c)^2 / c^2 - m^2 c^2 - p_0^2 P_\xi^2 - p_0^2 P_y^2]^{1/2} + (qB\ell/2)(\rho_0 + \xi)^2 \}, \quad (\text{V.2.19})$$

or, equivalently,

$$\tilde{K} = \lambda \ell p_0 \{ -(\rho_0 + \xi)[(p_t^0 p_0^{-1} c^{-1} + P_\tau)^2 - m^2 c^2 / p_0^2 - P_\xi^2 - P_y^2]^{1/2} + [qB\ell/(2p_0)](\rho_0 + \xi)^2 \}. \quad (\text{V.2.20})$$

Here

$$\lambda = (\ell p_0)^{-1}. \quad (\text{V.2.21})$$

Observe that there are the relations

$$p_0 = \gamma m v_0 = \gamma \beta m c, \quad (\text{V.2.22})$$

$$p_t^0 = -\gamma mc^2. \quad (\text{V.2.23})$$

It follows that there are the relations

$$m^2 c^2 / p_0^2 = m^2 c^2 / (\beta \gamma m c)^2 = 1 / (\beta \gamma)^2, \quad (\text{V.2.24})$$

$$p_t^0 / (p_0 c) = -(\gamma m c^2) / (\gamma \beta m c^2) = -1 / \beta. \quad (\text{V.2.25})$$

Consequently, \tilde{K} can also be written on the form

$$\tilde{K} = -(\rho_0 + \xi)[(-1/\beta + P_\tau)^2 - (\beta\gamma)^{-2} - P_\xi^2 - P_y^2]^{1/2} + (b/2)(\rho_0 + \xi)^2 \quad (\text{V.2.26})$$

where

$$b = qB\ell/p_0. \quad (\text{V.2.27})$$

We also observe that

$$1/\beta^2 - 1/(\beta\gamma)^2 = 1. \quad (\text{V.2.28})$$

It follows that \tilde{K} can also be written as

$$\tilde{K} = -(\rho_0 + \xi)[1 - 2P_\tau/\beta + P_\tau^2 - P_\xi^2 - P_y^2]^{1/2} + (b/2)(\rho_0 + \xi)^2. \quad (\text{V.2.29})$$

Finally, we take the limit $\rho_0 \rightarrow 0$ to obtain the limiting Hamiltonian

$$\tilde{K}^{\lim} = -\xi[1 - 2P_\tau/\beta + P_\tau^2 - P_\xi^2 - P_y^2]^{1/2} + (b/2)\xi^2. \quad (\text{V.2.30})$$

V.2.3 Design Trajectory

Evidently P_y and P_τ are integrals of motion and vanish on the design trajectory. Therefore the variables ξ, P_ξ on the *design trajectory* are governed by the Hamiltonian

$$\tilde{K}^{\text{dt}} = -\xi[1 - P_\xi^2]^{1/2} + (b/2)\xi^2. \quad (\text{V.2.31})$$

Correspondingly, the associated equations of motion for these variables on the design trajectory are given by the relations

$$\xi' = \partial \tilde{K}^{\text{dt}} / \partial P_\xi = \xi P_\xi [1 - P_\xi^2]^{-1/2}, \quad (\text{V.2.32})$$

$$P'_\xi = -\partial \tilde{K}^{\text{dt}} / \partial \xi = [1 - P_\xi^2]^{1/2} - b\xi. \quad (\text{V.2.33})$$

They have the particular solution

$$\xi = 0, \quad (\text{V.2.34})$$

$$P_\xi(\phi) = \sin \Delta, \quad (\text{V.2.35})$$

where

$$\Delta = \phi - \phi^{\text{in}}. \quad (\text{V.2.36})$$

We will take (2.34) through (2.36) to be the ξ, P_ξ results for the design trajectory. Note that the design trajectory does not depend on b . That is, it does not depend on the magnetic field.

As assumed earlier, and consistent with the full equations of motion associated with the full Hamiltonian (2.30), the remaining variables on the design trajectory vanish,

$$Y = P_y = \tau = P_\tau = 0. \quad (\text{V.2.37})$$

It follows, from (2.34) through (2.37), that all the variables save P_ξ are deviation variables.

V.2.4 Deviation Variables

To proceed further we wish to replace ξ and P_ξ by deviation variables, which we will call $\hat{\xi}$ and \hat{P}_ξ , so that all variables are deviation variables. This is simply done by making the definitions

$$\xi = \hat{\xi}, \quad (\text{V.2.38})$$

$$P_\xi = \sin \Delta + \hat{P}_\xi, \quad (\text{V.2.39})$$

and leaving all other variables in peace. The relations (2.38) and (2.39) are a canonical transformation and can be obtained from the F_2 generating function given by

$$F_2(\xi, \hat{P}_\xi) = \xi(\sin \Delta + \hat{P}_\xi). \quad (\text{V.2.40})$$

Indeed, employing the standard machinery (6.5.5) yields the results

$$P_\xi = \partial F_2 / \partial \xi = \sin \Delta + \hat{P}_\xi, \quad (\text{V.2.41})$$

$$\hat{\xi} = \partial F_2 / \partial \hat{P}_\xi = \xi, \quad (\text{V.2.42})$$

as desired.

V.2.5 Deviation Variable Hamiltonian

We may regard the deviation variables as new variables. Associated with the use of these new variables will be a new Hamiltonian H given by the relation

$$H = \tilde{K}^{\text{lim}} + \partial F_2 / \partial \phi = \tilde{K}^{\text{lim}} + \xi \cos \Delta = \tilde{K}^{\text{lim}} + \hat{\xi} \cos \Delta. \quad (\text{V.2.43})$$

Use of (2.43) yields the result

$$H = -\hat{\xi}[1 - 2P_\tau/\beta + P_\tau^2 - (\hat{P}_\xi + \sin \Delta)^2 - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi} \cos \Delta, \quad (\text{V.2.44})$$

or

$$H = -\hat{\xi}[1 - \sin^2 \Delta - 2P_\tau/\beta + P_\tau^2 - \hat{P}_\xi^2 - 2\hat{P}_\xi \sin \Delta - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi} \cos \Delta, \quad (\text{V.2.45})$$

or

$$H = -\hat{\xi}[\cos^2 \Delta - 2P_\tau/\beta + P_\tau^2 - \hat{P}_\xi^2 - 2\hat{P}_\xi \sin \Delta - P_y^2]^{1/2} + (b/2)\hat{\xi}^2 + \hat{\xi} \cos \Delta. \quad (\text{V.2.46})$$

V.2.6 Computation of Transfer Map

Our aim is to find the transfer map associated with H . According to Section 10.4, this entails expanding H in terms of homogeneous polynomials,

$$H = H_0 + H_1 + H_2 + H_3 + H_4 + \dots. \quad (\text{V.2.47})$$

So doing gives for H_0 through H_2 the results

$$H_0 = 0, \quad (\text{V.2.48})$$

$$H_1 = 0, \quad (\text{V.2.49})$$

$$H_2 = \hat{\xi}P_\tau/(\beta \cos \Delta) + \hat{\xi}\hat{P}_\xi \tan \Delta + (b/2)\hat{\xi}^2. \quad (\text{V.2.50})$$

Note that H_1 vanishes as expected. That is, the design trajectory is given by the relations (2.37) supplemented by the relations

$$\hat{\xi} = \hat{P}_\xi = 0. \quad (\text{V.2.51})$$

Linear Part of Transfer Map

The first step is to find \mathcal{R} , the linear part of the transfer map. This requires solving the equations of motion associated with H_2 . They read

$$\hat{\xi}' = \partial H_2 / \partial \hat{P}_\xi = \hat{\xi} \tan \Delta, \quad (\text{V.2.52})$$

$$\hat{P}_\xi' = -\partial H_2 / \partial \hat{\xi} = -\hat{P}_\xi \tan \Delta - P_\tau / (\beta \cos \Delta) - b\hat{\xi}, \quad (\text{V.2.53})$$

$$Y' = \partial H_2 / \partial P_y = 0, \quad (\text{V.2.54})$$

$$P_y' = -\partial H_2 / \partial Y = 0, \quad (\text{V.2.55})$$

$$\tau' = \partial H_2 / \partial P_\tau = \hat{\xi}/(\beta \cos \Delta), \quad (\text{V.2.56})$$

$$P_\tau' = -\partial H_2 / \partial \tau = 0. \quad (\text{V.2.57})$$

The solutions to (2.54), (2.55), and (2.57) can be written immediately,

$$Y(\phi) = Y^{\text{in}}, \quad (\text{V.2.58})$$

$$P_Y(\phi) = P_Y^{\text{in}}, \quad (\text{V.2.59})$$

$$P_\tau(\phi) = P_\tau^{\text{in}}. \quad (\text{V.2.60})$$

The solution to (2.52), which is less trivial, is

$$\hat{\xi}(\phi) = \hat{\xi}^{\text{in}} / \cos \Delta. \quad (\text{V.2.61})$$

The results (2.60) and (2.61) can now be inserted into (2.53) to yield the differential equation

$$\hat{P}_\xi' = -\hat{P}_\xi \tan \Delta - P_\tau^{\text{in}} / (\beta \cos \Delta) - b\hat{\xi}^{\text{in}} / \cos \Delta = -\hat{P}_\xi \tan \Delta - (b\hat{\xi}^{\text{in}} + P_\tau^{\text{in}} / \beta) / \cos \Delta. \quad (\text{V.2.62})$$

It has the solution

$$\hat{P}_\xi(\phi) = \hat{P}_\xi^{\text{in}} \cos \Delta - (b\hat{\xi}^{\text{in}} + P_\tau^{\text{in}} / \beta) \sin \Delta. \quad (\text{V.2.63})$$

Finally, insertion of (2.61) into (2.56) yields the differential equation

$$\tau' = \hat{\xi}^{\text{in}} / (\beta \cos^2 \Delta). \quad (\text{V.2.64})$$

It has the solution

$$\tau(\phi) = \tau^{\text{in}} + \hat{\xi}^{\text{in}}(1/\beta) \tan \Delta. \quad (\text{V.2.65})$$

From these solutions we can read off the matrix R associated with \mathcal{R} . From (2.58) through (2.61), (2.63), and (2.65) we see that there is the vector/matrix relation

$$\begin{pmatrix} \hat{\xi}(\phi) \\ \hat{P}_\xi(\phi) \\ Y(\phi) \\ P_y(\phi) \\ \tau(\phi) \\ P_\tau(\phi) \end{pmatrix} = \begin{pmatrix} 1/\cos \Delta & 0 & 0 & 0 & 0 \\ -b \sin \Delta & \cos \Delta & 0 & 0 & -(1/\beta) \sin \Delta \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ (1/\beta) \tan \Delta & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\xi}^{\text{in}} \\ \hat{P}_\xi^{\text{in}} \\ Y^{\text{in}} \\ P_y^{\text{in}} \\ \tau^{\text{in}} \\ P_\tau^{\text{in}} \end{pmatrix}. \quad (\text{V.2.66})$$

The matrix R is the matrix appearing in (2.66).

Nonlinear Parts of Transfer Map

To compute the nonlinear parts of the transfer map it is necessary to continue the expansion (2.47) begun in (2.48) through (2.50) to find $H_3, H_4 \dots$ and to then apply the machinery of Section 10.5 to find the associated $f_3, f_4 \dots$ For example, one finds from (2.46) and (2.47) the result

$$H_3 = . \quad (\text{V.2.67})$$

Exercises

V.2.1. Verify that the solutions given by (2.34) through (2.36) do indeed satisfy the differential equations (2.32) and (2.33).

V.2.2. Verify the expansion (2.48) through (2.50).

V.2.3. Verify that the solutions (2.58) through (2.61), (2.63), and (2.65) do indeed satisfy the differential equations (2.52) through (2.57).

V.2.4. Verify that R , the matrix appearing in (1.65), is symplectic.

V.2.5. Verify (2.67).

V.3 The Inhomogeneous Field Case

The work so far has dealt with the case of a constant magnetic field. We now consider the general case.

V.3.1 Vector Potential for the General Inhomogeneous Field Case

We begin by expanding the vector potential in the Poincaré-Coulomb gauge and in homogeneous polynomials employing Cartesian coordinates and Cartesian unit vectors. That is we write

$$\mathbf{A}(\mathbf{r}) = \mathbf{A}^{\text{min } 1}(\mathbf{r}) + \mathbf{A}^{\text{min } 2}(\mathbf{r}) + \mathbf{A}^{\text{min } 3}(\mathbf{r}) + \dots \quad (\text{V.3.1})$$

For $\mathbf{A}^{\min 1}(\mathbf{r})$ we use (2.9) to account for the constant part of the magnetic field and write

$$\mathbf{A}^{\min 1}(\mathbf{r}) = (B/2)(-x\mathbf{e}_z + z\mathbf{e}_x). \quad (\text{V.3.2})$$

For example, in the case of a magnetic monopole doublet, there is the result

$$\mathbf{A}^{\min 1}(\mathbf{r}) = [ga/(X_0^2 + Z_0^2 + a^2)^{3/2}](-z\mathbf{e}_x + x\mathbf{e}_z), \quad (\text{V.3.3})$$

and there is the relation

$$B = -2[ga/(X_0^2 + Z_0^2 + a^2)^{3/2}]. \quad (\text{V.3.4})$$

And, for the same example and again employing Cartesian coordinates and Cartesian unit vectors, there is the result

$$\begin{aligned} \mathbf{A}^{\min 2}(\mathbf{r}) &= [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &[(Z_0y^2 - Z_0z^2 - X_0xz)\mathbf{e}_x + (X_0yz - Z_0xy)\mathbf{e}_y + (X_0x^2 + Z_0xz - X_0y^2)\mathbf{e}_z]. \end{aligned} \quad (\text{V.3.5})$$

See (15.9.7) and (15.9.8) in Section 15.9. For other examples the $\mathbf{A}^{\min n}(\mathbf{r})$ with $n \geq 2$ will be different, but still homogeneous of degree n . We will continue to assume that the constant part of the magnetic field is of the form (2.5), and therefore (3.2) will always be assumed to hold.

V.3.2 Transition to Cylindrical Coordinates

Next, as in Exercise 1.5.4, introduce polar coordinates in the x, z plane by the relations

$$x = \rho \cos \phi, \quad (\text{V.3.6})$$

$$z = \rho \sin \phi.$$

That is, we will again employ the cylindrical coordinates ρ, y, ϕ and also the unit vectors $\mathbf{e}_\rho, \mathbf{e}_y, \mathbf{e}_\phi$ of Exercise 1.5.4. See (3.6) and (1.5.52) through (1.5.54). Let us express \mathbf{A} in terms of these cylindrical coordinates and unit vectors.

Begin, for example, with (3.2). With the aid of (3.6) and (1.5.53) the relation (3.2) can be rewritten in the form

$$\begin{aligned} \mathbf{A}^{\min 1}(\mathbf{r}) &= -(B/2)\rho(-\sin \phi \mathbf{e}_x + \cos \phi \mathbf{e}_z) \\ &= -(B/2)\rho\mathbf{e}_\phi. \end{aligned} \quad (\text{V.3.7})$$

Since $\mathbf{e}_\rho, \mathbf{e}_y, \mathbf{e}_\phi$ form an orthonormal triad, it follows from (1.5.44) and (3.2) that there are the results

$$A_\rho^{\min 1}(\mathbf{r}) = \mathbf{e}_\rho \cdot \mathbf{A}^{\min 1}(\mathbf{r}) = 0, \quad (\text{V.3.8})$$

$$A_y^{\min 1}(\mathbf{r}) = \mathbf{e}_y \cdot \mathbf{A}^{\min 1}(\mathbf{r}) = 0, \quad (\text{V.3.9})$$

$$A_\phi^{\min 1}(\mathbf{r}) = \mathbf{e}_\phi \cdot \mathbf{A}^{\min 1}(\mathbf{r}) = -(B/2)\rho, \quad (\text{V.3.10})$$

which are to be expected in accord with (2.2) through (2.4) and (2.9).

As a second illustrative example, let us work on (3.5), which would be the $\mathbf{A}^{\min 2}(\mathbf{r})$ in the case of a magnetic monopole doublet. With the aid of (3.6) it takes the form

$$\begin{aligned}\mathbf{A}^{\min 2}(\mathbf{r}) &= [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{[Z_0y^2 - \rho^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)]\mathbf{e}_x + [\rho y(X_0 \sin \phi - Z_0 \cos \phi)]\mathbf{e}_y \\ &+ [\rho^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0y^2]\mathbf{e}_z\}. \end{aligned}\quad (\text{V.3.11})$$

From (1.5.35), the definitions (1.5.44), and (3.11) there are the results

$$\begin{aligned}A_\rho^{\min 2}(\mathbf{r}) &= \mathbf{e}_\rho \cdot \mathbf{A}^{\min 2}(\mathbf{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{[\cos \phi \mathbf{e}_x + \sin \phi \mathbf{e}_z] \cdot [Z_0y^2\mathbf{e}_x - \rho^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)\mathbf{e}_x] \\ &+ [\cos \phi \mathbf{e}_x + \sin \phi \mathbf{e}_z] \cdot [\rho^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi)\mathbf{e}_z - X_0y^2\mathbf{e}_z]\} \\ &= [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{(\cos \phi)[Z_0y^2 - \rho^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ &+ (\sin \phi)[\rho^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0y^2]\}, \end{aligned}\quad (\text{V.3.12})$$

$$A_y^{\min 2}(\mathbf{r}) = \mathbf{e}_y \cdot \mathbf{A}^{\min 2}(\mathbf{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\rho y(X_0 \sin \phi - Z_0 \cos \phi)], \quad (\text{V.3.13})$$

$$\begin{aligned}A_\phi^{\min 2}(\mathbf{r}) &= \mathbf{e}_\phi \cdot \mathbf{A}^{\min 2}(\mathbf{r}) = [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{[-\sin \phi \mathbf{e}_x + \cos \phi \mathbf{e}_z] \cdot [Z_0y^2\mathbf{e}_x - \rho^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)\mathbf{e}_x] \\ &+ [-\sin \phi \mathbf{e}_x + \cos \phi \mathbf{e}_z] \cdot [\rho^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi)\mathbf{e}_z - X_0y^2\mathbf{e}_z]\} \\ &= [-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{-(\sin \phi)[Z_0y^2 - \rho^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ &+ (\cos \phi)[\rho^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0y^2]\}. \end{aligned}\quad (\text{V.3.14})$$

V.3.3 Dimensionless Variables and Limiting Vector Potential

We now make the substitutions (2.10) through (2.15) and take the limit $\rho_0 \rightarrow 0$ to obtain the limiting vector potential whose components we will denote by letters with breve marks $\check{}$ above. So doing yields for the constant part of the magnetic field the limiting result

$$\check{A}_\phi^{\min 1} = -\ell(B/2)\xi. \quad (\text{V.3.15})$$

And for the leading term of the nonconstant part of the magnetic field, again taking for illustrative purposes the magnetic monopole doublet example, there are the results

$$\begin{aligned}\check{A}_\rho^{\min 2} &= \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{(\cos \phi)[Z_0Y^2 - \xi^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ &+ (\sin \phi)[\xi^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0Y^2]\}, \end{aligned}\quad (\text{V.3.16})$$

$$\check{A}_y^{\min 2} = \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\xi Y(X_0 \sin \phi - Z_0 \cos \phi)], \quad (\text{V.3.17})$$

$$\begin{aligned}\check{A}_\phi^{\min 2} &= \ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ &\{-(\sin \phi)[Z_0Y^2 - \xi^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ &+ (\cos \phi)[\xi^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0Y^2]\}. \end{aligned}\quad (\text{V.3.18})$$

V.3.4 Computation of Limiting Hamiltonian in Dimensionless Variables

At this point we are again ready to compute \tilde{K}^{lim} , but this time with possible inhomogeneity in the magnetic field included. Let us introduce the notation

$$\check{\mathbf{A}}^{\text{min non}} = \check{\mathbf{A}}^{\text{min 2}} + \check{\mathbf{A}}^{\text{min 3}} + \check{\mathbf{A}}^{\text{min 4}} + \dots \quad (\text{V.3.19})$$

to denote the *nonlinear* part of the vector potential. We now find for the limiting Hamiltonian the result

$$\begin{aligned} \tilde{K}^{\text{lim}} = & -\xi[1 - 2P_\tau/\beta + P_\tau^2 - (P_\xi - \mathcal{A}_\rho^{\text{min non}})^2 - (P_y - \mathcal{A}_y^{\text{min non}})^2]^{1/2} \\ & +(b/2)\xi^2 - \xi\mathcal{A}_\phi^{\text{min non}} \end{aligned} \quad (\text{V.3.20})$$

where

$$\mathcal{A}_\rho^{\text{min non}} = (q/p_0)\check{\mathbf{A}}_\rho^{\text{min non}}, \quad (\text{V.3.21})$$

$$\mathcal{A}_y^{\text{min non}} = (q/p_0)\check{\mathbf{A}}_y^{\text{min non}}, \quad (\text{V.3.22})$$

$$\mathcal{A}_\phi^{\text{min non}} = (q/p_0)\check{\mathbf{A}}_\phi^{\text{min non}}. \quad (\text{V.3.23})$$

V.3.5 Deviation Variable Hamiltonian

Introduce, as before, deviation variables $(\hat{\xi}, \tau, Y; \hat{P}_\xi, P_\tau, P_y)$ with $\hat{\xi}$ and \hat{P}_ξ defined by (2.38) and (2.39). Doing so, and employing the rule (2.43), yields the new Hamiltonian

$$\begin{aligned} H = & -\hat{\xi}[1 - 2P_\tau/\beta + P_\tau^2 - (\sin \Delta + \hat{P}_\xi - \hat{\mathcal{A}}_\rho^{\text{min non}})^2 - (P_y - \hat{\mathcal{A}}_y^{\text{min non}})^2]^{1/2} \\ & +(b/2)\hat{\xi}^2 - \hat{\xi}\hat{\mathcal{A}}_\phi^{\text{min non}} + \hat{\xi} \cos \Delta. \end{aligned} \quad (\text{V.3.24})$$

Here we have used the notation $\hat{\mathcal{A}}_\rho^{\text{min non}}$ to indicate that the variable ξ in $\mathcal{A}_\rho^{\text{min non}}$ has been replaced by $\hat{\xi}$, etc. For example, in the case of a magnetic monopole doublet, there are the results

$$\begin{aligned} \hat{\mathcal{A}}_\rho^{\text{min 2}} = & (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ & \{(\cos \phi)[Z_0Y^2 - \hat{\xi}^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ & + (\sin \phi)[\hat{\xi}^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0Y^2]\}, \end{aligned} \quad (\text{V.3.25})$$

$$\hat{\mathcal{A}}_y^{\text{min 2}} = (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}][\hat{\xi}Y(X_0 \sin \phi - Z_0 \cos \phi)], \quad (\text{V.3.26})$$

$$\begin{aligned} \hat{\mathcal{A}}_\phi^{\text{min 2}} = & (q/p_0)\ell^2[-2ga/(X_0^2 + Z_0^2 + a^2)^{5/2}] \times \\ & \{-(\sin \phi)[Z_0Y^2 - \hat{\xi}^2(Z_0 \sin^2 \phi - X_0 \cos \phi \sin \phi)] \\ & + (\cos \phi)[\hat{\xi}^2(X_0 \cos^2 \phi + Z_0 \cos \phi \sin \phi) - X_0Y^2]\}. \end{aligned} \quad (\text{V.3.27})$$

V.3.6 Expansion of Deviation Variable Hamiltonian and Computation of Transfer Map

As done before in Subsection 2.6, we expand H in terms of homogeneous polynomials. Begin by observing that there is the relation

$$\begin{aligned} 1 - (\sin \Delta + \hat{P}_\xi - \hat{\mathcal{A}}_\rho^{\min \text{ non}})^2 = \\ \cos^2 \Delta - \hat{P}_\xi^2 - (\hat{\mathcal{A}}_\rho^{\min \text{ non}})^2 - 2(\sin \Delta) \hat{P}_\xi + 2(\sin \Delta) \hat{\mathcal{A}}_\rho^{\min \text{ non}} + 2\hat{P}_\xi \hat{\mathcal{A}}_\rho^{\min \text{ non}}. \end{aligned} \quad (\text{V.3.28})$$

Consequently H , as given by (3.24), can be rewritten in the form

$$\begin{aligned} H = & -\hat{\xi}[\cos^2 \Delta - 2P_\tau/\beta + P_\tau^2 - \hat{P}_\xi^2 - (\hat{\mathcal{A}}_\rho^{\min \text{ non}})^2 \\ & - 2(\sin \Delta) \hat{P}_\xi + 2(\sin \Delta) \hat{\mathcal{A}}_\rho^{\min \text{ non}} + 2\hat{P}_\xi \hat{\mathcal{A}}_\rho^{\min \text{ non}} \\ & - P_y^2 + 2P_y \hat{\mathcal{A}}_y^{\min \text{ non}} - (\hat{\mathcal{A}}_y^{\min \text{ non}})^2]^{1/2} \\ & +(b/2)\hat{\xi}^2 - \hat{\xi} \hat{\mathcal{A}}_\phi^{\min \text{ non}} + \hat{\xi} \cos \Delta. \end{aligned} \quad (\text{V.3.29})$$

We are now prepared to expand H in the form (2.47). Compare (2.46) and (3.29). Since $\hat{\mathcal{A}}_\rho^{\min \text{ non}}$ and $\hat{\mathcal{A}}_y^{\min \text{ non}}$ consist entirely of terms of degree 2 and higher, and $\hat{\xi} \hat{\mathcal{A}}_\phi^{\min \text{ non}}$ consists entirely of terms of degree 3 and higher, it follows that they make *no* contribution to H_0 through H_2 . {Note that the term of the form $[***]^{1/2}$ in (3.29) is multiplied by $\hat{\xi}$.} Therefore the H_0 , H_1 , and H_2 terms in the expansion are the *same* as those given by (2.48) through (2.50). Consequently the design orbit and \mathcal{R} , the linear part of the transfer map about the design orbit, are the same as those found earlier. That is, the design orbit does *not* depend on the magnetic field, and the linear part of the transfer map depends *only* on the uniform part of the magnetic field, described by $A_\phi^{\min 1}$ or b . The design orbit and the linear part of the transfer map do *not* depend on field inhomogeneities described by the $\mathbf{A}^{\min n}$ with $n \geq 2$. Field inhomogeneities play a role only in the calculation of the H_m with $m \geq 3$. Correspondingly, field inhomogeneities play a role in the transfer map only for the generators f_m with $m \geq 3$.

We close this subsection by computing, for example, the H_3 that occurs in the expansion of (3.29). We find the result

$$H_3 = . \quad (\text{V.3.30})$$

See Exercise 3.1. Note that (2.67) and (3.30) agree when there are no field inhomogeneities.

Exercises

V.3.1. Verify that, in the computation of the H_3 term that occurs in the expansion of (3.29), the terms * are of too high an order to play a role, and therefore may be neglected.

Bibliography

- [1] É. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Publishers (1998).

Appendix W

Smoothing for Harmonic Functions

W.1 Introduction

W.2 The Line in Two Space

Consider in x, y space the line $y = 0$ and suppose a potential $\psi_0(x)$ is specified on this line. Define its Fourier transform $\tilde{\psi}_0(k_x)$ by the rule

$$\tilde{\psi}_0(k_x) = [1/\sqrt{(2\pi)}] \int dx \exp(-ik_x x) \psi_0(x). \quad (\text{W.2.1})$$

Make the Ansatz

$$\psi(x, y) = [1/\sqrt{(2\pi)}] \int dk_x \exp(ik_x x) \exp(-ky) \tilde{\psi}_0(k_x) \quad (\text{W.2.2})$$

where

$$k = \sqrt{k_x^2} = |k_x|. \quad (\text{W.2.3})$$

Evidently this $\psi(x, y)$ is harmonic and vanishes as $y \rightarrow +\infty$. We also have the result

$$\psi(x, 0) = [1/\sqrt{(2\pi)}] \int dk_x \exp(ik_x x) \tilde{\psi}_0(k_x) = \psi_0(x). \quad (\text{W.2.4})$$

It follows that we have found the solution to Laplace's equation in the upper half plane $y \geq 0$ associated with the $y = 0$ boundary value $\psi_0(x)$.

Note that the operation defined by (2.2) is smoothing for $y > 0$. High spatial frequencies are suppressed by the factor $\exp(-ky)$, and this *exponential* suppression/damping is ever more effective the larger the value of y . The higher the y observation line is above the $y = 0$ line, the smoother $\psi(x, y)$ on this observation line becomes as a function of x .

We also observe, in passing, two facts. First, suppose ψ_0 , now to be called ψ_0^c , is a *constant* function,

$$\psi_0^c(x) = c. \quad (\text{W.2.5})$$

Then, by (2.1),

$$\tilde{\psi}_0^c(k_x) = c\sqrt{2\pi}\delta(k_x). \quad (\text{W.2.6})$$

It follows from (2.2) that there is the relation

$$\psi^c(x, y) = c. \quad (\text{W.2.7})$$

As expected, if ψ is constant on the boundary $y = 0$, it will have the same constant value in the upper half plane $y \geq 0$. Second, for any solution, there is the relation

$$\begin{aligned} \int dx \psi(x, y) &= [1/\sqrt{(2\pi)}] \int dk_x \exp(-ky) \tilde{\psi}_0(k_x) \int dx \exp(ik_x x) \\ &= [1/\sqrt{(2\pi)}] \int dk_x \exp(-ky) \tilde{\psi}_0(k_x) (2\pi) \delta(k_x) \\ &= \sqrt{(2\pi)} \tilde{\psi}_0(0) = \int dx \psi_0(x). \end{aligned} \quad (\text{W.2.8})$$

That is, the dx integral of $\psi(x, y)$ over any line of constant y is independent of y .

To further study smoothing in the case of a line, suppose ψ_0 , now to be called ψ_0^δ , is a delta function centered on the origin,

$$\psi_0^\delta(x) = \delta(x). \quad (\text{W.2.9})$$

Then, by (2.1),

$$\tilde{\psi}_0^\delta(k_x) = 1/\sqrt{(2\pi)}, \quad (\text{W.2.10})$$

and (2.2) takes the form

$$\psi^\delta(x, y) = [1/(2\pi)] \int dk_x \exp(ik_x x) \exp(-ky). \quad (\text{W.2.11})$$

This integral can be evaluated to give the result

$$\psi^\delta(x, y) = (1/\pi)[y/(x^2 + y^2)]. \quad (\text{W.2.12})$$

We next observe directly that, as expected, the function $\psi^\delta(x, y)$ given by (2.12) is harmonic. Define ρ by the rule

$$\rho = \sqrt{x^2 + y^2}. \quad (\text{W.2.13})$$

From 2-D potential theory we know that the function $\log(\rho)$ is harmonic. By the properties of the logarithm function there is the relation

$$\log(\rho^2) = 2 \log(\rho), \quad (\text{W.2.14})$$

and therefore the function $\log(\rho^2)$ is also harmonic. We next observe that the operators ∂_y and ∇^2 commute. It follows that the function $\partial_y \log(\rho^2)$ is also harmonic. Finally, there is the result

$$\partial_y \log(\rho^2) = \partial_y \log(x^2 + y^2) = 2y/(x^2 + y^2). \quad (\text{W.2.15})$$

Upon comparing (2.12) and (2.15) we see that $\psi^\delta(x, y)$ is indeed harmonic.

Let us now, with the aid of (2.12), illustrate the general behavior of $\psi^\delta(x, y)$. Figure 2.1 displays $\psi^\delta(x, y)$ as a function of x for various values of y . Figure 2.2 displays $\psi^\delta(x, y)$ as a function of y for various values of x .

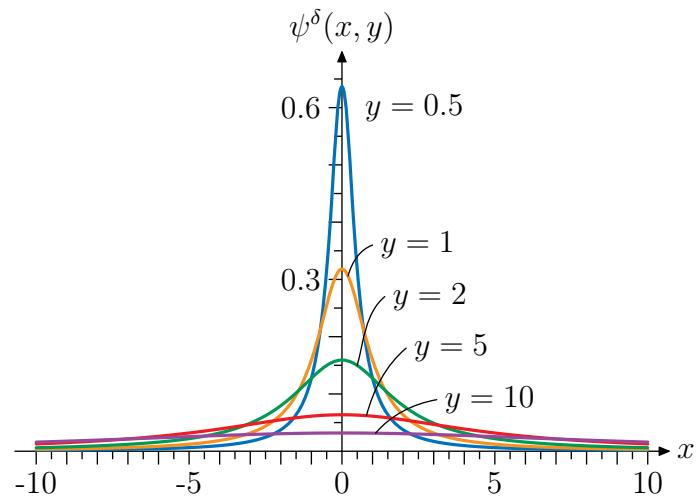


Figure W.2.1: The function $\psi^\delta(x, y)$ as a function of x for various values of y .

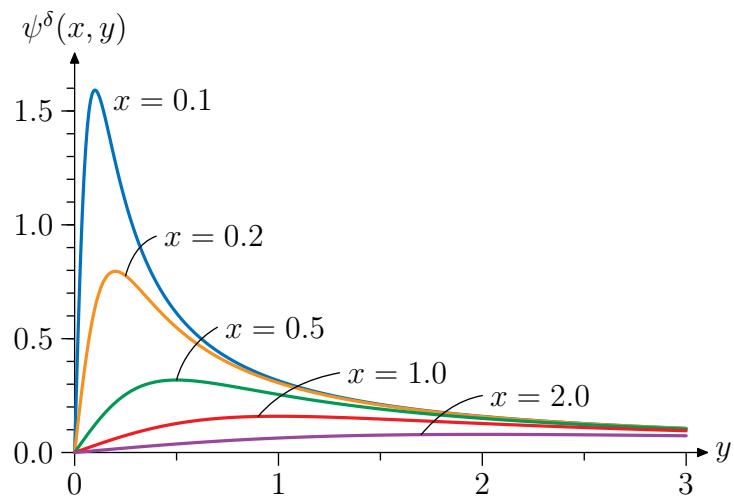


Figure W.2.2: The function $\psi^\delta(x, y)$ as a function of y for various values of x .

From Figure 2.1 we see that the delta function potential spike on the $y = 0$ line becomes an ever lower and broader bump on lines of increasing y . And we know from (2.8) that the weighted area under the bump remains the same for all y ,

$$\int dx \psi^\delta(x, y) = \int dx \delta(x) = 1. \quad (\text{W.2.16})$$

Thus the effect of a disturbance in the potential on the $y = 0$ line “decays” away as one moves to lines with successively larger values of y . Figure 2.2 illustrates this decay as a function of y for various values of x . Indeed, there are the expansions

$$\psi^\delta(x, y) = (1/\pi)(1/y)[1 - (x^2/y^2) + (x^2/y^2)^2 - \dots] \text{ for } x < y, \quad (\text{W.2.17})$$

$$\psi^\delta(x, y) = [1/(2\pi)](1/y) \text{ for } x = y, \quad (\text{W.2.18})$$

$$\psi^\delta(x, y) = (1/\pi)(y/x^2)[1 - (y^2/x^2) + (y^2/x^2)^2 - \dots] \text{ for } y < x. \quad (\text{W.2.19})$$

Evidently, as expected, the sequence of functions $\psi^\delta(x, y)$ for varying y converges to the delta function,

$$\lim_{y \rightarrow 0^+} \psi^\delta(x, y) = \delta(x). \quad (\text{W.2.20})$$

Finally, we see from (2.17) that $\psi^\delta(x, y)$ falls off as y^{-1} for fixed x and large y , and observe that the dimension of a line is 1. And, from (2.19), we see that $\psi^\delta(x, y)$ falls off as x^{-2} for fixed y and large x . For yet more insight, see Exercise 2.3.

What is the mechanism for this decay? In agreement with (2.8) and (2.16), this decay occurs entirely due to spreading. Indeed, the relations (2.5) and (2.7) illustrate that if the initial/boundary potential distribution is completely “spread out”, i.e. constant, then no decay occurs.

We have seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (2.12) to a delta function disturbance (2.9) we can derive the response to a general disturbance. With (2.12) in mind, define a kernel $G(x; x'; y)$ by the rule

$$G(x; x'; y) = (1/\pi)\{y/[y^2 + (x - x')^2]\}. \quad (\text{W.2.21})$$

Observe that a general disturbance $\psi_0(x)$ has the integral representation

$$\psi_0(x) = \int dx' \psi_0(x') \delta(x - x'). \quad (\text{W.2.22})$$

It follows that the response to $\psi_0(x)$ is given by the integral

$$\psi(x, y) = \int dx' \psi_0(x') G(x; x'; y). \quad (\text{W.2.23})$$

For what it’s worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(x; x'; y)$ is the probability that a random walk initiated at the point x, y will reach the point $x', 0$.

Exercises

W.2.1. Evaluate the integral (2.11) to verify the claim (2.12).

W.2.2. Evaluate directly the integral on the left side of (2.16) using (2.12). Verify (2.23) for the case (2.5).

W.2.3. The purpose of this exercise is to verify and employ an observation made by *Dan Abell*. Using (2.12), consider the *level curves* of ψ^δ having heights h by writing

$$\psi^\delta(x, y) = h. \quad (\text{W.2.24})$$

Show that so doing yields the relation

$$x^2 + [y - 1/(2\pi h)]^2 = [1/(2\pi h)]^2. \quad (\text{W.2.25})$$

Observe that the level curves (*equipotential lines*) are all circles, and that all the circles pass through the origin. Sketch them for yourself, and label them according to the values of h ! Employ this result to explain the features of Figures 2.1 and 2.2.

W.3 The Plane in Three Space

Consider in x, y, z space the plane $z = 0$ and suppose a potential $\psi_0(x, y)$ is specified on this plane. Define its Fourier transform $\tilde{\psi}_0(k_x, k_y)$ by the rule

$$\tilde{\psi}_0(k_x, k_y) = [1/(2\pi)] \int dx dy \exp(-ik_x x) \exp(-ik_y y) \psi_0(x, y). \quad (\text{W.3.1})$$

Make the Ansatz

$$\psi(x, y, z) = [1/(2\pi)] \int dk_x dk_y \exp(ik_x x) \exp(ik_y y) \exp(-kz) \tilde{\psi}_0(k_x, k_y) \quad (\text{W.3.2})$$

where

$$k = \sqrt{k_x^2 + k_y^2}. \quad (\text{W.3.3})$$

Evidently this $\psi(x, y, z)$ is harmonic and vanishes as $z \rightarrow +\infty$. We also have the result

$$\psi(x, y, 0) = [1/(2\pi)] \int dk_x dk_y \exp(ik_x x) \exp(ik_y y) \tilde{\psi}_0(k_x, k_y) = \psi_0(x, y). \quad (\text{W.3.4})$$

It follows that we have found the solution to Laplace's equation in the upper half space $z \geq 0$ associated with the $z = 0$ boundary value $\psi_0(x, y)$.

Note that the operation defined by (3.2) is smoothing for $z > 0$. High spatial frequencies are suppressed by the factor $\exp(-kz)$, and this *exponential* suppression/damping is ever more effective the larger the value of z . The higher the z observation plane is above the $z = 0$ plane, the smoother $\psi(x, y, z)$ on this observation plane becomes as a function of x and y .

We also observe, in passing, two facts. First, suppose ψ_0 , now to be called ψ_0^c , is a *constant* function,

$$\psi_0^c(x, y) = c. \quad (\text{W.3.5})$$

Then, by (3.1),

$$\tilde{\psi}_0^c(k_x) = c(2\pi)\delta(k_x)\delta(k_y). \quad (\text{W.3.6})$$

It follows from (3.2) that there is the relation

$$\psi^c(x, y, z) = c. \quad (\text{W.3.7})$$

As expected, if ψ is constant on the boundary $z = 0$, it will have the same constant value in the upper half space $z \geq 0$. Second, for any solution, there is the relation

$$\begin{aligned} \int dxdy \psi(x, y, z) &= [1/(2\pi)] \int dk_x dk_y \exp(-kz) \tilde{\psi}_0(k_x, k_y) \int dxdy \exp(ik_x x) \exp(ik_y y) \\ &= [1/(2\pi)] \int dk_x dk_y \exp(-kz) \tilde{\psi}_0(k_x, k_y) (2\pi)^2 \delta(k_x) \delta(k_y) \\ &= (2\pi) \tilde{\psi}_0(0, 0) = \int dxdy \psi_0(x, y). \end{aligned} \quad (\text{W.3.8})$$

That is, the $dxdy$ integral of $\psi(x, y, z)$ over any plane of constant z is independent of z .

To further study smoothing in the case of a plane, suppose ψ_0 , now to be called ψ_0^δ , is a delta function centered on the origin,

$$\psi_0^\delta(x, y) = \delta(x)\delta(y). \quad (\text{W.3.9})$$

Then, by (3.1),

$$\tilde{\psi}_0^\delta(k_x, k_y) = 1/(2\pi), \quad (\text{W.3.10})$$

and (3.2) takes the form

$$\psi^\delta(x, y, z) = [1/(2\pi)^2] \int dk_x dk_y \exp(ik_x x) \exp(ik_y y) \exp(-kz). \quad (\text{W.3.11})$$

Let work to evaluate this double integral. Introduce polar variables by writing

$$\begin{aligned} x &= \rho \cos(\theta), \\ y &= \rho \sin(\theta); \end{aligned} \quad (\text{W.3.12})$$

$$\begin{aligned} k_x &= k \cos(\phi), \\ k_y &= k \sin(\phi). \end{aligned} \quad (\text{W.3.13})$$

Then we have the relations

$$k_x x + k_y y = k\rho[\cos(\phi) \cos(\theta) + \sin(\phi) \sin(\theta)] = k\rho \cos(\phi - \theta), \quad (\text{W.3.14})$$

$$dk_x dk_y = kdkd\phi. \quad (\text{W.3.15})$$

Correspondingly, (3.11) takes the form

$$\psi^\delta(x, y, z) = [1/(2\pi)^2] \int_0^\infty k dk \exp(-kz) \int_0^{2\pi} d\phi \exp[ik\rho \cos(\phi - \theta)]. \quad (\text{W.3.16})$$

Next perform further manipulations. By periodicity we have the result

$$\int_0^{2\pi} d\phi \exp[ik\rho \cos(\phi - \theta)] = \int_0^{2\pi} d\phi \exp[ik\rho \cos(\phi)]. \quad (\text{W.3.17})$$

Also there is the result

$$\exp[ik\rho \cos(\phi)] = \cos[k\rho \cos(\phi)] + i \sin[k\rho \cos(\phi)]. \quad (\text{W.3.18})$$

Moreover, we recall the relations

$$\cos[k\rho \cos(\phi)] = J_0(k\rho) + 2 \sum_{k=1}^{\infty} (-1)^k J_{2k}(k\rho) \cos(2k\phi), \quad (\text{W.3.19})$$

$$\sin[k\rho \cos(\phi)] = 2 \sum_{k=0}^{\infty} (-1)^k J_{2k+1}(k\rho) \cos[(2k+1)\phi]. \quad (\text{W.3.20})$$

It follows that

$$\int_0^{2\pi} d\phi \cos[k\rho \cos(\phi)] = (2\pi)J_0(k\rho), \quad (\text{W.3.21})$$

$$\int_0^{2\pi} d\phi \sin[k\rho \cos(\phi)] = 0, \quad (\text{W.3.22})$$

and therefore

$$\int_0^{2\pi} d\phi \exp[ik\rho \cos(\phi)] = (2\pi)J_0(k\rho). \quad (\text{W.3.23})$$

Upon combining the fruits of our labor we find the pleasant result

$$\psi^\delta(x, y, z) = [1/(2\pi)] \int_0^\infty k dk J_0(k\rho) \exp(-kz). \quad (\text{W.3.24})$$

Note that $\psi^\delta(x, y, z)$ depends on x and y only through the rotationally invariant quantity ρ , as expected by axial symmetry about the z axis.

Yet more can be accomplished. There is the general Bessel function relation

$$\int_0^\infty t dt \exp(-at) J_0(bt) = a/(a^2 + b^2)^{3/2}. \quad (\text{W.3.25})$$

Consequently, we have the final result

$$\psi^\delta(x, y, z) = [1/(2\pi)][z/(z^2 + \rho^2)^{3/2}]. \quad (\text{W.3.26})$$

We next observe directly that, as expected, the function $\psi^\delta(x, y, z)$ given by (3.26) is harmonic. Define r by the rule

$$r = \sqrt{x^2 + y^2 + z^2} = \sqrt{z^2 + \rho^2}. \quad (\text{W.3.27})$$

From 3-D potential theory we know that the function $1/r$ is harmonic. We next observe that the operators ∂_z and ∇^2 commute. It follows that the function $\partial_z(1/r)$ is also harmonic. Finally, there is the result

$$\partial_z(1/r) = \partial_z(1/\sqrt{z^2 + \rho^2}) = z/(z^2 + \rho^2)^{3/2}. \quad (\text{W.3.28})$$

Upon comparing (3.26) and (3.28) we see that $\psi^\delta(x, y, z)$ is indeed harmonic.

Let us now, with the aid of (3.26), illustrate the general behavior of $\psi^\delta(x, y, z)$. Figure 3.1 displays $\psi^\delta(x, y, z)$ as a function of ρ for various values of z . Figure 3.2 displays $\psi^\delta(x, y, z)$ as a function of z for various values of ρ .

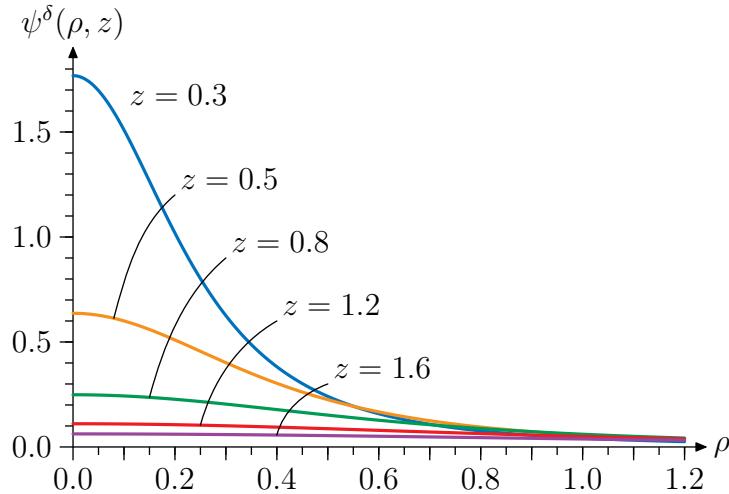


Figure W.3.1: The function $\psi^\delta(x, y, z) = \psi^\delta(\rho, z)$ as a function of ρ for various values of z .

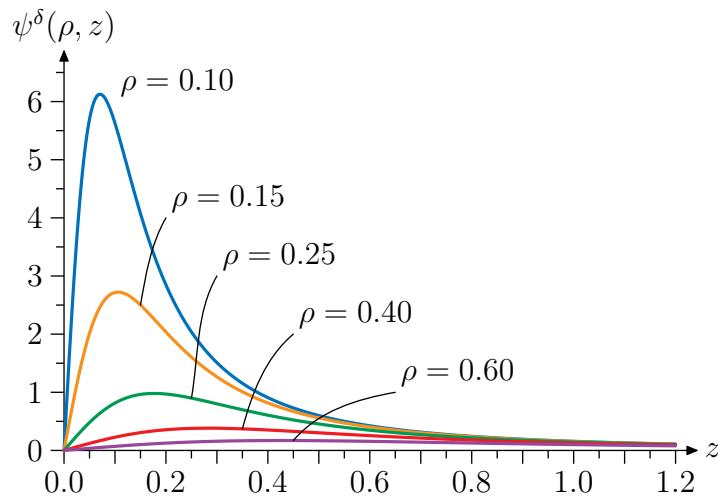


Figure W.3.2: The function $\psi^\delta(x, y, z) = \psi^\delta(\rho, z)$ as a function of z for various values of ρ .

From Figure 3.1 we see that the delta function potential spike in the $z = 0$ plane becomes an ever lower and broader bump in planes with increasing z . And we know from (3.8) that

the weighted area under the bump remains the same for all z ,

$$\int dx dy \psi^\delta(x, y, z) = \int dx dy \delta(x)\delta(y) = 1. \quad (\text{W.3.29})$$

Thus the effect of a disturbance in the potential in the $z = 0$ plane “decays” away as one moves to planes with successively larger values of z . Figure 3.2 illustrates this decay as a function of z for various values of ρ . Indeed, there are the expansions

$$\psi^\delta(x, y, z) = [1/(2\pi)](1/z^2)[1 - (3/2)(\rho^2/z^2) + (15/8)(\rho^2/z^2)^2 - \dots] \text{ for } \rho < z, \quad (\text{W.3.30})$$

$$\psi^\delta(x, y, z) = [1/(2\pi)](1/2^{3/2})(1/z^2) \text{ for } \rho = z, \quad (\text{W.3.31})$$

$$\psi^\delta(x, y, z) = [1/(2\pi)](z/\rho^3)[1 - (3/2)(z^2/\rho^2) + (15/8)(z^2/\rho^2)^2 - \dots] \text{ for } z < \rho. \quad (\text{W.3.32})$$

Evidently, as expected, the sequence of functions $\psi^\delta(x, y, z)$ for varying z converges to the delta function,

$$\lim_{z \rightarrow 0^+} \psi^\delta(x, y, z) = \delta(x)\delta(y). \quad (\text{W.3.33})$$

Finally, we see from (3.30) that $\psi^\delta(x, y, z)$ falls off as z^{-2} for fixed ρ and large z , and observe that the dimension of a plane is 2. And, from (3.32), we see that $\psi^\delta(x, y, z)$ falls off as ρ^{-3} for fixed z and large ρ . What is the mechanism for this decay? In agreement with (3.8) and (3.29), this decay occurs entirely due to spreading. Indeed, the relations (3.5) and (3.7) illustrate that if the initial/boundary potential distribution is completely “spread out”, i.e. constant, then no decay occurs.

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (3.26) to a delta function disturbance (3.9) we can derive the response to a general disturbance. With (3.26) in mind, define a kernel $G(x, y; x', y'; z)$ by the rule

$$G(x, y; x', y'; z) = [1/(2\pi)]\{z/[z^2 + (x - x')^2 + (y - y')^2]^{3/2}\}. \quad (\text{W.3.34})$$

Observe that a general disturbance $\psi_0(x, y)$ has the integral representation

$$\psi_0(x, y) = \int dx' dy' \psi_0(x', y') \delta(x - x') \delta(y - y'). \quad (\text{W.3.35})$$

It follows that the response to $\psi_0(x, y)$ is given by the integral

$$\psi(x, y, z) = \int dx' dy' \psi_0(x', y') G(x, y; x', y'; z). \quad (\text{W.3.36})$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(x, y; x', y'; z)$ is the probability that a random walk initiated at the point x, y, z will reach the point $x', y', 0$.

Exercises

W.3.1. Evaluate directly the integral on the left side of (3.29) using (3.26). Verify (3.36) for the case (3.5).

W.4 The Circle in Two Space

Consider in x, y space a circle of radius R centered on the origin, and suppose a potential ψ_R is specified on this circle. More specifically, employ the polar variables (3.9) so that

$$\psi(x, y) = \psi(\rho, \theta) \quad (\text{W.4.1})$$

and

$$\psi(R, \theta) = \psi_R(\theta). \quad (\text{W.4.2})$$

Define the angular Fourier transform of $\psi_R(\theta)$ by the rule

$$\tilde{\psi}_R(m) = ([1/(2\pi)] \int_0^{2\pi} d\theta \exp(-im\theta) \psi_R(\theta)). \quad (\text{W.4.3})$$

Make the Ansatz

$$\psi(\rho, \theta) = \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp(im\theta) \tilde{\psi}_R(m). \quad (\text{W.4.4})$$

Evidently this $\psi(\rho, \theta)$ is harmonic. We also have the result

$$\psi(R, \theta) = \sum_{m=-\infty}^{m=\infty} \exp(im\theta) \tilde{\psi}_R(m) = \psi_R(\theta). \quad (\text{W.4.5})$$

It follows that we have found the solution to Laplace's equation in the disk of radius R with the boundary value $\psi_R(\theta)$.

Note that the operation defined by (4.4) is smoothing for $(\rho/R) < 1$. We see that high angular frequencies are suppressed by the factor

$$(\rho/R)^{|m|} = \exp[|m| \log(\rho/R)], \quad (\text{W.4.6})$$

and observe that $\log(\rho/R) < 0$. This *exponential* suppression/damping is ever more effective the larger the value of R and/or the smaller the value of ρ . The larger the radius R of the boundary circle and/or the smaller the radius ρ of the observation circle, the smoother $\psi(\rho, \theta)$ becomes as a function of θ .

We also observe, in passing, three facts. First, suppose ψ_R , now to be called ψ_R^c , is a *constant* function,

$$\psi_R^c(\theta) = c. \quad (\text{W.4.7})$$

Then, by (4.3),

$$\tilde{\psi}_R^c(m) = c\delta_{m,0}. \quad (\text{W.4.8})$$

It follows from (4.4) that there is the relation

$$\psi^c(\rho, \theta) = c. \quad (\text{W.4.9})$$

As expected, if ψ is constant on the boundary $\rho = R$, it will have the same constant value inside the circle. Second, for any solution, there is the relation

$$\int_0^{2\pi} d\theta \psi(\rho, \theta) = (2\pi) \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \tilde{\psi}_R(m) \delta_{m,0} = (2\pi) \tilde{\psi}_R(0) = \int_0^{2\pi} d\theta \psi_R(\theta). \quad (\text{W.4.10})$$

That is, the $d\theta$ integral of $\psi(\rho, \theta)$ over any circle of constant ρ is independent of ρ . Third, we also see from (4.4) that there is the relation

$$\psi(0, \theta) = \tilde{\psi}_R(0) = [1/(2\pi)] \int_0^{2\pi} d\theta \psi_R(\theta), \quad (\text{W.4.11})$$

which shows that the average value of an harmonic function over a circle equals its value at the center of the circle, a result which in turn is a special case of the connection between Laplace and Monte Carlo.

To further study smoothing in the case of a circle in two space, suppose ψ_R , now to be called ψ_R^δ , is a delta function centered on $\theta = 0$,

$$\psi_R^\delta(\theta) = \delta(\theta). \quad (\text{W.4.12})$$

Then, by (4.3),

$$\tilde{\psi}_R^\delta(m) = 1/(2\pi), \quad (\text{W.4.13})$$

and (4.4) takes the form

$$\psi^\delta(\rho, \theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp(im\theta). \quad (\text{W.4.14})$$

Let us work to evaluate this sum. Introduce the simplifying notation

$$\lambda = \rho/R \quad (\text{W.4.15})$$

with the understanding that, for our purposes,

$$\lambda \in [0, 1]. \quad (\text{W.4.16})$$

Correspondingly, make the definition

$$\hat{\psi}^\delta(\lambda, \theta) = \psi^\delta(\rho, \theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} \lambda^{|m|} \exp(im\theta). \quad (\text{W.4.17})$$

Observe that there is the result

$$\sum_{m=-\infty}^{m=\infty} \lambda^{|m|} \exp(im\theta) = -1 + \sum_{m=0}^{m=\infty} \lambda^m \exp(im\theta) + \sum_{m=0}^{m=\infty} \lambda^m \exp(-im\theta). \quad (\text{W.4.18})$$

Each of the series appearing on the right side of (4.18) is a geometric series, and can therefore be evaluated. We find the results

$$\sum_{m=0}^{m=\infty} \lambda^m \exp(im\theta) = 1/[1 - \lambda \exp(i\theta)], \quad (\text{W.4.19})$$

$$\sum_{m=0}^{m=\infty} \lambda^m \exp(-im\theta) = 1/[1 - \lambda \exp(-i\theta)]. \quad (\text{W.4.20})$$

Consequently,

$$\hat{\psi}^\delta(\lambda, \theta) = [1/(2\pi)]\{-1 + 1/[1 - \lambda \exp(i\theta)] + 1/[1 - \lambda \exp(-i\theta)]\}. \quad (\text{W.4.21})$$

The three terms on the right side of (4.21) can be put over a common denominator. Doing so, and recalling that

$$\cos(\theta) = (1/2)[\exp(i\theta) + \exp(-i\theta)], \quad (\text{W.4.22})$$

give the final result

$$\hat{\psi}^\delta(\lambda, \theta) = [1/(2\pi)]\{[\lambda^{-1} - \lambda]/[\lambda^{-1} + \lambda - 2\cos(\theta)]\}. \quad (\text{W.4.23})$$

We know that, by construction, the function $\hat{\psi}^\delta(\lambda, \theta)$ is harmonic. See (4.14). We will next observe *directly* that the function $\hat{\psi}^\delta(\lambda, \theta)$, as given by (4.23), is harmonic. To do so we will exploit a fact about analytic functions. Suppose f is an *analytic* function of the complex variable $z = x + iy$. (Here z is *not* a Cartesian coordinate.) Define a function $u(x, y)$ by writing

$$u(x, y) = f(z) = f(x + iy). \quad (\text{W.4.24})$$

Then, by the chain rule, it follows that

$$(\partial_x)^2 u(x, y) = f''(z) \quad (\text{W.4.25})$$

and

$$(\partial_y)^2 u(x, y) = (i^2)f''(z) = -f''(z), \quad (\text{W.4.26})$$

from which it follows that

$$[(\partial_x)^2 + (\partial_y)^2]u(x, y) = 0; \quad (\text{W.4.27})$$

the function $u(x, y)$ defined by (4.24) is harmonic. Similarly the function $v(x, y)$ defined by

$$v(x, y) = f(\bar{z}) = f(x - iy) \quad (\text{W.4.28})$$

is also harmonic. Now look at the right sides of (4.19) and (4.20). They can be rewritten in the forms

$$1/[1 - \lambda \exp(i\theta)] = 1/[1 - (1/R)(x + iy)] = 1/[1 - (1/R)z], \quad (\text{W.4.29})$$

$$1/[1 - \lambda \exp(-i\theta)] = 1/[1 - (1/R)(x - iy)] = 1/[1 - (1/R)\bar{z}]. \quad (\text{W.4.30})$$

It follows that both these functions are harmonic. Finally, we see from (4.23) that $\hat{\psi}^\delta(\lambda, \theta)$ is the sum of a constant (which is a harmonic function) and multiples of the harmonic functions in (4.29) and (4.30). Therefore $\hat{\psi}^\delta(\lambda, \theta)$ is harmonic.

Let us now, with the aid of (4.23), illustrate the general behavior of $\hat{\psi}^\delta$. Figure 4.1 displays the function $\hat{\psi}^\delta(\lambda, \theta)$ as a function of $\theta \in (-\pi, \pi)$ for various values of $\lambda \in [0, 1]$. Note that there are the relations

$$(\lambda^{-1} + \lambda) > 2 \text{ for } \lambda \in (0, 1) \quad (\text{W.4.31})$$

and

$$(\lambda^{-1} + \lambda) = 2 \text{ for } \lambda = 1. \quad (\text{W.4.32})$$

From the figure two facts are evident:

- For $\lambda \simeq 1$, $\hat{\psi}^\delta(\lambda, \theta)$ is highly peaked about $\theta = 0$ and is small for $\theta \neq 0$.
- For $\lambda \simeq 0$, $\hat{\psi}^\delta(\lambda, \theta)$ is nearly 1. Indeed, by (4.14), there is the small λ expansion

$$\hat{\psi}^\delta(\lambda, \theta) = [1/(2\pi)][1 + 2\lambda \cos(\theta) + 2\lambda^2 \cos(2\theta) + \dots]. \quad (\text{W.4.33})$$

Also, again by (4.14), there is the integral relation

$$\int_{-\pi}^{\pi} d\theta \hat{\psi}^\delta(\lambda, \theta) = 1. \quad (\text{W.4.34})$$

[Note that (4.34) is a special case of (4.10).] Putting all these facts together, we conclude that the sequence of functions $\hat{\psi}^\delta(\lambda, \theta)$ for varying λ converges to the delta function,

$$\lim_{\lambda \rightarrow 1^-} \hat{\psi}^\delta(\lambda, \theta) = \delta(\theta). \quad (\text{W.4.35})$$

We see that the delta function spike about $\theta = 0$ on the circle $\rho = R$ (which corresponds to $\lambda = 1$) becomes an ever lower and broader bump with decreasing ρ . Thus the effect of a disturbance in the potential on the $\rho = R$ circle “decays” away as one moves to circles with successively smaller values of ρ . Finally we note that, in agreement with (4.10), the “decay” we have been observing is entirely due to spreading. Indeed, the relations (4.7) and (4.9) illustrate that if the initial/boundary potential distribution is completely “spread out”, i.e. constant, then no decay occurs.

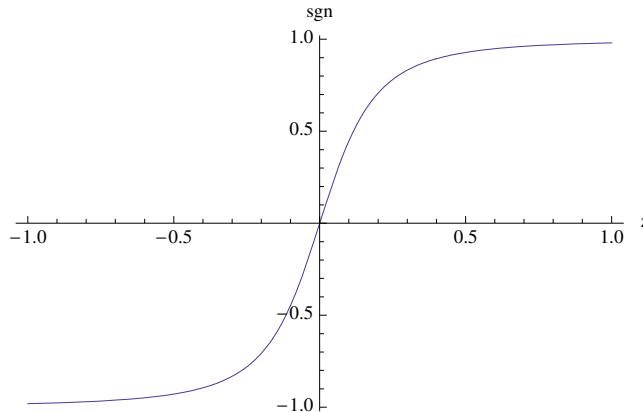


Figure W.4.1: (Place Holder) The function $\hat{\psi}^\delta(\lambda, \theta)$ as a function of θ for various values of λ .

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (4.23) to a delta function disturbance (4.12) we can derive the response to a general disturbance. With (4.23) in mind, define a kernel $G(\theta; \theta'; \lambda)$ by the rule

$$G(\theta; \theta'; \lambda) = [1/(2\pi)]\{[\lambda^{-1} - \lambda]/[\lambda^{-1} + \lambda - 2 \cos(\theta - \theta')]\}. \quad (\text{W.4.36})$$

Observe that a general disturbance $\psi_R(\theta)$ has the integral representation

$$\psi_R(\theta) = \int_{-\pi}^{\pi} d\theta' \psi_R(\theta') \delta(\theta - \theta'). \quad (\text{W.4.37})$$

It follows that the response to $\psi_R(\theta)$ is given by the integral

$$\psi(\rho, \theta) = \int_{-\pi}^{\pi} d\theta' \psi_R(\theta') G(\theta; \theta'; \lambda). \quad (\text{W.4.38})$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(\theta; \theta'; \lambda)$ is the probability that a random walk initiated at the point ρ, θ will reach the point R, θ' .

We close this section by making two calculations that will be of future use. First we will present a previous result in terms of the Cartesian coordinates x, y . From (4.33) we see that there is the result

$$\psi^\delta(x, y) = \psi^\delta(\rho, \theta) = [1/(2\pi)][1 + 2(\rho/R) \cos(\theta) + 2(\rho/R)^2 \cos(2\theta) + \dots]. \quad (\text{W.4.39})$$

There are also the relations

$$\rho \cos(\theta) = x, \quad (\text{W.4.40})$$

$$\rho^2 \cos(2\theta) = \rho^2 [\cos^2(\theta) - \sin^2(\theta)] = x^2 - y^2. \quad (\text{W.4.41})$$

It follows that there is the result

$$\psi^\delta(x, y) = [1/(2\pi)][1 + 2(1/R)x + 2(1/R^2)(x^2 - y^2) + \dots]. \quad (\text{W.4.42})$$

As a second complementary case, suppose ψ_R , now to be called ψ_R^Δ , is a delta function centered on $\theta = \pi/2$,

$$\psi_R^\Delta(\theta) = \delta(\theta - \pi/2). \quad (\text{W.4.43})$$

Then, by (4.3),

$$\tilde{\psi}_R^\Delta(m) = 1/(2\pi) \exp(-im\pi/2), \quad (\text{W.4.44})$$

and (4.4) takes the form

$$\begin{aligned} \psi^\Delta(x, y) &= \psi^\Delta(\rho, \theta) = [1/(2\pi)] \sum_{m=-\infty}^{m=\infty} (\rho/R)^{|m|} \exp[im(\theta - \pi/2)] \\ &= [1/(2\pi)\{1 + 2(\rho/R) \cos(\theta - \pi/2) + 2(\rho/R)^2 \cos[2(\theta - \pi/2)] + \dots\}] \\ &= [1/(2\pi)\{1 + 2(\rho/R) \sin(\theta) - 2(\rho/R)^2 \cos(2\theta) + \dots\}] \\ &= [1/(2\pi)\{1 + 2(1/R)y - 2(1/R^2)(x^2 - y^2) + \dots\}]. \end{aligned} \quad (\text{W.4.45})$$

Exercises

W.4.1. Verify the steps that led from (4.14) to (4.23).

W.4.2. Verify (4.29) and (4.30).

W.4.3. Verify (4.31).

W.4.4. Show that

$$\lim_{\lambda \rightarrow 1^-} \hat{\psi}^\delta(\lambda, \theta) = 0 \text{ for } \theta \neq 0. \quad (\text{W.4.46})$$

Show that

$$\lim_{\lambda \rightarrow 1^-} \hat{\psi}^\delta(\lambda, 0) = +\infty. \quad (\text{W.4.47})$$

W.4.5. Verify directly the relation (4.34) using (4.23). Verify (4.38) for the case (4.7).

W.5 The Circular Cylinder in Three Space

Consider in x, y, z space a circular cylinder of radius R centered on the z axis, and suppose a potential $\psi_R(\phi, z)$ is specified on this cylinder. [Here we have used the cylindrical coordinates ρ, ϕ , and z specified by the rules (15.2.12) through (15.2.16).] More specifically, write

$$\psi(x, y, z) = \psi(\rho, \phi, z) \quad (\text{W.5.1})$$

and

$$\psi(R, \phi, z) = \psi_R(\phi, z). \quad (\text{W.5.2})$$

Given the boundary potential $\psi_R(\phi, z)$, we wish to find the interior harmonic function (solution to Laplace's equation) $\psi(\rho, \phi, z)$ associated with this boundary potential.

From (15.3.7) we know that the most general ψ that is harmonic and finite within the cylinder is of the form

$$\psi(\rho, \phi, z) = \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dk G_m(k) \exp(ikz) \exp(im\phi) I_m(k\rho). \quad (\text{W.5.3})$$

Next define the *double Fourier transform* $\tilde{\psi}(R, m', k')$ of the boundary potential by the rule

$$\tilde{\psi}(R, m', k') = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dz \exp(-ik'z) \int_0^{2\pi} d\phi \exp(-im'\phi) \psi_R(\phi, z). \quad (\text{W.5.4})$$

See (17.2.2). Then we know from (17.2.5) that

$$G_m(k) = \tilde{\psi}(R, m, k) / I_m(kR). \quad (\text{W.5.5})$$

Consequently, the desired ψ is given by the relation

$$\psi(\rho, \phi, z) = \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk \exp(ikz) \tilde{\psi}(R, m, k) [I_m(k\rho) / I_m(kR)]. \quad (\text{W.5.6})$$

Let us verify that our criteria have been met. By construction the ψ given by (5.6) is a superposition of the functions $\exp(ikz) \exp(im\phi) I_m(k\rho)$ and therefore is harmonic. Also it has the property

$$\psi(R, \phi, z) = \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk \exp(ikz) \tilde{\psi}(R, m, k) = \psi_R(\phi, z). \quad (\text{W.5.7})$$

We have found the solution to Laplace's equation in the circular cylinder in three space having boundary $\rho = R$ and the boundary value $\psi_R(\phi, z)$.

We also observe, in passing, two facts. First, suppose ψ_R , now to be called ψ_R^c , is a *constant* function,

$$\psi_R^c(\phi, z) = c. \quad (\text{W.5.8})$$

Then, by (5.4),

$$\tilde{\psi}^c(R, m', k') = c\delta(k')\delta_{m',0}. \quad (\text{W.5.9})$$

It follows from (5.6) that there is the relation

$$\psi^c(\rho, \phi, z) = c. \quad (\text{W.5.10})$$

As expected, if ψ is constant on the cylinder boundary $\rho = R$, it will have the same constant value everywhere within the cylinder. Second, for any solution, there is the relation

$$\begin{aligned} \int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz \psi(\rho, \phi, z) &= (2\pi)^2 \sum_{m=-\infty}^{\infty} \delta_{m,0} \int_{-\infty}^{\infty} dk \delta(k) \tilde{\psi}(R, m, k) [I_m(k\rho)/I_m(kR)] \\ &= (2\pi)^2 \tilde{\psi}(R, 0, 0) = \int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz \psi_R(\phi, z). \end{aligned} \quad (\text{W.5.11})$$

That is, the $d\phi dz$ integral of $\psi(\rho, \phi, z)$ over any cylinder of constant ρ is independent of ρ .

We are now prepared to address the general subject of smoothing. We will examine the asymptotic behavior of the kernel $K(m, k; \rho, R)$, defined by the rule

$$K(m, k; \rho, R) = [I_m(k\rho)/I_m(kR)], \quad (\text{W.5.12})$$

that appears in (5.6). Note that, because

$$I_{-m}(w) = I_m(w) \quad (\text{W.5.13})$$

and

$$I_m(-w) = (-1)^m I_m(w), \quad (\text{W.5.14})$$

the kernel $K(m, k; \rho, R)$ is evidently an *even* function of both m and k . Therefore we only need consider the cases $m \geq 0$ and $k \geq 0$. Finally, by (15.3.11), we see that

$$K(m, 0; \rho, R) = (\rho/R)^{|m|}. \quad (\text{W.5.15})$$

For fixed m and large w the Bessel functions $I_m(w)$ have the asymptotic property

$$|I_m(w)| \simeq (1/\sqrt{2\pi w}) \exp(w) \text{ as } w \rightarrow \infty. \quad (\text{W.5.16})$$

Consequently, for fixed m , there is the asymptotic relation

$$K(m, k; \rho, R) \simeq (\sqrt{R/\rho}) \exp[k(\rho - R)] \text{ as } k \rightarrow \infty. \quad (\text{W.5.17})$$

We see that, for each fixed m , there is smoothing/damping in the longitudinal variable z when $\rho < R$. Note that this smoothing is analogous to that for the line in two space and the plane in three space.

For fixed w and large m the Bessel functions $I_m(w)$ have the asymptotic property

$$\begin{aligned} |I_m(w)| &\simeq (1/\sqrt{2\pi m})[(e|w|)/(2m)]^m \\ &\simeq (1/2)^m [\sqrt{2\pi m}(m/e)^m]^{-1} |w|^m \\ &\simeq (1/2)^m (1/m!) |w|^m \text{ as } m \rightarrow \infty. \end{aligned} \quad (\text{W.5.18})$$

Here we have used the Stirling large m approximation

$$m! \simeq \sqrt{2\pi m} (m/e)^m, \quad (\text{W.5.19})$$

which is actually already quite accurate for $m \geq 2$.¹ Consequently, for fixed k , there is the asymptotic relation

$$K(m, k; \rho, R) \simeq (\rho/R)^m \text{ as } m \rightarrow \infty. \quad (\text{W.5.20})$$

We see that, for each fixed k , there is smoothing/damping in the angular variable ϕ when $\rho < R$. Note that this smoothing is analogous to that for the circle in two space.

With regard to angular smoothing there is also the consideration that the angular Fourier transform filters out all angular Fourier modes save for the one of interest. Moreover, the disturbance in the angular Fourier mode of interest produced by an error in any given grid-point value is suppressed by $1/N$ where N is the number of sampling points used in the discrete angular Fourier transform.

What happens if k and m increase simultaneously? This is a more difficult question. We will explore the case were k and m are proportional,

$$k = \lambda m \quad (\text{W.5.21})$$

where λ is some proportionality constant having the dimensions of inverse length.

If $w = \tau m$, where τ is some proportionality constant, there is the uniform doubly asymptotic relation

$$I_m(\tau m) \simeq (1/\sqrt{2\pi m})[1/(1+\tau^2)^{1/4}] \exp(m\eta) \text{ as } m \rightarrow \infty. \quad (\text{W.5.22})$$

Here η is a function of τ given by the relation

$$\eta(\tau) = \sqrt{1+\tau^2} + \log[\tau/(1+\sqrt{1+\tau^2})]. \quad (\text{W.5.23})$$

We will use the assumption (5.21) and the result (5.22) to estimate $I_m(\lambda m \rho)$ and $I_m(\lambda m R)$, the numerator and denominator appearing in (5.9).

For the numerator define a quantity $\hat{\tau}$ by the rule

$$\hat{\tau} = \lambda \rho, \quad (\text{W.5.24})$$

and for the denominator define a quantity $\check{\tau}$ by the rule

$$\check{\tau} = \lambda R. \quad (\text{W.5.25})$$

¹Note that the final result in (5.18) also follows from retaining only the $\ell = 0$ term in (15.3.11).

(Note that both $\hat{\tau}$ and $\check{\tau}$ are dimensionless.) In terms of these definitions we have, as a consequence of (5.22), the large m results

$$I_m(\lambda m \rho) \simeq (1/\sqrt{2\pi m}) [1/(1 + \hat{\tau}^2)^{1/4}] \exp(m\hat{\eta}), \quad (\text{W.5.26})$$

$$I_m(\lambda m R) \simeq (1/\sqrt{2\pi m}) [1/(1 + \check{\tau}^2)^{1/4}] \exp(m\check{\eta}). \quad (\text{W.5.27})$$

Here we have used the notation

$$\hat{\eta} = \eta(\hat{\tau}), \quad (\text{W.5.28})$$

$$\check{\eta} = \eta(\check{\tau}). \quad (\text{W.5.29})$$

It follows that there is the large m result

$$K(m, \lambda m; \rho, R) \simeq [(1 + \check{\tau}^2)^{1/4}/(1 + \hat{\tau}^2)^{1/4}] \exp[m(\hat{\eta} - \check{\eta})] \text{ as } m \rightarrow \infty. \quad (\text{W.5.30})$$

We see that there is *exponential* smoothing/damping if

$$\hat{\eta} < \check{\eta}. \quad (\text{W.5.31})$$

When does the smoothing condition (5.31) hold? Let us examine the function $\eta(\tau)$ given by (5.23). Its behavior is displayed in Figure 5.1. Evidently, for $\tau \geq 0$, it appears to be *monotonically increasing*. This surmise is proved in Exercise 5.1. Consequently, (5.31) holds if $\hat{\tau} < \check{\tau}$ and hence $\rho < R$. We conclude, assuming $\rho < R$, that there is exponential smoothing/damping as one goes out in any direction from the origin in the $m k$ plane; and the damping rate depends on the direction. For example, Figure 5.2 displays $K(m, k; \rho, R)$ as function of m and k for the case $\rho = 2$ cm and $R = 2.5$ cm.

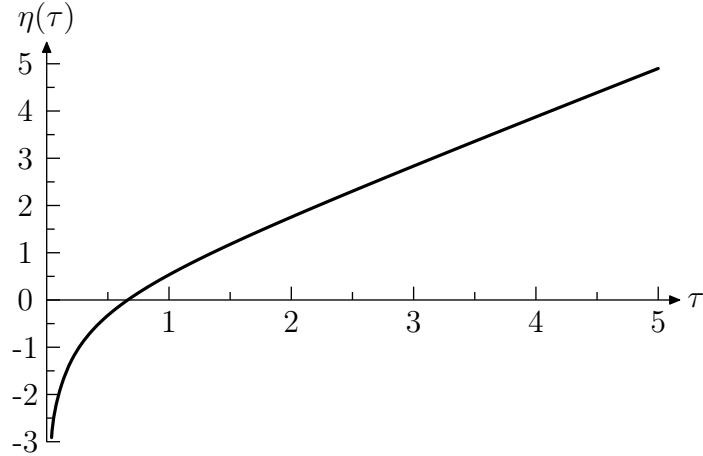


Figure W.5.1: The function $\eta(\tau)$. It appears to be monotonically increasing.

To further study smoothing in the case of a circular cylinder in three space, suppose ψ_R , now to be called ψ_R^δ , is a delta function centered on $(\phi, z) = (0, 0)$,

$$\psi_R^\delta(\phi, z) = \delta(\phi)\delta(z). \quad (\text{W.5.32})$$

Then, by (5.4),

$$\tilde{\psi}^\delta(R, m', k') = [1/(2\pi)]^2, \quad (\text{W.5.33})$$

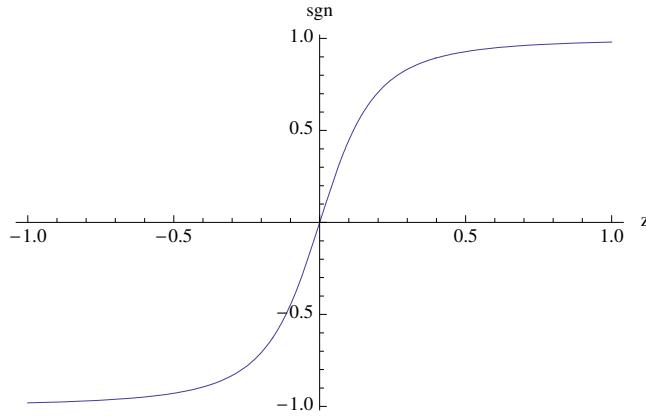


Figure W.5.2: (Place Holder) The kernel $K(m, k; \rho, R)$ as function of m and k for the case $\rho = 2$ cm and $R = 2.5$ cm. The quantity k has units of inverse centimeters.

and (5.6) takes the form

$$\psi^\delta(\rho, \phi, z) = [1/(2\pi)]^2 \sum_{m=-\infty}^{\infty} \exp(im\phi) \int_{-\infty}^{\infty} dk \exp(ikz) [I_m(k\rho)/I_m(kR)]. \quad (\text{W.5.34})$$

Our task now is to study the properties of $\psi^\delta(\rho, \phi, z)$. We begin by observing that, as a consequence of (5.11), there is the integral relation

$$\int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz \psi^\delta(\rho, \phi, z) = (2\pi)^2 \tilde{\psi}^\delta(R, 0, 0) = 1. \quad (\text{W.5.35})$$

To proceed further, and in analogy to what was done in previous sections for ψ^δ , it would be ideal if the representation (5.34) could be evaluated analytically in terms of known functions. However, this seems to be a difficult. What we can do is to evaluate (5.34) numerically for various values of ρ and R . Figure 5.3 shows $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 2$ cm and $R = 2.5$ cm. And Figure 5.4 shows $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 1$ cm and $R = 2.5$ cm. Evidently $\psi^\delta(\rho, \phi, z)$ falls off for large z . Moreover, it is smaller and less peaked about $(\phi, z) = (0, 0)$ for the smaller value of ρ . The effect of a disturbance in the potential at the point $(\phi, z) = (0, 0)$ “decays” away as one moves to cylinders with successively smaller values of ρ . Conversely, as ρ increases, the sequence of functions $\psi^\delta(\rho, \phi, z)$ for varying ρ converges to the delta function,

$$\lim_{\rho \rightarrow R^-} \psi^\delta(\rho, \phi, z) = \delta(\phi)\delta(z). \quad (\text{W.5.36})$$

Finally we note that, in agreement with (5.11), the “decay” we have been observing is entirely due to spreading. Indeed, the relations (5.8) and (5.10) illustrate that if the initial/boundary potential distribution is completely “spread out”, i.e. constant, then no decay occurs.

At this point we might wonder how fast the effect of a disturbance falls off as a function of z . Figure 5.5 shows $\psi^\delta(1, 0, z)$ as a function of z when $R = 2.5$ cm. We can also study the on-axis case $\rho = 0$ for which (5.34) takes the simpler form

$$\psi^\delta(0, \phi, z) = [1/(2\pi)]^2 \int_{-\infty}^{\infty} dk \exp(ikz) [1/I_0(kR)]. \quad (\text{W.5.37})$$

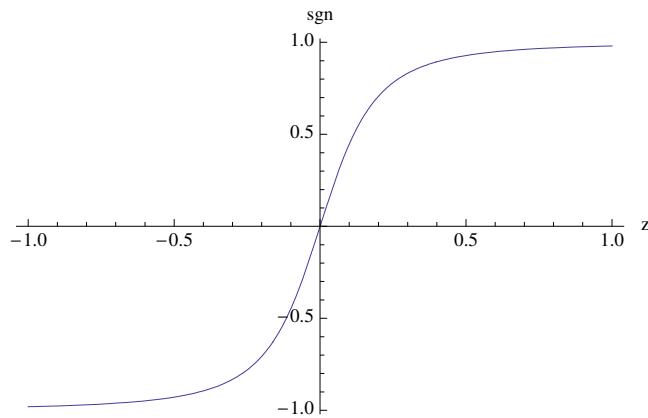


Figure W.5.3: (Place Holder) The function $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 2$ cm and $R = 2.5$ cm.

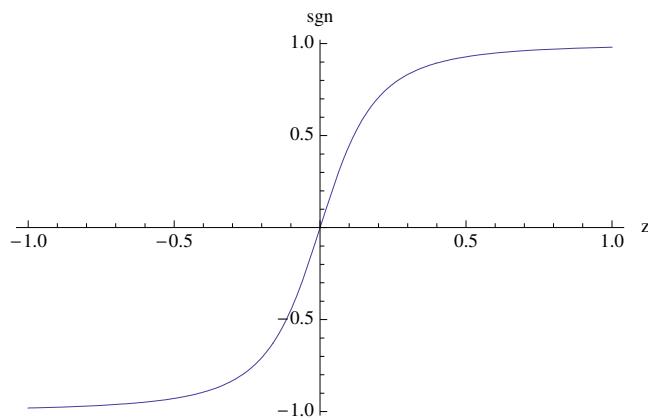


Figure W.5.4: (Place Holder) The function $\psi^\delta(\rho, \phi, z)$ as a function of ϕ and z when $\rho = 1$ cm and $R = 2.5$ cm.

Make the change of variables $\lambda = kR$. So doing brings (5.37) to the form

$$\psi^\delta(0, \phi, z) = [1/(2\pi)]^2 (1/R) \int_{-\infty}^{\infty} d\lambda \exp(i\lambda z/R) / I_0(\lambda). \quad (\text{W.5.38})$$

We have already studied the integral appearing on the right side of (5.38). Reference to (21.1.37) shows that there is the result

$$\psi^\delta(0, \phi, z) = [1/(2\pi)](1/R)F(z/R, 0). \quad (\text{W.5.39})$$

It follows from the work of Section 21.1.3, see Figure 21.1.2, that there is the asymptotic behavior

$$\psi^\delta(0, \phi, z) \propto \exp[-\pi|z|/(2R)] \text{ as } |z| \rightarrow \infty. \quad (\text{W.5.40})$$

When viewed from on axis the effect of a disturbance falls off exponentially. Reference to Figure 5.5 shows that there is a similar rapid fall off with z in the off-axis case.

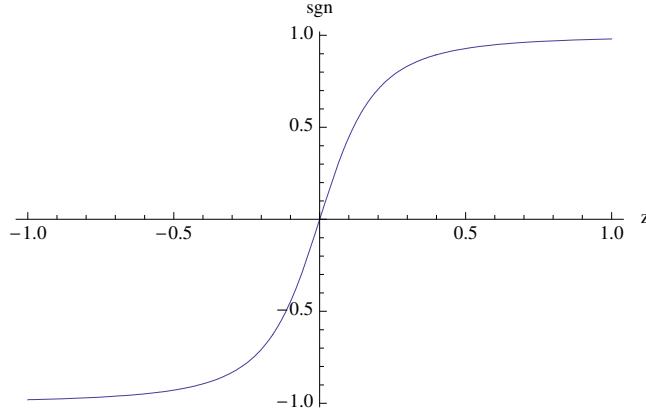


Figure W.5.5: (Place Holder) The function $\psi^\delta(1, 0, z)$ as a function of z when $R = 2.5$ cm.

We have again seen an example of how the effects of a local disturbance in the potential diminish with distance from the disturbance.

From the response (5.34) to a delta function disturbance (5.32) we can derive the response to a general disturbance. With (5.34) in mind, define a kernel $G(\phi, z; \phi', z'; \rho)$ by the rule

$$G(\phi, z; \phi', z'; \rho) = \psi^\delta(\rho, \phi - \phi', z - z'). \quad (\text{W.5.41})$$

Observe that a general disturbance $\psi_R(\phi, z)$ has the integral representation

$$\psi_R(\phi, z) = \int_{-\pi}^{\pi} d\phi' \int_{-\infty}^{\infty} dz' \psi_R(\phi', z') \delta(\phi - \phi') \delta(z - z'). \quad (\text{W.5.42})$$

It follows that the response to $\psi_R(\phi, z)$ is given by the integral

$$\psi(\rho, \phi, z) = \int_{-\pi}^{\pi} d\phi' \int_{-\infty}^{\infty} dz' \psi_R(\phi', z') G(\phi, z; \phi', z'; \rho). \quad (\text{W.5.43})$$

For what it's worth we remark that, according to the connection between Laplace and Monte Carlo, the quantity $G(\phi, z; \phi', z'; \rho)$ is the probability that a random walk initiated at the point ρ, ϕ, z will reach the point R, ϕ', z' .

Exercises

W.5.1. The purpose of this exercise is to prove that $\eta(\tau)$ is a monotonically increasing function of τ . We begin by observing that the $\sqrt{1+\tau^2}$ term just to the right of the equal sign in (5.23) is a monotonically increasing function of τ , and we know that the log function is a monotonically increasing function of its argument. It remains to be shown that the function $f(\tau)$ defined by

$$f(\tau) = \tau / (1 + \sqrt{1 + \tau^2}), \quad (\text{W.5.44})$$

the argument of the log function, is monotonically increasing. If this can be verified, then $\eta(\tau)$ is a monotonically increasing function of τ .

To complete the proof, show that

$$\begin{aligned} f'(\tau) &= 1/(1 + \sqrt{1 + \tau^2}) - (\tau^2 / \sqrt{1 + \tau^2}) / (1 + \sqrt{1 + \tau^2})^2 \\ &= [1/(1 + \sqrt{1 + \tau^2})^2][(1 + \sqrt{1 + \tau^2}) - (\tau^2 / \sqrt{1 + \tau^2})] \\ &= [1/(1 + \sqrt{1 + \tau^2})^2](1/\sqrt{1 + \tau^2})(\sqrt{1 + \tau^2} + 1 + \tau^2 - \tau^2) \\ &= [1/(1 + \sqrt{1 + \tau^2})^2](1/\sqrt{1 + \tau^2})(\sqrt{1 + \tau^2} + 1) \\ &= [1/(1 + \sqrt{1 + \tau^2})](1/\sqrt{1 + \tau^2}). \end{aligned} \quad (\text{W.5.45})$$

Evidently all factors on the far right side of (5.45) are positive, and therefore $f(\tau)$ is monotonically increasing.

W.5.2. Verify (5.15) given (15.3.11).

W.5.3. Verify (5.17) given (5.16).

W.5.4. Verify (5.20) given (5.18).

W.5.5. Verify (5.30) given (5.22).

W.5.6. Verify (5.38) given (5.37).

W.5.7. Compare the fall off in the case of a plane in three space given by (3.32) with the fall off in the case of a circular cylinder in three space given (5.40). Explain why the fall off is so much more rapid in the case of a cylinder.

W.6 The Ellipse in Two Space

For the ellipse in two space let us employ the coordinates given by (17.4.1) and (17.4.2) and illustrated in Figure 17.4.2. Then, upon writing the relation

$$\psi(x, y) = \psi(u, v), \quad (\text{W.6.1})$$

we find from (17.4.12) that

$$\nabla^2 \psi = (1/f^2)[\cosh^2(u) - \cos^2(v)]^{-1}[(\partial_u)^2 + (\partial_v)^2]\psi. \quad (\text{W.6.2})$$

Consequently, ψ will be harmonic provided it satisfies the relation

$$[(\partial_u)^2 + (\partial_v)^2]\psi = 0. \quad (\text{W.6.3})$$

In analogy with (17.4.35) through (17.4.37) let us define functions $c_n(v)$ and $s_n(v)$ by the rules

$$c_0(v) = 1/\sqrt{2}, \quad (\text{W.6.4})$$

$$c_n(v) = \cos(nv) \text{ for } n \geq 1, \quad (\text{W.6.5})$$

$$s_0(v) = 0, \quad (\text{W.6.6})$$

$$s_n(v) = \sin(nv) \text{ for } n \geq 1. \quad (\text{W.6.7})$$

Evidently they form a complete orthogonal set and are normalized so that

$$\int_0^{2\pi} dv c_m(v) c_n(v) = \pi \delta_{mn}, \quad (\text{W.6.8})$$

$$\int_0^{2\pi} dv s_m(v) s_n(v) = \pi \delta_{mn}, \quad (\text{W.6.9})$$

$$\int_0^{2\pi} dv c_m(v) s_n(v) = 0. \quad (\text{W.6.10})$$

Also, in analogy with (17.4.70) and (7.4.71), let us define functions $C_n(u)$ and $S_n(u)$ by the rules

$$C_n(u) = c_n(iu), \quad (\text{W.6.11})$$

$$S_n(u) = -is_n(iu). \quad (\text{W.6.12})$$

In view of (6.4) through (6.7) we have the results

$$C_0(u) = 1/\sqrt{2}, \quad (\text{W.6.13})$$

$$C_n(u) = \cosh(nu) \text{ for } n \geq 1, \quad (\text{W.6.14})$$

$$S_0(u) = 0, \quad (\text{W.6.15})$$

$$S_n(u) = \sinh(nu) \text{ for } n \geq 1. \quad (\text{W.6.16})$$

Evidently the functions $C_n(u)$ and $S_n(u)$ are entire functions of u , and the functions $c_n(v)$ and $s_n(v)$ are entire functions of v . However, they are not entire functions of x and y because of the singularities described in Exercise 17.4.2.

It is easily verified that functions $\psi_n^c(u, v)$ and $\psi_n^s(u, v)$ of the form

$$\psi_n^c(u, v) \propto C_n(u)c_n(v), \quad (\text{W.6.17})$$

$$\psi_n^s(u, v) \propto S_n(u)s_n(v) \quad (\text{W.6.18})$$

satisfy (6.3) and are therefore harmonic functions. We claim that they are also polynomial, and therefore entire analytic, functions of x and y . For example, there are the relations

$$C_0(u)c_0(v) = 1/2, \quad (\text{W.6.19})$$

$$S_0(u)S_0(v) = 0, \quad (W.6.20)$$

$$C_1(u)C_1(v) = \cosh(u)\cos(v) = x/f, \quad (W.6.21)$$

$$S_1(u)S_1(v) = \sinh(u)\sin(v) = y/f, \quad (W.6.22)$$

$$\begin{aligned} C_2(u)C_2(v) &= \cosh(2u)\cos(2v) = (1/2)\cosh(2u)\cos(2v) + (1/2)\cosh(2u)\cos(2v) \\ &= (1/2)[2\cosh^2(u)-1][2\cos^2(v)-1] \\ &\quad -(1/2)[2\sinh^2(u)+1][2\sin^2(v)-1] \\ &= (1/2)\{4\cosh^2(u)\cos^2(v)-2[\cosh^2(u)+\cos^2(v)]+1\} \\ &\quad -(1/2)\{4\sinh^2(u)\sin^2(v)-2[\sinh^2(u)-\sin^2(v)]-1\} \\ &= 2\cosh^2(u)\cos^2(v)-2\sinh^2(u)\sin^2(v) \\ &\quad -\cosh^2(u)+\sinh^2(u)-\cos^2(v)-\sin^2(v)+1/2+1/2 \\ &= 2\cosh^2(u)\cos^2(v)-2\sinh^2(u)\sin^2(v)-1 \\ &= 2(x^2-y^2)/f^2-1. \end{aligned} \quad (W.6.23)$$

$$\begin{aligned} S_2(u)S_2(v) &= \sinh(2u)\sin(2v) = 4\sinh(u)\cosh(u)\sin(v)\cos(v) \\ &= 4xy/f^2. \end{aligned} \quad (W.6.24)$$

For a general proof for all n , see Exercise 6.3.

With the above background in mind, consider the ellipse $u = U$ and suppose a potential $\psi_U(v)$ is specified on this ellipse. Since the functions $c_n(v)$ and $s_n(v)$ form a complete set, we may make the expansion

$$\psi_U(v) = \sum_{n=0}^{\infty} \tilde{\psi}_U^c(n)c_n(v) + \sum_{n=1}^{\infty} \tilde{\psi}_U^s(n)s_n(v) \quad (W.6.25)$$

with

$$\tilde{\psi}_U^c(n) = (1/\pi) \int_0^{2\pi} dv c_n(v) \psi_U(v), \quad (W.6.26)$$

$$\tilde{\psi}_U^s(n) = (1/\pi) \int_0^{2\pi} dv s_n(v) \psi_U(v). \quad (W.6.27)$$

Now make the Ansatz

$$\psi(u, v) = \sum_{n=0}^{\infty} \tilde{\psi}_U^c(n)[C_n(u)/C_n(U)]c_n(v) + \sum_{n=1}^{\infty} \tilde{\psi}_U^s(n)[S_n(u)/S_n(U)]s_n(v). \quad (W.6.28)$$

By construction $\psi(u, v)$ is a harmonic function, and also has the property

$$\psi(U, v) = \psi_U(v). \quad (W.6.29)$$

We have found the solution to Laplace's equation in the ellipse having boundary $u = U$ and the boundary value $\psi_U(v)$.

Note that the operation (6.28) is smoothing for $u < U$. Indeed, according (6.14) and (6.16), there are the asymptotic results

$$C_n(u) \propto \exp(nu) \text{ as } n \rightarrow \infty, \quad (\text{W.6.30})$$

$$S_n(u) \propto \exp(nu) \text{ as } n \rightarrow \infty. \quad (\text{W.6.31})$$

It follows that there are the asymptotic results

$$[C_n(u)/C_n(U)] \propto \exp[-n(U-u)] \text{ as } n \rightarrow \infty, \quad (\text{W.6.32})$$

$$[S_n(u)/S_n(U)] \propto \exp[-n(U-u)] \text{ as } n \rightarrow \infty. \quad (\text{W.6.33})$$

Consequently, there is exponential smoothing provided $u < U$.

We also observe, in passing, two facts. First, suppose ψ_U , now to be called ψ_U^d , is a constant function with value d ,

$$\psi_U^d(v) = d. \quad (\text{W.6.34})$$

Then, by (6.26) and (6.27),

$$\tilde{\psi}_U^{dc}(n) = d\sqrt{2}\delta_{n,0}, \quad (\text{W.6.35})$$

$$\tilde{\psi}_U^{ds}(n) = 0. \quad (\text{W.6.36})$$

It follows from (6.28) that there is the relation

$$\psi^d(u, v) = d. \quad (\text{W.6.37})$$

As expected, if ψ is constant on the boundary $u = U$, it will have the same constant value inside the ellipse. Second, for any solution, we find from (6.28) that there is the relation

$$\int_0^{2\pi} dv \psi(u, v) = (\pi\sqrt{2}) \sum_{n=0}^{\infty} \tilde{\psi}_U^c(n) [C_n(u)/C_n(U)] \delta_{n,0} = (\pi\sqrt{2}) \tilde{\psi}_U^c(0) = \int_0^{2\pi} dv \psi_U(v). \quad (\text{W.6.38})$$

That is, the dv integral of $\psi(u, v)$ over any ellipse of constant u is independent of u .

To further study smoothing in the case of an ellipse in two space, suppose ψ_U , now to be called ψ_U^δ , is a delta function centered on $v = 0$,

$$\psi_U^\delta(v) = \delta(v). \quad (\text{W.6.39})$$

Then, by (6.26) and (6.27),

$$\tilde{\psi}_U^{\delta c}(0) = 1/(\pi\sqrt{2}), \quad (\text{W.6.40})$$

$$\tilde{\psi}_U^{\delta c}(n) = 1/\pi \text{ for } n \geq 1, \quad (\text{W.6.41})$$

$$\tilde{\psi}_U^{\delta s}(n) = 0, \quad (\text{W.6.42})$$

and (6.28) takes the form

$$\begin{aligned} \psi^\delta(u, v) &= \sum_{n=0}^{\infty} \tilde{\psi}_U^{\delta c}(n) [1/C_n(U)] C_n(u) c_n(v) \\ &= [1/(\pi\sqrt{2})] [1/C_0(U)] C_0(u) c_0(v) + (1/\pi) \sum_{n=1}^{\infty} [1/C_n(U)] C_n(u) c_n(v). \\ &= [1/(2\pi)] + (1/\pi) \sum_{n=1}^{\infty} [1/\cosh(nU)] \cosh(nu) \cos(nv). \end{aligned} \quad (\text{W.6.43})$$

Figure 6.1 shows $\psi^\delta(u, v)$ as a function of v for various values of u . For this example $U = 0.5$. We see that the delta function spike about $v = 0$ on the ellipse $u = U$ becomes an ever lower and broader bump with decreasing u . Thus the effect of a disturbance in the potential on the $u = U$ ellipse “decays” away as one moves to ellipses with successively smaller values of u . Finally we note that, in agreement with (6.38), the “decay” we have been observing is entirely due to spreading. Indeed, the relations (6.34) and (6.37) illustrate that if the initial/boundary potential distribution is completely “spread out”, i.e. constant, then no decay occurs.

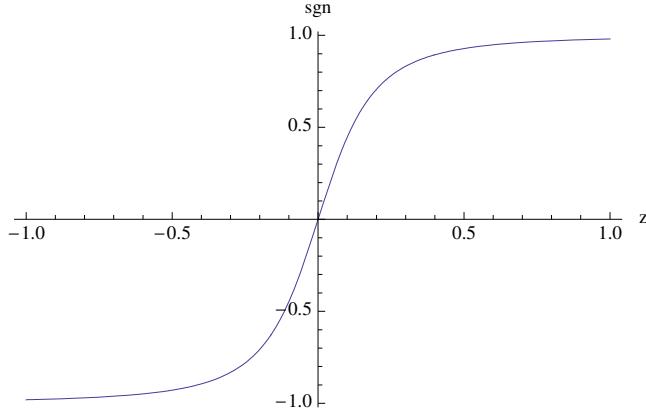


Figure W.6.1: (Place Holder) The function $\psi^\delta(u, v)$ as a function of v for various values of u when $U = 0.5$ and therefore $\tanh(U) = 0.46 \dots$.

For future use let us examine the behavior of $\psi^\delta(u, v)$ about the origin. Evidently there is the expansion

$$\begin{aligned}
\psi^\delta(u, v) &= \sum_{n=0}^2 \tilde{\psi}_U^{\delta c}(n)[1/C_n(U)]C_n(u)c_n(v) + \dots \\
&= [1/(\pi\sqrt{2})][1/C_0(U)]C_0(u)c_0(v) + (1/\pi)[1/C_1(U)]C_1(u)c_1(v) \\
&\quad + (1/\pi)[1/C_2(U)]C_2(u)c_2(v) + \dots \\
&= [1/(\pi\sqrt{2})][1/\sqrt{2}] + (1/\pi)[1/C_1(U)](x/f) \\
&\quad + (1/\pi)[1/C_2(U)][2(x^2 - y^2)/f^2 - 1] + \dots \\
&= 1/(2\pi) + (1/\pi)[1/\cosh(U)](x/f) \\
&\quad + (1/\pi)[1/\cosh(2U)][2(x^2 - y^2)/f^2 - 1] + \dots. \tag{W.6.44}
\end{aligned}$$

Next, suppose ψ_U , now to be called ψ_U^Δ , is a delta function centered on $v = \pi/2$,

$$\psi_U^\Delta(v) = \delta(v - \pi/2). \tag{W.6.45}$$

Then, by (6.26) and (6.27), the first few Fourier coefficients results are

$$\tilde{\psi}_U^{\Delta c}(0) = 1/(\pi\sqrt{2}), \tag{W.6.46}$$

$$\tilde{\psi}_U^{\Delta c}(1) = 0, \tag{W.6.47}$$

$$\tilde{\psi}_U^{\Delta c}(2) = -1/\pi; \quad (\text{W.6.48})$$

$$\tilde{\psi}_U^{\Delta s}(0) = 0, \quad (\text{W.6.49})$$

$$\tilde{\psi}_U^{\Delta s}(1) = 1/\pi, \quad (\text{W.6.50})$$

$$\tilde{\psi}_U^{\Delta s}(2) = 0. \quad (\text{W.6.51})$$

Correspondingly, (6.28) now takes the form

$$\begin{aligned} \psi^\Delta(u, v) &= \sum_{n=0}^2 \tilde{\psi}_U^{\Delta c}(n)[C_n(u)/C_n(U)]c_n(v) + \dots \\ &\quad + \sum_{n=1}^2 \tilde{\psi}_U^{\Delta s}(n)[S_n(u)/S_n(U)]s_n(v) + \dots \\ &= [1/(\pi\sqrt{2})][1/C_0(U)]C_0(u)c_0(v) - (1/\pi)[1/C_2(U)]C_2(u)c_2(v) + \dots \\ &\quad + (1/\pi)[1/S_1(U)]S_1(u)s_1(v) + \dots \\ &= 1/(2\pi) + (1/\pi)[1/\sinh(U)](y/f) \\ &\quad - (1/\pi)[1/\cosh(2U)][2(x^2 - y^2)/f^2 - 1] + \dots \end{aligned} \quad (\text{W.6.52})$$

Let us compare the terms in ψ^δ given by (6.44) with the terms in ψ^Δ given by (6.52). In particular, let us begin by making the comparison

$$[1/\cosh(U)](x/f) \text{ versus } [1/\sinh(U)](y/f). \quad (\text{W.6.53})$$

For ψ^δ the delta function disturbance in the boundary potential is made at the point $(x, y) = (x^\delta, 0)$ with

$$x^\delta = f \cosh(U). \quad (\text{W.6.54})$$

See (17.4.1) and Figure 17.4.2. And for ψ^Δ the delta function disturbance in the boundary potential is made at the point $(x, y) = (0, y^\Delta)$ with

$$y^\Delta = f \sinh(U). \quad (\text{W.6.55})$$

See (17.4.2). With this observation in mind, we see that the comparison (6.53) can be rewritten in the form

$$x/x^\delta \text{ versus } y/y^\Delta. \quad (\text{W.6.56})$$

Now suppose the bounding ellipse $u = U$ has been chosen so that

$$y^\Delta < x^\delta. \quad (\text{W.6.57})$$

See Figure 17.4.3. Then, according to (6.56), near the origin the effect of a disturbance at $(x^\delta, 0)$ is diminished from the effect of a disturbance at $(0, y^\Delta)$ by a factor of

$$y^\Delta/x^\delta = \tanh(U). \quad (\text{W.6.58})$$

By contrast, comparison of (4.42) and (4.45) shows, as expected, there is no such effect in the case of a circular boundary. Our findings are in accord with the expectation described

in Section 17.4.1 to the effect that, for wigglers or dipoles with small gaps and wide pole faces, use of a cylinder with elliptical cross section should give improved error insensitivity.

Let us also examine the next higher-order (and non constant) terms in ψ^δ and ψ^Δ which, according to (6.44) and (6.52), are

$$\pm (1/\pi)[1/\cosh(2U)][2(x^2 - y^2)/f^2]. \quad (\text{W.6.59})$$

Note that there is the relation

$$f^2 \cosh(2U) = f^2[\cosh^2(U) + \sinh^2(U)] = (x^\delta)^2 + (y^\Delta)^2. \quad (\text{W.6.60})$$

Thus, (6.59) can also be written in the form

$$\pm (1/\pi)(x^2 - y^2)/\{(1/2)[(x^\delta)^2 + (y^\Delta)^2]\}. \quad (\text{W.6.61})$$

The comparable term for the circle in two-space case, that given in (4.42) or (4.49), is

$$\pm (1/\pi)(x^2 - y^2)/R^2. \quad (\text{W.6.62})$$

Thus, to contrast the use of a circle with the use of an ellipse, we should make the comparison

$$R^2 \text{ versus } \{(1/2)[(x^\delta)^2 + (y^\Delta)^2]\}. \quad (\text{W.6.63})$$

Moreover, when contrasting the use of a circle to the use of an ellipse, it is reasonable to presume that the ellipse just contains the circle so that

$$y^\Delta = R. \quad (\text{W.6.64})$$

In this case (6.63) becomes

$$(1/2)(y^\Delta)^2 \text{ versus } (1/2)(x^\delta)^2. \quad (\text{W.6.65})$$

In view of (6.57) the left term in the comparison (6.65) is smaller than the term on the right. Correspondingly, the denominator in (6.61) is larger than that in (6.62) thereby again illustrating that the use of a cylinder with elliptical cross section should give improved error insensitivity.

Exercises

W.6.1. Verify that the functions (6.17) and (6.18) are harmonic.

W.6.2. Since deriving the result (6.23) involved considerable algebra, it is useful to check a few specific cases. Consider the points $(x, y) = (0, 0)$ and $(x, y) = (\pm f, 0)$ for which $(u, v) = (0, \pi/2 \text{ or } 3\pi/2)$ and $(u, v) = (0, 0 \text{ or } \pi)$. See Figure 17.4.3. Verify that (6.23) holds at these points.

W.6.3. The purpose of this exercise is to prove the claim that, for all n , functions ψ_n^c and ψ_n^s of the form (6.17) and (6.18) are polynomial functions of x and y . Recall (17.4.7). Show that from this relation it follows that

$$(x + iy)^n / f^n = [\cosh(w)]^n. \quad (\text{W.6.66})$$

Next verify the expansion

$$\begin{aligned} [\cosh(w)]^n &= (1/2^n)\{\exp(w) + \exp(-w)\}^n \\ &= (1/2^n)\{\exp(nw) + n\exp[(n-2)w] + [n(n-1)/2!] \exp[(n-4)w] \\ &\quad + \cdots + [n(n-1)/2!] \exp[-(n-4)w] + n\exp[-(n-2)w] + \exp(-nw)\}. \end{aligned} \quad (\text{W.6.67})$$

Show that the terms in this expansion can be combined to yield the result

$$[\cosh(w)]^n = (1/2^{n-1})\{\cosh(nw) + n\cosh[(n-2)w] + [n(n-1)/2!] \cosh[(n-4)w] + \cdots\}. \quad (\text{W.6.68})$$

Verify that the last term on the right side of (6.68) is

$$(1/2)^n \binom{n}{n/2} \text{ if } n \text{ is even,} \quad (\text{W.6.69})$$

and is

$$(1/2)^{n-1} \binom{n}{(n-1)/2} \cosh(w) \text{ if } n \text{ is odd.} \quad (\text{W.6.70})$$

Next verify that

$$\begin{aligned} \cosh(mw) &= \cosh(mu + imv) = \cosh(mu) \cosh(imv) + \sinh(mu) \sinh(imv) \\ &= \cosh(mu) \cos(mv) + i \sinh(mu) \sin(mv), \end{aligned} \quad (\text{W.6.71})$$

from which it follows that

$$\cosh(mu) \cos(mv) = \Re[\cosh(mw)], \quad (\text{W.6.72})$$

$$\sinh(mu) \sin(mv) = \Im[\cosh(mw)]. \quad (\text{W.6.73})$$

Using the results found so far, take real and imaginary parts to rewrite (6.66) in the form

$$(1/2^{n-1}) \cosh(nu) \cos(nv) = \Re[(x + iy)^n / f^n] - (1/2^{n-1}) \Re\{n \cosh[(n-2)w] + \cdots\}, \quad (\text{W.6.74})$$

$$(1/2^{n-1}) \sinh(nu) \sin(nv) = \Im[(x + iy)^n / f^n] - (1/2^{n-1}) \Im\{n \cosh[(n-2)w] + \cdots\}. \quad (\text{W.6.75})$$

Conclude that $\cosh(nu) \cos(nv)$ is a polynomial in x and y provided the same is true of $\cosh(mu) \cos(mv)$ for $m = n-2, n-4, \dots$. Make a similar conclusion for $\sinh(nu) \sin(nv)$. Finally prove by induction, starting with (6.19) through (6.22), the claim stated at the beginning of this exercise.

W.6.4. Use some of the machinery of Exercise 6.3 above to produce an easy derivation of the relations (6.23) and (6.24).

W.7 The Elliptical Cylinder in Three Space**W.8 The Rectangle in Two Space****W.9 The Rectangular Cylinder in Three Space****W.10 The Sphere in Three Space**

Higher angular modes are suppressed by the exponential factor $(r/R)^\ell = \exp[\ell \log(r/R)]$. Note that $\log(r/R) < 0$.

W.11 The Ellipsoid in Three Space

Bibliography

- [1] F. Olver, D. Lozier, R. Boisvert, and C. Clark, Editors, *NIST Handbook of Mathematical Functions*, Cambridge (2010). For properties of Bessel functions, see Chapter 10 at the Web site <http://dlmf.nist.gov/>.

Appendix X

Lie Algebraic Theory of Light Optics

Overview

Need text here. Version 2/16/2024

X.1 Hamiltonian Formulation

Consider the optical system illustrated schematically in Figure 1.1. A ray originates at the general *initial* point P^i with spatial coordinate \mathbf{r}^i and moves in an initial direction specified by the unit vector $\hat{\mathbf{s}}^i$. After passing through an optical device it arrives at the *final* point P^f with spatial coordinate \mathbf{r}^f and moves in a final direction specified by the unit vector $\hat{\mathbf{s}}^f$. Given the initial quantities $(\mathbf{r}^i, \hat{\mathbf{s}}^i)$, the fundamental problem of geometrical optics is to determine the final quantities $(\mathbf{r}^f, \hat{\mathbf{s}}^f)$ and to design an optical device in such a way that the relation between the initial and final ray quantities has various desired properties.

Suppose the z coordinates of the initial and final points P^i and P^f are held fixed. In some instances the planes $z = z^i$ and $z = z^f$ can be viewed as object and image planes, respectively. But in other cases they simply serve as convenient reference planes. Further, suppose the general light ray from P^i to P^f is parameterized using z as an *independent/time-like* variable. That is, the path of a general ray is described by specifying the two functions $x(z)$ and $y(z)$. Then the element of path length ds along a ray is given by the expression

$$ds = [(dz)^2 + (dx)^2 + (dy)^2]^{1/2} = [1 + (x')^2 + (y')^2]^{1/2} dz. \quad (\text{X.1.1})$$

Here a prime denotes the differentiation d/dz . Consequently the *optical* path length A along a ray from P^i to P^f is given by the integral

$$\begin{aligned} A &= \int_{P^i}^{P^f} c dt = \int_{P^i}^{P^f} c(dt/ds) ds = \int_{P^i}^{P^f} c(n/c) ds \\ &= \int_{P^i}^{P^f} n ds = \int_{z^i}^{z^f} n(x, y, z) [1 + (x')^2 + (y')^2]^{1/2} dz. \end{aligned} \quad (\text{X.1.2})$$

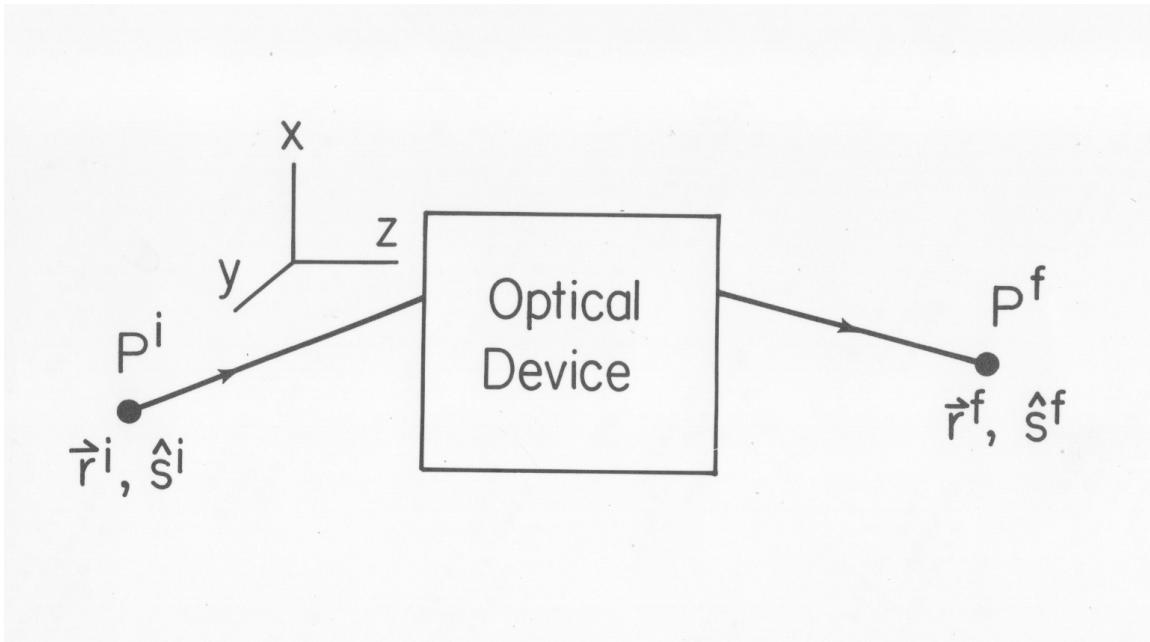


Figure X.1.1: Optical system consisting of an optical device preceded and followed by simple transit. A ray originates at P^i with *initial* location \vec{r}^i and *initial* direction \hat{s}^i , and terminates at P^f with *final* location \vec{r}^f and *final* direction \hat{s}^f

Here the function $n(x, y, z) = n(\mathbf{r})$ specifies the index of refraction at each point before and after the optical device and in the device itself.

Fermat's principle requires that A be an extremum for the path of an actual ray. Therefore the ray path satisfies the Euler-Lagrange equations

$$d/dz(\partial L/\partial x') - \partial L/\partial x = 0, \quad (\text{X.1.3})$$

$$d/dz(\partial L/\partial y') - \partial L/\partial y = 0, \quad (\text{X.1.4})$$

with a Lagrangian L given by the expression

$$L = n(x, y, z)[1 + (x')^2 + (y')^2]^{1/2}. \quad (\text{X.1.5})$$

To proceed further, it is useful to pass from a Lagrangian formulation to a Hamiltonian formulation. Introduce two momenta p_x and p_y conjugate to the coordinates x and y by the rule

$$p_x = \partial L/\partial x', \quad (\text{X.1.6})$$

$$p_y = \partial L/\partial y', \quad (\text{X.1.7})$$

with the explicit results that

$$p_x = n(\mathbf{r})x'/[1 + (x')^2 + (y')^2]^{1/2}, \quad (\text{X.1.8})$$

$$p_y = n(\mathbf{r})y'/[1 + (x')^2 + (y')^2]^{1/2}. \quad (\text{X.1.9})$$

The Hamiltonian H is defined in terms of the Lagrangian L by the Legendre transformation

$$H(x, y, p_x, p_y; z) = p_x x' + p_y y' - L. \quad (\text{X.1.10})$$

It follows from (1.5) through (1.10) that in our case H is given by the relation

$$H = -[n^2(\mathbf{r}) - p_x^2 - p_y^2]^{1/2}. \quad (\text{X.1.11})$$

Let \mathbf{q} be a two-component vector with entries $q_x = x$ and $q_y = y$, and let \mathbf{p} be a two-component vector with entries p_x and p_y . Evidently, a ray leaving the initial point P^i is characterized by the quantities z^i , \mathbf{q}^i , and \mathbf{p}^i . The quantity \mathbf{q}^i specifies the initial point of origin on the plane $z = z^i$ and, according to (1.8) and (1.9), \mathbf{p}^i describes the initial direction of the ray. Similarly, \mathbf{q}^f and \mathbf{p}^f characterize the ray as it arrives at the final point P^f in the plane $z = z^f$. Finally, the relation between the initial conditions \mathbf{q}^i and \mathbf{p}^i and the final conditions \mathbf{q}^f and \mathbf{p}^f is given by following from $z = z^i$ to $z = z^f$ a trajectory $\mathbf{q}(z)$, $\mathbf{p}(z)$ governed by the Hamiltonian H .

At this point it is convenient to introduce a four-component vector \mathbf{w} with entries \mathbf{q} , \mathbf{p} :

$$\mathbf{w} = (w_1, w_2, w_3, w_4) = (q_x, p_x, q_y, p_y). \quad (\text{X.1.12})$$

Also, let \mathbf{w}^i and \mathbf{w}^f denote initial and final values of \mathbf{w} . The fact that initial conditions determine the final conditions can be expressed in terms of a functional relationship or mapping \mathcal{M} . This relationship can be defined formally by writing the expression

$$\mathbf{w}^f = \mathcal{M}\mathbf{w}^i. \quad (\text{X.1.13})$$

Hamilton's equations of motion for the canonical variables \mathbf{q} and \mathbf{p} read

$$q'_\alpha = \partial H / \partial p_\alpha =: -H : q_\alpha, \quad (\text{X.1.14})$$

$$p'_\alpha = -\partial H / \partial q_\alpha =: -H : p_\alpha. \quad (\text{X.1.15})$$

Correspondingly, there is an equation of motion for \mathcal{M} given by the relation

$$\mathcal{M}' = \mathcal{M} : -H : \quad (\text{X.1.16})$$

with the initial condition

$$\mathcal{M}|_{z=z^i} = \mathcal{I} \quad (\text{X.1.17})$$

where \mathcal{I} is the identity map. Recall Subsection 10.1.1.

As has been seen, Fermat's principle is equivalent to the statement that the initial conditions \mathbf{w}^i and the final conditions \mathbf{w}^f are related by following a trajectory governed by a Hamiltonian, namely the Hamiltonian (1.11). From the work of Subsection 6.4.1 this statement is equivalent in turn to the statement that \mathcal{M} is a *symplectic* map.

Let us recapitulate briefly for the light optics context some of what we have learned about symplectic maps. Let M be the Jacobian matrix associated with the mapping \mathcal{M} . It is defined by the relation

$$M_{\alpha\beta} = \partial w_\alpha^f / \partial w_\beta^i, \quad (\text{X.1.18})$$

and describes how small changes in \mathbf{w}^i produce small changes in \mathbf{w}^f . Also, let J be the four-by-four matrix defined by the equation

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \quad (\text{X.1.19})$$

Then from earlier work we know that M satisfies the matrix equation

$$M^T JM = J. \quad (\text{X.1.20})$$

Equation (1.20) is the condition that M be a *symplectic* matrix (and in the context of optics is sometimes called the *lens equation*). Correspondingly, as described earlier, a map \mathcal{M} whose Jacobian matrix M is symplectic is said to be a *symplectic map*. Note that, as indicated by the notation $M(\mathbf{w}^i; z^i, z^f)$, the matrix M depends in general on the variables \mathbf{w}^i , z^i , and z^f . Observe, however, that the right side of (1.20), namely the matrix J given by (1.19), does not depend on these variables and is in fact a constant matrix. The requirement that (1.20) holds for all values of \mathbf{w}^i , z^i , and z^f places strong restrictions on the nature of symplectic maps. These restrictions were first studied by Hamilton (in the context of light optics!) and led to the introduction/invention of *characteristic/generating* functions to describe and manage symplectic maps. In this appendix we will see, in the context of light optics, how Lie methods can also be used for this purpose.

Exercises

X.1.1. Verify that H as given by (1.11) is indeed the Hamiltonian associated with the Lagrangian L given by (1.5).

X.1.2. Recall Liouville's theorem. See Subsection 6.8.1. Google the word *etendue*. Work out the consequences of Liouville's theorem when applied to the case of light optics.

X.2 Assumption of Axial Symmetry and Lie-algebraic Consequences

X.2.1 Preliminaries

Although framed in the context of light optics, the discussion so far is applicable to general Hamiltonian systems having a four-dimensional phase space. We are dealing with symplectic maps \mathcal{M} whose linear parts M about any trajectory/ray are elements of $Sp(4, \mathbb{R})$. We now turn to the specific case of the optical Hamiltonian (1.11). Moreover, at this point we also assume that the optical device has *axial/rotational symmetry* about the z axis. Introduce the definitions

$$q^2 = (q_x)^2 + (q_y)^2 = \mathbf{q} \cdot \mathbf{q}, \quad (\text{X.2.1})$$

$$p^2 = (p_x)^2 + (p_y)^2 = \mathbf{p} \cdot \mathbf{p}, \quad (\text{X.2.2})$$

$$\mathbf{p} \cdot \mathbf{q} = p_x q_x + p_y q_y. \quad (\text{X.2.3})$$

(Note that this notation can be misleading since, for example, q^2 is not the square of any Cartesian coordinate.) From (1.11) we see that the optical Hamiltonian depends on \mathbf{p} only through the quantity p^2 . To enforce axial symmetry, we assume that $n(\mathbf{r})$ is of the functional form

$$n(\mathbf{r}) = \hat{n}(q^2, z) \quad (\text{X.2.4})$$

so that the index of refraction also has axial symmetry. Now imagine that H as given by (1.11) and the assumption (2.4) is expanded in a power series in the components of \mathbf{q} and \mathbf{p} . By the assumption of axial symmetry, such an expansion must be of the form

$$H = H_0 + H_2 + H_4 + H_6 + \dots \quad (\text{X.2.5})$$

where the H_m are homogeneous polynomials of degree m in the components of \mathbf{q} and \mathbf{p} . That is, only *even* powers can occur. Indeed, the H_m depend only on powers of q^2 and p^2 and products of these powers. Since only even powers of the components of \mathbf{w} can occur in H , it follows that the z axis, $\mathbf{w}(z) = \mathbf{0}$, is a trajectory/ray for the equations of motion generated by H , and the phase-space origin is a fixed point of \mathcal{M} : $\mathcal{M}\mathbf{0} = \mathbf{0}$.

Let L_z denote the second degree homogeneous polynomial

$$L_z = (\mathbf{q} \times \mathbf{p}) \cdot \mathbf{e}_z = q_x p_y - q_y p_x \quad (\text{X.2.6})$$

and let \mathcal{L}_z be the associated Lie operator

$$\mathcal{L}_z =: L_z : . \quad (\text{X.2.7})$$

Then, as a Lie-algebraic expression of the condition of axial symmetry for the quantities q^2 , p^2 , and $(\mathbf{p} \cdot \mathbf{q})$, we have the relations

$$\mathcal{L}_z q^2 = \mathcal{L}_z p^2 = \mathcal{L}_z (\mathbf{p} \cdot \mathbf{q}) = 0. \quad (\text{X.2.8})$$

And, as a Lie-algebraic expression of the condition of axial symmetry for H , we have the relations

$$\mathcal{L}_z H_{2n} = 0 \quad (\text{X.2.9})$$

and

$$\mathcal{L}_z H = 0. \quad (\text{X.2.10})$$

Using (2.10) we find that

$$0 = \mathcal{L}_z H =: L_z : H = [L_z, H] \Leftrightarrow dL_z/dz = 0, \quad (\text{X.2.11})$$

from which it follows that there is the relation

$$L_z^f = L_z^i, \quad (\text{X.2.12})$$

which takes the explicit form

$$q_x^f p_y^f - q_y^f p_x^f = q_x^i p_y^i - q_y^i p_x^i. \quad (\text{X.2.13})$$

That is, L_z is an integral of motion.

We note, for future use, that there is the relation

$$(L_z)^2 = (q_x p_y - q_y p_x)^2 = q_x^2 p_y^2 - 2q_x p_x q_y p_y + p_x^2 q_y^2 = p^2 q^2 - (\mathbf{p} \cdot \mathbf{q})^2, \quad (\text{X.2.14})$$

and the relation

$$\begin{aligned} (\mathcal{L}_z)^\dagger &= :L_z:^\dagger = (:q_x p_y - q_y p_x:)^\dagger =:q_x p_y:^\dagger - :q_y p_x:^\dagger \\ &= :q_y p_x: - :q_x p_y: = -\mathcal{L}_z. \end{aligned} \quad (\text{X.2.15})$$

For the steps made in obtaining the latter relation, recall (7.3.16) through (7.3.18).

X.2.2 What Generators Can Occur and Their Relation to Aberrations

Next we invoke a Lie algebraic fact: It can be shown that the solution of (1.16) involves only the ingredients of H and quantities that can be formed by taking Poisson brackets of the ingredients of H . See Chapter 10. It follows that, under the assumption of axial symmetry, the solution to (1.16) must be of the form

$$\begin{aligned} \mathcal{M} &= \exp(:f_2^c:) \exp(:f_2^a:) \exp(:f_4:) \exp(:f_6:) \exp(:f_8:) \cdots \\ &= \mathcal{R} \exp(:f_4:) \exp(:f_6:) \exp(:f_8:) \cdots. \end{aligned} \quad (\text{X.2.16})$$

That is, only the f_m with *even* m can occur in the factored product representation of \mathcal{M} . Moreover, all the f_{2n} must satisfy the axial symmetry (rotational invariance) relation

$$\mathcal{L}_z f_{2n} = 0. \quad (\text{X.2.17})$$

X.2.2.1 Quadratic Generators and Paraxial Optics

Also, there is the Poisson bracket relation

$$[q^2, p^2] = 4\mathbf{q} \cdot \mathbf{p} = 4\mathbf{p} \cdot \mathbf{q}. \quad (\text{X.2.18})$$

Therefore f_2^c and f_2^a , and hence \mathcal{R} , can depend only on the quantities

$$q^2, \mathbf{p} \cdot \mathbf{q}, \text{ and } p^2. \quad (\text{X.2.19})$$

(We note, as can be easily verified, that $p^2 - q^2$ and $\mathbf{p} \cdot \mathbf{q}$ are f_2^a polynomials, and $p^2 + q^2$ is an f_2^c polynomial.)

Relation of Quadratic Generators to Some Simple Paraxial Elements and Systems

The two simple paraxial elements are

$$\mathcal{R}_{d/n} = \exp[-(d/n) : p^2/2 :] \quad (\text{X.2.20})$$

and

$$\mathcal{R}_f = \exp(-f^{-1} : q^2/2 :). \quad (\text{X.2.21})$$

With phase-space coordinates listed in the order (q_1, p_1, q_2, p_2) , their associated matrices are

$$R_{d/n} = \begin{pmatrix} 1 & d/n & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d/n \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (\text{X.2.22})$$

and

$$R_f = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/f & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1/f & 1 \end{pmatrix}. \quad (\text{X.2.23})$$

Thus $\mathcal{R}_{d/n}$ describes (in paraxial approximation) *transit/drift* through a medium having refractive index n and length d . And \mathcal{R}_f describes (in paraxial approximation) the effect of a *thin lens* of focal length f . Here it assumed that readers have enough familiarity with Gaussian (paraxial) geometric optics to be aware of the concepts of drift spaces and thin lenses, and will recognize that the matrices $R_{d/n}$ and R_f describe their effects in linear approximation and under the assumption of axial symmetry. In Section X.6 formulas for all the f_{2n} from f_2 through f_8 will be derived and presented (under the assumption of axial symmetry) including results for thick lens with aspheric surfaces.

There are two two-element systems that are of interest. The first is a “burning glass” or spot-forming system. Ideally, it takes in rays parallel to the z axis (and therefore also parallel to each other), and focuses them to a common spot on the z axis. That is, all initial conditions with $\mathbf{p}^{in} = 0$ and \mathbf{q}^{in} arbitrary are sent to final conditions with \mathbf{p}^{fin} arbitrary and $\mathbf{q}^{fin} = 0$. With regard to position space \mathbf{q} , the burning glass system focuses/gathers rays with arbitrary \mathbf{q}^{in} and sends them to $\mathbf{q}^{fin} = 0$. In linear (paraxial) approximation it is described by a map of the form

$$\mathcal{R} = \exp(-f^{-1} : q^2/2 :) \exp(-f : p^2/2 :), \quad (\text{X.2.24})$$

which is the map for a thin lens with focal length f (and in paraxial approximation) followed by the map for a drift/transit (we assume the medium i air or vacuum so that $n = 1$) over a distance of length f , also in paraxial approximation. Its associated matrix is

$$R = R_d|_{d=f} R_f = \begin{pmatrix} 0 & f & 0 & 0 \\ -1/f & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/f & 1 \end{pmatrix}. \quad (\text{X.2.25})$$

Note that the mark of a spot-forming system (in paraxial approximation) is that

$$R_{11} = R_{33} = 0. \quad (\text{X.2.26})$$

The second system of interest, which is the “reverse” of the first, is a “search-light” system. Ideally, it takes in diverging rays from a point source on the z axis and produces

outgoing rays parallel to the z axis (and parallel to each other). That is, all initial conditions with \mathbf{p}^{in} arbitrary and $\mathbf{q}^{in} = 0$ are sent to final conditions with $\mathbf{p}^{fin} = 0$ and \mathbf{q}^{fin} arbitrary. With regard to momentum space \mathbf{p} , the search-light system “focuses” /gathers rays with arbitrary \mathbf{p}^{in} and sends them to $\mathbf{p}^{fin} = 0$. In linear (paraxial) approximation it is described by a map of the form

$$\mathcal{R} = \exp(-f : p^2/2 :) \exp(-f^{-1} : q^2/2 :). \quad (\text{X.2.27})$$

In paraxial approximation the system consists of paraxial transit through a drift space (with $n = 1$) of length f followed by a thin lens with focal length f , also in paraxial approximation. Its associated matrix is

$$R = R_f R_d|_{d=f} = \begin{pmatrix} 1 & f & 0 & 0 \\ -1/f & 0 & 0 & 0 \\ 0 & 0 & 1 & f \\ 0 & 0 & -1/f & 0 \end{pmatrix}. \quad (\text{X.2.28})$$

The mark of a search-light system (in paraxial approximation) is that

$$R_{22} = R_{44} = 0. \quad (\text{X.2.29})$$

A particular three-element system that is of special interest is a simple imaging system. In linear (paraxial) approximation it is described by a map of the form

$$\mathcal{R} = \exp(-d_1 : p^2/2 :) \exp(-f^{-1} : q^2/2 :) \exp(-d_2 : p^2/2 :). \quad (\text{X.2.30})$$

The system consists of transit through a drift space of length d_1 followed by a thin lens with focal length f followed again by transit through a drift space of length d_2 , all in paraxial approximation. The associated matrix for the system is

$$R = R_{d_2} R_f R_{d_1} = \begin{pmatrix} m & 0 & 0 & 0 \\ -1/f & 1/m & 0 & 0 \\ 0 & 0 & m & 0 \\ 0 & 0 & -1/f & 1/m \end{pmatrix}. \quad (\text{X.2.31})$$

Here we have imposed the imaging condition

$$(1/d_1) + (1/d_2) = (1/f) \quad (\text{X.2.32})$$

and find that the magnification m is given by

$$m = -d_2/d_1. \quad (\text{X.2.33})$$

Note that the mark of an imaging system (in paraxial approximation and when in focus) is that

$$R_{12} = R_{34} = 0. \quad (\text{X.2.34})$$

X.2.2.2 Quartic Generators and Their Relation to Third-Order Aberrations in Various Systems

Similarly, f_4 can depend only on the six quantities $(p^2)^2$, $p^2(\mathbf{p} \cdot \mathbf{q})$, $(\mathbf{p} \cdot \mathbf{q})^2$, p^2q^2 , $(\mathbf{p} \cdot \mathbf{q})q^2$, and $(q^2)^2$. That is, under the assumption of axial symmetry, we may write

$$f_4 = A(p^2)^2 + Bp^2(\mathbf{p} \cdot \mathbf{q}) + C(\mathbf{p} \cdot \mathbf{q})^2 + Dp^2q^2 + E(\mathbf{p} \cdot \mathbf{q})q^2 + F(q^2)^2 \quad (\text{X.2.35})$$

where the coefficients A through F are to be determined. Note that all these ingredients of f_4 are rotationally invariant [as required by (2.17)] and are powers and products of powers of the ingredients for f_2 . Finally, we note that the ingredients of all the f_{2n} for $2n \geq 4$ are axially symmetric and are powers and products of powers of the ingredients for f_2 . This result will be studied further in the next section.

For an imaging system it can be demonstrated that the polynomials in (2.19), shown multiplied by the coefficients A through E , are related to the *Seidel* aberrations called spherical aberration, coma, astigmatism, curvature of field, and distortion:¹

$$A(p^2)^2 \leftrightarrow \text{spherical aberration}, \quad (\text{X.2.36})$$

$$Bp^2(\mathbf{p} \cdot \mathbf{q}) \leftrightarrow \text{coma}, \quad (\text{X.2.37})$$

$$C(\mathbf{p} \cdot \mathbf{q})^2 \leftrightarrow \text{astigmatism}, \quad (\text{X.2.38})$$

$$Dp^2q^2 \leftrightarrow \text{curvature of field}, \quad (\text{X.2.39})$$

$$E(\mathbf{p} \cdot \mathbf{q})q^2 \leftrightarrow \text{distortion}, \quad (\text{X.2.40})$$

$$F(q^2)^2 \leftrightarrow \text{pocus}. \quad (\text{X.2.41})$$

It can be shown that the *net* “pocus” aberration of an imaging system does not affect its ability to form images. (It does not affect the positions where rays arrive on the image plane, but only affects their arrival directions.) Therefore it is less commonly discussed, and indeed its name has been coined only recently. But it can be important for other systems.

The most offensive third-order aberration for a spot-forming system is spherical aberration, which corresponds to the f_4 for this system being primarily of the form $f_4 \sim (p^2)^2$. It damages the ability to focus/gather \mathbf{q} values.

¹Philipp Ludwig von Seidel (1821-1896) classified and described the possible third-order geometric aberrations for axially symmetric optical systems. Subsequent extensive work on aberrations was done by many others including definitive work by Karl Schwarzschild (1873-1916). This work took into account the symplectic condition by employing $F_1(q, Q)$ generating functions, called *point* characteristic functions in the optics literature, and $F_4(p, P)$ generating functions, called *angular* characteristic functions in the optics literature. [$F_1(q, Q)$ generating functions are not applicable to imaging systems when object and image planes are employed, but are applicable when object and some aperture planes are employed. $F_4(p, P)$ generating functions are applicable to imaging systems when object and image planes are employed. See Exercise 6.7.20. It can be shown that, despite the plethora of generating function types, there is no one generating function type that works for (is compatible with) all symplectic maps.] After his optics work, in 1915 while serving in the army in World War I, Schwarzschild (in his spare time) did his celebrated work in General Relativity to find the Schwarzschild metric, the first exact solution to the Einstein field equations and the key ingredient for the understanding of black holes.

The most offensive third-order aberration for a search-light system is pocus, which corresponds to the f_4 for this system being primarily of the form $f_4 \sim (q^2)^2$. It damages the ability to “focus”/gather \mathbf{p} values.² Hence the letter p in its name.

Because net pocus does not affect imaging, it might be viewed as being magical: hence another explanation for the whimsical name *pocus* because of its association with the incantation *hocus-pocus*. Moreover, as we will see, pocus is the *only* third-order aberration that can be controllably “injected” into a system by employing lenses with aspherical surfaces. And as will be further seen, pocus then morphs into other aberrations in the course of paraxial transit and paraxial lensing, and these aberrations can be used to cancel other existing aberrations.

We also remark that the use of the name *astigmatism* can be confusing. In the present context it refers to a particular third-order aberration. With regard to vision, the name astigmatism usually refers to a possible lack of axial symmetry of a visual system in terms of its paraxial (linear) properties, and it is this defect (as well as focal length) that eye glasses are principally designed to correct.

X.2.2.3 Aberrations of Degree 5 and Higher

X.2.3 Equivariance

What fundamentally is going on in our basic assumption of axial/rotational symmetry? Mathematicians have a name for it. It is called *equivariance*. They would say that \mathcal{M} of the form (2.16) with the f_{2n} satisfying (2.17) is an example of an *equivariant symplectic* map. Here is the set up: We know that \mathcal{M} maps a certain space Γ (in our case phase space) into itself. Let \mathcal{S} be another map whose domain and range are also Γ . (Often the action of \mathcal{S} is viewed as some kind of *symmetry* operation.) Then the product maps \mathcal{SM} and \mathcal{MS} are well defined and both map Γ into itself. Now suppose

$$\mathcal{SM} = \mathcal{MS}. \quad (\text{X.2.42})$$

(In words, one may first transform Γ using \mathcal{S} and then act on the results using \mathcal{M} , or vice versa. Both possibilities give the same final result.) If (2.26) holds, \mathcal{M} is said to be equivariant. And if \mathcal{M} is also symplectic, in which case Γ = phase space, then \mathcal{M} is said to be an equivariant symplectic map.

How does this definition work out in our case? Let \mathcal{O} be the map

$$\mathcal{O}(\phi) = \exp(-\phi \mathcal{L}_z). \quad (\text{X.2.43})$$

It is easily verified, as the notation suggests, that \mathcal{O} produces rotations in phase space about the z axis by angle ϕ . For example, there are the relations

$$\mathcal{O}(\phi)q_x = \cos(\phi)q_x - \sin(\phi)q_y, \quad (\text{X.2.44})$$

²Analogous considerations apply in charged-particle magnetic optics when one seeks to design a system that produces beams which, when emitted into vacuum, travel large distances with very little spreading. In such systems lensing is provided by quadrupole magnets, and their fringe fields, as an unavoidable consequence of the Maxwell equations, produce pocus-like aberrations. These aberrations are corrected with the use of octupole magnets.

$$\mathcal{O}(\phi)q_y = \sin(\phi)q_x + \cos(\phi)q_y, \quad (\text{X.2.45})$$

with analogous relations for p_x, p_y . That is, \mathcal{O} acts on configuration space as the group $SO(2)$, and is extended to phase space as a lift from configuration space. (See Exercise 6.5.5.) Next observe, from the relations (2.17), it follows that

$$\mathcal{O}(\phi)\mathcal{M} = \mathcal{M}\mathcal{O}(\phi). \quad (\text{X.2.46})$$

Consequently, when acting on phase space, one may first act by rotating and then by \mathcal{M} , or vice versa. Both possibilities give the same result.³ Evidently in our case (2.26) holds with the role of \mathcal{S} being played by \mathcal{O} ; and therefore \mathcal{M} is an equivariant symplectic map.

Suppose there are multiple symmetry operations acting on Γ , and each is invertible. They may be used to form a group. If there is a continuous family of such operations, they may form a continuous group or be used to form a continuous group. Suppose Γ = phase space and the action of this group on Γ is symplectic. Suppose also that \mathcal{M} is symplectic. Then, by Noether's theorem, we may expect that \mathcal{M} will have an integral. (See Subsection 27.15.1.) In our case the role of \mathcal{S} is played by \mathcal{O} . Moreover, it follows from the definitions (2.7) and (2.27) that the $\mathcal{O}(\phi)$ are symplectic and form a continuous group. Therefore we expect that \mathcal{M} will have an integral. Indeed, for the case at hand, there is the result (2.12). The function L_z is an integral.

Curiously, in this example, one may turn matters around. We have already observed that \mathcal{O} is a symplectic map. Also the map \mathcal{M} acts on phase space. Therefore, one may interpret (2.30) to also say that \mathcal{O} is an equivariant symplectic map with \mathcal{M} playing the role of a symmetry operation! And what is the corresponding integral for \mathcal{O} ? From (2.7), (2.10), and (2.27) we see that

$$\mathcal{O}H = H. \quad (\text{X.2.47})$$

That is H , which generates \mathcal{M} , is an integral for \mathcal{O} .

Exercises

X.2.1. Verify (2.8) through (2.10).

X.2.2. Suppose that f_{2m} and g_{2n} have axial symmetry,

$$\mathcal{L}_z f_{2m} = 0 \text{ and } \mathcal{L}_z g_{2n} = 0. \quad (\text{X.2.48})$$

Show that then both their ordinary and Lie products,

$$e_{2m+2n} = f_{2m}g_{2n} \text{ and } h_{2m+2n-2} = [f_{2m}, g_{2n}], \quad (\text{X.2.49})$$

have axial symmetry.

X.2.3. Verify that (2.16) and (2.17) follow from (2.9) and (2.33) and the work of Chapter 10.

X.2.4. Verify (2.18).

³Note also that $\mathcal{O} \neq \mathcal{M}$ for any choice of the allowed f_{2n} .

X.2.5. Verify (2.28) and (2.29).

X.2.6. Verify (2.30).

X.2.7. Verify from the definitions (2.7) and (2.27) that the $\mathcal{O}(\phi)$ are symplectic maps.

X.2.8. Suppose we present a general/arbitrary f_4 in terms of monomials by writing

$$f_4 = \sum_{|k|=4} f(k_1, k_2, k_3, k_4) q_x^{k_1} p_x^{k_2} q_y^{k_3} p_y^{k_4} \quad (\text{X.2.50})$$

where the coefficients $f(k_1, k_2, k_3, k_4)$ are arbitrary and we have used the notation

$$|k| = k_1 + k_2 + k_3 + k_4. \quad (\text{X.2.51})$$

Under the assumption of axial symmetry, as exemplified by (2.17), there will be relations among the various $f(k_1, k_2, k_3, k_4)$ so that (2.19) also holds. Therefore, as a consequence of (2.19), there will be relations among the $f(k_1, k_2, k_3, k_4)$, and there will be relations that determine the $f(k_1, k_2, k_3, k_4)$ in terms of the quantities A through F . Part of your task is to show that, in particular, among them will be the relations

$$A = f(0, 4, 0, 0), \quad (\text{X.2.52})$$

$$B = f(1, 3, 0, 0), \quad (\text{X.2.53})$$

$$C = (1/2)f(1, 1, 1, 1), \quad (\text{X.2.54})$$

$$D = f(0, 2, 2, 0), \quad (\text{X.2.55})$$

$$E = f(3, 1, 0, 0), \quad (\text{X.2.56})$$

$$F = f(4, 0, 0, 0). \quad (\text{X.2.57})$$

To begin, consider the ingredients of (2.19). Verify that they have the monomial decompositions

$$(p^2)^2 = [(p_x)^2 + (p_y)^2]^2 = (p_x)^4 + 2(p_x)^2(p_y)^2 + (p_y)^4, \quad (\text{X.2.58})$$

$$p^2(\mathbf{p} \cdot \mathbf{q}) = [(p_x)^2 + (p_y)^2](p_x q_x + p_y q_y) = q_x(p_x)^3 + (p_x)^2 q_y p_y + q_x p_x (p_y)^2 + q_y (p_y)^3, \quad (\text{X.2.59})$$

$$(\mathbf{p} \cdot \mathbf{q})^2 = (p_x q_x + p_y q_y)^2 = (q_x)^2(p_x)^2 + 2q_x p_x q_y p_y + (q_y)^2(p_y)^2, \quad (\text{X.2.60})$$

$$p^2 q^2 = [(p_x)^2 + (p_y)^2][(q_x)^2 + (q_y)^2] = (q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2, \quad (\text{X.2.61})$$

$$(\mathbf{p} \cdot \mathbf{q}) q^2 = (p_x q_x + p_y q_y)[(q_x)^2 + (q_y)^2] = (q_x)^3 p_x + q_x p_x (q_y)^2 + q_x (q_y)^2 p_y + (q_y)^3 p_y, \quad (\text{X.2.62})$$

$$(q^2)^2 = [(q_x)^2 + (q_y)^2]^2 = (q_x)^4 + 2(q_x)^2(q_y)^2 + (q_y)^4. \quad (\text{X.2.63})$$

Now equate coefficients of like terms: Verify from (2.19), (2.33), and (2.42) that there are the results

$$f(0, 4, 0, 0) = A, \quad f(0, 2, 0, 2) = 2A, \quad f(0, 0, 0, 4) = A; \quad (\text{X.2.64})$$

Verify from (2.19), (2.33), and (2.43) that there are the results

$$f(1, 3, 0, 0) = B, \quad f(2, 0, 1, 1) = B, \quad f(1, 1, 0, 2) = B, \quad f(0, 0, 1, 3) = B; \quad (\text{X.2.65})$$

Verify from (2.19), (2.33), (2.44), and (2.45) that there are the results

$$f(2, 2, 0, 0) = C + D, \quad f(0, 0, 2, 2) = C + D, \quad (\text{X.2.66})$$

$$f(1, 1, 1, 1) = 2C, \quad (\text{X.2.67})$$

$$f(0, 2, 2, 0) = D, \quad f(2, 0, 0, 2) = D; \quad (\text{X.2.68})$$

Verify from (2.19), (2.33), and (2.46) that there are the results

$$f(3, 1, 0, 0) = E, \quad f(1, 1, 2, 0) = E, \quad f(1, 0, 2, 1) = E, \quad f(0, 0, 3, 1) = E; \quad (\text{X.2.69})$$

Verify from (2.19), (2.33), and (2.47) that there are the results

$$f(4, 0, 0, 0) = F, \quad f(2, 0, 2, 0) = 2F, \quad f(0, 0, 4, 0) = F. \quad (\text{X.2.70})$$

Using (2.48) through (2.54), verify (2.36) through (2.41).

X.2.9. Verify (2.15). Your other task for this exercise is to verify (2.14). According to (2.45) and (2.44) there are the relations

$$\begin{aligned} p^2 q^2 &= [(p_x)^2 + (p_y)^2][(q_x)^2 + (q_y)^2] = (q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2, \\ (\mathbf{p} \cdot \mathbf{q})^2 &= (p_x q_x + p_y q_y)^2 = (q_x)^2(p_x)^2 + 2q_x p_x q_y p_y + (q_y)^2(p_y)^2. \end{aligned}$$

Show that

$$p^2 q^2 - (\mathbf{p} \cdot \mathbf{q})^2 = (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 - 2q_x p_x q_y p_y = (L_z)^2.$$

X.2.10. With regard to the imaging condition (2.31), define quantities δ_1 and δ_2 by the relations

$$d_1 = f + \delta_1 \text{ and } d_2 = f + \delta_2. \quad (\text{X.2.71})$$

Verify that the imaging condition (2.31) is equivalent to the condition

$$\delta_1 \delta_2 = f^2, \quad (\text{X.2.72})$$

which is Newton's condition for imaging.

X.2.11. Using a collection of thin lenses and drifts with $d > 0$, design a system that in paraxial approximation has an associated matrix R of the form

$$R = \begin{pmatrix} \cos \theta & \sin \theta & 0 & 0 \\ -\sin \theta & \cos \theta & 0 & 0 \\ 0 & 0 & \cos \theta & \sin \theta \\ 0 & 0 & -\sin \theta & \cos \theta \end{pmatrix}. \quad (\text{X.2.73})$$

X.2.12. Using a collection of thin lenses and drifts with $d > 0$, design a system that in paraxial approximation has an associated matrix R of the form

$$R = \begin{pmatrix} m & 0 & 0 & 0 \\ 0 & 1/m & 0 & 0 \\ 0 & 0 & m & 0 \\ 0 & 0 & 0 & 1/m \end{pmatrix} \quad (\text{X.2.74})$$

with $m > 0$. Verify that \mathcal{R} given by

$$\mathcal{R} = \exp(: f_2 :) \quad (\text{X.2.75})$$

with f_2 of the form

$$f_2 = -\sigma(\mathbf{p} \cdot \mathbf{q}) \quad (\text{X.2.76})$$

has an R of the form (2.71). But unfortunately there is no simple element that has an f_2 of the form (2.73). The only simple elements are those given by (2.19) and (2.20). In fact, even (2.20) is an idealization. There are no such things as *infinitesimally* thin lenses, only lenses with some finite thickness. Also, although there are both focusing and defocusing lenses, there are no single elements that behave like drifts with either sign of d . But fortunately thick lenses and positive length drifts suffice to achieve a wide variety of linear symplectic maps. It is an interesting exercise to see just what linear symplectic maps can be produced by combinations of thick lenses and drift maps with positive d .

X.3 Lie-Algebraic Decomposition of Polynomials

X.3.1 Introduction of $sp(2, \mathbb{R})$

We will see that in the case of axial symmetry we can employ a particular $Sp(2, \mathbb{R})$ subgroup of $Sp(4, \mathbb{R})$, thereby simplifying/organizing many computations and results. In particular, we will see that the use of the associated $sp(2, \mathbb{R})$ facilitates the construction and labeling of the axially symmetric f_{2n} .

Make the definitions

$$\mathcal{L}_+ =: -p^2/2 :, \quad (\text{X.3.1})$$

$$\mathcal{L}_0 =: (\mathbf{p} \cdot \mathbf{q})/2 :, \quad (\text{X.3.2})$$

$$\mathcal{L}_- =: q^2/2 : . \quad (\text{X.3.3})$$

From (7.3.16) through (7.3.18) we see that

$$(\mathcal{L}_+)^{\dagger} =: -p_x^2/2 - p_y^2/2 :^{\dagger} =: q_x^2/2 + q_y^2/2 := \mathcal{L}_-, \quad (\text{X.3.4})$$

$$(\mathcal{L}_0)^{\dagger} =: (\mathbf{p} \cdot \mathbf{q})/2 :^{\dagger} = (1/2) : p_x q_x + p_y q_y :^{\dagger} = (1/2) : p_x q_x + p_y q_y := \mathcal{L}_0, \quad (\text{X.3.5})$$

$$(\mathcal{L}_-)^{\dagger} =: q_x^2/2 + q_y^2/2 :^{\dagger} =: -p_x^2/2 - p_y^2/2 := \mathcal{L}_+. \quad (\text{X.3.6})$$

Note that the operators \mathcal{L}_0 and $(\mathcal{L}_+ + \mathcal{L}_-)$ are Hermitian, and therefore are operators of the form $: f_2^a :$. And the operator $(\mathcal{L}_+ - \mathcal{L}_-)$ is antiHermitian, and therefore is an operator of the form $: f_2^c :$.

The Lie operators (3.1) through (3.3) obey the commutation rules

$$\{\mathcal{L}_+, \mathcal{L}_-\} = 2\mathcal{L}_0, \quad (\text{X.3.7})$$

$$\{\mathcal{L}_0, \mathcal{L}_+\} = \mathcal{L}_+, \quad (\text{X.3.8})$$

$$\{\mathcal{L}_0, \mathcal{L}_-\} = -\mathcal{L}_-. \quad (\text{X.3.9})$$

According to Exercise 27.5.5 these commutation rules are a variant of the commutation rules for $sp(2, \mathbb{R})$. Only some labeling and normalizations have been changed.

The rules (3.7) through (3.9) are also the commutation rules for $su(2)$ in its raising and lowering operator form. [Recall that $su(2)$ and $sp(2, \mathbb{R})$ are equivalent over the complex field, and therefore a relation of this form between them should not be a surprise. See Subsection 3.7.6.] Also note, as a consequence of (5.3.14) and (2.8), that \mathcal{L}_z commutes with \mathcal{L}_\pm and \mathcal{L}_0 ,

$$\{\mathcal{L}_z, \mathcal{L}_\pm\} = \{\mathcal{L}_z, \mathcal{L}_0\} = 0. \quad (\text{X.3.10})$$

How does the $sp(2, \mathbb{R})$ we have defined appear as a subalgebra of $sp(4, \mathbb{R})$? Review Sections 27.4 and 27.5 and Exercise 27.5.5. There it is shown that there are two *commuting* $sp(2, \mathbb{R})$ subalgebras of $sp(4, \mathbb{R})$, and they are added together in “locked/ganged” fashion [only analogous elements are added, see (27.5.55) through (27.5.57)] to produce the $sp(2, \mathbb{R})$ we have defined. There it is also shown that \mathcal{L}_z is an element of $sp(4, \mathbb{R})$ that commutes, as we have already seen, with the $sp(2, \mathbb{R})$ we have defined.⁴

X.3.2 Fourth Degree Homogeneous Polynomials

Define the *fourth* degree homogeneous polynomials ${}^4\chi_0^0(\mathbf{w})$ and ${}^4\chi_m^2(\mathbf{w})$ by the rules

$${}^4\chi_0^0 = (L_z)^2 = q_x^2 p_y^2 - 2q_x p_x q_y p_y + p_x^2 q_y^2 = p^2 q^2 - (\mathbf{p} \cdot \mathbf{q})^2; \quad (\text{X.3.11})$$

$${}^4\chi_2^2 = (p^2)^2, \quad (\text{X.3.12})$$

$${}^4\chi_1^2 = 2p^2(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.13})$$

$${}^4\chi_0^2 = (2/3)^{1/2}[p^2 q^2 + 2(\mathbf{p} \cdot \mathbf{q})^2], \quad (\text{X.3.14})$$

$${}^4\chi_{-1}^2 = 2q^2(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.15})$$

$${}^4\chi_{-2}^2 = (q^2)^2. \quad (\text{X.3.16})$$

(Note that all these polynomials are axially symmetric.) Then it can be verified that there are the operator results

$$\mathcal{L}_\pm {}^4\chi_0^0 = 0, \quad (\text{X.3.17})$$

$$\mathcal{L}_0 {}^4\chi_0^0 = 0; \quad (\text{X.3.18})$$

$$\mathcal{L}_+ {}^4\chi_m^2 = [(2-m)(3+m)]^{1/2} {}^4\chi_{m+1}^2, \quad (\text{X.3.19})$$

$$\mathcal{L}_- {}^4\chi_m^2 = [(2+m)(3-m)]^{1/2} {}^4\chi_{m-1}^2, \quad (\text{X.3.20})$$

$$\mathcal{L}_0 {}^4\chi_m^2 = m {}^4\chi_m^2. \quad (\text{X.3.21})$$

Note that, in a manner *identical* to that found in the subject of quantum-mechanical angular momentum [which amounts to a study of the representations of $su(2)$], the operator \mathcal{L}_0 extracts the m value, and the operators \mathcal{L}_+ and \mathcal{L}_- raise and lower m values, respectively. In particular, the results (3.17) through (3.21) can be written in the form

$$\mathcal{L}_+ {}^n\chi_m^j = [(j-m)(j+m+1)]^{1/2} {}^n\chi_{m+1}^j, \quad (\text{X.3.22})$$

⁴There \mathcal{L}_z is called \mathcal{J}_z .

$$\mathcal{L}_- {}^n\chi_m^j = [(j+m)(j-m+1)]^{1/2} {}^n\chi_{m-1}^j, \quad (\text{X.3.23})$$

$$\mathcal{L}_0 {}^n\chi_m^j = m {}^n\chi_m^j, \quad (\text{X.3.24})$$

with j , which (as will become evident subsequently) plays the role of “spin”, having the values $j = 0$ or $j = 2$.⁵ Thus, under the action of \mathcal{L}_\pm and \mathcal{L}_0 , ${}^4\chi_0^0$ behaves as a singlet and the $5 = 2j + 1$ with $j = 2$ quantities ${}^4\chi_m^2$ behave as a quintuplet. Lastly, n denotes the degree of the homogeneous polynomial, with $n = 4$ in this case. This similarity arises because the underlying Lie algebra/group theory is the same here and in the treatment of quantum-mechanical angular momentum.

The relations (3.11) through (3.16) can be inverted to express the ingredients of (2.19) in terms of the ${}^4\chi_0^0$ and the ${}^4\chi_m^2$. Doing so gives the results

$$(p^2)^2 = {}^4\chi_2^2, \quad (\text{X.3.25})$$

$$p^2(\mathbf{p} \cdot \mathbf{q}) = (1/2) {}^4\chi_1^2, \quad (\text{X.3.26})$$

$$(\mathbf{p} \cdot \mathbf{q})^2 = -(1/3) {}^4\chi_0^0 + (1/6)^{1/2} {}^4\chi_0^2, \quad (\text{X.3.27})$$

$$p^2 q^2 = (2/3) {}^4\chi_0^0 + (1/6)^{1/2} {}^4\chi_0^2, \quad (\text{X.3.28})$$

$$(\mathbf{p} \cdot \mathbf{q}) q^2 = (1/2) {}^4\chi_{-1}^2, \quad (\text{X.3.29})$$

$$(q^2)^2 = {}^4\chi_{-2}^2. \quad (\text{X.3.30})$$

See Exercise 3.4. Correspondingly, we observe that f_4 decomposes into a singlet spanned by ${}^4\chi_0^0$ and a quintuplet spanned by the ${}^4\chi_m^2$. That is, we may write

$$f_4 = {}^4c_0^0 {}^4\chi_0^0 + \sum_{m=-2}^2 {}^4c_m^2 {}^4\chi_m^2 \quad (\text{X.3.31})$$

where ${}^4c_0^0$ and the ${}^4c_m^2$ are uniquely defined coefficients.⁶

Let us compare the presentations (2.19) and (3.25). Upon equating like powers we obtain the the relations

$$A(p^2)^2 = {}^4c_2^2 {}^4\chi_2^2 \Leftrightarrow A = {}^4c_2^2 \Leftrightarrow {}^4c_2^2 = A, \quad (\text{X.3.32})$$

$$B p^2(\mathbf{p} \cdot \mathbf{q}) = {}^4c_1^2 {}^4\chi_1^2 \Leftrightarrow B = 2 {}^4c_1^2 \Leftrightarrow {}^4c_1^2 = B/2, \quad (\text{X.3.33})$$

$$C(\mathbf{p} \cdot \mathbf{q})^2 + D p^2 q^2 = {}^4c_0^0 {}^4\chi_0^0 + {}^4c_0^2 {}^4\chi_0^2, \quad (\text{X.3.34})$$

$$E(\mathbf{p} \cdot \mathbf{q}) q^2 = {}^4c_{-1}^2 {}^4\chi_{-1}^2 \Leftrightarrow E = 2 {}^4c_{-1}^2 \Leftrightarrow {}^4c_{-1}^2 = E/2, \quad (\text{X.3.35})$$

$$F(q^2)^2 = {}^4c_{-2}^2 {}^4\chi_{-2}^2 \Leftrightarrow F = {}^4c_{-2}^2 \Leftrightarrow {}^4c_{-2}^2 = F. \quad (\text{X.3.36})$$

⁵We have placed the word *spin* in quotation marks because here j does not arise in the context of physical rotations, but rather in this instance is an aspect of the symplectic Lie algebra $sp(2, \mathbb{R})$. Note also that, because of the $(j - m)$ term on the right side of (3.22), the raising operation terminates when $m = j$. Similarly, the lowering operation terminates when $m = -j$.

⁶It is interesting to note that, while in the case of $su(2)$ the construction of representations involves the mathematical use of two-variable polynomials, these variables otherwise play no direct physical role. By contrast, in the construction/representation of symplectic maps, polynomials in the phase-space variables play a direct physical role and have specific physical interpretations.

Further equating of like terms in (3.28) yields the results

$$C = -{}^4c_0^0 + 2(2/3)^{1/2} {}^4c_0^2, \quad (\text{X.3.37})$$

$$D = {}^4c_0^0 + (2/3)^{1/2} {}^4c_0^2. \quad (\text{X.3.38})$$

See Exercise 3.4. Finally, the relations (3.31) and (3.32) can be inverted to yield the relations

$${}^4c_0^0 = (1/3)(-C + 2D), \quad (\text{X.3.39})$$

$${}^4c_0^2 = (1/6)^{1/2}(C + D). \quad (\text{X.3.40})$$

If we make use of (2.36) through (2.41), and the results of the previous paragraph, we can also express the c_m^j in terms of various $f(k_1, k_2, k_3, k_4)$. So doing gives the results

$$\begin{aligned} {}^4c_0^0 &= (1/3)(-C + 2D) = (1/3)[-(1/2)f(1, 1, 1, 1) + 2f(0, 2, 2, 0)] \\ &= (1/6)[-f(1, 1, 1, 1) + 4f(0, 2, 2, 0)]; \end{aligned} \quad (\text{X.3.41})$$

$${}^4c_0^2 = A = f(0, 4, 0, 0), \quad (\text{X.3.42})$$

$${}^4c_1^2 = B/2 = (1/2)f(1, 3, 0, 0), \quad (\text{X.3.43})$$

$$\begin{aligned} {}^4c_0^2 &= (1/6)^{1/2}(C + D) = (1/6)^{1/2}[(1/2)f(1, 1, 1, 1) + f(0, 2, 2, 0)] \\ &= (1/2)(1/6)^{1/2}[f(1, 1, 1, 1) + 2f(0, 2, 2, 0)], \end{aligned} \quad (\text{X.3.44})$$

$${}^4c_{-1}^2 = E/2 = (1/2)f(3, 1, 0, 0), \quad (\text{X.3.45})$$

$${}^4c_{-2}^2 = F = f(4, 0, 0, 0). \quad (\text{X.3.46})$$

We interrupt our discussion to remark that to aid communication it would be nice to have names for ${}^4\chi_0^0$ and the ${}^4\chi_m^2$. We coin the following names:

$${}^4\chi_0^0 = (L_z)^2 = [p^2q^2 - (\mathbf{p} \cdot \mathbf{q})^2] \text{ is the } Petzval \text{ } \chi, \quad (\text{X.3.47})$$

$${}^4\chi_2^2 = (p^2)^2 \text{ is the spherical aberration } \chi, \quad (\text{X.3.48})$$

$${}^4\chi_1^2 = 2p^2(\mathbf{p} \cdot \mathbf{q}) \text{ is the coma } \chi, \quad (\text{X.3.49})$$

$${}^4\chi_0^2 = (2/3)^{1/2}[p^2q^2 + 2(\mathbf{p} \cdot \mathbf{q})^2] \text{ is the } Katarina \text{ } \chi, \quad (\text{X.3.50})$$

$${}^4\chi_{-1}^2 = 2q^2(\mathbf{p} \cdot \mathbf{q}) \text{ is the distortion } \chi, \quad (\text{X.3.51})$$

$${}^4\chi_{-2}^2 = (q^2)^2 \text{ is the pocus } \chi. \quad (\text{X.3.52})$$

We have called ${}^4\chi_0^0$ the *Petzval* in honor of Joseph Maximilian Petzval (1807-1891).⁷ As we will learn, it has special properties. Next observe that ${}^4\chi_0^0$ and ${}^4\chi_0^2$ both have $m = 0$ and

⁷Petzval was a German-Hungarian mathematician/physicist who, among other things, performed and oversaw early analytic work (directed by himself and involving 8 artillery gunners and 3 corporals, serving as computers, supplied by Archduke Louis of Austria) on geometric aberrations. He also designed the Petzval lens system, which was the first lens system to be designed using analytical methods and was very important in the development of high quality portrait photography. Such lens systems are still for sale today, and are valued for their high speed and the pleasant bokeh/background that they produce in the out-of focus area around and behind the subject due to an exquisite mix of aberrations that come into play at points far from the paraxial focal point. Opera glasses were another of his inventions.

In 1859 his home was broken into and his manuscripts, a result of many years of research, were destroyed. His most refined book on optics, lost with his manuscripts, would never appear in print. Much of what we do know of his work is found in the written descriptions of others who had some familiarity with some of his work. Also, apparently Petzval independently discovered Laplace transforms while working on differential equations. But the name *Laplace* for these transforms, also of course discovered by Laplace, was coined by Poincaré and they have been so called ever since.

both are linear combinations of astigmatism and curvature of field. See (2.38) and (2.39). We may think of them as *partners*. Consequently we have called ${}^4\chi_0^2$ the *Katarina* in honor of Petzval's wife. Finally the names *spherical aberration* and *coma* are consistent with (2.36) and (2.37), and the names *distortion* and *pocus* are consistent with (2.40) and (2.41).

To continue our discussion, we will see that there is another way of obtaining explicit formulas for the ${}^n c_m^j$ in terms of f_4 that is both instructive and convenient. Let $\langle *, * \rangle$ denote the scalar product defined in Section 7.3. See (7.3.1) through (7.3.9). Upon employing this scalar product we find, as will be seen subsequently, the normalization and orthogonality results

$$\langle {}^4\chi_0^0, {}^4\chi_0^0 \rangle = 12, \quad (\text{X.3.53})$$

$$\langle {}^4\chi_m^2, {}^4\chi_{m'}^2 \rangle = 64\delta_{mm'}, \quad (\text{X.3.54})$$

$$\langle {}^4\chi_m^2, {}^4\chi_0^0 \rangle = 0. \quad (\text{X.3.55})$$

Consequently, there are the formulas

$${}^4c_0^0 = (1/12)\langle {}^4\chi_0^0, f_4 \rangle, \quad (\text{X.3.56})$$

$${}^4c_m^2 = (1/64)\langle {}^4\chi_m^2, f_4 \rangle. \quad (\text{X.3.57})$$

In view of (3.31) and (3.53) through (3.57) we may say that ${}^4c_0^0$ is the amount of Petzval in f_4 . From (3.39) we see that the amount of Petzval in f_4 vanishes when

$$-C + 2D = 0. \quad (\text{X.3.58})$$

And according to (3.40) the quantity ${}^4c_0^2$, the amount of Katarina in f_4 , vanishes when

$$C + D = 0. \quad (\text{X.3.59})$$

X.3.3 Second Degree Homogeneous Polynomials

Let us proceed further with the Lie-algebraic decomposition of polynomials. The polynomials q^2 , p^2 , and $\mathbf{p} \cdot \mathbf{q}$ that go into making up f_2 may be labelled according to a scheme that is analogous to that used in (3.12) through (3.16). Define the axially-symmetric *second* degree homogeneous polynomials ${}^2\chi_m^1(\mathbf{w})$ by the rules

$${}^2\chi_1^1 = p^2, \quad (\text{X.3.60})$$

$${}^2\chi_0^1 = (2)^{1/2}(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.61})$$

$${}^2\chi_{-1}^1 = q^2. \quad (\text{X.3.62})$$

Then we find the operator results

$$\mathcal{L}_+ {}^2\chi_m^1 = [(1-m)(2+m)]^{1/2} {}^2\chi_{m+1}^1, \quad (\text{X.3.63})$$

$$\mathcal{L}_- {}^2\chi_m^1 = [(1+m)(2-m)]^{1/2} {}^2\chi_{m-1}^1, \quad (\text{X.3.64})$$

$$\mathcal{L}_0 {}^2\chi_m^1 = m {}^2\chi_m^1. \quad (\text{X.3.65})$$

Consequently the rules (3.22) through (3.24) continue to hold with, in this case, $j = 1$ and $n = 2$.

We may also define an axially-symmetric second degree homogeneous polynomial ${}^2\psi_0^0(\mathbf{w})$ by the rule

$${}^2\psi_0^0 = L_z. \quad (\text{X.3.66})$$

It meets our requirement for axial symmetry because it satisfies

$$\mathcal{L}_z {}^2\psi_0^0 =: L_z : L_z = 0. \quad (\text{X.3.67})$$

It also satisfies the relation

$$\mathcal{L}_0 {}^2\psi_0^0 = (1/2) : (\mathbf{p} \cdot \mathbf{q}) : L_z = (1/2)[(\mathbf{p} \cdot \mathbf{q}), L_z] = -(1/2)\mathcal{L}_z(\mathbf{p} \cdot \mathbf{q}) = 0. \quad (\text{X.3.68})$$

Similarly, there are the relations

$$\mathcal{L}_{\pm} {}^2\psi_0^0 = 0. \quad (\text{X.3.69})$$

(The polynomial ${}^2\psi_0^0$ is therefore entitled to carry the index values $j = 0$ and $m = 0$.) According to our previous discussion, $L_z = {}^2\psi_0^0$ does not appear as an odd power in f_2 or any other f_{2n} . Only the ${}^2\chi_m^1$ can appear in the f_{2n} . That is why we have used the symbol ψ for it rather than χ .

However that is not the end of the matter. According to (2.14) L_z^2 depends on p^2 , q^2 , and $(\mathbf{p} \cdot \mathbf{q})$, which are allowed ingredients for the f_{2n} . Therefore, although *odd* powers of L_z are not allowed to appear in the f_{2n} , *even* powers are allowed as, for example, in ${}^4\chi_0^0$. See (3.11).⁸

At this point it may be observed that there are the relations

$${}^4\chi_0^0 = (4/3)^{1/2}[({}^2\chi_1^1)({}^2\chi_{-1}^1) - ({}^2\chi_0^1)^2]; \quad (\text{X.3.70})$$

$${}^4\chi_2^2 = ({}^2\chi_1^1)^2, \quad (\text{X.3.71})$$

$${}^4\chi_1^2 = (2)^{1/2}({}^2\chi_1^1)({}^2\chi_0^1), \quad (\text{X.3.72})$$

$${}^4\chi_0^2 = (2/3)^{1/2}[({}^2\chi_1^1)({}^2\chi_{-1}^1) + ({}^2\chi_0^1)^2], \quad (\text{X.3.73})$$

$${}^4\chi_{-1}^2 = (2)^{1/2}({}^2\chi_{-1}^1)({}^2\chi_0^1), \quad (\text{X.3.74})$$

$${}^4\chi_{-2}^2 = ({}^2\chi_{-1}^1)^2. \quad (\text{X.3.75})$$

These relations are examples of the Clebsch-Gordan series for $sp(2, \mathbb{R})$, and are identical to the relations in quantum mechanics for coupling spin 1 and spin 1 to achieve net spin 0 or net spin 2. Agreement between the Clebsch-Gordan series for $sp(2, \mathbb{R})$ and $su(2)$ is to be expected because the commutation rules (3.7) through (3.9) for $sp(2, \mathbb{R})$ are the same as those for $su(2)$ in raising and lowering operator form, and the “state” relations (3.22) through (3.24) are the same in both cases.

⁸Odd (as well as even) powers of L_z are allowed in the f_n for the magnetic optics case of a solenoid, which also has axial symmetry. See Section 16.2.

X.3.4 Sixth and Eighth Degree Homogeneous Polynomials

The homogeneous polynomials that go into making f_6 , f_8 , etc., may be classified in similar fashion. One finds, for example, that f_6 decomposes into a triplet and a septuplet given by the relations

$${}^6\chi_m^1 = ({}^4\chi_0^0)({}^2\chi_m^1); \quad (\text{X.3.76})$$

$${}^6\chi_3^3 = (p^2)^3, \quad (\text{X.3.77})$$

$${}^6\chi_2^3 = (6)^{1/2}(p^2)^2(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.78})$$

$${}^6\chi_1^3 = (3/5)^{1/2}[4p^2(\mathbf{p} \cdot \mathbf{q})^2 + (p^2)^2q^2], \quad (\text{X.3.79})$$

$${}^6\chi_0^3 = (4/5)^{1/2}[2(\mathbf{p} \cdot \mathbf{q})^3 + 3(\mathbf{p} \cdot \mathbf{q})p^2q^2], \quad (\text{X.3.80})$$

$${}^6\chi_{-1}^3 = (3/5)^{1/2}[4q^2(\mathbf{p} \cdot \mathbf{q})^2 + (q^2)^2p^2], \quad (\text{X.3.81})$$

$${}^6\chi_{-2}^3 = (6)^{1/2}(q^2)^2(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.82})$$

$${}^6\chi_{-3}^3 = (q^2)^3. \quad (\text{X.3.83})$$

Consequently we may present a general (axially symmetric) f_6 in the form

$$f_6 = \sum_{m=-1}^1 {}^6c_m^1 {}^6\chi_m^1 + \sum_{m=-3}^3 {}^6c_m^3 {}^6\chi_m^3 \quad (\text{X.3.84})$$

where the ${}^6c_m^1$ and ${}^6c_m^3$ are uniquely defined coefficients.

Similarly, f_8 decomposes into a singlet, a quintuplet, and a 9-tuplet given by the relations

$${}^8\chi_0^0 = ({}^4\chi_0^0)^2; \quad (\text{X.3.85})$$

$${}^8\chi_m^2 = ({}^4\chi_0^0)({}^4\chi_m^2); \quad (\text{X.3.86})$$

$${}^8\chi_4^4 = ({}^2\chi_1^1)^4 = (p^2)^4, \quad (\text{X.3.87})$$

$${}^8\chi_3^4 = (8)^{1/2}(p^2)^3(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.88})$$

$${}^8\chi_2^4 = (4/7)^{1/2}[(p^2)^3q^2 + 6(p^2)^2(\mathbf{p} \cdot \mathbf{q})^2], \quad (\text{X.3.89})$$

$${}^8\chi_1^4 = (8/7)^{1/2}[3(p^2)^2(\mathbf{p} \cdot \mathbf{q})q^2 + 4(p^2)(\mathbf{p} \cdot \mathbf{q})^3], \quad (\text{X.3.90})$$

$${}^8\chi_0^4 = (2/35)^{1/2}[24p^2q^2(\mathbf{p} \cdot \mathbf{q})^2 + 3(q^2)^2(p^2)^2 + 8(\mathbf{p} \cdot \mathbf{q})^4], \quad (\text{X.3.91})$$

$${}^8\chi_{-1}^4 = (8/7)^{1/2}[3(q^2)^2(\mathbf{p} \cdot \mathbf{q})p^2 + 4(q^2)(\mathbf{p} \cdot \mathbf{q})^3], \quad (\text{X.3.92})$$

$${}^8\chi_{-2}^4 = (4/7)^{1/2}[(q^2)^3p^2 + 6(q^2)^2(\mathbf{p} \cdot \mathbf{q})^2], \quad (\text{X.3.93})$$

$${}^8\chi_{-3}^4 = (8)^{1/2}(q^2)^3(\mathbf{p} \cdot \mathbf{q}), \quad (\text{X.3.94})$$

$${}^8\chi_{-4}^4 = ({}^2\chi_{-1}^1)^4 = (q^2)^4. \quad (\text{X.3.95})$$

Consequently we may present a general (axially symmetric) f_8 in the form

$$f_8 = {}^8c_0^0 {}^8\chi_0^0 + \sum_{m=-2}^2 {}^8c_m^2 {}^8\chi_m^2 + \sum_{m=-4}^4 {}^8c_m^4 {}^8\chi_m^4 \quad (\text{X.3.96})$$

where ${}^8c_0^0$ and the ${}^8c_m^2$ and the ${}^8c_m^4$ are uniquely defined coefficients.

It can be checked, as implied by the presentations (3.25), (3.72), and (3.84), that the decompositions given above are exhaustive. That is, the various ${}^n\chi_m^j$ span each (axially symmetric) f_{2n} . Moreover, the ${}^n\chi_m^j$ have been defined in such a way that one has the general relations (3.22) through (3.24).

X.3.5 Proof of Orthogonality and Definition/Use of the Quadratic Casimir Operator

The aim of this subsection is to prove the scalar product relations

$$\langle {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle = N(n, j) \delta_{n'n} \delta_{j'j} \delta_{m'm} \quad (\text{X.3.97})$$

where the $N(n, j)$ are normalization constants *independent* of m .⁹ One of the tools for doing so will be the quadratic Casimir operator for our realization of $sp(2, \mathbb{R})$.

That (3.85) should contain the delta function $\delta_{n'n}$ is obvious because, by definition, $\langle *, * \rangle$ is zero for unlike monomial pairs, and hence vanishes for homogeneous polynomials of different degrees. See Subsection 7.3.1 and Exercise 7.3.25.

Let us next verify the $\delta_{m'm}$ factor. Consider the quantity $\langle {}^{n'}\chi_{m'}^{j'}, \mathcal{L}_0 {}^n\chi_m^j \rangle$. With the aid of (3.24) we may write

$$\langle {}^{n'}\chi_{m'}^{j'}, \mathcal{L}_0 {}^n\chi_m^j \rangle = m \langle {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle. \quad (\text{X.3.98})$$

But, with the aid of (3.5) and (3.24), we may also write

$$\langle {}^{n'}\chi_{m'}^{j'}, \mathcal{L}_0 {}^n\chi_m^j \rangle = \langle (\mathcal{L}_0)^\dagger {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle = \langle \mathcal{L}_0 {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle = m' \langle {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle. \quad (\text{X.3.99})$$

It follows that

$$(m' - m) \langle {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle = 0 \quad (\text{X.3.100})$$

from which we conclude that

$$\langle {}^{n'}\chi_{m'}^{j'}, {}^n\chi_m^j \rangle = 0 \text{ when } m' \neq m. \quad (\text{X.3.101})$$

To see that $N(n, j)$ is (as the notation asserts) independent of m , consider the quantity $\langle \mathcal{L}_+ {}^n\chi_m^j, \mathcal{L}_+ {}^n\chi_m^j \rangle$. Making use of (3.22) gives the result

$$\langle \mathcal{L}_+ {}^n\chi_m^j, \mathcal{L}_+ {}^n\chi_m^j \rangle = [(j-m)(j+m+1)]^{1/2} [(j-m)(j+m+1)]^{1/2} \langle {}^n\chi_{m+1}^j, {}^n\chi_{m+1}^j \rangle. \quad (\text{X.3.102})$$

From (3.4) we conclude that

$$\langle \mathcal{L}_+ {}^n\chi_m^j, \mathcal{L}_+ {}^n\chi_m^j \rangle = \langle {}^n\chi_m^j, (\mathcal{L}_+)^{\dagger} \mathcal{L}_+ {}^n\chi_m^j \rangle = \langle {}^n\chi_m^j, \mathcal{L}_- \mathcal{L}_+ {}^n\chi_m^j \rangle. \quad (\text{X.3.103})$$

But from (3.22) and (3.23) we see that

$$\begin{aligned} \mathcal{L}_- \mathcal{L}_+ {}^n\chi_m^j &= [(j-m)(j+m+1)]^{1/2} \mathcal{L}_- {}^n\chi_{m+1}^j \\ &= [(j-m)(j+m+1)]^{1/2} [(j+m+1)(j-m)]^{1/2} {}^n\chi_m^j \end{aligned} \quad (\text{X.3.104})$$

so that

$$\langle \mathcal{L}_+ {}^n\chi_m^j, \mathcal{L}_+ {}^n\chi_m^j \rangle = [(j-m)(j+m+1)]^{1/2} [(j+m+1)(j-m)]^{1/2} \langle {}^n\chi_m^j, {}^n\chi_m^j \rangle. \quad (\text{X.3.105})$$

Upon comparing (3.90) with (3.93) we see that there is the relation

$$\begin{aligned} &[(j-m)(j+m+1)]^{1/2} [(j-m)(j+m+1)]^{1/2} \langle {}^n\chi_{m+1}^j, {}^n\chi_{m+1}^j \rangle = \\ &[(j-m)(j+m+1)]^{1/2} [(j+m+1)(j-m)]^{1/2} \langle {}^n\chi_m^j, {}^n\chi_m^j \rangle. \end{aligned} \quad (\text{X.3.106})$$

⁹To simplify notation, in this subsection the quantities n and n' are *even* integers.

Therefore, as long as $[(j-m)(j+m+1)] \neq 0$ (which will be true if raising of m is possible), we have found the relation

$$\langle {}^n\chi_{m+1}^j, {}^n\chi_{m+1}^j \rangle = \langle {}^n\chi_m^j, {}^n\chi_m^j \rangle. \quad (\text{X.3.107})$$

With this result we can repeatedly increase m starting from $m = -j$ to obtain the result

$$\langle {}^n\chi_m^j, {}^n\chi_m^j \rangle = \langle {}^n\chi_{-j}^j, {}^n\chi_{-j}^j \rangle \text{ for } m \in [-j, j], \quad (\text{X.3.108})$$

which verifies that N does not depend on m .

The last step is to verify the $\delta_{j'j}$ factor in (3.85). This can be done with the aid of the quadratic Casimir operator \mathcal{C}_2 for our realization of $sp(2, \mathbb{R})$. For the purposes of this appendix, it is defined by the rule

$$\mathcal{C}_2 = (\mathcal{L}_+\mathcal{L}_- + \mathcal{L}_-\mathcal{L}_+ + 2\mathcal{L}_0^2)/2. \quad (\text{X.3.109})$$

(For a discussion of Casimir operators, see Section 27.11.) It follows from (3.4) through (3.6) that \mathcal{C}_2 is Hermitian,

$$\mathcal{C}_2^\dagger = \mathcal{C}_2. \quad (\text{X.3.110})$$

And it follows from (3.7) through (3.9) that \mathcal{C}_2 commutes with all the $sp(2, \mathbb{R})$ generators, \mathcal{L}_\pm and \mathcal{L}_0 ,

$$\{\mathcal{C}_2, \mathcal{L}_\pm\} = \{\mathcal{C}_2, \mathcal{L}_0\} = 0, \quad (\text{X.3.111})$$

as expected for a Casimir operator.

Let us compute $\mathcal{C}_2 {}^n\chi_j^j$. Evidently,

$$\mathcal{C}_2 {}^n\chi_j^j = (1/2)(\mathcal{L}_+\mathcal{L}_- + \mathcal{L}_-\mathcal{L}_+ + 2\mathcal{L}_0^2) {}^n\chi_j^j. \quad (\text{X.3.112})$$

Evaluate each of the three terms appearing in (3.100). For $\mathcal{L}_0^2 {}^n\chi_j^j$ there is the result

$$\mathcal{L}_0^2 {}^n\chi_j^j = j^2 {}^n\chi_j^j. \quad (\text{X.3.113})$$

Recall (3.24). For $(1/2)\mathcal{L}_-\mathcal{L}_+ {}^n\chi_j^j$ there is the result

$$(1/2)\mathcal{L}_-\mathcal{L}_+ {}^n\chi_j^j = 0. \quad (\text{X.3.114})$$

Recall (3.22) evaluated for $m = j$. Finally, for $(1/2)\mathcal{L}_+\mathcal{L}_- {}^n\chi_j^j$, there is the result

$$(1/2)\mathcal{L}_+\mathcal{L}_- {}^n\chi_j^j = (1/2)\mathcal{L}_+(2j)^{1/2} {}^n\chi_{j-1}^j = (1/2)(2j)^{1/2}(2j)^{1/2} {}^n\chi_j^j = j {}^n\chi_j^j. \quad (\text{X.3.115})$$

Recall (3.22) and (3.23). Consequently, as in the analogous quantum-mechanical calculation, there is the net result

$$\mathcal{C}_2 {}^n\chi_j^j = (j + j^2) {}^n\chi_j^j = [j(j+1)] {}^n\chi_j^j. \quad (\text{X.3.116})$$

To continue our exploration, multiply both sides of (3.104) by \mathcal{L}_-^ℓ where $\ell = j - m$. So doing gives the result

$$[j(j+1)]\mathcal{L}_-^\ell {}^n\chi_j^j = \mathcal{L}_-^\ell \mathcal{C}_2 {}^n\chi_j^j = \mathcal{C}_2 \mathcal{L}_-^\ell {}^n\chi_j^j \quad (\text{X.3.117})$$

where (3.99) has been used. But from (repeated, if necessary) use of (3.23) it follows that there is a relation of the form

$$\mathcal{L}_-^{\ell} {}^n\chi_j^j = \lambda(j, \ell) {}^n\chi_m^j \quad (\text{X.3.118})$$

where $\lambda(j, \ell)$ is a non-vanishing proportionality constant. Combining (3.105) and (3.106) gives the final result

$$\mathcal{C}_2 {}^n\chi_m^j = [j(j+1)] {}^n\chi_m^j. \quad (\text{X.3.119})$$

We are now prepared to verify the $\delta_{j'j}$ factor in (3.85). Using (3.107) gives the relation

$$\langle {}^n\chi_m^{j'}, \mathcal{C}_2 {}^n\chi_m^j \rangle = [j(j+1)] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = [(j+1/2)^2 - 1/4] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle. \quad (\text{X.3.120})$$

But, from (3.98), we also have the relation

$$\langle {}^n\chi_m^{j'}, \mathcal{C}_2 {}^n\chi_m^j \rangle = \langle \mathcal{C}_2 {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = [(j'+1/2)^2 - 1/4] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle. \quad (\text{X.3.121})$$

Upon combining (3.108) and (3.109) we see that

$$[(j'+1/2)^2 - (j+1/2)^2] \langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = 0. \quad (\text{X.3.122})$$

It is easily verified that

$$[(j'+1/2)^2 - (j+1/2)^2] = 0 \Leftrightarrow j' = j \text{ or } j' + j = -1. \quad (\text{X.3.123})$$

The second possibility on the right side of (3.111) cannot occur since we are only working with nonnegative values of j and j' . We conclude that for our purposes the factor

$$[(j'+1/2)^2 - (j+1/2)^2] \neq 0 \text{ for } j' \neq j, \quad (\text{X.3.124})$$

and therefore (3.110) requires that

$$\langle {}^n\chi_m^{j'}, {}^n\chi_m^j \rangle = 0 \text{ for } j' \neq j. \quad (\text{X.3.125})$$

What remains is to evaluate/specify the $N(n, j)$. Presumably there is a relatively simple formula that does so. But for our purposes the Table below suffices for values of n and j of present interest.

Table X.3.1: Some Values of $N(n, j)$.

n	$N(n, 0)$	$N(n, 1)$	$N(n, 2)$	$N(n, 3)$	$N(n, 4)$
2	*	4	*	*	*
4	12	*	64	*	*
6	*	160	*	2304	*
8	?	*	?	*	?

Exercises

X.3.1. Verify (3.17) through (3.21) and (3.51) through (3.53).

X.3.2. Verify (3.35) through (3.40).

X.3.3. Review Exercise 2.9. Equation (3.11) provides the monomial decomposition for ${}^4\chi_0^0$. Verify that the relations below provide the monomial decompositions for the ${}^4\chi_m^2$:

$${}^4\chi_2^2 = (p^2)^2 = (p_x)^4 + 2(p_x)^2(p_y)^2 + (p_y)^4, \quad (\text{X.3.126})$$

$${}^4\chi_1^2 = 2p^2(\mathbf{p} \cdot \mathbf{q}) = 2q_x(p_x)^3 + 2(p_x)^2q_y p_y + 2q_x p_x(p_y)^2 + 2q_y(p_y)^3, \quad (\text{X.3.127})$$

$$\begin{aligned} {}^4\chi_0^2 &= (2/3)^{1/2}[p^2 q^2 + 2(\mathbf{p} \cdot \mathbf{q})^2] \\ &= (2/3)^{1/2}\{[(p_x)^2 + (p_y)^2][(q_x)^2 + (q_y)^2] + 2(q_x p_x + q_y p_y)^2\} \\ &= (2/3)^{1/2}[(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2 \\ &\quad + 2(q_x)^2(p_x)^2 + 4q_x p_x q_y p_y + 2(q_y)^2(p_y)^2] \\ &= (2/3)^{1/2}[3(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + 3(q_y)^2(p_y)^2 + 4q_x p_x q_y p_y], \end{aligned} \quad (\text{X.3.128})$$

$${}^4\chi_{-1}^2 = 2(\mathbf{p} \cdot \mathbf{q})q^2 = 2(q_x)^3 p_x + 2q_x p_x(q_y)^2 + 2(q_x)^2 q_y p_y + 2(q_y)^3 p_y, \quad (\text{X.3.129})$$

$${}^4\chi_{-2}^2 = (q^2)^2 = (q_x)^4 + 2(q_x)^2(q_y)^2 + (q_y)^4. \quad (\text{X.3.130})$$

X.3.4. The aim of this exercise is to verify the relations (3.25) through (3.30). Most of them can be read off directly from (3.11) through (3.16). Only two cannot. Equation (3.11) reads

$${}^4\chi_0^0 = p^2 q^2 - (\mathbf{p} \cdot \mathbf{q})^2, \quad (\text{X.3.131})$$

and from (3.14) we have

$$(3/2)^{1/2} {}^4\chi_0^2 = p^2 q^2 + 2(\mathbf{p} \cdot \mathbf{q})^2. \quad (\text{X.3.132})$$

Thus Petzval and Katarina are linear combinations of astigmatism and curvature of field. Conversely, astigmatism and curvature of field are linear combinations of Petzval and Katarina. See (3.27) and (3.28). Verify that (3.27) and (3.28) follow from (3.131) and (3.132).

X.3.5. The aim of this exercise is to verify the relations (3.31) through (3.34). Review the results (2.42) through (2.47), (3.11), and (3.114) through (3.118). Verify that

$$\begin{aligned} C(\mathbf{p} \cdot \mathbf{q})^2 + Dp^2q^2 &= C[(q_x)^2(p_x)^2 + 2q_x p_x q_y p_y + (q_y)^2(p_y)^2] \\ &\quad + D[(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + (q_y)^2(p_y)^2] \\ &= (C + D)[(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2] \\ &\quad + 2Cq_x p_x q_y p_y + D[(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \end{aligned} \quad (\text{X.3.133})$$

Verify that

$$\begin{aligned} {}^4c_0^0 {}^4\chi_0^0 + {}^4c_0^2 {}^4\chi_0^2 &= {}^4c_0^0 [q_x^2 p_y^2 - 2q_x p_x q_y p_y + p_x^2 q_y^2] \\ &\quad + {}^4c_0^2 (2/3)^{1/2} [3(q_x)^2(p_x)^2 + (p_x)^2(q_y)^2 + (q_x)^2(p_y)^2 + 3(q_y)^2(p_y)^2 + 4q_x p_x q_y p_y] \\ &= (6)^{1/2} {}^4c_0^2 [(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2] + [-2 {}^4c_0^0 + 4(2/3)^{1/2} {}^4c_0^2] q_x p_x q_y p_y \\ &\quad + [{}^4c_0^0 + (2/3)^{1/2} {}^4c_0^2][(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \end{aligned} \quad (\text{X.3.134})$$

Since the monomials appearing on the right sides of (3.119) and (3.120) are linearly independent, we may equate the coefficients of like terms. In particular, we may equate the coefficients of the polynomials

$$[(q_x)^2(p_x)^2 + (q_y)^2(p_y)^2], q_x p_x q_y p_y, \text{ and } [(p_x)^2(q_y)^2 + (q_x)^2(p_y)^2]. \quad (\text{X.3.135})$$

Conclude that (3.28), when combined with (3.119) and (3.120), implies the relations

$${}^4c_0^2 = 6^{-1/2}(C + D) \Leftrightarrow C + D = (6)^{1/2} {}^4c_0^2, \quad (\text{X.3.136})$$

$$2C = [-2 {}^4c_0^0 + 4(2/3)^{1/2} {}^4c_0^2] \Leftrightarrow C = [-{}^4c_0^0 + 2(2/3)^{1/2} {}^4c_0^2], \quad (\text{X.3.137})$$

$$D = [{}^4c_0^0 + (2/3)^{1/2} {}^4c_0^2], \quad (\text{X.3.138})$$

from which it follows that

$${}^4c_0^0 = (1/3)(-C + 2D) \Leftrightarrow (-C + 2D) = 3 {}^4c_0^0. \quad (\text{X.3.139})$$

Observe that (3.123) and (3.124) agree with the claims (3.31) and (3.32). Also, observe that taking the sum of (3.31) and (3.32) produces (3.34), which agrees with (3.122). Finally, verify that (3.33) follows from (3.31) and (3.32).

X.3.6. Observe that the operators \mathcal{L}_\pm and \mathcal{L}_0 are derivations. Work out their effects on ${}^4\chi_0^0$ and the ${}^4\chi_m^2$ using (3.11) and (3.58) through (3.63) and (3.51) through (3.53). Verify that your results agree with (3.17) through (3.21).

X.3.7. Review the derivation of the result (3.95) obtained with the use of the operator \mathcal{L}_+ . Using the operator \mathcal{L}_- in a similar way, derive the relation

$$\langle {}^n\chi_{m-1}^j, {}^n\chi_{m-1}^j \rangle = \langle {}^n\chi_m^j, {}^n\chi_m^j \rangle. \quad (\text{X.3.140})$$

X.3.8. According to (3.97), the different ${}^n\chi_m^j$ are orthogonal. Show that

$$\langle (\mathbf{p} \cdot \mathbf{q})^2, p^2 q^2 \rangle = 8 \neq 0 \quad (\text{X.3.141})$$

so that astigmatism and curvature of field are *not* orthogonal. Hint: Use (3.27), (3.28), and (3.53) through (3.55). What about the other Seidel aberrations?

X.3.9. Verify that the ${}^n\chi_m^j$ given by (3.11) through (3.16), (3.48) through (3.50), (3.64) through (3.71), and (3.73) through (3.83) all obey the rules (3.22) through (3.24).

X.3.10. Verify the relations

$$\mathcal{L}_+ =: -p^2/2 := -(1/2) : {}^2\chi_1^1 :, \quad (\text{X.3.142})$$

$$\mathcal{L}_0 =: (\mathbf{p} \cdot \mathbf{q})/2 := (1/8)^{1/2} : {}^2\chi_0^1 :, \quad (\text{X.3.143})$$

$$\mathcal{L}_- =: q^2/2 := (1/2) : {}^2\chi_{-1}^1 :. \quad (\text{X.3.144})$$

X.3.11. Verify (3.98) and (3.99).

X.3.12. Verify the correctness of the entries in Table 3.1.

X.3.13. The purpose of this exercise is to ponder a conversation Poisson, Jacobi, Clebsch, Gordan, Casimir, Petzval, and Lie might have should they meet for dinner. According to Jacobi there is the homomorphism (5.3.14) between the Poisson bracket Lie algebra of phase-space functions and the commutator Lie algebra of Lie operators. According to (3.97) the Casimir \mathcal{C}_2 is defined by

$$\mathcal{C}_2 = (\mathcal{L}_+ \mathcal{L}_- + \mathcal{L}_- \mathcal{L}_+ + 2\mathcal{L}_0^2)/2. \quad (\text{X.3.145})$$

Use, in (3.130), the relations (3.127) through (3.129) to verify that we may also write

$$\mathcal{C}_2 = -(1/8)[(: {}^2\chi_1^1 :)(:{}^2\chi_{-1}^1 :)+(: {}^2\chi_{-1}^1 :)(:{}^2\chi_1^1 :)-2(: {}^2\chi_0^1 :)^2]. \quad (\text{X.3.146})$$

According to (3.58) there is for the Petzval polynomial ${}^4\chi_0^0$ the Clebsch-Gordan relation

$$\begin{aligned} {}^4\chi_0^0 &= (4/3)^{1/2}[({}^2\chi_1^1)({}^2\chi_{-1}^1) - ({}^2\chi_0^1)^2] \\ &= (1/3)^{1/2}[({}^2\chi_1^1)({}^2\chi_{-1}^1) + ({}^2\chi_{-1}^1)({}^2\chi_1^1) - 2({}^2\chi_0^1)^2]. \end{aligned} \quad (\text{X.3.147})$$

Why, apart from different overall multiplicative constants and some colons, are the ingredients in (3.131) and (3.132) the same? And what does this coincidence have to do with the properties (3.99) of the Casimir and the properties (3.17) and (3.18) of the Petzval polynomial?

X.4 Applications of Multiplet Decompositions

It has been shown that the various f_{2n} can be decomposed into multiplets with members ${}^n\chi_m^j$. What is this decomposition good for? Suppose an optical system is composed of N elements, and let \mathcal{M}_i be the optical transfer map for the i 'th element. Then the net optical transfer map \mathcal{M} for the entire system can be written as the product

$$\mathcal{M}_{\text{net}} = \mathcal{M}_1 \mathcal{M}_2 \cdots \mathcal{M}_N. \quad (\text{X.4.1})$$

Next observe that each of the \mathcal{M}_i has a factorization of the form (2.16). Suppose further that the various f 's for the various \mathcal{M}_i are all known. Then the only problem involved in computing \mathcal{M}_{net} is that of combining a collection of known maps and writing the result in factorized form.

The general problem of combining/concatenating maps has been treated in Section 8.4. Let us briefly summarize the consequences of this treatment for the present discussion. Suppose \mathcal{M}_f and \mathcal{M}_g are two optical transfer maps written in factored product form,

$$\mathcal{M}_f = \mathcal{R}_f \exp(: f_4 :) \exp(: f_6 :) \exp(: f_8 :) \cdots, \quad (\text{X.4.2})$$

$$\mathcal{M}_g = \mathcal{R}_g \exp(: g_4 :) \exp(: g_6 :) \exp(: g_8 :) \cdots. \quad (\text{X.4.3})$$

Let \mathcal{M}_h be their product,

$$\mathcal{M}_h = \mathcal{M}_f \mathcal{M}_g, \quad (\text{X.4.4})$$

and write \mathcal{M}_h in the factorized form

$$\mathcal{M}_h = \mathcal{R}_h \exp(: h_4 :) \exp(: h_6 :) \exp(: h_8 :) \cdots. \quad (\text{X.4.5})$$

Employ (4.2) and (4.3) in (4.4), and judiciously insert factors of $\mathcal{I} = \mathcal{R}_g \mathcal{R}_g^{-1}$, to write

$$\begin{aligned} \mathcal{M}_h &= \mathcal{R}_f \exp(: f_4 :) \exp(: f_6 :) \exp(: f_8 :) \cdots \times \mathcal{R}_g \exp(: g_4 :) \exp(: g_6 :) \exp(: g_8 :) \cdots \\ &= \mathcal{R}_f \mathcal{R}_g \mathcal{R}_g^{-1} \exp(: f_4 :) \mathcal{R}_g \mathcal{R}_g^{-1} \exp(: f_6 :) \mathcal{R}_g \mathcal{R}_g^{-1} \exp(: f_8 :) \mathcal{R}_g \cdots \\ &\quad \times \exp(: g_4 :) \exp(: g_6 :) \exp(: g_8 :) \cdots. \end{aligned} \quad (\text{X.4.6})$$

Next manipulate and make the definition

$$\mathcal{R}_g^{-1} \exp(: f_{2n} :) \mathcal{R}_g = \exp[\mathcal{R}_g^{-1} : f_{2n} : \mathcal{R}_g] = \exp(: \mathcal{R}_g^{-1} f_{2n} :) = \exp(: f_{2n}^{\text{tr}} :). \quad (\text{X.4.7})$$

Here the *transformed* polynomials f_{2n}^{tr} are defined in terms of the original f_{2n} by the relations

$$f_{2n}^{\text{tr}} = \mathcal{R}_g^{-1} f_{2n}, \quad (\text{X.4.8})$$

from which it follows that

$$f_{2n}^{\text{tr}}(\mathbf{w}) = f_{2n}[(R^g)^{-1} \mathbf{w}] \quad (\text{X.4.9})$$

where R^g is the matrix associated with \mathcal{R}_g . [See (8.2.26).] Upon combining the results of our manipulation and definition we conclude that

$$\mathcal{M}_h = \mathcal{R}_f \mathcal{R}_g \exp(: f_4^{\text{tr}} :) \exp(: f_6^{\text{tr}} :) \exp(: f_8^{\text{tr}} :) \cdots \times \exp(: g_4 :) \exp(: g_6 :) \exp(: g_8 :) \cdots. \quad (\text{X.4.10})$$

Now compare (4.5) and (4.10) to conclude that

$$\mathcal{R}_h = \mathcal{R}_f \mathcal{R}_g, \quad (\text{X.4.11})$$

and

$$\begin{aligned} \exp(: h_4 :) \exp(: h_6 :) \exp(: h_8 :) \cdots &= \\ \exp(: f_4^{\text{tr}} :) \exp(: f_6^{\text{tr}} :) \exp(: f_8^{\text{tr}} :) \cdots \times \exp(: g_4 :) \exp(: g_6 :) \exp(: g_8 :) \cdots. \end{aligned} \quad (\text{X.4.12})$$

It follows from (4.11) that the associated matrices obey the relation

$$R^h = R^g R^f. \quad (\text{X.4.13})$$

And, from the work of Section 8.4, we know that (4.12) yields the aberration generator relations

$$h_4 = f_4^{\text{tr}} + g_4, \quad (\text{X.4.14})$$

$$h_6 = f_6^{\text{tr}} + g_6 + [f_4^{\text{tr}}, g_4]/2, \quad (\text{X.4.15})$$

$$h_8 = f_8^{\text{tr}} + g_8 + [f_6^{\text{tr}}, g_6] - [f_4^{\text{tr}}, [f_4^{\text{tr}}, g_4]]/6 + [g_4, [g_4, f_4^{\text{tr}}]]/3, \text{ etc.} \quad (\text{X.4.16})$$

Inspection of (4.14) through (4.16) shows that the determination of the aberration generators involves carrying out the transformations (4.9) and the evaluation of certain Poisson brackets. It is these two tasks that may, in some cases, be simplified by multiplet decomposition.

Observe that any \mathcal{R}_g^{-1} is generated by \mathcal{L}_{\pm} and \mathcal{L}_0 . It follows from (3.22) through (3.24) that any \mathcal{R}_g^{-1} acting on any element of a given multiplet must give a result in the *same* multiplet. Specifically, one must have results of the form

$$\mathcal{R}_g^{-1} {}^{2n} \chi_m^j = \sum_{m'=-j}^j D_{m'm}^j [(R^g)^{-1}] {}^{2n} \chi_{m'}^j. \quad (\text{X.4.17})$$

Here the $D_{m'm}^j [(R^g)^{-1}]$ are the transformation functions associated with the symplectic matrices $(R^g)^{-1}$ and are the analytic continuation from $SU(2)$ to $Sp(2, \mathbb{R})$ of the $SU(2)$ Wigner functions (which are entire so that unique analytic continuation is always well defined). Special cases of the relations (4.17) are the results

$$\mathcal{R}_g^{-1} {}^{2n} \chi_0^0 = {}^{2n} \chi_0^0 \text{ for } 2n = 4 \text{ or } 8, \quad (\text{X.4.18})$$

which are immediately evident consequences of (3.17) and (3.18), and analogous relations for ${}^8 \chi_0^0$.

The relations (4.17) in analytic form may or may not be computationally useful.¹⁰ However, they do show what contributes to what. Suppose, for example, that f_4 is given by (3.25), and similarly g_4 is given by

$$g_4 = {}^4 d_0^0 {}^4 \chi_0^0 + \sum_{m=-2}^2 {}^4 d_m^2 {}^4 \chi_m^2. \quad (\text{X.4.19})$$

¹⁰What is essentially involved here is the operation (4.9). It can be realized numerically utilizing the efficient and fast algorithm described in Section 39.9. The operation (4.9) is of direct/immediate use if one wishes to compute the f_{2n}^{tr} . It is of intermediate use if one wishes to compute the $D_{m'm}^j$ using (4.42).

Then, from the definition (4.8) and using (4.17) and (4.18), we find that

$$\begin{aligned}
f_4^{\text{tr}} &= \mathcal{R}_g^{-1} f_4 = \mathcal{R}_g^{-1} [{}^4c_0^0 {}^4\chi_0^0 + \sum_{m=-2}^2 {}^4c_m^2 {}^4\chi_m^2] = \\
&= {}^4c_0^0 {}^4\chi_0^0 + \sum_{m'=-2}^2 \left\{ \sum_{m=-2}^2 D_{mm'}^2 [(R^g)^{-1}] {}^4c_m^2 \right\} {}^4\chi_{m'}^2 \\
&= {}^4c_0^0 {}^4\chi_0^0 + \sum_{m=-2}^2 \left\{ \sum_{m'=-2}^2 D_{mm'}^2 [(R^g)^{-1}] {}^4c_{m'}^2 \right\} {}^4\chi_m^2 \\
&= {}^4c_0^0 {}^4\chi_0^0 + \sum_{m=-2}^2 {}^4e_m^2 {}^4\chi_m^2
\end{aligned} \tag{X.4.20}$$

where the coefficients ${}^4e_m^2$ are given by the relation

$${}^4e_m^2 = \sum_{m'=-2}^2 D_{mm'}^2 [(R^g)^{-1}] {}^4c_{m'}^2. \tag{X.4.21}$$

All the ingredients are now available for insertion into (4.14) to yield the result

$$h_4 = f_4^{\text{tr}} + g_4 = ({}^4c_0^0 + {}^4d_0^0) {}^4\chi_0^0 + \sum_{m=-2}^2 ({}^4e_m^2 + {}^4d_m^2) {}^4\chi_m^2. \tag{X.4.22}$$

We see that the singlet contributions (the Petzval contributions) to h_4 are *purely additive*, and depend *only* on the singlet content of f_4 and g_4 . Similarly, the quintuplet content of h_4 depends *only* on the quintuplet content of f_4 and g_4 , although in a somewhat more complicated way: The quintuplet content of f_4 first has to be transformed by \mathcal{R}_g^{-1} before its addition to the quintuplet content of g_4 to yield the net quintuplet content for h_4 .

The discussion so far has dealt with the combining of third-order aberrations as described by (4.14). Now look at (4.15) which describes how fifth-order aberrations combine/arise. Evidently, by an argument similar to that made for third-order aberrations, the triplet and septuplet components of f_6 and g_6 contribute separately and independently to the triplet and septuplet components, respectively, of h_6 . Moreover, according to (4.15), there is a contribution arising from third-order aberrations due to the Poisson bracket term. We will discuss Poisson bracket terms shortly,

We end this part of the discussion with the observation that there are some similarities in the computation of h_8 and the computation of h_4 . From (3.73) and (3.84) we see that eighth-order polynomials of the form f_8 may also have a singlet content ${}^8\chi_0^0$, which may be viewed as a higher-order Petzval. And these singlet contributions to h_8 will also be purely additive. Moreover, there are quintuplet, and 9-tuple components of f_8 and g_8 that contribute separately and independently to the quintuplet, and 9-tuple components, respectively, of h_8 .

What remains is to study the Poisson bracket terms in the expressions (4.15), (4.16) etc. for h_6 , h_8 , etc. These terms describe how lower-order aberrations combine (feed up) to

produce higher-order aberrations.¹¹ We will begin with the Poisson bracket term $[f_4^{\text{tr}}, g_4]$ in (4.15). Before going into specifics, there are two general observations. First, the ordinary product and the Lie product (the Poisson bracket) of any two axially symmetric polynomials must also be axially symmetric. See Exercise 2.3. Second, there is the rule (7.6.14) relating the degree of a Poisson bracket to the degrees of its ingredients. From these observations it follows, for example, that $[f_4^{\text{tr}}, g_4]$ must be some linear combination of the ${}^6\chi_m^1$ and the ${}^6\chi_m^3$.

Let us now be more specific. From (4.19) and (4.20) we see that $[f_4^{\text{tr}}, g_4]$ is some linear combination of the Poisson brackets

$$[{}^4\chi_0^0, {}^4\chi_0^0], \quad (\text{X.4.23})$$

$$[{}^4\chi_0^0, {}^4\chi_m^2], \quad (\text{X.4.24})$$

$$[{}^4\chi_m^2, {}^4\chi_{m'}^2]. \quad (\text{X.4.25})$$

Evidently the Poisson bracket term (4.23) vanishes due to antisymmetry. We will soon see that the Poisson bracket terms (4.24) vanish due to axial symmetry. All that remains are the Poisson brackets (4.25). It follows that the only feed-up terms contributing to h_6 arise from quintuplet terms in f_4 interacting with quintuplet terms in g_4 . There are no feed up terms arising from a singlet term interacting with a singlet term, nor from the interaction of singlet and quintuplet terms.

To see that (4.24) vanishes, observe that

$$[{}^4\chi_0^0, {}^4\chi_m^2] = [(L_z)^2, {}^4\chi_m^2] = 2L_z[L_z, {}^4\chi_m^2] = 2L_z : L_z : {}^4\chi_m^2 = 2L_z \mathcal{L}_z {}^4\chi_m^2 = 0. \quad (\text{X.4.26})$$

Here we have used (3.11), the derivation property (1.7.7), and the axial symmetry of the ${}^4\chi_m^2$. We remark that in fact ${}^4\chi_0^0$ must have a vanishing Poisson bracket with *any* axially symmetric f_{2n} . See Exercise 4.3.

We now study the remaining quantities (4.25). Define polynomials $\theta_{m,m'}^{22}$ by the rule

$$\theta_{m,m'}^{22} = [{}^4\chi_m^2, {}^4\chi_{m'}^2]. \quad (\text{X.4.27})$$

Then, since \mathcal{L}_0 and \mathcal{L}_{\pm} are derivations with respect to the Poisson bracket Lie product, recall (5.3.9), we find the results

$$\mathcal{L}_0 \theta_{m,m'}^{22} = [\mathcal{L}_0 {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_0 {}^4\chi_{m'}^2] = (m+m')\theta_{m,m'}^{22}, \quad (\text{X.4.28})$$

$$\begin{aligned} \mathcal{L}_+ \theta_{m,m'}^{22} &= [\mathcal{L}_+ {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_+ {}^4\chi_{m'}^2] \\ &= [(2-m)(2+m+1)]^{1/2} [{}^4\chi_{m+1}^2, {}^4\chi_{m'}^2] + [(2-m')(2+m'+1)]^{1/2} [{}^4\chi_m^2, {}^4\chi_{m'+1}^2] \\ &= [(2-m)(2+m+1)]^{1/2} \theta_{m+1,m'}^{22} + [(2-m')(2+m'+1)]^{1/2} \theta_{m,m'+1}^{22}, \end{aligned} \quad (\text{X.4.29})$$

¹¹Such terms are called *secondary/induced/extrinsic* aberrations in the optics literature, and are distinguished from *primary/intrinsic* aberrations, the aberrations that are produced directly by the optical elements themselves. Thus, in relations of the form (4.15), (4.16), etc., the terms involving Poisson brackets are secondary, and those that do not are called primary. Actually, the terms primary/intrinsic are somewhat misleading. In reality, with a physically meaningful definition of aberrations, as a consequence of the symplectic condition any given element with axial symmetry can produce aberrations of any odd order 3, 5, 7, etc. See Subsubsection 2.2.3.

$$\begin{aligned}
\mathcal{L}_- \theta_{m,m'}^{22} &= [\mathcal{L}_- {}^4\chi_m^2, {}^4\chi_{m'}^2] + [{}^4\chi_m^2, \mathcal{L}_- {}^4\chi_{m'}^2] \\
&= [(2+m)(2-m+1)]^{1/2} [{}^4\chi_{m-1}^2, {}^4\chi_{m'}^2] + [(2+m')(2-m'+1)]^{1/2} [{}^4\chi_m^2, {}^4\chi_{m'-1}^2] \\
&= [(2+m)(2-m+1)]^{1/2} \theta_{m-1,m'}^{22} + [(2+m')(2-m'+1)]^{1/2} \theta_{m,m'-1}^{22}. \quad (\text{X.4.30})
\end{aligned}$$

Inspection of the relations (4.28) through (4.30) shows that they are analogous to the behavior of the (tensor) product of two $j = 2$ entities. Therefore all the standard Clebsch-Gordan and Wigner-Eckart $su(2)$ machinery is available. Also, following earlier reasoning, all the entries in (4.25) must be some linear combination of the ${}^6\chi_m^1$ and the ${}^6\chi_m^3$. Consequently there must be relations of the form

$$\begin{aligned}
\sum_{m_1 m_2} C(22j; m_1, m_2, m) [{}^4\chi_{m_1}^2, {}^4\chi_{m_2}^2] &= \sum_{m_1 m_2} C(22j; m_1, m_2, m) \theta_{m_1 m_2}^{22} \\
&= \delta_{j1} \alpha(221) {}^6\chi_m^1 + \delta_{j3} \alpha(223) {}^6\chi_m^3, \quad (\text{X.4.31})
\end{aligned}$$

where the coefficients C are $su(2)$ Clebsch-Gordan coefficients and the coefficients $\alpha(221)$ and $\alpha(223)$ are to be determined. See Exercise 4.4. The relations (4.31) can be inverted using the completeness properties of the Clebsch-Gordan coefficients to yield the final result

$$\begin{aligned}
[{}^4\chi_m^2, {}^4\chi_{m'}^2] &= \alpha(221) C(221; m, m', m+m') {}^6\chi_{m+m'}^1 \\
&\quad + \alpha(223) C(223; m, m', m+m') {}^6\chi_{m+m'}^3, \quad (\text{X.4.32})
\end{aligned}$$

where the coefficients α are seen to play the role of the reduced matrix elements that occur in applications of the Wigner-Eckart theorem. Again see Exercise 4.4. The needed Clebsch-Gordan coefficients are listed in Tables 4.1 and 4.2 below.

Table X.4.1: Some values of $C(221; m, m', m + m')$ and $C(223; m, m', m + m')$

m	m'	$m + m'$	$C(221; * * *)$	$C(223; * * *)$
2	2	4	0	0
2	1	3	0	$\sqrt{1/2}$
2	0	2	0	$\sqrt{1/2}$
2	-1	1	$\sqrt{1/5}$	$\sqrt{3/10}$
2	-2	0	$\sqrt{2/5}$	$\sqrt{1/10}$
1	2	3	0	$-\sqrt{1/2}$
1	1	2	0	0
1	0	1	$-\sqrt{3/10}$	$\sqrt{1/5}$
1	-1	0	$-\sqrt{1/10}$	$\sqrt{2/5}$
1	-2	-1	$\sqrt{1/5}$	$\sqrt{3/10}$
0	2	2	0	$-\sqrt{1/2}$
0	1	1	$\sqrt{3/10}$	$-\sqrt{1/5}$
0	0	0	0	0

Table X.4.2: Remaining values of $C(221; m, m', m + m')$ and $C(223; m, m', m + m')$

m	m'	$m + m'$	$C(221; * * *)$	$C(223; * * *)$
0	-1	-1	$-\sqrt{3/10}$	$\sqrt{1/5}$
0	-2	-2	0	$\sqrt{1/2}$
-1	2	1	$-\sqrt{1/5}$	$-\sqrt{3/10}$
-1	1	0	$\sqrt{1/10}$	$-\sqrt{2/5}$
-1	0	-1	$\sqrt{3/10}$	$-\sqrt{1/5}$
-1	-1	-2	0	0
-1	-2	-3	0	$\sqrt{1/2}$
-2	2	0	$-\sqrt{2/5}$	$-\sqrt{1/10}$
-2	1	-1	$-\sqrt{1/5}$	$-\sqrt{3/10}$
-2	0	-2	0	$-\sqrt{1/2}$
-2	-1	-3	0	$-\sqrt{1/2}$
-2	-2	-4	0	0

What remains is to find $\alpha(221)$ and $\alpha(223)$. An easy computation gives the result

$$[{}^4\chi_2^2, {}^4\chi_{-2}^2] = -(384/25)^{1/2} {}^6\chi_0^1 - (64/5)^{1/2} {}^6\chi_0^3. \quad (\text{X.4.33})$$

For the same j and m values use of (4.32) gives the result

$$\begin{aligned} [{}^4\chi_2^2, {}^4\chi_{-2}^2] &= \alpha(221) C(221; 2, -2, 0) {}^6\chi_0^1 \\ &\quad + \alpha(223) C(223; 2, -2, 0) {}^6\chi_0^3, \end{aligned} \quad (\text{X.4.34})$$

Upon comparing (4.33) and (4.34) and with the knowledge that ${}^6\chi_0^1$ and ${}^6\chi_0^3$ are linearly independent, see (3.85), we conclude that

$$\alpha(221) C(221; 2, -2, 0) = -(384/25)^{1/2}, \quad (\text{X.4.35})$$

$$\alpha(221) C(223; 2, -2, 0) = -(64/5)^{1/2}. \quad (\text{X.4.36})$$

According to Table 4.1 the Clebsch-Gordan coefficients associated with (4.35) and (4.36) are given by the relations

$$C(221; 2, -2, 0) = \sqrt{2/5}, \quad (\text{X.4.37})$$

$$C(223; 2, -2, 0) = \sqrt{1/10}. \quad (\text{X.4.38})$$

It follows from (4.35) through (4.38) that the constants $\alpha(221)$ and $\alpha(223)$ have the values

$$\alpha(221) = -(384/25)^{1/2}/(2/5)^{1/2} = -(192/5)^{1/2}, \quad (\text{X.4.39})$$

$$\alpha(223) = -(64/5)^{1/2}/(1/10)^{1/2} = -(128)^{1/2}. \quad (\text{X.4.40})$$

Correspondingly, (4.32) takes the final form

$$\begin{aligned} [{}^4\chi_m^2, {}^4\chi_{m'}^2] &= -(192/5)^{1/2} C(221; m, m', m + m') {}^6\chi_{m+m'}^1 \\ &\quad - (128)^{1/2} C(223; m, m', m + m') {}^6\chi_{m+m'}^3. \end{aligned} \quad (\text{X.4.41})$$

Upon reflection, we see that what has been illustrated is that the evaluation of Poisson brackets can be carried out in general in terms of $su(2)$ Clebsch-Gordan coefficients and a few simply computed numbers analogous to reduced matrix elements.

The discussion of the combining of fifth-order aberrations, and the feed-up effect of lower-order aberrations to contribute to fifth-order aberrations, is now complete. Moreover, it is clear from (4.16) and Poisson bracket relations analogous to (4.41) that the the combining of seventh and still higher-order aberrations, and the feed-up effect of lower-order aberrations to contribute to these higher-order aberrations, follow a similar pattern. All that is needed is the computation of the f_{2n}^{tr} and various single and multiple Poisson brackets. Finally we remark that, although the existence and knowledge of explicit formulas, such as (4.41), for Poisson brackets may be illuminating, they are not required for actual numerical computation. Their rapid numerical evaluation may be performed using the methods described in Section 39.8.

Exercises

X.4.1. Show that (4.14) through (4.16) are special cases of (8.4.32), (8.4.34), and (8.4.36).

X.4.2. Show, using (3.85) and (4.17), that there is the relation

$$D_{m'm}^j[(R^g)^{-1}] = [1/N(n, j)] \langle {}^n\chi_{m'}^j, \mathcal{R}_g^{-1} {}^n\chi_m^j \rangle \text{ for } n = 2, 4, \dots \quad (\text{X.4.42})$$

X.4.3. Suppose that some f_{2n} is axially symmetric, and therefore satisfies (2.17). Verify that

$$[{}^4\chi_0^0, f_{2n}] = [(L_z)^2, f_{2n}] = 2L_z[L_z, f_{2n}] = 2L_z : L_z : f_{2n} = 2L_z \mathcal{L}_z f_{2n} = 0. \quad (\text{X.4.43})$$

Verify also that $:{}^4\chi_0^0:$ and $:f_{2n}:$ commute,

$$\{:{}^4\chi_0^0:, :f_{2n}: \} =: [{}^4\chi_0^0, f_{2n}] := 0. \quad (\text{X.4.44})$$

X.4.4. Verify (4.33).

X.4.5. The purpose of this exercise is to establish (4.31) and its inverse (4.32). We have seen from (4.28) through (4.30) that the behavior of the $\theta_{m_1 m_2}^{22}$ is analogous to the behavior of the product of two $j = 2$ entities. From the Quantum Theory of angular momentum, we know that two entities of spin 2 can be combined/coupled to produce entities of spins 0, 1, 2, 3, and 4. That is what the left side of (4.31) seeks to do for the values $j = 0, 1, 2, 3, 4$. The right side of (4.31) states the expected results for these same j values. The expected results seem sensible for the cases $j = 1$ and $j = 3$. But what about the cases $j = 0, 2, 4$? Do we expect the left side of (4.31) to actually *vanish* in these cases as the right side states? We do. It can be shown that the $su(2)$ Clebsch-Gordan coefficients have the symmetry property

$$C(j_1 j_2 j; m_1, m_2, m_1 + m_2) = (-1)^{j_1 + j_2 - j} C(j_2 j_1 j; m_2, m_1, m_1 + m_2). \quad (\text{X.4.45})$$

A special case of (4.43) is the relation

$$C(22j; m_1, m_2, m_1 + m_2) = (-1)^{-j} C(22j; m_2, m_1, m_1 + m_2). \quad (\text{X.4.46})$$

That is, these C values are *even* under the interchange of m_1 and m_2 for even values of j , and *odd* under the interchange for odd values of j . But from the antisymmetry property of the Poisson bracket we know that

$$\theta_{m_1 m_2}^{22} = -\theta_{m_2 m_1}^{22}. \quad (\text{X.4.47})$$

Verify, therefore, that there must be the result

$$\sum_{m_1 m_2} C(22j; m_1, m_2, m) \theta_{m_1 m_2}^{22} = 0 \text{ for } j = 0, 2, 4. \quad (\text{X.4.48})$$

And this desired result follows simply from symmetry considerations alone without any additional information.

It can be shown that the $su(2)$ Clebsch-Gordan coefficients satisfy the *completeness relation*

$$\sum_j C(j_1 j_2 j; m_1, m_2, m_1 + m_2) C(j_1 j_2 j; m'_1, m'_2, m'_1 + m'_2) = \delta_{m_1 m'_1} \delta_{m_2 m'_2}. \quad (\text{X.4.49})$$

Use this result to derive (4.32) from (4.31).

Finally note that, because of the described symmetry properties of the Clebsch-Gordan coefficients and Poisson brackets, both sides of (4.32) are antisymmetric under the interchange of m and m' , as desired, and the Clebsch-Gordan coefficients involved in (4.32) vanish when $m = m'$. Scan the entries of Tables 4.1 and 4.2 to verify that the listed C values do indeed have these advertised symmetry properties.

X.4.6. Using (3.85) and (4.32) show that there is the relation

$$\begin{aligned} \langle {}^6\chi_{m+m'}^j, [{}^4\chi_m^2, {}^4\chi_{m'}^2] \rangle &= -(192/5)^{1/2} C(221; m, m', m + m') N(6, 1) \delta_{j1} \\ &\quad - (128)^{1/2} C(223; m, m', m + m') N(6, 3) \delta_{j3}. \end{aligned} \quad (\text{X.4.50})$$

X.5 Wave Aberrations

In Section 1 we learned that in geometric light-ray optics the initial and final ray variables, w^i and w^f , are related by a symplectic map \mathcal{M} . And in Subsection 2.2 we saw that in the case of axial symmetry \mathcal{M} can be written in the form (2.16). Moreover the f_{2n} must be linear combinations of the polynomials listed in (2.19), (2.35), and Subsection 3.4, etc. Thus, in the case of axial symmetry, a knowledge of the coefficients of these polynomials appearing in the f_{2n} is equivalent to a knowledge of the relation between w^f and w^i .¹² For future convenience we introduce the notation

$$w^i = (q_x, p_x, q_y, p_y) \text{ and } w^f = (Q_x, P_x, Q_y, P_y). \quad (\text{X.5.1})$$

Recall (1.12).

We know that in reality light consists of oscillating and propagating electromagnetic disturbances. Describing these disturbances in detail as they propagate through an optical system is very complicated. There are, of course, diffraction effects that occur because optical systems have finite apertures, and perhaps stops to remove offensive rays whose aberrations are too severe. But also light does not simply refract and forward propagate when entering and exiting lenses. At each interface going into and exiting a lens there are reflections as well as refractions. And these reflected disturbances, as they propagate backwards, both reflect again to propagate forward as well as refract to continue propagating backward, etc. Thus each lens and air space is filled with forward and backward propagating disturbances. In the design and construction of cheaper optical systems these reflections are ignored, but

¹²Although, for simplicity, we have made the assumption of axial asymmetry, we needn't do so. Without this assumption, there will be generators f_m with both even and odd m , and the f_m may depend on w in an arbitrary way. However, there will be no f_1 since it is still assumed that coordinates have been defined in such a way that \mathcal{M} sends the origin into itself.

they can be annoying. They can even be seriously disturbing in cheap eyeglasses. In more expensive systems lenses are coated to minimize reflections.

Although a detailed description of light propagating in optical systems is challenging, one can hope to construct a hybrid ray/wave description that would at least model some wave features of interest. The first step is to ignore the vector/polarization nature of electromagnetism to replace a two-vectors (\mathbf{E} and \mathbf{B}) theory by a scalar theory. The second is to ignore reflections and probably also absorption. The third is to assign a *phase* to each ray based on an assumed initial phase at some initial plane and the optical path length of the ray to some final plane or perhaps surface that it intercepts. Let us call this phase ϕ . If the propagating light is assumed to be monochromatic with angular frequency ω [time dependence of the form $\sin(\omega t)$], then we may write

$$\phi = [(\omega/c)A] \quad (\text{X.5.2})$$

where A is the optical path length given by (1.2). We also observe that A depends on the optical path, and that an optical path is determined by the initial conditions (q, p) . Therefore we may write

$$A = A(q, p). \quad (\text{X.5.3})$$

Correspondingly, the phase ϕ depends on (q, p) ,

$$\phi(q, p) = [(\omega/c)A(q, p)]. \quad (\text{X.5.4})$$

It is this dependence that we would like to know.

To our great pleasure we will find that this dependence can be obtained from a knowledge of \mathcal{M} and hence, in the case of axial symmetry, from a knowledge of the f_{2n} . Thus, the f_{2n} specify not only geometric light-ray optics, but also the *phase* in our approximate hybrid ray/wave description.

To begin our exploration, review Subsubsection 6.5.2.2. From there, in the context of general Lagrange/Hamiltonian mechanics, we find the relation

$$F_2(q, P, t) = \sum_k P_k Q_k - \int_{t^i}^t L \, d\tau. \quad (\text{X.5.5})$$

(See *.) There we also find the relations

$$\partial F_2 / \partial q_j = p_j, \quad \partial F_2 / \partial P_j = Q_j. \quad (\text{X.5.6})$$

Note that these latter relations determine $F_2(q, P, t)$ up to an additive constant.

Now rewrite (5.5) in the form

$$\int_{t^i}^t L \, d\tau = \sum_k P_k Q_k - F_2(q, P, t). \quad (\text{X.5.7})$$

It is a relation in the context of general Lagrangian/Hamiltonian mechanics. In the context of geometric light-ray optics, the role of time t is played by the coordinate z and the Lagrangian

L is given by (1.5). Therefore, in view of (1.2) and (5.5), in the case of geometric light-ray optics there is the relation

$$A(q, p) = \int_{z^i}^{z^f} L dz = \sum_k P_k Q_k - F_2(q, P, z^f). \quad (\text{X.5.8})$$

At this point we might rejoice to have found a formula for $A(q, p)$. But we must also admit that the formula is seriously tangled since it contains both old and new variables. To be satisfied, we must still disentangle it to find all quantities solely in terms of the variables q, p .

Some examples are useful to see how this disentanglement goes.

Preliminary Calculations

$$F_2(q, P) = aq^2 + 2b\mathbf{q} \cdot \mathbf{P} + cP^2 \quad (\text{X.5.9})$$

$$Q_i = \partial F_2 / \partial P_i = 2bq_i + 2cP_i \quad (\text{X.5.10})$$

$$p_i = \partial F_2 / \partial q_i = 2aq_i + 2bP_i \quad (\text{X.5.11})$$

$$P_i = (p_i - 2aq_i) / (2b) = q_i[-(a/b)] + p_i[1/(2b)] \quad (\text{X.5.12})$$

$$\begin{aligned} Q_i &= 2bq_i + 2cP_i = 2bq_i + 2c[(p_i - 2aq_i)/(2b)] \\ &= q_i[2b - 2a(c/b)] + p_i[(c/b)] \end{aligned} \quad (\text{X.5.13})$$

$$\begin{pmatrix} Q_i \\ P_i \end{pmatrix} = M \begin{pmatrix} q_i \\ p_i \end{pmatrix}. \quad (\text{X.5.14})$$

$$M = \begin{pmatrix} 2b - 2a(c/b) & c/b \\ -(a/b) & 1/(2b) \end{pmatrix} \quad (\text{X.5.15})$$

$$\det(M) = 1 - ac/b^2 + ac/b^2 = 1 \quad (\text{X.5.16})$$

$$\begin{pmatrix} Q_1 \\ P_1 \\ Q_2 \\ P_2 \end{pmatrix} = R \begin{pmatrix} q_1 \\ p_1 \\ q_2 \\ p_2 \end{pmatrix}. \quad (\text{X.5.17})$$

$$R = \begin{pmatrix} M & O \\ O & M \end{pmatrix} \quad (\text{X.5.18})$$

Thin Lens Case

$$R_f = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/f & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1/f & 1 \end{pmatrix}. \quad (\text{X.5.19})$$

$$M = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \quad (\text{X.5.20})$$

$$1/(2b) = 1 \Leftrightarrow b = (1/2). \quad (\text{X.5.21})$$

$$c/b = 0 \Leftrightarrow c = 0. \quad (\text{X.5.22})$$

$$a/b = 1/f \Leftrightarrow a = [1/(2f)]. \quad (\text{X.5.23})$$

$$2b - 2a(c/b) = 2b = 1, \text{ which is consistent with } M_{11} = 1. \quad (\text{X.5.24})$$

$$F_2(q, P) = aq^2 + 2b\mathbf{q} \cdot \mathbf{P} + cP^2 = [1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{P}. \quad (\text{X.5.25})$$

$$\mathbf{P} = (-1/f)\mathbf{q} + \mathbf{p}. \quad (\text{X.5.26})$$

$$\begin{aligned} F_2(q, P) &= [1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{P} = [1/(2f)]q^2 + \mathbf{q} \cdot [(-1/f)\mathbf{q} + \mathbf{p}] = \\ &= -[1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{p} \end{aligned} \quad (\text{X.5.27})$$

$$\mathbf{Q} = \mathbf{q}, \quad (\text{X.5.28})$$

$$\mathbf{Q} \cdot \mathbf{P} = (-1/f)q^2 + \mathbf{q} \cdot \mathbf{p} \quad (\text{X.5.29})$$

$$\begin{aligned} A(q, p) &= \mathbf{Q} \cdot \mathbf{P} - F_2(q, P, z^f) \\ &= [(-1/f)q^2 + \mathbf{q} \cdot \mathbf{p}] - \{-[1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{p}\} = -[1/(2f)]q^2. \end{aligned} \quad (\text{X.5.30})$$

$$R = R_d|_{d=f} R_f = \begin{pmatrix} 0 & f & 0 & 0 \\ -1/f & 1 & 0 & 0 \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/f & 1 \end{pmatrix}. \quad (\text{X.5.31})$$

$$M = \begin{pmatrix} 0 & f \\ -1/f & 1 \end{pmatrix} \quad (\text{X.5.32})$$

$$1/(2b) = 1 \Leftrightarrow b = 1/2 \quad (\text{X.5.33})$$

$$M|_{b=1/2} = \begin{pmatrix} 1 - 4ac & 2c \\ -2a & 1 \end{pmatrix} \quad (\text{X.5.34})$$

$$1 - 4ac = 0 \Leftrightarrow ac = 1/4 \quad (\text{X.5.35})$$

$$M = \begin{pmatrix} 0 & 2c \\ -2a & 1 \end{pmatrix} \quad (\text{X.5.36})$$

$$2c = f \text{ and } 2a = 1/f \Leftrightarrow 4ac = 1 \quad (\text{X.5.37})$$

Therefore, we have

$$b = 1/2, \quad a = 1/(2f), \quad c = f/2. \quad (\text{X.5.38})$$

And

$$F_2(q, P) = [1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{P} + (f/2)P^2 \quad (\text{X.5.39})$$

Also

$$P_i = p_i - (1/f)q_i \quad (\text{X.5.40})$$

so that

$$\begin{aligned} F_2(q, P)|_{p_i=0} &= [1/(2f)]q^2 - (1/f)\mathbf{q} \cdot \mathbf{q} + (f/2)(1/f)^2q^2 \\ &= q^2\{[1/(2f)] - (1/f) + [1/(2f)]\} = 0 \end{aligned} \quad (\text{X.5.41})$$

$$\begin{aligned} F_2(q, P)|_{\mathbf{P}=\mathbf{p}-(1/f)\mathbf{q}} &= [1/(2f)]q^2 + [\mathbf{q} \cdot \mathbf{P} + (f/2)P^2]|_{\mathbf{P}=\mathbf{p}-(1/f)\mathbf{q}} \\ &= [1/(2f)]q^2 + \mathbf{q} \cdot [\mathbf{p} - (1/f)\mathbf{q}] + (f/2)[\mathbf{p} - (1/f)\mathbf{q}] \cdot [\mathbf{p} - (1/f)\mathbf{q}] \\ &= [1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{p} - (1/f)q^2 + (f/2)p^2 - \mathbf{p} \cdot \mathbf{q} + [1/(2f)]q^2 \\ &= (f/2)p^2. \end{aligned} \quad (\text{X.5.42})$$

$$\begin{aligned} (f/2)p^2 &= (f/2)[\mathbf{P} + (1/f)\mathbf{q}] \cdot [\mathbf{P} + (1/f)\mathbf{q}] \\ &= (f/2)[P^2 + (2/f)\mathbf{q} \cdot \mathbf{P} + (1/f)^2q^2] \\ &= [1/(2f)]q^2 + \mathbf{q} \cdot \mathbf{P} + (f/2)P^2 \end{aligned} \quad (\text{X.5.43})$$

And

$$Q_i = fp_i \quad (\text{X.5.44})$$

So that

$$\mathbf{Q} \cdot \mathbf{P} = f\mathbf{p} \cdot [\mathbf{p} - (1/f)\mathbf{q}] = fp^2 - \mathbf{q} \cdot \mathbf{p} \quad (\text{X.5.45})$$

And

$$\begin{aligned} A(q, p) &= \sum_k P_k Q_k - F_2(q, P) = fp^2 - \mathbf{q} \cdot \mathbf{p} - (f/2)p^2 \\ &= (f/2)p^2 - \mathbf{q} \cdot \mathbf{p} \end{aligned} \quad (\text{X.5.46})$$

$$\begin{aligned} [f^2 + (\mathbf{q} - \mathbf{Q}) \cdot (\mathbf{q} - \mathbf{Q})]^{1/2} &= [f^2 + (\mathbf{q} - f\mathbf{p}) \cdot (\mathbf{q} - f\mathbf{p})]^{1/2} = \\ [f^2 + q^2 - 2f\mathbf{p} \cdot \mathbf{q} + f^2 p^2]^{1/2} &= f[1 + (q^2 - 2f\mathbf{p} \cdot \mathbf{q} + f^2 p^2)/f^2]^{1/2} = \\ f[1 + (1/2)(q^2 - 2f\mathbf{p} \cdot \mathbf{q} + f^2 p^2)/f^2 + \dots] &= \\ f + [1/(2f)]q^2 - \mathbf{p} \cdot \mathbf{q} + (f/2)p^2 + \dots. & \end{aligned} \quad (\text{X.5.47})$$

X.6 Maps/Lie Generators for Continuous Systems

X.7 Maps/Lie Generators for Discontinuous Systems

This section presents formulas for the generators of maps associated with lens surfaces. The derivation and content of these generators is given in the *Foundations* paper. In accord with the assumption of axial symmetry, we require that the surfaces to be employed are surfaces of revolution about the z axis, and therefore depend on x and y only through the variable $x^2 + y^2 = q^2$. We also require that they be *analytic* functions of the variable q^2 , and pass through the origin. Such surfaces are described by relations of the form

$$z = \beta_2(q^2) + \beta_4(q^2)^2 + \beta_6(q^2)^3 + \beta_8(q^2)^4 + \dots \quad (\text{X.7.1})$$

If only β_2 is nonvanishing, the surface is a parabola of revolution opening either to the left or the right depending on the sign of β_2 .

Often in optics surfaces are part of the surface of a *sphere*: Consider a sphere of radius r centered on the origin. Move it to the *left* (direction of decreasing z) while keeping its center on the z axis so that its *right* surface passes through the origin and curves to the *left*. Its surface in the vicinity of the origin is given by

$$z = -r + (r^2 - q^2)^{1/2} = -r + r[1 - (q^2/r^2)]^{1/2}. \quad (\text{X.7.2})$$

Recall the Taylor expansion

$$\begin{aligned} (1 - \lambda)^{1/2} &= 1 + [0 - (1/2)][\lambda] + [0 - (1/2)][1 - (1/2)][\lambda^2/2!] \\ &+ [0 - (1/2)][1 - (1/2)][2 - (1/2)][\lambda^3/3!] \\ &+ [0 - (1/2)][1 - (1/2)][2 - (1/2)][3 - (1/2)][\lambda^4/4!] + \dots \\ &= 1 - [(1/2)]\lambda - [(1/2)(1/2)(1/2)]\lambda^2 \\ &- [(1/2)(1/2)(3/2)(1/6)]\lambda^3 - [(1/2)(1/2)(3/2)(5/2)(1/24)]\lambda^4 + \dots \\ &= 1 - (1/2)\lambda - (1/8)\lambda^2 - (1/16)\lambda^3 - (5/128)\lambda^4 - \dots. \end{aligned} \quad (\text{X.7.3})$$

It follows that for the case (7.2) there is the expansion

$$\begin{aligned} z &= -r + r[1 - (1/2)(q^2/r^2) - (1/8)(q^2/r^2)^2 - (1/16)(q^2/r^2)^3 - (5/128)(q^2/r^2)^4 \dots] \\ &= [-1/(2r)]q^2 + [-1/(8r^3)](q^2)^2 + [-1/(16r^5)](q^2)^3 + [-5/(128r^7)](q^2)^4 + \dots \end{aligned} \quad (\text{X.7.4})$$

so that for such a spherical surface

$$\begin{aligned} \beta_2 &= -1/(2r), \quad \beta_4 = -1/(8r^3) = (\beta_2)^3, \\ \beta_6 &= -1/(16r^5) = 2(\beta_2)^5, \quad \beta_8 = -5/(128r^7) = 5(\beta_2)^7, \dots \end{aligned} \quad (\text{X.7.5})$$

Next again consider a sphere of radius r centered on the origin. Move it to the *right* (direction of increasing z) while keeping its center on the z axis so that its *left* surface passes through the origin and curves to the *right*. Its surface in the vicinity of the origin is given by

$$z = r - (r^2 - q^2)^{1/2} = r - r[1 - (q^2/r^2)]^{1/2}. \quad (\text{X.7.6})$$

It follows that for the case (7.6) there is the expansion

$$\begin{aligned} z &= r - r[1 - (1/2)(q^2/r^2) - (1/8)(q^2/r^2)^2 - (1/16)(q^2/r^2)^3 - (5/128)(q^2/r^2)^4 \dots] \\ &= [1/(2r)]q^2 + [1/(8r^3)](q^2)^2 + [1/(16r^5)](q^2)^3 + [5/(128r^7)](q^2)^4 + \dots \end{aligned} \quad (\text{X.7.7})$$

so that for such a spherical surface

$$\begin{aligned} \beta_2 &= 1/(2r), \quad \beta_4 = 1/(8r^3) = (\beta_2)^3, \\ \beta_6 &= 1/(16r^5) = 2(\beta_2)^5, \quad \beta_8 = 5/(128r^7) = 5(\beta_2)^7, \dots \end{aligned} \quad (\text{X.7.8})$$

It is assumed that a light ray crosses a surface from a medium having index of refraction n^- into a medium having index n^+ . The map associated with crossing such a surface is factorized in the form

$$\mathcal{M} = \exp(: f_2 :) \exp(: f_4 :) \exp(: f_6 :) \dots. \quad (\text{X.7.9})$$

Formulas for the f_n are available through f_8 . Below are the expressions for f_2 and f_4 :

$$f_2 = \beta_2(n^- - n^+)q^2; \quad (\text{X.7.10})$$

one can also write

$$f_2 = -[1/(2f)]q^2 \quad (\text{X.7.11})$$

with the *focal length* f given by

$$1/f = -2\beta_2(n^- - n^+). \quad (\text{X.7.12})$$

$$\begin{aligned} \text{Surface injects } f_4 &= -\beta_2^3[(n^- - n^+)/n^-]\{n^-[2 - (\beta_4/\beta_2^3)] - 2n^+\}(q^2)^2 \\ &\quad + 2\beta_2^2[(n^- - n^+)/n^-]q^2(\mathbf{p} \cdot \mathbf{q}) \\ &\quad + \beta_2[(n^- - n^+)/(2n^-n^+)]q^2p^2. \end{aligned} \quad (\text{X.7.13})$$

$$\begin{aligned} \text{Equivalently, } f_4 = & -\beta_2^3[(n^- - n^+)/n^-]\{n^-[2 - (\beta_4/\beta_2^3)] - 2n^+\}^4\chi_{-2}^2 \\ & + \beta_2^2[(n^- - n^+)/n^-]^4\chi_{-1}^2 \\ & + \beta_2[(n^- - n^+)/2(n^-n^+)][(2/3)^4\chi_0^0 + (1/6)^{1/2}4\chi_0^2]. \end{aligned} \quad (\text{X.7.14})$$

no: spherical aberration, coma, and astigmatism

yes: pocus, distortion, curvature of field \Rightarrow Petzval ${}^4\chi_0^0$ and Katarina ${}^4\chi_0^2$ combination.

Petzval and Katarina? Yes, but *only* in the curvature of field combination

$$p^2q^2 = (2/3)^4\chi_0^0 + (1/6)^{1/2}4\chi_0^2.$$

Only pocus is adjustable (depends on β_4) by making surface suitably aspherical.

Drifting injects $f_4 \sim (p^2)^2$.

Spherical aberration ${}^4\chi_2^2 = (p^2)^2$.

Pocus ${}^4\chi_{-2}^2 = (q^2)^2$.

$$\exp(- : g_2 :) \exp(: g_4 :) \exp(: g_2 :) = \exp(: \exp(- : g_2 :) g_4 :) = \exp(: g_4^{tr} :).$$

$$g_4^{tr} = \exp(- : g_2 :) g_4.$$

$$\exp(\lambda : q^2 :) = \exp(2\lambda\mathcal{L}_-) \Rightarrow \text{lensing produces lower } m \text{ values.}$$

$$\exp(\lambda : p^2 :) = \exp(-2\lambda\mathcal{L}_+) \Rightarrow \text{drifting produces higher } m \text{ values}$$

NOTICE: The first term in (X.7.14) is proportional to β_2^3 and the second is proportional to β_2^2 . And the pocus and distortion they produce become Katarina and coma and spherical aberration under the action of drifts. Therefore, if we wish to diminish third-order aberrations, the use of many weak lenses appears better than a few strong ones.

X.8 Three Sample Designs

X.8.1 Aberration Corrected Spot-Forming System

X.8.2 Aberration Corrected Doublet Imaging System

In this subsection we will illustrate how some of our earlier results can be used to design a doublet imaging system that is free of all third-order aberrations and four fifth-order aberrations. This system is illustrated schematically in Figure 7.1 below. Subsequently we will use it to design a doublet imaging system that is corrected to be free of all third-order aberrations and four fifth-order aberrations.

The system consists of four surfaces separated by drift spaces either in air or two possibly different refractive media. Between the object plane and *Surface S*¹ there is a *left-side* drift space (in air) of on-axis length D_L . Surfaces *S*¹ and *S*², with an on-axis separation of thickness t_L , constitute a first lens made of a medium with refractive index n_L . Surfaces *S*³ and *S*⁴, with an on-axis separation of thickness t_R , constitute a second lens made of a medium with refractive index n_R . Between surfaces *S*² and *S*³ there is a drift space (in air) of on-axis length D . Finally, between *S*⁴ and the image plane there is a *right-side* drift space (in air) of on-axis length D_R . The surfaces *S*¹ and *S*² will be chosen so that (in paraxial approximation) the first lens is converging, and the surfaces *S*³ and *S*⁴ will be chosen so that (in paraxial approximation) the second lens is diverging.

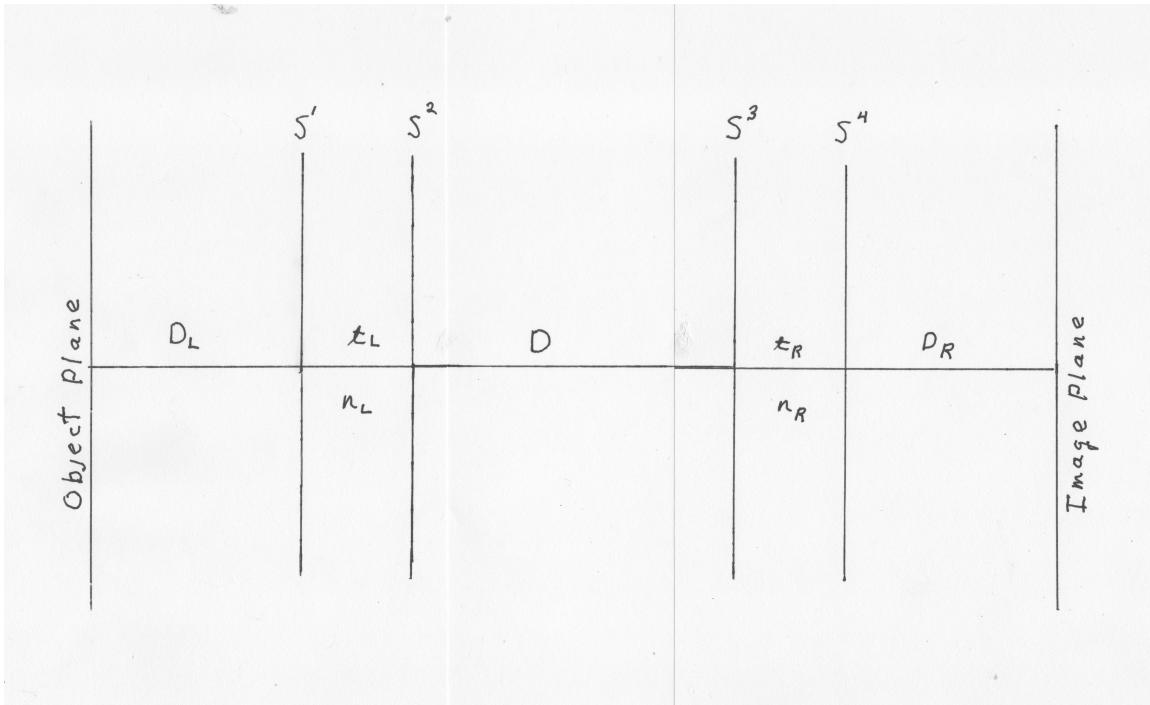


Figure X.8.1: Schematic layout of doublet imaging system that is free of all third-order aberrations and (potentially) four fifth-order aberrations. The object plane is on the left and the image plane is on the right.

Elimination of Petzval

We begin our design with the requirement that the net (third-order) Petzval aberration vanish. From our earlier work we know that only the maps associated with the surfaces S^1 through S^4 contribute to the Petzval, and their contributions are additive. According to *, passage through a surface S from a medium with refraction index n^- to a medium with refraction index n^+ makes a contribution to the Petzval coefficient given by the relation

$$\text{contribution} = \beta_2[(1/n^+) - (1/n^-)]. \quad (\text{X.8.1})$$

Here β_2 is the quadratic parameter of the surface.¹³ Consequently, for the four surfaces, we find the Petzval coefficient contributions to be as follows:

$$\text{For } S^1, n^- = 1 \text{ and } n^+ = n_L \Rightarrow \text{contribution} = \beta_2^1[(1/n_L) - 1], \quad (\text{X.8.2})$$

$$\text{For } S^2, n^- = n_L \text{ and } n^+ = 1 \Rightarrow \text{contribution} = \beta_2^2[1 - (1/n_L)], \quad (\text{X.8.3})$$

$$\text{For } S^3, n^- = 1 \text{ and } n^+ = n_R \Rightarrow \text{contribution} = \beta_2^3[(1/n_R) - 1], \quad (\text{X.8.4})$$

¹³Unlike the third-order aberrations associated with the ${}^4\chi_m^2$, the Petzval aberration (associated with ${}^4\chi_0^0$) is independent of the quartic parameter β_4 of surfaces. We may view parameters that govern paraxial behavior as being “paraxial” parameters. Consequently, the β_2 parameters, as well as indices of refraction and lengths, are paraxial parameters. By contrast, the β_4 , β_6 , etc. have no effect on paraxial behavior and therefore are not paraxial parameters. The Petzval is different from other third-order aberrations in that it is governed by paraxial parameters, and is independent of the β_4 parameters.

$$\text{For } S^4, n^- = n_R \text{ and } n^+ = 1 \Rightarrow \text{contribution} = \beta_2^4[1 - (1/n_R)]. \quad (\text{X.8.5})$$

Here the quantity β_2^j is the quadratic parameter for the j^{th} surface. The net Petzval coefficient is the sum of these terms, and we require that it vanish,

$$\text{Net Petzval coefficient} = (\beta_2^1 - \beta_2^2)[(1/n_L) - 1] + (\beta_2^3 - \beta_2^4)[(1/n_R) - 1] = 0. \quad (\text{X.8.6})$$

[In the optics literature the quantity on the right side of (8.6) is an example of what is called called the Petzval *sum*.] There are several ways to satisfy (7.6). For simplicity, we specify that

$$n_L = n_R = n. \quad (\text{X.8.7})$$

Also we specify that

$$\beta_2^1 > 0 \text{ and } \beta_2^2 = -\beta_2^1 \quad (\text{X.8.8})$$

so that the first lens is (symmetrically) biconvex and converging. And we specify that

$$\beta_2^3 < 0 \text{ and } \beta_2^4 = -\beta_2^3 \quad (\text{X.8.9})$$

so that the second lens is (symmetrically) biconcave and diverging. (Our intuition, which can be checked, is that minimizing the curvatures of all lens surfaces by making lenses symmetrical, which essentially amounts to sharing power equally between leading and trailing lens surfaces save for finite lens thickness effects, should on average help minimize aberrations.) With these specifications the requirement (7.6) takes the simpler form

$$\text{Net Petzval coefficient} = 2\beta_2^1[(1/n) - 1] + 2\beta_2^3[(1/n) - 1] = 0, \quad (\text{X.8.10})$$

and we see that (7.10) is satisfied providing

$$\beta_2^3 = -\beta_2^1. \quad (\text{X.8.11})$$

See Figure 7.2 where these specifications and the requirements (7.7) through (7.9) and (7.11) are depicted graphically. We conclude that, in order to be Petzval aberration free, a system must have both focusing and defocusing elements.

Paraxial Properties

The next design step is to examine, in the paraxial approximation, the optical transfer map associated with the drifts and lenses depicted in Figure 7.2. These maps can all be written as products of maps of the form $\exp(: f_2 :)$. Listed below are the f_2 polynomials for the various items depicted in Figure 7.2.

$$\text{Drift space of length } d \text{ in air: } f_2 = -(d/2)p^2. \quad (\text{X.8.12})$$

Here d takes the values

$$d = D_L, d = D, \text{ and } d = D_R. \quad (\text{X.8.13})$$

$$\text{Drift space of length } d \text{ in medium with index } n: f_2 = -[d/(2n)]p^2. \quad (\text{X.8.14})$$

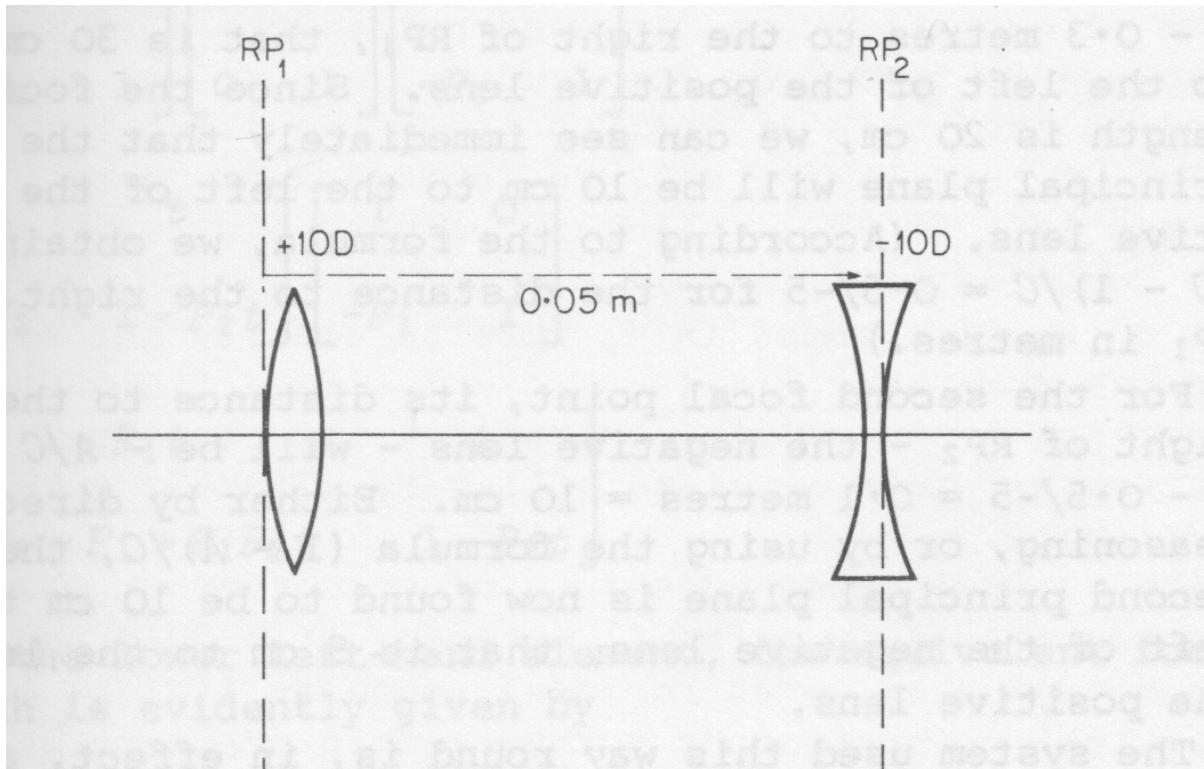


Figure X.8.2: Less schematic layout of imaging doublet system that is free of all third-order aberrations and four fifth-order aberrations. The object plane is on the far left and the image plane is on the far right (so that both are not visible), and only the shapes of the various lens surfaces and the lens thicknesses and spacings are illustrated. The *reference plane* RP_1 is at the beginning of the convex lens, and RP_2 is at the end of the concave lens. In the thin-lens approximation the convex and concave lenses have powers of $1/f = 10 \times 10^{-2}$ and $1/f = -10 \times 10^{-2}$, respectively.

Here we assume that both lenses in the doublet have on-axis thickness t so that

$$d = t. \quad (\text{X.8.15})$$

According to (*) the f_2 associated with passage through a surface S from a medium with refraction index n^- to a medium with refraction index n^+ is given by the relation

$$f_2 = \beta_2(n^- - n^+)q^2. \quad (\text{X.8.16})$$

Here again β_2 is the quadratic parameter for the surface. Therefore, for the surfaces S^1 through S^4 , there are the following general results:

$$\text{For } S^1, n^- = 1 \text{ and } n^+ = n_L \Rightarrow f_2 = \beta_2^1(1 - n_L)q^2, \quad (\text{X.8.17})$$

$$\text{For } S^2, n^- = n_L \text{ and } n^+ = 1 \Rightarrow f_2 = \beta_2^2(n_L - 1)q^2, \quad (\text{X.8.18})$$

$$\text{For } S^3, n^- = 1 \text{ and } n^+ = n_R \Rightarrow f_2 = \beta_2^3(1 - n_R)q^2, \quad (\text{X.8.19})$$

$$\text{For } S^4, n^- = n_R \text{ and } n^+ = 1 \Rightarrow f_2 = \beta_2^4(n_R - 1)q^2. \quad (\text{X.8.20})$$

We will use these general results for the specific cases described by (7.7) through (7.9), (7.11), (7.13), and (7.15).

We are now prepared to compute \mathcal{R} , the linear part of the transfer map for the optical system illustrated in Figure 7.2. Based on the results summarized in *, it is given by the product

$$\begin{aligned} \mathcal{R} = & \exp[-(D_L/2) : p^2 :] \exp[\beta_2^1(1 - n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[\beta_2^1(1 - n) : q^2 :] \times \\ & \exp[-(D/2) : p^2 :] \exp[-\beta_2^1(1 - n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[-\beta_2^1(1 - n) : q^2 :] \times \\ & \exp[-(D_R/2) : p^2 :]. \end{aligned} \quad (\text{X.8.21})$$

Let R be the matrix associated with \mathcal{R} . Since only the Lie operators $:p^2:$ and $:q^2:$ appear in \mathcal{R} , and since these operators map the pairs q_x, p_x and q_y, p_y into themselves and in the same way, it follows that R must be of the block form

$$R = \begin{pmatrix} G & O \\ O & G \end{pmatrix} \quad (\text{X.8.22})$$

where each block is 2×2 , the block G is symplectic, and the block O is the zero matrix. Therefore, for the computation of R , there is the simplification of only needing to work with various 2×2 matrices corresponding to the various $\exp(:f_2:)$. Let us list these matrices, call them K : For f_2 of the form (7.12) there is the correspondence

$$f_2 = -(d/2)p^2 \leftrightarrow K = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}. \quad (\text{X.8.23})$$

For f_2 of the form (7.14) there is the correspondence

$$f_2 = -[(d/(2n))p^2] \leftrightarrow K = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix}. \quad (\text{X.8.24})$$

For f_2 of the form (7.16) there is the correspondence

$$f_2 = \beta_2(n^- - n^+)q^2 \leftrightarrow K = \begin{pmatrix} 1 & 0 \\ 2\beta_2(n^- - n^+) & 1 \end{pmatrix}. \quad (\text{X.8.25})$$

Since (7.21) has nine factors, it follows that the G associated with \mathcal{R} is the product of nine 2×2 matrices of the form (7.23) through (7.25). We will eventually compute this G numerically. But we will first make some preliminary observations/calculations.

Map \mathcal{R} for the System and Map \mathcal{R}' for the Device

In practical applications, we may imagine that most of the parameter values in (7.21) are fixed save for D_L and D_R , which could be fairly readily adjusted to achieve imaging and the desired magnification. This circumstance suggests that we should understand the nature of the map that these leading and trailing drifts surround. That is, we are interested in the map \mathcal{R}' defined by the product

$$\begin{aligned} \mathcal{R}' = & \exp[\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[\beta_2^1(1-n) : q^2 :] \times \\ & \exp[-(D/2) : p^2 :] \exp[-\beta_2^1(1-n) : q^2 :] \exp\{-[t/(2n)] : p^2 :\} \exp[-\beta_2^1(1-n) : q^2 :]. \end{aligned} \quad (\text{X.8.26})$$

Compare (7.21) and (7.26). That is, we have the relation

$$\mathcal{R} = \exp[-(D_L/2) : p^2 :] \mathcal{R}' \exp[-(D_R/2) : p^2 :]. \quad (\text{X.8.27})$$

Put another way, we may view \mathcal{R}' as being the linear part of the map for the *optical device* and \mathcal{R} as being the linear part of the map for the complete *optical system*.

Normal Form

What would we like to know about \mathcal{R}' or, equivalently, the matrices R' and G' ? We will see that it is possible to associate with \mathcal{R}' a kind of *normal form*. Our discussion will be equivalent to the usual *principal plane* analysis.

Suppose, as a mathematical trick, we consider the map \mathcal{R}'' defined by relation

$$\mathcal{R}'' = \exp[(d_L/2) : p^2 :] \mathcal{R}' \exp[(d_R/2) : p^2 :]. \quad (\text{X.8.28})$$

We have “sandwiched” \mathcal{R}' between two *negative* length drift maps where d_L and d_R are to be determined. Let us compute the matrix G'' associated with \mathcal{R}'' . With the aid of (7.23)

we see that it is given by the relation

$$\begin{aligned}
G'' &= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} G' \begin{pmatrix} 1 & -d_L \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G'_{11} & G'_{12} \\ G'_{21} & G'_{22} \end{pmatrix} \begin{pmatrix} 1 & -d_L \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & -d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G'_{11} & -G'_{11}d_L + G'_{12} \\ G'_{21} & -G'_{21}d_L + G'_{22} \end{pmatrix} \\
&= \begin{pmatrix} G'_{11} - d_R G'_{21} & -G'_{11}d_L + G'_{12} - d_R(-G'_{21}d_L + G'_{22}) \\ G'_{21} & -G'_{21}d_L + G'_{22} \end{pmatrix} \\
&= \begin{pmatrix} G''_{11} & G''_{12} \\ G''_{21} & G''_{22} \end{pmatrix}.
\end{aligned} \tag{X.8.29}$$

Upon comparing the last two lines in (7.29) we see that

$$G''_{21} = G'_{21}. \tag{X.8.30}$$

Next we seek values of d_L and d_R such that

$$1 = G''_{11} = G'_{11} - d_R G'_{21} \Rightarrow d_R = (G'_{11} - 1)/G'_{21}, \tag{X.8.31}$$

$$1 = G''_{22} = -G'_{21}d_L + G'_{22} \Rightarrow d_L = (G'_{22} - 1)/G'_{21}. \tag{X.8.32}$$

We see that the goals $G''_{11} = 1$ and $G''_{22} = 1$ can be achieved provided

$$G'_{21} \neq 0. \tag{X.8.33}$$

For the values of d_R and d_L given by (7.31) and (7.32) we find that

$$\begin{aligned}
G''_{12} &= -G'_{11}d_L + G'_{12} - d_R(-G'_{21}d_L + G'_{22}) \\
&= -G'_{11}(G'_{22} - 1)/G'_{21} + G'_{12} - [(G'_{11} - 1)/G'_{21}]\{-G'_{21}[(G'_{22} - 1)/G'_{21}] + G'_{22}\} \\
&= -G'_{11}(G'_{22} - 1)/G'_{21} + G'_{12} - [(G'_{11} - 1)/G'_{21}] \\
&= [-G'_{11}(G'_{22} - 1) + G'_{12}G'_{21} - (G'_{11} - 1)]/G'_{21} \\
&= [-G'_{11}G'_{22} + G'_{12}G'_{21} + 1]/G'_{21} \\
&= [-\det(G') + 1]/G'_{21} \\
&= 0.
\end{aligned} \tag{X.8.34}$$

Here we have used the fact that G' is symplectic.¹⁴ We have verified the remarkable result that there is a (unique) choice for the pair d_L, d_R such that G'' takes the simple/normal form

$$G'' = \begin{pmatrix} 1 & 0 \\ G'_{21} & 1 \end{pmatrix}. \tag{X.8.35}$$

¹⁴As rewarding as the messy calculation (7.34) ultimately proved to be, it is/was actually not necessary. Once (7.31) through (7.33) are established, the relation $G''_{12} = 0$ must hold in order for G'' to be symplectic, which we already know to be the case.

Upon solving (7.29) for G' , we find the result

$$G' = \begin{pmatrix} 1 & d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & d_L \\ 0 & 1 \end{pmatrix}. \quad (\text{X.8.36})$$

We conclude that, in paraxial approximation, the device acts like a thin lens preceded by a drift of length d_L and followed by a drift of length d_R . And the thin lens has a focal length f given by

$$1/f = -G'_{21} \Leftrightarrow f = -1/(G'_{21}). \quad (\text{X.8.37})$$

That is, (7.35) can be rewritten in the form

$$G'' = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \quad (\text{X.8.38})$$

with f given by (7.37). For the problem at hand we will eventually find that

$$f > 0. \quad (\text{X.8.39})$$

From (7.35) we see that \mathcal{R}'' has the Lie form

$$\mathcal{R}'' = \exp[(G'_{21}/2) : q^2 :]. \quad (\text{X.8.40})$$

Correspondingly, from (7.28) and (7.40), we see that \mathcal{R}' has the factorization

$$\mathcal{R}' = \exp[-(d_L/2) : p^2 :] \exp[(G'_{21}/2) : q^2 :] \exp[-(d_R/2) : p^2 :].$$

Note that this result is consistent with (7.35) and (7.36).

Imaging Condition and Computation of Magnification

Let us use the normal form for \mathcal{R}' and the associated matrix G' to discuss the possible imaging and magnification properties of \mathcal{R} . We will do this by working with the associated matrix G . According to (7.27), it is given by the product

$$\begin{aligned} G &= \begin{pmatrix} 1 & D_R \\ 0 & 1 \end{pmatrix} G' \begin{pmatrix} 1 & D_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & d_L \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} G'' \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix}. \end{aligned} \quad (\text{X.8.41})$$

At this point, to simplify continuation of this calculation, it is convenient to define *effective* drift lengths D_L^e and D_R^e by the rules

$$D_L^e = D_L + d_L, \quad (\text{X.8.42})$$

$$D_R^e = D_R + d_R, \quad (\text{X.8.43})$$

With these definitions we can move the calculation (7.41) forward to find the result

$$\begin{aligned} G &= \begin{pmatrix} 1 & D_R + d_R \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L + d_L \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R^e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L^e \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & D_R^e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & D_L^e \\ -1/f & -D_L^e/f + 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 - D_R^e/f & D_L^e - D_R^e D_L^e/f + D_R^e \\ -1/f & -D_L^e/f + 1 \end{pmatrix}. \end{aligned} \quad (\text{X.8.44})$$

Suppose further we attempt to select D_L^e and D_R^e in such a way that

$$G_{12} = 0 \quad (\text{X.8.45})$$

so that R is imaging. Enforcing the relation (7.45) produces the chain of relations

$$\begin{aligned} 0 &= D_L^e - D_R^e D_L^e/f + D_R^e \Leftrightarrow \\ 0 &= 1/D_R^e - 1/f + 1/D_L^e \Leftrightarrow \\ 1/D_L^e + 1/D_R^e &= 1/f. \end{aligned} \quad (\text{X.8.46})$$

Note that the last line of (7.46) is the familiar elementary imaging condition except that it involves effective quantities.

When (7.45) is enforced, (7.44) takes the form

$$G = \begin{pmatrix} 1 - D_R^e/f & 0 \\ -1/f & -D_L^e/f + 1 \end{pmatrix}. \quad (\text{X.8.47})$$

As a sanity check, let us verify that this G is symplectic. We find that for this G

$$\begin{aligned} G_{11}G_{22} &= (1 - D_R^e/f)(1 - D_L^e/f) = 1 - (D_R^e/f + D_L^e/f) + (D_R^e/f)(D_L^e/f) \\ &= 1 - (D_L^e D_R^e/f)(1/D_L^e + 1/D_R^e) + (D_R^e D_L^e/f)(1/f) \\ &= 1 - (D_L^e D_R^e/f)(1/D_L^e + 1/D_R^e - 1/f) = 1 \end{aligned} \quad (\text{X.8.48})$$

as expected. [Here we have used the last line of (7.46).] It follows that (7.47) can be rewritten in the form

$$G = \begin{pmatrix} m & 0 \\ -1/f & 1/m \end{pmatrix} \quad (\text{X.8.49})$$

where m is the magnification given by the upper left entry in (7.47),

$$\begin{aligned} m &= -(D_R^e/f - 1) \\ &= -(D_R/f + d_R/f - 1). \end{aligned} \quad (\text{X.8.50})$$

From (7.50) it is evident that, for sufficiently large values of D_R , m is negative (the image is inverted as expected) and can become large in magnitude as $D_R \rightarrow \infty$ providing physically possible values of D_L can be found such that (7.46) is satisfied. Expressing the imaging condition (7.46) in terms of D_L and D_R yields the result

$$1/(D_L + d_L) + 1/(D_R + d_R) = 1/f. \quad (\text{X.8.51})$$

In the limit $D_R \rightarrow \infty$ we find from (7.51) that

$$D_L \rightarrow f - d_L. \quad (\text{X.8.52})$$

The right side of (7.52) is a physically possible value for D_L provided

$$f - d_L \geq 0 \Leftrightarrow f \geq d_L. \quad (\text{X.8.53})$$

We conclude that

$$m \rightarrow -\infty \text{ as } D_R \rightarrow +\infty \quad (\text{X.8.54})$$

provided (7.53) holds. In this case the magnification can be made arbitrarily large in magnitude.

How small can the magnification be? Can it be made vanishingly small for physical values of D_L and D_R ? From the second line of (7.50) we see that

$$m = 0 \Leftrightarrow D_R = f - d_R \quad (\text{X.8.55})$$

so that D_R is non-negative (physically possible) provided

$$f - d_R \geq 0 \Leftrightarrow f \geq d_R. \quad (\text{X.8.56})$$

But does the D_R given by (7.55) lead to a physical value of D_L when employed in (7.51)? Inserting the right side of (7.55) into (7.51) yields the relation

$$1/(D_L + d_L) + 1/f = 1/f \Rightarrow 1/(D_L + d_L) = 0 \Rightarrow D_L = +\infty. \quad (\text{X.8.57})$$

We conclude that $D_R \rightarrow (f - d_R)$ and $D_L \rightarrow +\infty$ is consistent with imaging and results in vanishing magnification. This conclusion is valid provided (7.56) holds.

Thin-Lens Approximation

Eventually we will want to compute to compute G' and hence R' . According to (7.26) this computation involves the product of 7 matrices, and is therefore best done numerically. But before doing so it would be useful to have an approximate result to get some feeling for the expected nature of the exact result. This can be done by treating the two lenses in the thin-lens approximation. That is, we will set $t = 0$ in (7.26). When this is done, what remains is to compute the map $\bar{\mathcal{R}}'$ defined by the product

$$\bar{\mathcal{R}}' = \exp[2\beta_2^1(1-n) : q^2 :] \exp[-(D/2) : p^2 :] \exp[-2\beta_2^1(1-n) : q^2 :]. \quad (\text{X.8.58})$$

Since the Lie transformation $\exp[2\beta_2^1(1-n) : q^2 :]$ is that for a thin lens, let us make the correspondence

$$\exp[2\beta_2^1(1-n) : q^2 :] \leftrightarrow \begin{pmatrix} 1 & 0 \\ -1/F & 1 \end{pmatrix} \quad (\text{X.8.59})$$

where F is the focal length of the first lens in the thin-lens approximation. So doing yields the relation

$$-1/F = 4\beta_2^1(1-n). \quad (\text{X.8.60})$$

Note, according to our prescription that the first lens in the doublet be focusing, recall (7.8), it follows that

$$F > 0. \quad (\text{X.8.61})$$

Now, to compute the matrix \bar{G}' associated with $\bar{\mathcal{R}'}$, we only need to compute the matrix product

$$\begin{aligned} \bar{G}' &= \begin{pmatrix} 1 & 0 \\ 1/F & 1 \end{pmatrix} \begin{pmatrix} 1 & D \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/F & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 1/F & 1 \end{pmatrix} \begin{pmatrix} 1-D/F & D \\ -1/F & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1-D/F & D \\ -D/F^2 & 1+D/F \end{pmatrix}. \end{aligned} \quad (\text{X.8.62})$$

What can we conclude in the thin-lens approximation? Let us apply the normal-form procedure to \bar{G}' . First, in analogy to (7.35) and from (7.62), we see that

$$\bar{G}'' = \begin{pmatrix} 1 & 0 \\ \bar{G}'_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -D/F^2 & 1 \end{pmatrix}. \quad (\text{X.8.63})$$

Also we know that $D > 0$ and therefore

$$\bar{G}'_{21} = -D/F^2 < 0. \quad (\text{X.8.64})$$

That is, in the thin-lens approximation, the net effect of the doublet is *focusing*.¹⁵ In analogy to (7.37) we make the definition

$$1/\bar{f} = -\bar{G}'_{21} \Leftrightarrow \bar{f} = -1/(\bar{G}'_{21}) = F^2/D. \quad (\text{X.8.65})$$

That is, (7.63) can be rewritten in the form

$$G'' = \begin{pmatrix} 1 & 0 \\ -1/\bar{f} & 1 \end{pmatrix} \quad (\text{X.8.66})$$

with \bar{f} given by (7.65). Moreover, from (7.31) and (7.32) we find that

$$\bar{d}_L = (\bar{G}'_{22} - 1)/\bar{G}'_{21} = (D/F)/(-D/F^2) = -F \quad (\text{X.8.67})$$

¹⁵Note that this conclusion holds no matter the sign F . This observation is the essence of *strong focussing* in Accelerator Physics applications. Although only obtained here in the thin-lens approximation, there are analogous results for thick lenses and thick magnetic elements.

$$\bar{d}_R = (\bar{G}'_{11} - 1)/\bar{G}'_{21} = -(D/F)/(-D/F^2) = F. \quad (\text{X.8.68})$$

Let us use these thin-lens results to examine what magnifications can be achieved in the thin-lens approximation. According to (7.54) the magnification can be made arbitrarily large in magnitude providing (7.53) holds. But, in the thin-lens approximation, we find that

$$\bar{f} - \bar{d}_L = \bar{f} + F > 0 \quad (\text{X.8.69})$$

because both \bar{f} and F are positive. Thus, if the thin-lens approximation is to be believed, the magnification can be made arbitrarily large in magnitude. What about making the magnification arbitrarily small? According to (7.55) the magnification can be made to vanish providing (7.56) holds. But, in the thin-lens approximation we find

$$\bar{f} - \bar{d}_R = F^2/D - F. \quad (\text{X.8.70})$$

Therefore

$$\bar{f} - \bar{d}_R \geq 0 \Leftrightarrow F^2/D - F \geq 0 \Leftrightarrow F/D - 1 \geq 0 \Leftrightarrow F \geq D \Leftrightarrow D \leq F. \quad (\text{X.8.71})$$

The inequality on the far right side of (7.71) provides a design criterion: D must not be too large. And if this criterion is met and the thin-lens approximation is to be believed, then arbitrarily small magnification can also be achieved.

Case for which Both Lenses Have Finite Thickness

We have verified that a doublet can be designed to have satisfactory paraxial performance in the thin-lens approximation. The next step is to verify that a doublet system can be found that has satisfactory paraxial performance when the two lenses have finite thickness. As an example, we suppose the doublet has the parameter values

$$\beta_2^1 = \text{to be determined by fitting}, \quad (\text{X.8.72})$$

$$n = 1.5, \quad (\text{X.8.73})$$

$$t = 0.75, \quad (\text{X.8.74})$$

$$D = 4.75, \quad (\text{X.8.75})$$

and that the remaining β_2^j are given by (7.8), (7.9), and (7.11).

Fitting the Focal Length

We now compute G' while varying β_2^1 to achieve some desired value for the focal length as given by (7.37). For example, suppose we wish/aim to have

$$f = 20 \Leftrightarrow G'_{21} = -0.05. \quad (\text{X.8.76})$$

For the aim ($G'_{21} = -0.05$) we find for the doublet that

$$G' = \begin{pmatrix} 4.69171E - 01 & 5.74703E + 00 \\ -5.00000E - 02 & 1.51895E + 00 \end{pmatrix} \quad (\text{X.8.77})$$

and

$$\beta_2^1 = 5.0003711901875095E - 02. \quad (\text{X.8.78})$$

Verification that Any Magnification can be Achieved

From (7.77) we see that

$$-1/f = G'_{21} = -5.00000E - 02 \Leftrightarrow f = 20 \text{ as desired,} \quad (\text{X.8.79})$$

$$d_L = (G'_{22} - 1)/G'_{21} = (1.51895 - 1.0)/(-.05) = -10.3790424, \quad (\text{X.8.80})$$

$$d_R = (G'_{11} - 1)/G'_{21} = (0.469171 - 1.0)/(-.05) = 10.6165777. \quad (\text{X.8.81})$$

Consequently,

$$f - d_L = 20 + 10.3790424 = 30.3790424 > 0 \quad (\text{X.8.82})$$

and

$$f - d_R = 20 - 10.6165777 = 9.3834224 > 0. \quad (\text{X.8.83})$$

We conclude that for this doublet there are physical/positive values of D_L and D_R for which any desired (negative) value of m can be achieved for the full system.

Selection of Magnification m and Determination of Lengths D_L and D_R

From (7.47) and (7.49) we see that

$$-D_L^e/f + 1 = 1/m \quad (\text{X.8.84})$$

from which it follows that

$$D_L = f[1 - (1/m)] - d_L. \quad (\text{X.8.85})$$

And from (7.50) we see that

$$D_R = f(1 - m) - d_R. \quad (\text{X.8.86})$$

Let us now select some value for m . For example, suppose we select the value

$$m = -0.5 = -1/2. \quad (\text{X.8.87})$$

Then D_L and D_R have the values

$$D_L = 20[1 + 2] + 10.3790424 = 70.3790424 \quad (\text{X.8.88})$$

and

$$D_R = 20(3/2) - 10.6165777 = 19.3834223. \quad (\text{X.8.89})$$

And, for the full system consisting of the device plus leading and trailing drifts, we find that the matrix R associated with the full linear map \mathcal{R} has the related matrix G given by

$$G = \begin{pmatrix} -5.00000E - 01 & 0.00000E + 00 \\ -5.00000E - 02 & -2.00000E + 00 \end{pmatrix}. \quad (\text{X.8.90})$$

Evidently, the full map \mathcal{M} is imaging in paraxial approximation because $G_{12} = 0$, and the magnification is

$$m = G_{11} = -0.50, \quad (\text{X.8.91})$$

as desired.

Vanishing of Petzval

What can be said about the net third-order aberrations of this system consisting of the doublet plus leading and trailing drifts when each of the elements of the system is treated through third order? The matrix R and the f_4 entries for \mathcal{M} are listed below.

Exhibit 7.1: \mathcal{M} for system with spherical lenses and corrected Petzval.

matrix for map is :

```
-5.00000E-01 -6.39488E-14  0.00000E+00  0.00000E+00
-5.00000E-02 -2.00000E+00  0.00000E+00  0.00000E+00
 0.00000E+00  0.00000E+00 -5.00000E-01 -6.39488E-14
 0.00000E+00  0.00000E+00 -5.00000E-02 -2.00000E+00
```

nonzero elements in generating polynomial are :

```
f( 84)=f( 40 00 00 )=-1.13076444191277E-03
f( 85)=f( 31 00 00 )= 0.10631214174624
f( 90)=f( 22 00 00 )= -3.8106320655559
f( 95)=f( 20 20 00 )=-2.26152888382554E-03
f( 96)=f( 20 11 00 )= 0.10631214174624
f( 99)=f( 20 02 00 )= -1.2702106885186
f(105)=f( 13 00 00 )= 61.310392578458
f(110)=f( 11 20 00 )= 0.10631214174624
f(111)=f( 11 11 00 )= -5.0808427540746
f(114)=f( 11 02 00 )= 61.310392578458
f(140)=f( 04 00 00 )= -375.30889070384
f(145)=f( 02 20 00 )= -1.2702106885186
f(146)=f( 02 11 00 )= 61.310392578458
f(149)=f( 02 02 00 )= -750.61778140769
f(175)=f( 00 40 00 )=-1.13076444191277E-03
f(176)=f( 00 31 00 )= 0.10631214174624
f(179)=f( 00 22 00 )= -3.8106320655559
f(185)=f( 00 13 00 )= 61.310392578458
f(195)=f( 00 04 00 )= -375.30889070384
```

Calculation shows that for this f_4

$$\langle {}^4\chi_0^0, f_4 \rangle = 0, \quad (\text{X.8.92})$$

thereby indicating that the Petzval aberration indeed vanishes as desired. Alternatively, using (2.42), (2.43), and (3.23), we expect that

$$0 = (2C - 4D) = f(1, 1, 1, 1) - 4f(0, 2, 2, 0) = 0 \Leftrightarrow f(111) - 4f(145) = 0. \quad (\text{X.8.93})$$

Examination of the values for $f(111)$ and $f(145)$ in the list of f_4 values above shows that the relation on the right side of (7.93) is indeed satisfied.

Evidently many of the f_4 entries listed above are nonzero. Calculation shows that there are the results

$$\langle {}^4\chi_2^2, f_4 \rangle = *, \quad (\text{X.8.94})$$

$$\langle {}^4\chi_1^2, f_4 \rangle = *, \quad (\text{X.8.95})$$

$$\langle {}^4\chi_0^2, f_4 \rangle = *, \quad (\text{X.8.96})$$

$$\langle {}^4\chi_{-1}^2, f_4 \rangle = *, \quad (\text{X.8.97})$$

$$\langle {}^4\chi_{-2}^2, f_4 \rangle = *. \quad (\text{X.8.98})$$

In view of (3.15) the scalar product results (7.92) and (7.94) through (7.98) specify f_4 completely because of the assumption/imposition of axial symmetry.

We have examined a particular system which is specified by the parameter values given by (7.73) through (7.75), (7.8), (7.9), (7.11) and the requirements (7.76) [which led to (7.78)] and (7.91). For this system we have verified that the third-order Petzval aberration vanishes, as desired. But upon reflection we see that, no matter what parameter values and requirements are imposed, the third-order Petzval aberration will vanish as long as (7.6) is satisfied.

Elimination of Remaining Third-Order Aberrations

Ideally we would like to have vanishing values for *all* the coefficients A through E appearing in (2.30) through (2.34). There are five such coefficients, but we have already caused the Petzval combination ($C - 2D$) to vanish thereby leaving four more goals to be met. We also observe that there are four surfaces S^1 through S^4 for which values β_4^1 through β_4^4 can be assigned. Can they be set in such a way that all the f_4 save for the F terms vanish? We will find that the answer is *yes*, but the matter is subtle.

For a spherical surface of radius r there is the relation

$$\beta_4 = (\beta_2)^3. \quad (\text{X.8.99})$$

See (7.5) and (7.8). For the calculation that produced the f_4 in Exhibit 7.1, the β_4^j values were set in such a way that the relation (8.99) was satisfied for all four surfaces. Equivalently, the surfaces were assumed to be spherical.

But what happens if the β_4^j values are instead set to zero? In this case, because all surfaces are now parabolas of revolution through fourth order, one might hope that third-order aberrations would be reduced. Below are the relevant scalar products for the f_4 found in this case:

$$\langle {}^4\chi_0^0, f_4 \rangle = 0, \quad (\text{X.8.100})$$

$$\langle {}^4\chi_2^2, f_4 \rangle = *, \quad (\text{X.8.101})$$

$$\langle {}^4\chi_1^2, f_4 \rangle = *, \quad (\text{X.8.102})$$

$$\langle {}^4\chi_0^2, f_4 \rangle = *, \quad (\text{X.8.103})$$

$$\langle {}^4\chi_{-1}^2, f_4 \rangle = *, \quad (\text{X.8.104})$$

$$\langle {}^4\chi_{-2}^2, f_4 \rangle = *. \quad (\text{X.8.105})$$

Looking at (8.100), we see that the Petzval still vanishes as before. But this is not surprising since we know that the Petzval is independent of β_4 , and the condition (7.6) is still met because the paraxial parameters have not been changed. What about the remaining entries?

Comparison of the entries in (7.94) through (7.98) with those in (7.103) through (7.107) shows that the latter are still sizable despite all surfaces being parabolic. Why is this? First of all, surface maps are not the only source of aberrations. As we see from $*$, transfer maps for drifts involve ${}^4\chi_2^2, {}^6\chi_3^3 \dots$ generators. And these generators can turn into ${}^4\chi_m^2, {}^6\chi_m^3 \dots$ generators under the action of $\exp(: f_2 :)$ maps that occur/act in the course of concatenation. Second, inspection of $*$ for example, shows that for a surface map f_4 generator there is the expansion

$$f_4 = *{}^4\chi_0^0 + *{}^4\chi_0^2 + *{}^4\chi_{-1}^2 + *{}^4\chi_{-2}^2, \quad (\text{X.8.106})$$

and only the ${}^4\chi_{-2}^2$ term depends on β_4 but does not vanish when $\beta_4 = 0$. See Exercise *. Evidently there are non-vanishing ${}^4\chi_m^2$ terms even when $\beta_4 = 0$.

What to do now that a simple strategy has been explored and found wanting? Since there are four goals to be achieved and four β_4^j parameters to be set, we might try varying the β_4^j to meet the goals

$$\langle {}^4\chi_m^2, f_4 \rangle = 0 \text{ for } m = 2, 1, 0, -1. \quad (\text{X.8.107})$$

Experience shows that this strategy succeeds, as hoped. Below are values for the f_4 when the β_4^j are optimally set. Evidently all entries *vanish* save for the coefficients of $q_x^4, q_x^2q_y^2$, and q_y^4 . Examine the entries for $f(84), f(95)$, and $f(175)$. Note that $f(95) = 2f(84) = 2f(175)$, and therefore $q_x^4, q_x^2q_y^2$, and q_y^4 appear in the *pocus* combination ${}^4\chi_{-2}^2 = (q^2)^2$. All third-order aberrations that affect image formation have been caused to vanish. This was accomplished by selecting the values

$$\beta_4^1 = -2.19365745146874E - 02, \quad (\text{X.8.108})$$

$$\beta_4^2 = -2.53461805590275E - 02, \quad (\text{X.8.109})$$

$$\beta_4^3 = -5.23769203858872E - 02, \quad (\text{X.8.110})$$

$$\beta_4^4 = -4.30353218952918E - 02. \quad (\text{X.8.111})$$

It can be shown that the calculation leading to the satisfaction of (8.107), after the Petzval has already been eliminated, amounts to a linear fitting operation, and consequently the values (8.108) through (8.111) are unique.

Exhibit 7.2: \mathcal{M} when the β_4^j are given the values (8.108) through (8.111).

matrix for map is :

```
-5.00000E-01 -6.39488E-14 0.00000E+00 0.00000E+00
-5.00000E-02 -2.00000E+00 0.00000E+00 0.00000E+00
0.00000E+00 0.00000E+00 -5.00000E-01 -6.39488E-14
0.00000E+00 0.00000E+00 -5.00000E-02 -2.00000E+00
```

nonzero elements in generating polynomial are :

```
f( 84)=f( 40 00 00 )=-1.73560018907580E-05
f( 85)=f( 31 00 00 )= 1.23620731706797E-15
f( 90)=f( 22 00 00 )=-4.32362479152459E-14
```

```

f( 95)=f( 20 20 00 )=-3.47120037815021E-05
f( 96)=f( 20 11 00 )= 5.36896915814822E-16
f( 99)=f( 20 02 00 )=-5.55805401702969E-15
f(105)=f( 13 00 00 )= 6.72684130620382E-13
f(110)=f( 11 20 00 )= 5.25621213220973E-16
f(111)=f( 11 11 00 )=-2.17187379192296E-14
f(114)=f( 11 02 00 )= 2.17381668221606E-13
f(140)=f( 04 00 00 )=-3.93240995322230E-12
f(145)=f( 02 20 00 )=-5.29090660172926E-15
f(146)=f( 02 11 00 )= 2.10609307771392E-13
f(149)=f( 02 02 00 )=-2.00106597958438E-12
f(175)=f( 00 40 00 )=-1.73560018907580E-05
f(176)=f( 00 31 00 )= 1.23620731706797E-15
f(179)=f( 00 22 00 )=-4.32362479152459E-14
f(185)=f( 00 13 00 )= 6.72684130620382E-13
f(195)=f( 00 04 00 )=-3.93240995322230E-12

```

Elimination of First Four Leading Fifth-Order Aberrations

The values of the β_6^j are also available to be set thereby attempting to cause four fifth-order aberrations to vanish. On the assumption that it is the sixth-order monomials with the highest powers of p_x and p_y that are the most damaging for image formation, we may attempt to cause the coefficients of ${}^6\chi_3^3$, ${}^6\chi_2^3$, ${}^6\chi_1^3$, and ${}^6\chi_0^3$ in f_6 to vanish by a suitable choice of the β_6^j . This goal can also be achieved. Below are the entries in f_6 for two cases. In the first case the β_6^j are set to zero. In the second case the β_6^1 through β_6^4 are set to cause the coefficients of ${}^6\chi_3^3$ through ${}^6\chi_0^3$ in f_6 to vanish. [In both cases the β_4^j are set to the values (7.110) through (7.113)]. These β_6^j have the values

$$\beta_6^1 = , \quad (\text{X.8.112})$$

$$\beta_6^2 = , \quad (\text{X.8.113})$$

$$\beta_6^3 = , \quad (\text{X.8.114})$$

$$\beta_6^4 = . \quad (\text{X.8.115})$$

Entries in f_6 when the $\beta_6^j = 0$.

Entries in f_6 when the β_6^1 through β_6^4 have the values (7.114) through (7.117).

Elimination of Remaining Fifth-Order Aberrations

Corrector Strengths Depend on Magnification

Ray Traces

X.8.3 Aberration Corrected Hubble and James Webb Telescopes

Exercises

X.9 Inclusion of Chromatic Effects

X.10 Possibly Complementary Approaches

X.10.1 The Constant Index Case

Suppose the lenses of an optical system are all made of constant (not graded) index materials. In that case computers can numerically trace a large number of rays through the system in a very short time simply by invoking Snell's law at each interface.

In that case fairly rapid ray traces can also be performed using Lie methods. The interface maps \mathcal{S} described in the Technical Report *Foundations of a Lie* ... can be computed in milliseconds using equations (7.46a) through (7.46d) of that paper and combined in further milliseconds with surrounding transit maps (6.10) to yield the f_m displayed in the first line of equation (2.44) below:

$$\begin{aligned} w_\alpha^f &= \mathcal{M}w_\alpha^i = \{\exp(: f_2 :) \exp(: f_4 :) \exp(: f_6 :) \exp(: f_8 :) \dots\} w_\alpha^i \\ &= g_1^\alpha(\mathbf{w}^i) + g_3^\alpha(\mathbf{w}^i) + g_5^\alpha(\mathbf{w}^i) + g_7^\alpha(\mathbf{w}^i) \dots \end{aligned} \quad (\text{X.10.1})$$

Then the homogeneous polynomials g_m^α displayed in the second line of (2.44) can be found in a few milliseconds more. The terms $g_1^\alpha(\mathbf{w}^i)$ describe the paraxial behavior of the optical system, and the terms $g_3^\alpha(\mathbf{w}^i)$, $g_5^\alpha(\mathbf{w}^i)$, ... describe ever higher degree departures from paraxial behavior. [Note that the g_m^α are *not* independent because of the symplectic condition (1.16). Therefore they are ill suited for use in optimization/fitting procedures. By contrast, the f_{2n} are independent, and any choice for them is consistent with the symplectic condition.] Finally, these polynomials $g_m^\alpha(\mathbf{w}^i)$ can be evaluated rapidly for any collection of initial conditions \mathbf{w}^i to find the associated final conditions \mathbf{w}^f . And if ray coordinates are desired at intermediate positions, they may be found by performing ray traces at intermediate positions as the full end-to-end map \mathcal{M} is being built up.

Of course, this use of Lie methods presumes that the series in the second line of (2.44) is convergent and that to good approximation terms beyond some degree can be neglected. But this can be checked using the Snell's law ray trace. For example, let $w_\alpha^{fs\text{rt}}$ denote the result of a Snell's law ray trace for some initial condition \mathbf{w}^i and let $w_\alpha^{f[7]}$ be the associated through seventh order result

$$w_\alpha^{f[7]} = g_1^\alpha(\mathbf{w}^i) + g_3^\alpha(\mathbf{w}^i) + g_5^\alpha(\mathbf{w}^i) + g_7^\alpha(\mathbf{w}^i). \quad (\text{X.10.2})$$

Then we expect the result

$$w_{\alpha}^{f[7]} = w_{\alpha}^{fs\text{slt}} + O(|\mathbf{w}^i|^9). \quad (\text{X.10.3})$$

The validity of (2.46) can be verified by comparing $w_{\alpha}^{f[7]}$ and $w_{\alpha}^{fs\text{slt}}$ for a variety of initial conditions \mathbf{w}^i as $\mathbf{w}^i \rightarrow 0$.

Assuming that the series in the second line of (2.44) is convergent and that to good approximation terms beyond some degree can be neglected, it might be illuminating/interesting to monitor the Lie generators f_{2n} during the course of a ray-trace-driven design/optimization process to observe, from a Lie perspective, what is being accomplished during the process. If it is observed, for example, that some particular f_{2n} or set of f_{2n} is being driven to zero, then one might experiment with including their values as part of a merit function.

In some settings, at least in the context of magnetic optics, it is useful to replace an optimization process by a *fitting* process in which several parameters are varied to drive several or all “offensive” Lie generators to zero. (To verify that some set of f_m is offensive, one can perform Lie algebraic ray traces with some of the f_m set to zero to see what effect that has on the \mathbf{w}^f so computed. Note that so doing does not violate the symplectic condition.) For example, it is possible to design a complete third-order achromat in which the strengths of three quadrupoles, three sextuples, and eight octupoles are varied to set/remove various f_m in a particular basis, and all remaining f_m are cancelled by repetitive symmetry. (In magnetic optics parlance, an achromat bends a charged-particle beam, but otherwise acts as the identity map.) In so doing, 203 conditions are met. (Magnetic optical systems generally do not have axial symmetry and, therefore, there are many more aberrations to be corrected.) It is unlikely that this goal could have been achieved with an optimization program.

It is also possible to carry out procedures in which fitting loops are inside an optimization loop. This was done in connection with some octupole-corrected Los Alamos charged-particle beam projects.

X.10.2 The Graded Index Case

Suppose one wishes to employ lenses made of graded index material. (Something analogous is always the case in magnetic optics since magnetic fields are position dependent.) In that case one ray-tracing possibility is to employ direct numerical integration of the equations of motion (1.34) and (1.35) associated with H , perhaps with the aid of symplectic integrators to ensure maintenance of the symplectic condition. But this process is slow if many rays are to be traced with high accuracy. Moreover, extraction of aberration data from ray data, if desired, is subject to the numerical errors associated with high-order numerical differentiation.

A second option is to employ Lie methods. Suppose all the graded-index lenses have *flat faces*. In that case there are equations of motion for the Lie generators f_m that can be integrated to yield a Lie representation for the end-to-end \mathcal{M} . (See Chapter 10 of *Lie Methods for Nonlinear ...*) And, once \mathcal{M} is found in Lie form, numerous ray traces can be carried out rapidly. (And fitting/optimization can be carried out using both the values of the Lie generators and ray-trace results.) Figures 9.1 and 9.2 illustrate ray traces for two charged-particle beam devices carried out in this fashion.

The treatment of graded-index lenses with curved faces is more complicated, but some

progress has also been made in handling this problem. And again, once \mathcal{M} is found in Lie form, numerous ray traces can be carried out rapidly. And fitting/optimization can again be carried out using both the values of the Lie generators and ray-trace results.

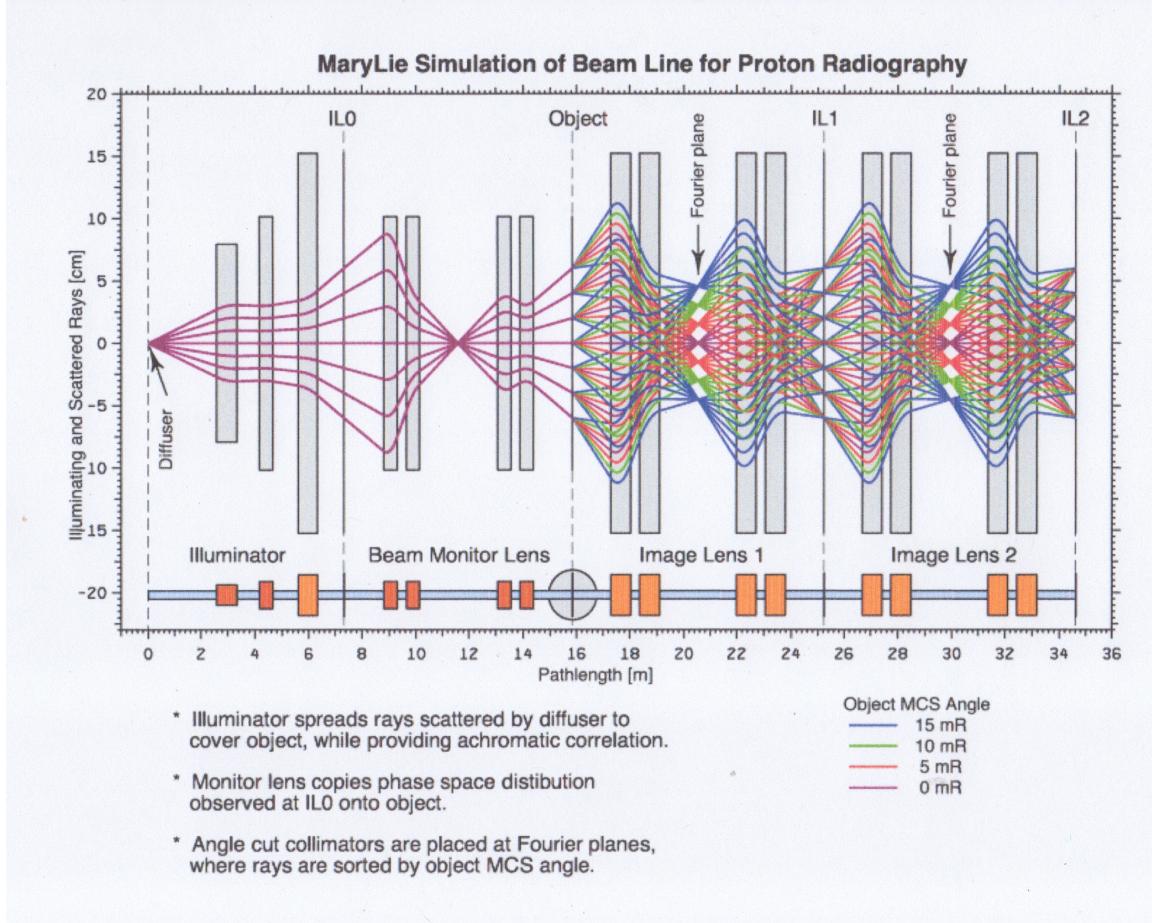


Figure X.10.1: Lie Algebraically Designed Magnetic Optical System for Fast Dynamic (Nanosecond) Imaging of Dense Objects Using High-Energy Proton Beams.

There are several iron-free cases for which the generalized gradients and their derivatives (up to some high order) can be found analytically, or numerically with certified accuracy, for specified current sources or specified rare-earth cobalt (REC) material distributions.[?, ?, 4] This information can then be used to compute transfer maps to high order including extended fringe-field effects. Such cases may be viewed as exactly soluble, and provide a useful guide for magnet design. For example, for rectangular coils on the surface of a cylinder, Bassetti and Biscare find the following analytic formulas: For solenoids $G_0^{[1]}(z) = g[F_0(z^+) - F_0(z^-)]$, dipoles $G_1^{[0]}(z) = g[F_1(z^+) - F_1(z^-)]$, and quadrupoles $G_2^{[0]}(z) = g[F_2(z^+) - F_2(z^-)]$; where $F_0(t) = f_1(t)$, $F_1(t) = 2f_1(t) - f_3(t)$, $F_2(t) = 9f_1(t) - 8f_3(t) + 3f_5(t)$, with $f_n(t) = t^n/(a^2 + t^2)^{n/2}$ and $z^\pm = z \pm L/2$. These results have to be normalized by factors depending on the currents in the windings. In principle exact results are also available when iron is present provided it is not saturated (B and H are assumed to be linearly related) and the

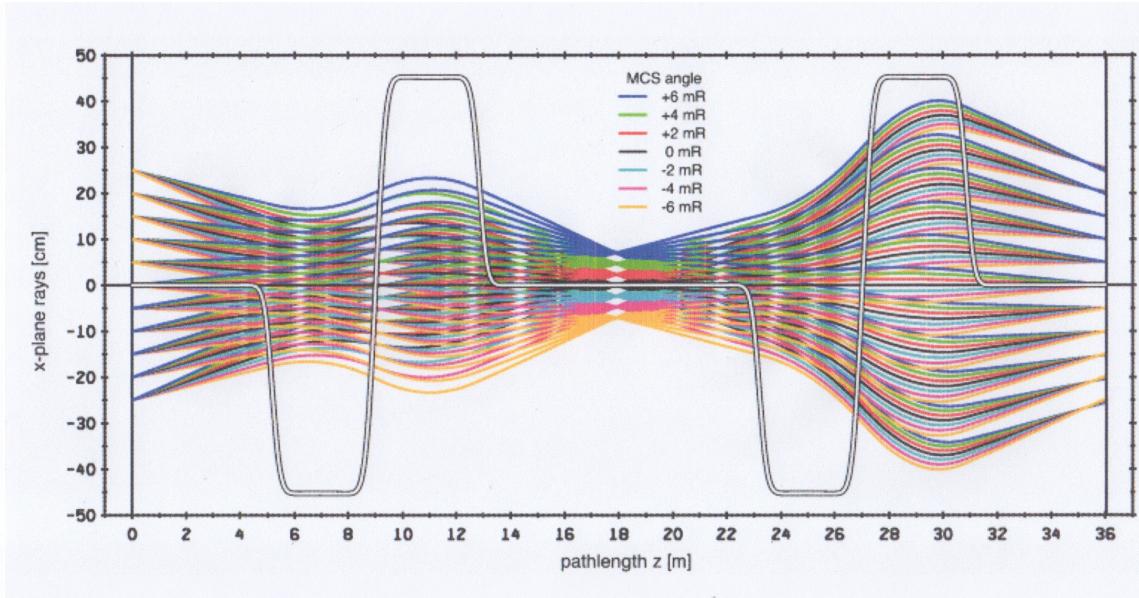


Figure X.10.2: Ray Trace of Soft-Edge Lie Algebraically Designed Super Lens for 50 GeV protons.

current/REC sources are completely surrounded by iron including fringe-field regions. In this case fringe fields fall off exponentially.

The general case when iron is present is more difficult to treat. Then usually the available information is in the form of field data at various discrete points obtained either by measurement or the use of 3-D finite element codes. If only on-axis or mid-plane field data is employed, one may try to fit it with some assumed analytical form, often taken to fall off exponentially. Examples for use with on-axis data include the Enge functions [?], of the form $1/(1 + e^{S_k(z)})$, with $S_k(z)$ being a polynomial of order k in z . From such fits one infers 3-D fields and from these fields seeks to compute transfer maps. So doing requires repeated differentiation of the fitting function. However it is well known to those working in the field of numerical analysis that differentiation amplifies the unavoidable errors present either because of imperfect fitting or noise in the field data. Use of these methods is therefore not expected to be reliable for calculations beyond first or perhaps second order.

Surface methods (circular cylinder) This problem of noise and its amplification by differentiation can be overcome to some extent by the use of surface data sufficiently far from the design orbit. Suppose, in the case of cylindrical geometry, that the normal field component $B_\rho(R, \phi, z)$ is known on the surface $\rho = R$ of a virtual circular cylinder (in practice, the values of the field on the surface are obtained by interpolation of the field data on nearby grid points). The cylinder, long enough to include the fringe regions, is to be contained within the aperture of the magnetic device and aligned with the z -axis. This surface information is sufficient to uniquely determine the full magnetic field within the surface. From $B_\rho(R, \phi, z)$ we compute the quantities $\hat{B}_{\rho, m, s}(k) = \int_{-\infty}^{\infty} dz e^{-ikz} \int_0^{2\pi} d\phi \sin(m\phi) B_\rho(R, \phi, z) / (2\pi)^2$ with a similar expression for $\hat{B}_{\rho, m, c}(k)$ with sin replaced by cos. Then the $G_{m, \alpha}^{[n]}(z)$ are given in

terms of surface data by the relations

$$G_{m,\alpha}^{[n]}(z) = \frac{i^n}{2^m} \int_{-\infty}^{\infty} dk \frac{k^{n+m-1}}{I'_m(kR)} \hat{B}_{\rho,m,\alpha}(k) e^{ikz}. \quad (\text{X.10.4})$$

Bibliography

- [1] D. Dilworth, *Lens Design*, IOP Publishing Ltd (2018). See the Web site <http://iopscience.iop.org/book/978-0-7503-1611-8>
- [2] A. Dragt, “Lie algebraic theory of geometrical optics and optical aberrations”, *J. Opt. Soc. Am.* **72** p 372 (1982).
- [3] A. Dragt and E. Forest, *Foundations of a Lie Algebraic Theory of Geometrical Optics*, University of Maryland Physics Department Technical Report (1985).
- [4] A. Dragt, E. Forest, and K. B. Wolf, “Foundations of a Lie algebraic theory of geometrical optics”, published in *Lie Methods in Optics*, p 105, J. S. Mondragón and K. B. Wolf, edit., Springer-Verlag (1986).
- [5] A. Torre, *Linear Ray and Wave Optics in Phase Space: Bridging Ray and Wave Optics via the Wigner Phase-Space Picture*, Elsevier Science (2005).
- [6] M. Born and E. Wolf, *Principles of Optics*, Sixth (Corrected) Edition, Pergamon Press (1984).

Appendix Y

Relation between the Classical Poisson Bracket Lie Algebra and the Quantum Commutator-Based Lie Algebra

Overview

This appendix explores the relation between the classical Poisson bracket Lie algebra and the quantum commutator-based Lie algebra. Lie methods are used to construct bases for each. It is found that the basis in the quantum case coincides with the Weyl basis. Next a natural correspondence is set up between the classical and quantum bases. Finally, these bases are used to determine the structure constants for the classical and quantum Lie algebras. It is found that many, *but not all*, of the structure constants for the two Lie algebras are the same. In particular, it is found that the classical Lie algebra is a contraction of the quantum Lie algebra. Conversely, the quantum Lie algebra is a deformation of the classical Lie algebra.

Y.1 Classical Polynomial Basis

For introductory simplicity, work with a two-dimensional phase space with variables q and p . (Higher-dimensional cases can be treated in a similar manner.) Let f and g be any functions of the phase-space variables. Introduce a *classical mechanical* Lie product $[*, *]_{\text{cm}}$ among such functions by use of the Poisson bracket,

$$[f, g]_{\text{cm}} = (\partial f / \partial q)(\partial g / \partial p) - (\partial f / \partial p)(\partial g / \partial q). \quad (\text{Y.1.1})$$

Given any phase-space function f , define an associated *Lie operator*, denoted by $: f :$, by the rule

$$: f : g = [f, g]_{\text{cm}}. \quad (\text{Y.1.2})$$

In view of (1.1) and (1.2), this is the usual definition of $: f :$, but presented with a slightly different notation.

Introduce basis monomials $a_{r-s,s}$ by the rule

$$a_{r-s,s}(q,p) = [(r-s)!/r!] : -p^2/2 :^s q^r. \quad (\text{Y.1.3})$$

So doing yields, for the first few monomials, the results

$$a_{00} = 1; \quad (\text{Y.1.4})$$

$$a_{10} = q, \quad (\text{Y.1.5})$$

$$a_{01} = p; \quad (\text{Y.1.6})$$

$$a_{20} = q^2, \quad (\text{Y.1.7})$$

$$a_{11} = qp, \quad (\text{Y.1.8})$$

$$a_{02} = p^2; \quad (\text{Y.1.9})$$

$$a_{30} = q^3, \quad (\text{Y.1.10})$$

$$a_{21} = q^2p, \quad (\text{Y.1.11})$$

$$a_{12} = qp^2, \quad (\text{Y.1.12})$$

$$a_{03} = p^3; \quad (\text{Y.1.13})$$

$$a_{40} = q^4, \quad (\text{Y.1.14})$$

$$a_{31} = q^3p, \quad (\text{Y.1.15})$$

$$a_{22} = q^2p^2, \quad (\text{Y.1.16})$$

$$a_{13} = qp^3, \quad (\text{Y.1.17})$$

$$a_{04} = p^4. \quad (\text{Y.1.18})$$

The degree of a monomial a_{rs} is given by the sum $(r+s)$, and we refer to the monomials of a fixed degree as forming a *ladder*. Within a ladder and up to multiplicative factors, $: -p^2/2 :$ acts on a_{rs} as an operator that lowers r and raises s . Indeed, there is the recursion relation

$$: -p^2/2 : a_{r,s} = r a_{r-1,s+1}. \quad (\text{Y.1.19})$$

Similarly, within a ladder and up to multiplicative factors, $: q^2/2 :$ acts as an operator that raises r and lowers s .

Y.2 Quantum Polynomial Basis

Let Q and P be the quantum-mechanical counterparts of q and p . They obey the commutation rule

$$\{Q, P\} = QP - PQ = i\hbar I \quad (\text{Y.2.1})$$

where \hbar is the reduced Planck's constant $h/(2\pi)$, and which we will eventually view as an adjustable parameter in Section 4. Suppose $F(Q, P)$ and $G(Q, P)$ are any polynomial functions of Q and P with some ordering rule for products of Q 's and P 's. Given these functions, define a *quantum mechanical* Lie product $[*, *]_{\text{qm}}$ by the rule

$$[F, G]_{\text{qm}} = (i\hbar)^{-1}\{F, G\}. \quad (\text{Y.2.2})$$

Here we note two facts: First, $[*, *]_{\text{qm}}$ is a Lie product because the commutator is a Lie product. Second, if F and G are Hermitian, $[F, G]_{\text{qm}}$ will also be Hermitian. Indeed, from the definition (2.2) there is the relation

$$[F, G]_{\text{qm}}^\dagger = -(i\hbar)^{-1}\{F, G\}^\dagger. \quad (\text{Y.2.3})$$

But, there is also the relation

$$\{F, G\}^\dagger = (FG - GF)^\dagger = (FG)^\dagger - (GF)^\dagger = G^\dagger F^\dagger - F^\dagger G^\dagger = GF - FG = -\{F, G\}. \quad (\text{Y.2.4})$$

Consequently, there is the advertised result

$$[F, G]_{\text{qm}}^\dagger = [F, G]_{\text{qm}}. \quad (\text{Y.2.5})$$

Within the quantum mechanical context, define an operator $: -P^2/2 :$ by the rule

$$: -P^2/2 : G = [-P^2/2, G]_{\text{qm}}. \quad (\text{Y.2.6})$$

Powers of $: -P^2/2 :$ are defined by the rules

$$: -P^2/2 :^0 G = G, \quad (\text{Y.2.7})$$

$$: -P^2/2 :^2 G = [-P^2/2, [-P^2/2, G]_{\text{qm}}]_{\text{qm}}, \text{ etc.} \quad (\text{Y.2.8})$$

Next, in analogy to the construction (1.3), use $: -P^2/2 :$ to define polynomials $A_{r-s,s}$ by the rule

$$A_{r-s,s}(Q, P) = [(r-s)!/r!] : -P^2/2 :^s Q^r. \quad (\text{Y.2.9})$$

Here use is to be made of the relation (2.1) to evaluate the commutators that occur, and we presume that

$$Q^0 = I. \quad (\text{Y.2.10})$$

Doing so gives, for the first few values of r and s , the results

$$A_{00} = I; \quad (\text{Y.2.11})$$

$$A_{10} = Q, \quad (\text{Y.2.12})$$

$$A_{01} = P; \quad (\text{Y.2.13})$$

$$A_{20} = Q^2, \quad (\text{Y.2.14})$$

$$A_{11} = (QP + PQ)/2, \quad (\text{Y.2.15})$$

$$A_{02} = P^2; \quad (\text{Y.2.16})$$

$$A_{30} = Q^3, \quad (\text{Y.2.17})$$

$$A_{21} = (Q^2P + PQ^2)/2, \quad (\text{Y.2.18})$$

$$A_{12} = (QP^2 + P^2Q)/2, \quad (\text{Y.2.19})$$

$$A_{03} = P^3; \quad (\text{Y.2.20})$$

$$A_{40} = Q^4, \quad (\text{Y.2.21})$$

$$A_{31} = (Q^3P + PQ^3)/2, \quad (\text{Y.2.22})$$

$$A_{22} = (QP + PQ)^2/6 + (Q^2P^2 + P^2Q^2)/6, \quad (\text{Y.2.23})$$

$$A_{13} = (QP^3 + P^3Q)/2, \quad (\text{Y.2.24})$$

$$A_{04} = P^4. \quad (\text{Y.2.25})$$

Note that the A_{rs} are Hermitian, as expected. Also, the A_{rs} are polynomials in the *Weyl* basis. That is, products of Q 's and P 's are Weyl ordered. Moreover, this ordering of constituents has not been achieved by demanding/imposing permutation symmetry, but instead arises *naturally* from a Lie algebraic procedure.

The degree of a polynomial A_{rs} is again given by the sum $(r+s)$, and we again refer to the polynomials of a fixed degree as forming a ladder. Within a ladder and up to multiplicative factors, $: -P^2/2 :$ acts on A_{rs} as an operator that lowers r and raises s . Indeed, there is the recursion relation

$$: -P^2/2 : A_{r,s} = rA_{r-1,s+1}. \quad (\text{Y.2.26})$$

Similarly, within a ladder and up to multiplicative factors, $: Q^2/2 :$ acts as an operator that raises r and lowers s .

Y.3 A Natural Correspondence between Classical and Quantum Bases

How can we set up a natural correspondence between the classical and quantum bases? In so doing, by linearity, we will also set up a natural correspondence between the classical Lie algebra of phase-space functions with Lie product $[*, *]_{\text{cm}}$ and the quantum Lie algebra of polynomials in Q and P with Lie product $[*, *]_{\text{qm}}$. We will call these Lie algebras \mathcal{L}_{cm} and \mathcal{L}_{qm} .

First, it is natural to set up the correspondences

$$1 \leftrightarrow I, \quad (\text{Y.3.1})$$

$$q^n \leftrightarrow Q^n, \quad (\text{Y.3.2})$$

$$p^n \leftrightarrow P^n. \quad (\text{Y.3.3})$$

But then, because of the Lie algebraic similarity of the definitions (1.3) and (2.9), it is also natural to set up the correspondences

$$a_{rs}(q, p) \leftrightarrow A_{rs}(Q, P), \quad (\text{Y.3.4})$$

for which (3.1) through (3.3) are special cases.

Y.4 Relation between the Lie Algebras \mathcal{L}_{cm} and \mathcal{L}_{qm}

The Lie algebra \mathcal{L}_{cm} has structure constants $c_{jk;rs}^{tu}$ defined by the relation

$$[a_{jk}, a_{rs}]_{\text{cm}} = \sum_{tu} c_{jk;rs}^{tu} a_{tu}. \quad (\text{Y.4.1})$$

Similarly, the Lie algebra \mathcal{L}_{qm} has structure constants $C_{jk;rs}^{tu}$ defined by the relation

$$[A_{jk}, A_{rs}]_{\text{qm}} = \sum_{tu} C_{jk;rs}^{tu} A_{tu}. \quad (\text{Y.4.2})$$

How do the structure constants $c_{jk;rs}^{tu}$ and $C_{jk;rs}^{tu}$ compare? They are not all the same, and therefore the two Lie algebras \mathcal{L}_{cm} and \mathcal{L}_{qm} are not manifestly the same.¹ However, *many* of the structure constants are the same. In particular, there are the equalities

$$C_{jk;rs}^{tu} = c_{jk;rs}^{tu} \text{ for } j + k \leq 2. \quad (\text{Y.4.3})$$

One consequence of (4.3) is that the subalgebras generated by classical and quantum polynomials of degree ≤ 2 are identical, and are in fact the Lie algebra $isp(2, \mathbb{R})$, the Lie algebra of the *inhomogeneous* symplectic group in two dimensions over the real field.²

Another consequence is the special role played by the a_{jk} and the A_{jk} with $j + k = 2$. Both generate the Lie algebra for $sp(2, \mathbb{R})$. We have already seen that p^2 and q^2 , and their quantum counterparts, act as raising and lowering operators within ladders. Recall (1.19)

¹In the very early editions of Dirac's classic text *The Principles of Quantum Mechanics* he proceeded as if these two Lie algebras were the same. This misconception was removed by more careful wording in later editions.

²In light wave optics the quantity $\lambda/(2\pi)$, where λ is the wavelength, plays the role of \hbar . Consequently, Fourier optics results can be read off from paraxial ray optics results. We also remark that although the subalgebras are identical, the underlying groups are not identical. In the classical case the group is the inhomogeneous symplectic group in two dimensions over the real field. In the quantum case the group is the inhomogeneous metaplectic group in two dimensions over the real field, which is a two-fold cover of the inhomogeneous symplectic group in two dimensions over the real field.

and (2.26). What can be said about the action of a_{11} and its quantum counterpart A_{11} ? Simple calculation gives the result

$$:a_{11}:a_{rs} = [qp, q^r p^s]_{\text{cm}} = (s - r)a_{rs}. \quad (\text{Y.4.4})$$

Thus the a_{rs} are eigenfunctions of the operator $:a_{11}:$ with eigenvalues $(s - r)$. It follows from (4.3) that there are the analogous quantum results

$$:A_{11}:A_{rs} = (s - r)A_{rs}. \quad (\text{Y.4.5})$$

The A_{rs} are eigenfunctions of the operator $:A_{11}:$ with eigenvalues $(s - r)$.

The first differences between \mathcal{L}_{cm} and \mathcal{L}_{qm} results occur at degree 4. There are the classical relations

$$[a_{03}, a_{30}]_{\text{cm}} = -9a_{22} \quad (\text{Y.4.6})$$

and

$$[a_{12}, a_{21}]_{\text{cm}} = -3a_{22}. \quad (\text{Y.4.7})$$

By contrast, there are the quantum relations

$$[A_{03}, A_{30}]_{\text{qm}} = -9A_{22} + (3/2)\hbar^2 A_{00}, \quad (\text{Y.4.8})$$

and

$$[A_{12}, A_{21}]_{\text{qm}} = -3A_{22} - (1/2)\hbar^2 A_{00}. \quad (\text{Y.4.9})$$

The Lie products of all other degree 3 polynomials yield degree 4 results that are identical in the classical and quantum cases.

It can be verified that there are the relations

$$\lim_{\hbar \rightarrow 0} C_{jk;rs}^{tu} = c_{jk;rs}^{tu}, \quad (\text{Y.4.10})$$

and consequently there is the limiting correspondence

$$[a_{jk}, a_{rs}]_{\text{cm}} \leftrightarrow \lim_{\hbar \rightarrow 0} [A_{jk}, A_{rs}]_{\text{qm}}. \quad (\text{Y.4.11})$$

Thus, \mathcal{L}_{cm} is a *contraction* of \mathcal{L}_{qm} and, conversely, \mathcal{L}_{qm} is a *deformation* of \mathcal{L}_{cm} .

It is notable that the differences between the classical and quantum cases actually involve \hbar^2 and not \hbar itself.³ Thus, the relation (4.10) may be replaced by the stronger result

$$C_{jk;rs}^{tu} = c_{jk;rs}^{tu} + O(\hbar^2). \quad (\text{Y.4.12})$$

Moreover note, as inspection of (4.8) and (4.9) illustrates, that the $O(\hbar^2)$ terms involve only lower-degree polynomials than those that occur classically.

We also observe that the nature of the terms that can occur in a Lie product can be inferred from the Clebsch-Gordan series for the symplectic group. We have already seen that in the case of a two-dimensional phase space the basis elements A_{jk} have well-defined

³It follows that wave-optics aberration results, up to corrections of order $[\lambda/(2\pi)]^2$, can be read off from ray-optics aberration results.

transformation properties under the action of the symplectic group Lie algebra $sp(2, \mathbb{R})$. Consequently, their Lie products must also have well-defined transformation properties under the action of $sp(2, \mathbb{R})$. For example, the A_{jk} with $j + k = d$ belong to an irreducible representation of $sp(2, \mathbb{R})$ that behaves like “spin” $d/2$.⁴ In the cases of (4.8) and (4.9), the ingredients of the Lie products on the left sides carry the representation $3/2$. According to Clebsch and Gordan, two spin $3/2$ objects can combine to produce objects of spins $3, 2, 1$, and 0 . Since the Lie product is antisymmetric under the interchange of its ingredients, the spins 3 and 1 are ruled out by symmetry considerations. What possibly remain are spins 2 and 0 . The two terms that occur on the right sides of (4.8) and (4.9), namely A_{22} and A_{00} , have spins 2 and 0 , respectively.

Y.5 Historical Comment

Introduce the notation

$$z = (z_1, z_2) = (q, p) \quad (\text{Y.5.1})$$

and

$$Z = (Z_1, Z_2) = (Q, P). \quad (\text{Y.5.2})$$

Then we have the results

$$[z_j, z_k]_{cm} = J_{jk} \mathbf{1} \quad (\text{Y.5.3})$$

and

$$[Z_j, Z_k]_{qm} = J_{jk} I. \quad (\text{Y.5.4})$$

Max Born (1882-1970) was one of the founders of Quantum Mechanics to recognize early on the fundamental importance of the relation (5.4), which may be viewed as introducing Lie-algebraic concepts into Quantum Mechanics. See Figure 5.1 below. He also made important contributions to several other fields including Light Optics, and might have been interested in the contents of Appendix X because of its Lie-algebraic approach to Light Optics.⁵

Exercises

Y.5.1. Verify the results (1.4) through (1.19) and the results (2.11) through (2.26).

Y.5.2. Verify the results (4.3) through (4.11).

Y.5.3. Determine the effect of $:q^2/2:$ on a_{rs} and the effect of $:Q^2/2:$ on A_{rs} .

⁴Here we use the fact that the Lie algebras $sp(2, \mathbb{R})$ and $su(2)$ are equivalent over the complex field, and therefore their Clebsch-Gordan series are essentially the same. In the case of phase spaces of dimension $2n$, one must know the Clebsch-Gordan series for $sp(2n, \mathbb{R})$.

⁵Born was also the thesis advisor of numerous prominent physicists, including J. Robert Oppenheimer, and the grandfather of Olivia Newton-John.



Figure Y.5.1: The gravestone of Max and Hedwig Born.

Bibliography

- [1] B. Hall, *Quantum Theory for Mathematicians*, Springer (2013).
- [2] T. Curtright, D. Fairlie, and C. Zachos, *A Concise Treatise on Quantum Mechanics in Phase Space*, World Scientific (2014).

