The SIAM 100-Digit

CHALLENGE

A Study in High-Accuracy Numerical Computing

Folkmar Bornemann
Technische Universität München
Munich, Germany

Dirk Laurie
University of Stellenbosch
Stellenbosch, South Africa

Stan Wagon
Macalester College
St. Paul, Minnesota

Jörg Waldvogel
Swiss Federal Institute of Technology (ETH) Zurich
Zurich, Switzerland

With a Foreword by David H. Bailey

siam
Society for Industrial and Applied Mathematics
Philadelphia

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

The SIAM 100-digit challenge : a study in high-accuracy numerical computing / Folkmar Bornemann ... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-89871-561-X (pbk.)
1. Numerical analysis. I. Bornemann, Folkmar, 1967-
QA297.S4782 2004
518—dc22 2004049139

siam is a registered trademark.

Chapter 4

Think Globally, Act Locally

Stan Wagon

In order to find all common points, which are the solutions of our nonlinear equations, we will (in general) have to do neither more nor less than map out the full zero contours of both functions. Note further that the zero contours will (in general) consist of an unknown number of disjoint closed curves. How can we ever hope to know when we have found all such disjoint pieces?

— W. H. Press et al. (*Numerical Recipes* [PTVF92])

Problem 4

What is the global minimum of the function

$$e^{\sin(50x)} + \sin(60e^y) + \sin(70\sin x) + \sin(\sin(80y)) - \sin(10(x+y)) + (x^2+y^2)/4\ ?$$

4.1 A First Look

Let $f(x, y)$ denote the given function. On a global scale, f is dominated by the quadratic term $(x^2 + y^2)/4$, since the values of the other five summands lie in the intervals $[1/e, e]$, $[-1, 1]$, $[-1, 1]$, $[-\sin 1, \sin 1]$, and $[-1, 1]$, respectively. Thus the overall graph of f resembles a familiar paraboloid (Figure 4.1). This indicates that the minimum is near $(0, 0)$. But a closer look shows the complexity introduced by the trigonometric and exponential functions. Indeed, as we shall see later, there are 2720 critical points inside the square $[-1, 1] \times [-1, 1]$. From this first look, we see that a solution to the problem breaks into three steps:

1. Find a bounded region that contains the minimum.
2. Identify the rough location of the lowest point in that region.
3. Zoom in closer to pinpoint the minimum to high precision.

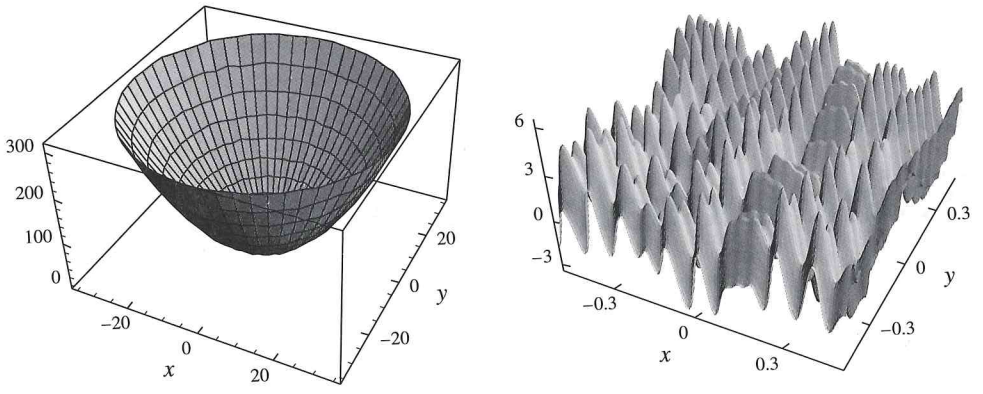

Figure 4.1. *Two views of the graph of $f(x, y)$. The scale of the left-hand view masks the complexity introduced by trigonometric and exponential terms. It is this complexity that makes finding the global minimum a challenge.*

Step 1 is easy. A quick computation using a grid of modest size yields the information that the function can be as small as -3.24. For example, one can look at the 2601 values obtained by letting x and y range from -0.25 to 0.25 in steps of 0.01.

A *Mathematica* Session

For later convenience, we define `f` so that it accepts as inputs either the two numeric arguments or a single list of length 2.

```
f[x_, y_] := e^Sin[50 x] + Sin[60 e^y] + Sin[70 Sin[x]] + Sin[Sin[80 y]] -
    Sin[10 (x + y)] + (x^2 + y^2)/4;
f[{x_, y_}] := f[x, y];

Min[Table[f[x, y], {x, -0.25, 0.25, 0.01}, {y, -0.25, 0.25, 0.01}]]
-3.246455170851875
```

This upper bound on the minimum tells us that the global minimum must lie inside the circle of radius 1 centered at the origin, since outside that circle the quadratic and exponential terms are at least $1/e + 1/4$ and the four sines are at least $-3 - \sin 1$, for a total above -3.23. And step 3 is easy once one is near the minimum: standard optimization algorithms, or root-finding algorithms on the gradient of f, can be used to locate the minimum very precisely. Once we are close, there is no problem getting several hundred digits. The main problem is in step 2: How can we pinpoint a region that contains the correct critical point, and is small enough that it contains no other? Figure 4.2, a contour plot of $f(x, y)$ with the contour lines suppressed, shows what we are up against.

In fact, a slightly finer grid search will succeed in locating the proper minimum; several teams used such a search together with estimates based on the partial derivatives of f to show that the search was fine enough to guarantee capture of the answer. But we will

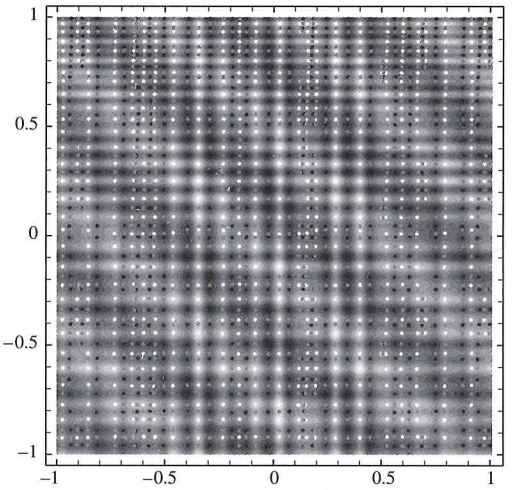

Figure 4.2. *A density plot of $f(x, y)$ with increasing values going from black to white. The local minima are shown as black dots (693), the local maxima as white dots (667), and the saddles (1360) as gray dots. There are 2720 critical points in this square, computed by the method of §4.4.*

not discuss such methods here, focusing instead on general algorithms that do not require an analysis of the objective function.

4.2 Speedy Evolution

While a grid search can be effective for this particular problem (because the critical points do have a rough lattice structure), it is easy to design a more general search procedure that uses randomness to achieve good coverage of the domain. An effective way to organize the search is to use ideas related to evolutionary algorithms. For each point in the current generation, n random points are introduced, and the n best results of each generation (and its parents) are used to form the new generation. The scale that governs the generation of new points shrinks as the algorithm proceeds.

Algorithm 4.1. Evolutionary Search to Minimize a Function.
Inputs: $f(x, y)$, the objective function;
R, the search rectangle;
n, the number of children for each parent, and the number of points in the new generation;
ϵ, a bound on the absolute error in the location of the minimum of f in R;
s, a scaling factor for shrinking the search domain.

Output: An upper bound to the minimum of f in R, and an approximation to its location.

Notation: `parents` = current generation of sample points, `fvals` = f-values for current generation.

Step 1: Initialize: Let z be the center of R; `parents` = $\{z\}$; `fvals` = $\{f(z)\}$;
$\{h_1, h_2\}$ = side-lengths of R.

Step 2: The main loop:
While $\min(h_1, h_2) > \epsilon$,
For each $p \in$ `parents`, let its children consist of n random points in a rectangle around p; get these points by using uniform random x and y chosen from $[-h_1, h_1]$ and $[-h_2, h_2]$, respectively;
Let `newfvals` be the f-values on the set of all children;
Form `fvals` $\cup$ `newfvals`, and use the n lowest values to determine the points from the children and the previous parents that will survive;
Let `parents` be this set of n points; let `fvals` be the set of corresponding f-values;
Let $h_i = s \cdot h_i$ for $i = 1, 2$.

Step 3: Return the smallest value in `fvals`, and the corresponding parent.

This algorithm is nicely simple and can be programmed in just a few lines in a traditional numerical computing environment. It generalizes with no problem to functions from $\mathbb{R}^n$ to $\mathbb{R}$. The tolerance is set to 10^{-6} in the *Mathematica* code that follows because the typical quadratic behavior at the minimum means that it should give about 12 digits of precision of the function value. Note that it is possible for some children to live outside the given rectangle, and therefore the final answer might be outside the rectangle. If that is an issue, then one can add a line to restrict the children to be inside the initial rectangle.

A *Mathematica* Session

Some experimentation will be needed to find an appropriate value of n. We use $n = 50$ here, though this algorithm will generally get the correct digits even with n as low as 30.

```
h = 1; gen = {f[#], #}&/@{{0, 0}};

While[h > 10^-6,
  new = Flatten[Table[#[[2]] + Table[h(2 Random[] - 1), {2}], {50}]&/@gen,
      1]; gen = Take[Sort[Join[gen, {f[#], #}&/@new]], 50];
  h = h/2];

gen[[1]]
{-3.30686864747396, {-0.02440308163632728, 0.2106124431628402}}
```

Figures 4.3 and 4.4 use a logarithmic scale to show the steady convergence of the points in each generation to the answer. Figure 4.4 shows how the points swarm into the correct one from all directions.

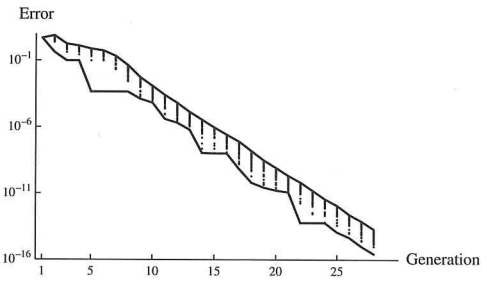

Figure 4.3. *The base-10 logarithm of the f-value error for each point in each generation of a genetic search, in a run with $n = 50$ and tolerance of 10^{-8} in the x–y coordinates. The convergence to the correct result is nicely steady.*

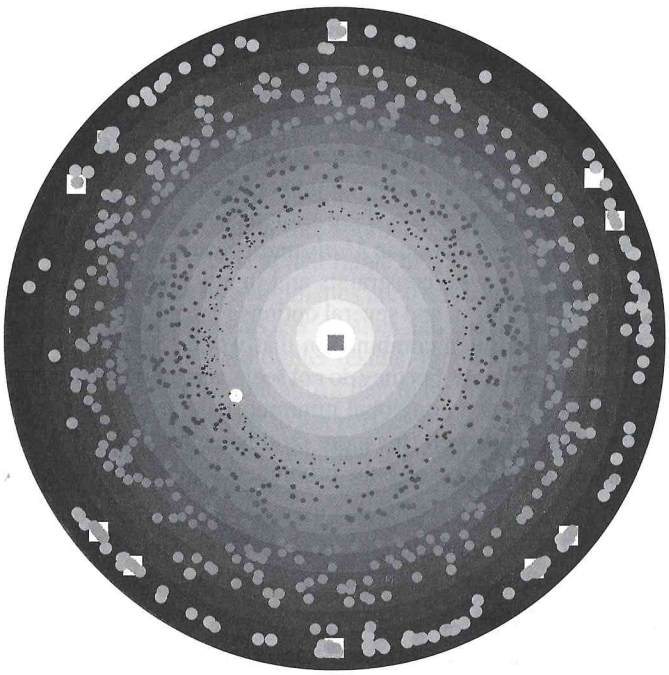

Figure 4.4. *Survival of the fittest: the results of a genetic algorithm with a high working precision and accuracy goal of 10^{-13}; the computation runs through 45 generations with 50 members of each. The scale is logarithmic, with each gray ring representing another power of 10 away from the true minimum location (center). Points of the same shade correspond to the same generation, so the innermost black points are the last generation (28th). They are within 10^{-6} of the true best; the point with the lowest f-value (which is not the point closest to the answer!) is the one atop a white disc. The outermost points—the first generation—have f-values between -1.3 and 4.2. But by the 9th generation all points are in the vicinity of the correct critical point. The white squares mark the location of the 20 lowest local minima; the clustering near those points at the early generations is visible.*

For more confidence one would compute a sequence of values as n increases. When $n = 70$, the algorithm solves the problem with high probability (992 successes in 1000 runs). This approach has the virtue of being easy to program and very fast, and so long as the resolution is fine enough to find the correct local minimum, it will converge to the answer even if one requests many digits (and uses appropriate working precision). Of course, it comes with no guarantee, and repeated trials will usually be necessary to gain confidence in the correctness of the result (the next section will present an algorithm that eliminates the uncertainty inherent in a random search). Searches such as this have limitations—a move to higher dimensions puts stress on the number of search points; a microscopic downward spike would be impossible to find—but such issues are problematic for just about any optimization method. Given the speed and minimal programming effort, this genetic approach is a fine tool for an initial investigation into an optimization problem.

An obvious way to speed things up is to switch at some point to a method that is designed to quickly converge to the nearest local minimum. It can be hard to decide, in an automated fashion, when one is close enough to the right minimum. However, for the problem at hand, switching to either a minimization routine (Brent's method is popular) or, better, a root-finder that finds the zero of the gradient speeds things up a lot if one wants a large number of digits. Once we have a seed that we believe in, whether from a random search or a more reliable interval algorithm as in §4.3, using Newton's method on the gradient is very fast and yields 10,000 digits in under a minute (see Appendix B).

If one wants to use a random search, there are some canned algorithms that can be used. For example, the `NMinimize` function in *Mathematica* can solve a variety of optimization problems, including problems with constraints and in any number of variables. But there are several methods available for it, and several options for each, so, as with all complicated software, it can be tricky finding a combination that works. Simulated annealing and the Nelder–Mead method are available, but they do not seem very good on this sort of continuous problem (they tend to end up at a nonglobal minimum). An evolutionary algorithm called *differential evolution* does work well (for more on this method, see §5.2) and will reliably (85 successes in 100 random trials) solve the problem at hand when called in the following form.

A *Mathematica* Session

```
NMinimize[{f[x, y], x^2 + y^2 ≤ 1}, {x, y},
  Method → {"DifferentialEvolution", "SearchPoints" → 250}]

{-3.3.3068686474752402,
  {x → -0.024403079743737212, y → 0.21061242727591697}}
```

This gives 14 digits of the answer and causes the objective function to be evaluated only about 6000 times (the simpler evolutionary algorithm presented earlier can get 15 digits with about 120,000 function evaluations).

Another approach is to use a more comprehensive global optimization package. One such, based on random search and statistically based reasoning, is *MathOptimizer*, a commercial package for *Mathematica* developed by Janos Pintér.[26]

[26] `http://www.dal.ca/~jdpinter/m_f_m.html`

4.3 Interval Arithmetic

Let R be the square $[-1, 1] \times [-1, 1]$, which we know must contain the answer. Random search methods can end up at a local but nonglobal minimum; the best algorithm will be one that is guaranteed to find the global minimum. Such an algorithm arises from a subdivision process. Repeatedly subdivide R into smaller rectangles, retaining only those rectangles that have a chance of containing the global minimum. The identification of these subrectangles can be done by estimating the size of the function and its derivatives on the rectangle. This point of view is really one of interval arithmetic. In interval arithmetic, the basic objects are closed intervals $[a, b]$ on the real line; extended arithmetic is generally used, so that endpoints can be $\pm\infty$. One can design algorithms so that elementary functions can be applied to such intervals (or, in our case, pairs of intervals), and the result is an interval that contains the image of the function on the domain in question. However, the resulting interval is not simply the interval from the minimum to the maximum of the function. Rather, it is an enclosing interval, defined so as to make its computation easy and fast; it will generally be larger than the smallest interval containing all the f-values.

For example, it is easy to determine an enclosing interval for $\sin([a, b])$: just check whether $[a, b]$ contains a real number of the form $\frac{\pi}{2} + 2n\pi$ ($n \in \mathbb{Z}$). If so, the upper end of the enclosing interval is 1; if not, it is simply $\max(\sin a, \sin b)$. Finding the lower end is similar. The exponential function is monotonic, so one need look only at the endpoints. For sums, one follows a worst-case methodology and adds the left ends and then the right ends. So if $g(x) = \sin x + \cos x$, then this approach will not yield a very tight result, as $[-2, 2]$ is only a loose bound on the actual range of g, $[-\sqrt{2}, \sqrt{2}]$. This phenomenon is known as *dependence*, since the two summands are treated as being independent when they are not. Nevertheless, as the intervals get smaller and the various pieces become monotonic, the interval computations yield tighter results. Products are similar to sums, but with several cases. There are several technicalities that make the implementation of an interval arithmetic system tedious and difficult to get perfectly correct: one critical point is that one must always round outwards. But *Mathematica* has interval arithmetic built in, and there are packages available for other languages, such as INTPAKX[27] for Maple, INTLAB[28] for MATLAB, and the public-domain SMATH library[29] for C. Thus one can use a variety of environments to design algorithms that, assuming the interval arithmetic and the ordinary arithmetic are implemented properly, will give digits that are verified to be correct. For the algorithms we present here, we will assume that a comprehensive interval package is available.

One simple application of these ideas is to the estimation exercise used in §4.1 to ascertain that the square R contains the global minimum. The interval output of f applied to the half-plane $-\infty < x \leqslant -1$ is $[-3.223, \infty]$, and the same is true if the input region—the half-plane—is rotated 90°, 180°, or 270° around the origin. This means that in these four regions—the complement of R—the function is greater than -3.23, and so the regions can be ignored. Here is how to do this in *Mathematica*, which has interval arithmetic implemented for elementary functions.

[27] http://www.mapleapps.com/powertools/interval/Interval.shtml
[28] http://www.ti3.tu-harburg.de/~rump/intlab/
[29] http://interval.sourceforge.net/interval/C/smathlib/README.html

A *Mathematica* Session

```
f[Interval[{-∞, -1.}], Interval[{-∞, ∞}]]
Interval[{-3.223591543636455, ∞}]
```

Now we can design an algorithm to solve Problem 4 as follows (this approach is based on the solution of the Wolfram Research team, the only team to use interval methods on this problem, and is one of the basic algorithms of interval arithmetic; see [Han92, Chap. 9] and [Kea96, §5.1]). We start with R and the knowledge that the minimum is less than -3.24. We then repeatedly subdivide R, retaining only those subrectangles T that have a chance of containing the global minimum. That is determined by checking the following three conditions. Throughout the interval discussion we use the notation $h[T]$ to refer to an enclosing interval for $\{h(t) : t \in T\}$.

(a) $f[T]$ is an interval whose left end is less than or equal to the current upper bound on the absolute minimum.

(b) $f_x[T]$ is an interval whose left end is negative and right end is positive.

(c) $f_y[T]$ is an interval whose left end is negative and right end is positive.

For (a), we have to keep track of the current upper bound. It is natural to try condition (a) by itself; such a simple approach will get one quickly into the region of the lowest minimum, but the number of intervals then blows up because the flat nature of the function near the minimum makes it difficult to get sufficiently tight enclosing intervals for the f-values to discard them. The other two conditions arise from the fact that the minimum occurs at a critical point and leads to an algorithm that is more aggressive in discarding intervals. Conditions (b) and (c) are easy to implement (the partial derivatives are similar to f in complexity) and the subdivision process then converges quickly to the answer for the problem at hand. While a finer subdivision might sometimes be appropriate, simply dividing each rectangle into four congruent subrectangles is adequate.

In the implementation we must be careful, as it is important to always improve the current upper bound as soon as that is possible. In the algorithm that follows, this improvement is done for the entire round of same-sized rectangles when a_1 is updated. The algorithm is given here for the plane, but nothing new is needed to apply it to n-space; later in this section we will present a more sophisticated algorithm for use in all dimensions.

Algorithm 4.2. Using Intervals to Minimize a Function in a Rectangle.
Assumptions: An interval arithmetic implementation that works for f, f_x, and f_y is available.

Inputs: R, the starting rectangle;
$f(x, y)$, a continuously differentiable function on R;
ϵ, a bound on the absolute error for the final approximation to the lowest f-value on R;
b, an upper bound on the lowest f-value on R, obtained perhaps by a preliminary search;
$i_{\max}$, a bound on the number of subdivisions.

Output: Interval approximations to the location of the minimum and the f-value there, the latter being an interval of size less than ϵ (or a warning that the maximum number of subdivisions has been reached). If the global minimum occurs more than once, then more than one solution will be returned.

Notation: $\mathcal{R}$ is a set of rectangles that might contain the global minimum; a_0 and a_1 are lower and upper bounds, respectively, on the f-value sought, and an *interior* rectangle is one that lies in the interior of R.

Step 1: Initialize: Let $\mathcal{R} = \{R\}$, $i = 0$; $a_0 = -\infty$, $a_1 = b$.
Step 2: The main loop:
While $a_1 - a_0 > \epsilon$ and $i < i_{\max}$:
Let $i = i + 1$;
Let $\mathcal{R}$ = the set of all rectangles that arise by uniformly dividing each rectangle in $\mathcal{R}$ into 4 rectangles;
Let $a_1 = \min(a_1, \min_{T \in \mathcal{R}}(f[T]))$;
Check size: Delete from $\mathcal{R}$ any rectangle T for which the left end of $f[T]$ is not less than a_1;
Check the gradient: Delete from $\mathcal{R}$ any interior rectangle T for which $f_x[T]$ does not contain 0 or $f_y[T]$ does not contain 0;
Let $a_0 = \min_{T \in \mathcal{R}}(f[T])$.
Step 3: Return the centers of the rectangles in $\mathcal{R}$ and of the f-interval for the rectangles in $\mathcal{R}$.

Appendix C.5.3 contains a bare-bones implementation in *Mathematica*, in the simplest form necessary to determine 10 digits of the minimum (both the checking of the number of subdivisions and the distinction involving interior squares are suppressed). Appendix C.4.3 also has code that does the same thing using the INTLAB package for MATLAB. Moreover, there are self-contained packages available that are devoted to the use of interval algorithms in global optimization; two prominent ones are COCONUT[30] and GlobSol,[31] both of which solve this problem with no difficulty. The *Mathematica* version of the algorithm takes under two seconds to solve Problem 4 completely as follows.

A *Mathematica* Session

```
LowestCriticalPoint[f[x, y], {x, -1, 1}, {y, -1, 1}, -3.24, 10^-9]
{{-3.306868647912808, -3.3068686470376663},
  {{-0.024403079696639973, 0.21061242715950385}}}
```

A more comprehensive routine, with many bells and whistles, is available at the web page for this book. That program includes an option for monitoring the progress of the algorithm (including the generation of graphics showing the candidate rectangles) and for getting high accuracy. A run with the tolerance ϵ set to 10^{-12} takes only a few seconds (all timings in this chapter are on a Macintosh G4 laptop with a 1 GHz CPU) and uses

[30] http://www.mat.univie.ac.at/~neum/glopt/coconut/branch.html
[31] http://www.mscs.mu.edu/~globsol/

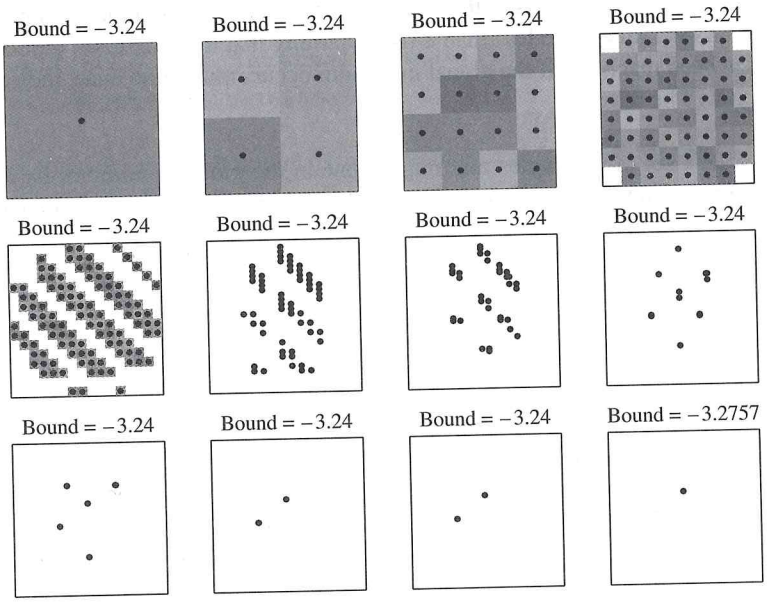

Figure 4.5. *The first 12 iterations of the interval subdivision algorithm. After 47 rounds, 12 digits of the answer are known.*

47 subdivision rounds (iterations of the main While loop). The total number of rectangles examined is 1372, the total number of interval function evaluations is 2210, and the numbers of candidate rectangles after each subdivision step are:

$$4, 16, 64, 240, 440, 232, 136, 48, 24, 12, 12, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,$$
$$4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4$$

Thus we see that at the third and later subdivisions, some subrectangles were discarded, and after 11 rounds only a single rectangle remained: it was then subdivided 35 times to obtain more accuracy. The interval returned by this computation is $-3.30687\frac{49}{56}$, which determines 12 digits of the answer (its midpoint is correct to 16 digits). Figure 4.5 shows the candidate rectangles at the end of each of the first 12 subdivision steps.

As with the random search of Algorithm 4.1, an obvious way to improve this basic algorithm is to switch to a rootfinder on the gradient once one has the minimum narrowed down to a small region. But there are many other ways to improve and extend this basic interval subdivision process, and we shall discuss those later in §§4.5 and 4.6.

4.4 Calculus

A natural idea for this problem is to use the most elementary ideas of calculus and try to find all the critical points—the roots of $f_x = f_y = 0$—in the rectangle R; then the smallest f-value among them is the global minimum. While this is not the most efficient way to

attack an optimization problem—most of the critical points are irrelevant, and the interval method of Algorithm 4.2 is more efficient at focusing on a region of interest—we present it here because it works, and it is a good general method for finding the zeros of a pair of functions in a rectangle. Moreover, if the objective function were one for which an interval implementation was not easily available, then the method of this section would be a useful alternative.

The partial derivatives are simple enough, but finding all the zeros of two nonlinear functions in a rectangle is a nontrivial matter (see the quote starting this chapter). Yet it can be done for the problem at hand by a very simple method (§4.6 will address the issue of verifying correctness of the list of critical points). The main idea, if one is seeking the roots of $f = g = 0$, is to generate a plot of the zero-set of f and use the data in the plot to help find the zero [SW97]. If a good contour-generating program is available, one can use it to approximate the zero-set of f and then one can design an algorithm to find all zeros in a rectangle as follows.

Algorithm 4.3. Using Contours to Approximate the Roots of $f = g = 0$.
Assumptions: The availability of a contour-drawing program that allows access to the points of the curves.

Inputs: f and g, continuous functions from $\mathbb{R}^2$ to $\mathbb{R}$;
R, the rectangular domain of interest;
r, the grid-resolution of the contour plot that will be used in Step 1.

Output: Numerical approximations to each pair (x, y) that is a root of $f = g = 0$ and lies in R.

Notation: s will contain points that serve as seeds to a root-finder.

Step 1: Generate the curves corresponding to $f = 0$. This is most easily done by a contour plot, set to compute just the zero-level curves, and using r for the resolution in each dimension.

Step 2: For each contour curve, evaluate g on the points that approximate the curve; whenever there is a sign change, put the point just scanned into s.

Step 3: For each point in s, use Newton's method (or a secant method in the case of nondifferentiable functions) to determine a root. Discard the root if it lies outside R.

Step 4: If the number of roots differs from the number of seeds, restart the algorithm with a finer resolution.

Step 5: Return the roots.

The use of a contour routine was given in Step 1 because that is quite simple and is adequate to the task at hand. However, there are more geometric (and faster) methods to compute an approximation to a contour curve that could be used. Such path-following algorithms have seen use in the study of bifurcation of differential equations. See [DKK91] and references therein to the AUTO program.

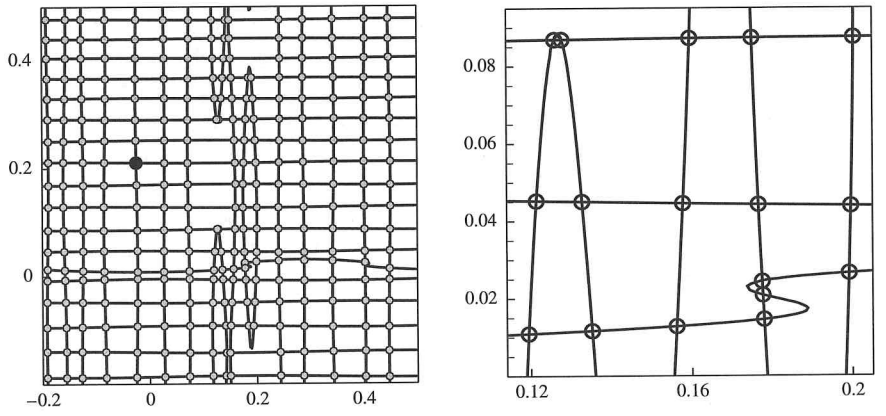

Figure 4.6. *A view of 290 solutions to $f_x = 0$ (vertical curves) and $f_y = 0$ (horizontal), where $f(x, y)$ is the objective function of Problem 4. The black point is the location of the global minimum. The close-up shows that the resolution was adequate for these curves.*

Implementing Algorithm 4.3 is quite simple provided one can get access to the data approximating the curves $f(x, y) = 0$. One might leave Step 4 for a manual check, as opposed to making it part of the program. Figure 4.6, which includes the zero-sets for both functions, shows the result for f_x and f_y on a small section of the plane, where a resolution of 450 was used (this refers to the grid size used by the contour algorithm); five seconds were required to compute all 290 roots. The close-up on the right shows that the resolution was good enough to catch all the intersections.

To solve the problem over all of the square $R = [-1, 1]^2$ it makes sense to work on subsquares to avoid the memory consumption of a very finely resolved contour plot. Therefore, we break R into 64 subsquares, compute all the critical points in each, and keep track of the one that gives the smallest value of f. Doing this takes under a minute and yields 2720 critical points. The one of interest is near $(-0.02, 0.21)$, where the f-value is near -3.3, the answer to the question. Newton's method can then be used to quickly obtain the function value at this critical point to more digits.

Manual examination of the contour plots on the subsquares to confirm that the resolution setting is adequate is not an efficient method for checking the results. A better way to gain confidence in the answer is to run the search multiple times, with increasing resolution, checking for stability; results of such work are in Table 4.1. While even the lowest resolution was enough to find the global minimum, quite high resolution is needed to catch all 2720 critical points in S. Note how, in some cases, 2720 seeds were found but they did not lead to the intended roots, and only 2719 roots were found. The high resolution needed to get 2720 roots (160×160 on each of 64 small squares) justifies the use of the subsquare approach. It is also a simple matter to use the second-derivative test to classify the critical points: 667 are maxima, 693 are minima, and 1360 are saddle points.

Note that the number of extrema ($693 + 667$) coincides with the number of saddles (1360). In some sense this is a coincidence, as it fails for, say, a rectangle that contains but

Table 4.1. *A resolution of* 160×160 *on each of 64 subsquares is needed to find all 2720 critical points in* $[-1, 1]^2$.

Resolution	Number of seeds	Number of roots
20	2642	2286
40	2700	2699
60	2716	2712
80	2714	2711
100	2718	2716
120	2720	2719
140	2720	2719
160	2720	2720
180	2720	2720
200	2720	2719
220	2720	2720
240	2720	2720
260	2720	2719
280	2720	2720
300	2720	2720
320	2720	2720
340	2720	2720
360	2720	2720
380	2720	2720
400	2720	2720

a single critical point; it fails also for the rectangle $[-0.999, 0.999]^2$, which has $692 + 667$ extrema and 1357 saddles. Yet this phenomenon is related to the interesting subject known as Morse theory, which provides formulas for

$$\text{number of maxima} + \text{number of minima} - \text{number of saddles}$$

on a closed surface, such as a sphere or torus. These formulas are related to the Euler characteristic; on a torus, this number equals 0 [Mil63]. So let's make our function toroidal by taking f on $R = [-1, 1]^2$ and reflecting the function in each of the four sides, and continuing this reflection process throughout the plane. This gives us a function—call it g—that is doubly periodic on the whole plane, with a fundamental region that is $[-1, 3]^2$; thus g can be viewed as a function on a torus. Each critical point of f in R generates four copies of itself for g. But we have the complication that we have introduced new critical points on the boundary.

To understand the boundary, consider marching along the four edges of R and looking at each one-dimensional extremum: there will be the same number of maxima as minima (we assume none occur at a corner, and no inflection points occur). Moreover, because of

the symmetry of the reflection, each such maximum will be either a saddle or a maximum for g, and each minimum will be either a minimum or a saddle for g. If we can enumerate all these occurrences, then we can use the toroidal formula for g to deduce the situation for f in R. But a couple of surprising things happen because of the nature of $f(x, y)$. First, f_x is positive on the left and right edges of R, and f_y is positive on the top and negative on the bottom. This means that all the maxima on the left edge become saddles and all the minima stay minima, with similar results holding for the other edges. We also need to know the numbers of extrema on the border. These are easily computed and turn out to be: top: 24 maxima, 24 minima; bottom: same; left: 29 maxima, 30 minima; right: same. These facts about f combine to show that the toroidal formula applies verbatim to f restricted to R, thus explaining the coincidence. More important, this analysis relies only on a one-dimensional computation and so provides supporting evidence that the counts obtained by the contour method (693, 667, and 1360) are correct.

Figure 4.7 shows another example that arose from a system of differential equations [Col95]. Finding the equilibrium points of the autonomous system

$$x' = 2y\cos\left(y^2\right)\cos(2x) - \cos y, \quad y' = 2\sin(y^2)\sin(2x) - \sin x,$$

is the same as finding all zeros of the right-hand sides. A contour resolution of 60 is sufficient, and the computation of all 73 zeros takes under a second.

The total amount of programming needed to implement Algorithm 4.3 is modest, provided one has a good way of accessing the data in the contour plot. In *Mathematica*, `Cases[ContourPlot[***],_Line,∞]` does the job.

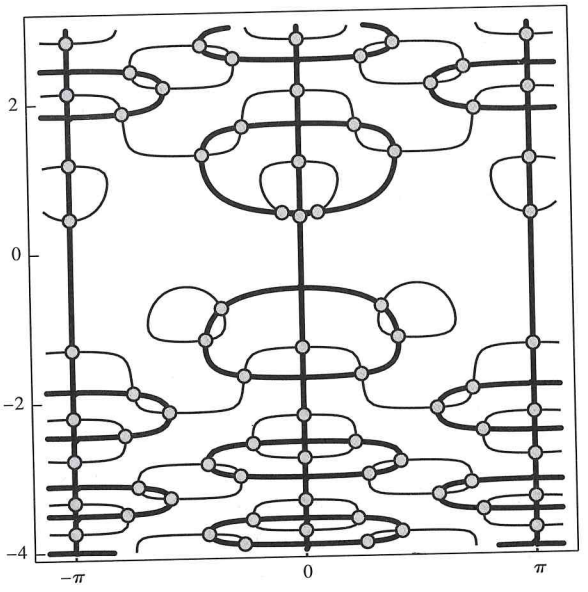

Figure 4.7. *This complicated example of simultaneous equations has 73 solutions in the rectangle shown.*

A *Mathematica* Session

The following is complete code for generating the zeros for the example in Figure 4.7. For a complete routine there are, as always, other issues to deal with: one should eliminate duplicates in the final result, since it is possible that different seeds will converge to the same zero.

```
g1[x_, y_] := 2 y Cos[y^2] Cos[2 x] - Cos[y];

g2[x_, y_] := 2 Sin[y^2] Sin[2x] - Sin[x];

{a, b, c, d} = {-3.45, 3.45, -4, 3};

contourdata = Map[First, Cases[Graphics[
        ContourPlot[g1[x, y], {x, a, b}, {y, c, d}, Contours → {0}, PlotPoints → 60]],
      _Line, ∞]];
seeds =
  Flatten[
    Map[
      #[[1 + Flatten[Position[Rest[ss = Sign[Apply[g2, #, 2]]] * Rest[RotateRight[ss]],
              -1]]]]&, contourdata], 1];
roots =
  Select[Map[{x, y}/.FindRoot[{g1[x, y] == 0, g2[x, y] == 0}, {x, #[[1]]}, {y, #[[2]]}]&,
      seeds], a < #[[1]] < b ∧ c < #[[2]] < d&];

Length[roots]
73
```

This technique also applies to finding all zeros of a complex function $f(z)$ in a rectangle by just applying the method to the system $\mathrm{Re} f = \mathrm{Im} f = 0$. The contour technique can fail badly if the zero-curves for f and g are tangent to each other (multiple zeros), since the approximations to the contours probably fail to have the necessary crossings. The technique is also limited to the plane. In §4.6 we will discuss a slower but more reliable method that addresses both of these problems.

4.5 Newton's Method for Intervals

The interval algorithm of §4.3 solves the problem nicely, but we can improve Algorithm 4.2 in several ways. Extensions of the algorithm so that it is faster and applies to functions in higher dimensions is an active area of research. Here we will only outline two ideas, one easy, one subtle, that can be used to improve it. The first, which might be called *opportunistic evaluation,* is quite simple: right after the subdivision, evaluate the objective function at the rectangle centers. This pointwise functional evaluation might pick up a value that can be used to improve the current best. Then the new current best is used in the size-checking step. This might not lead to great improvement in cases where we start with a good approximation to the answer (such as the -3.24 we had found), but if such an approximation is not available, these extra evaluations will help.

The second improvement relies on a root-finding algorithm called the interval Newton method, a beautiful and important idea originated by R. E. Moore in 1966 [Moo66]. We will

use the term *boxes* for intervals in $\mathbb{R}^n$. The starting point is the Newton operator on boxes, defined as follows. Suppose $F:\mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable; then for a box X in n-space, $N(X)$ is defined to be $m - J^{-1} \cdot F(m)$, where m is the center of X and J is the interval Jacobian $F'[X] = (\partial F_i/\partial x_j[X])_{ij}$. Here J^{-1} is obtained by the standard arithmetic interval operations that arise in the computation of a matrix inverse. If $n = 1$, then $N(X)$ is simply $m - F(m)/F'[X]$, where m is the midpoint of X; note that $N(X) = [-\infty, \infty]$ if $0 \in F'[X]$. This definition leads to several important properties that can be used to design an algorithm to find roots. Here are the main points in the case of one dimension, which is much simpler than the general case.

Theorem 4.1 *Suppose $F:\mathbb{R} \to \mathbb{R}$ is a continuously differentiable function, and $X = [a, b]$ is a finite interval. Then:*

(a) *If r is a zero of F in X, then r lies in $N(X)$.*

(b) *If $N(X) \subseteq X$, then F has a zero in $N(X)$.*

(c) *If $N(X) \subseteq X$, then F has at most one zero in X.*

Proof. (a) If $F'(r) = 0$, then $N(X) = [-\infty, \infty]$; otherwise, the mean-value theorem yields $c \in X$ so that $F(r) - F(m) = F'(c)(r - m)$, and this implies that $r = m - F(m)/F'(c) \in N(X)$.

(b) The hypothesis implies that $F'(x) \neq 0$ on X. Then the mean-value theorem yields values c_i in X so that $m - F(m)/F'(c_1) - a = -F(a)/F'(c_1)$ and $b - (m - F(m)/F'(c_2)) = F(b)/F'(c_2)$; because $N(X) \subseteq [a, b]$, the product of the left sides is positive. But $F'(c_1)$ and $F'(c_2)$ have the same sign, and so the product $F(a)F(b)$ must be negative and F therefore has a zero in $[a, b]$.

(c) If there were two zeros, then the derivative would vanish between them, causing $N(X)$ to become infinite. □

This theorem is very powerful since one can use it to design a simple subdivision algorithm to find zeros. Just subdivide the given interval into subintervals and check, for each subinterval, whether the Newton condition—$N(X) \subseteq X$—holds. If it does, then we know from (c) that there is one and only one zero in X. If it fails in the form $N(X) \cap X = \emptyset$, that is also good, for (a) tells us that there are no zeros in X. If neither situation applies, subdivide and try again. When the Newton condition does hold, we can just iterate the N operator. If we are close enough to the zero, the algorithm converges (see [Kea96, Thm. 1.14]), and in fact will converge quadratically, as does the traditional Newton root-finding method. But some bad things can happen: we might not be close enough for convergence (the exact condition for this depends on, among other things, the tightness of the interval approximation J to the inverse of F'); or there are zeros for which the Newton condition will never hold. Think of the process as a queue: intervals are removed from the queue while, sometimes, their subdivisions are added to the queue, or the interval is added to a list of answers. If the queue becomes empty, we are done. If a max-iteration counter is exceeded and the queue is not empty, then there are some unresolved intervals. This will happen with multiple roots, such as occurs with $x^2 = 0$, for in such cases $N(X) = [-\infty, \infty]$.

For an application of the interval Newton method to a one-dimensional root-finding problem that arises in Problem 8 of the SIAM 100-Digit Challenge, see §8.3.2.

In higher dimensions, parts (a) and (c) remain true, but one must use a variation of the Newton operator to get (b), which is important to guarantee that a zero exists. One approach is by a preconditioning matrix (see [Neu90, Thm. 5.1.7]). Another approach, which we shall follow here, uses the Krawczyk operator. For more information on this important variation to Newton's method, see [Kea96, Neu90]. Let F, m, and J be as defined for the Newton operator. Consider $P(x) = x - YF(x)$, where Y is some type of approximation to J^{-1}, the interval matrix that is $F'[X]^{-1}$. Two natural choices for Y are $F'(m)^{-1}$, or the inverse of the matrix of midpoints of the intervals in the matrix J. The latter is faster, since J has to be computed anyway, and that is what we shall use. Then the Krawczyk operator is defined to be $K(X) = m - YF(m) + (I - YJ)(X - m)$, where I is the $n \times n$ identity matrix (the numeric part of this computation should be done in an interval environment, with m replaced by a small interval around m). To understand the rationale behind the K operator, first recall the mean-value theorem. For a continuously differentiable $F : \mathbb{R}^n \to \mathbb{R}^n$, the mean-value theorem takes the following form: Given x and y in a box X, there are points $c_1, c_2, \ldots, c_n \in X$ so that $F(y) - F(x) = (\nabla F_i(c_i))(y - x)$, where $(\nabla F_i(c_i))$ denotes an $n \times n$ matrix with i indexing the rows.

Now $K(X)$ can be viewed as a "mean-value extension" of P in the following sense: if $P'[X]$ is an interval enclosure of P' on X, the mean-value theorem implies that the box $Q = P(m) + P'[X](X - m)$ contains $P(X)$, which here denotes the exact image, $\{P(x) : x \in X\}$. But because $P'(x) = I - YF'(x)$, we may take $I - YJ$ to be the enclosure. Then Q becomes precisely $K(X)$ and we have proved that $K(X)$ contains $P(X)$. This smooths out the n-dimensional theory nicely as follows.

Theorem 4.2 *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable function, X is a finite box in $\mathbb{R}^n$, $J = F'[X]$ is a componentwise interval enclosure, and Y is the inverse of the matrix of midpoints of the intervals in J; Y is assumed to be nonsingular. Let K be the corresponding Krawczyk operator. Then:*

(a) *If r is a root of $F = 0$ in X, then r lies in $K(X)$.*

(b) *If $K(X) \subseteq X$, then F has a zero in $K(X)$.*

(c) *If $K(X) \subset X$, then F has at most one zero in X.*

Proof. (a) $K(X)$ contains $P(X)$, and so $P(r) \in K(X)$. But $P(r) = r - YF(r) = r$.

(b) Since $P(X) \subseteq K(X) \subseteq X$, P is a contraction mapping, and so the Brouwer fixed-point theorem for continuous functions implies that P has a fixed point in X, which means that F has a zero in X.

(c) This one is subtle. Suppose x and y are distinct roots of $F = 0$ in X. Then the mean-value theorem tells us that $F(x) - F(y) = (\nabla F_i(c_i))(x - y)$ for some points $c_i \in X$. It follows that the matrix is singular, and therefore J contains a singular matrix. However, it is not immediately obvious that this contradicts the Krawczyk condition. Yet it does. For a complete proof, see [Neu90, Thm 5.1.8] and note the hypothesis of strict containment here. □

This theorem immediately gives a local root-finding algorithm, one that will come into play in §4.6. If the Krawczyk condition—$K(X) \subset X$—holds, then we know there is one and only one root in X, and that root lies in $K(X)$. Simply iterate K. When we have a small enough interval, we know the root to the desired accuracy.

Here is how root-finding can help in optimization. Once we have the problem reduced to a single box, as happens after 11 subdivisions in the problem at hand, we can check the Krawczyk condition. If it holds, we can use the Newton–Krawczyk iteration to zoom into the unique critical point in the box, and that will give us the answer more quickly than repeated subdivisions (analogous to the advantage Newton's method provides over bisection in one dimension). For the problem at hand and a tolerance of 10^{-6} for the location of the minimum, using this idea (but not opportunistic evaluation) gives a speedup of about 10%.

To be precise, add the following step to the interval method of Algorithm 4.2 right after the gradient check:

> If $\mathcal{R}$ contains only one rectangle X, compute $K(X)$.
>
> If $K(X) \cap X = \emptyset$, then there is no critical point, and the minimum is on the border; do nothing.
>
> If $K(X) \subset X$, then iterate the K operator starting with $K(X)$ until the desired tolerance is reached.
>
> Use the last rectangle to set a_0 and a_1 (the loop will then terminate).

A *Mathematica* Session

An implementation of this extension is available at the web page for this book. Here is how that code would be used to get 100 digits of the answer, using interval arithmetic throughout. The switch to root-finding occurs after the 12th round, and then convergence is very fast. The output shown is the center of an interval of length less than 10^{-102}.

```
IntervalMinimize[f[x, y], {x, -1, 1}, {y, -1, 1}, -3.24,
  AccuracyGoal → 102, WorkingPrecision → 110]
-3.30686864747523728007611377089851565716648236147628821750129308550
   30919983788829503582548807528349918619 3
```

The use of interval methods in global optimization is a well-developed area, with techniques that go well beyond the brief introduction given here (see [Han92, Kea96]; Hansen reports success on a wide variety of problems, including a 10-dimensional one). Nevertheless, the fact that the very simplest ideas give a verified answer to Problem 4 in a few seconds and with only a few lines of code shows how powerful these ideas are. And it is noteworthy that interval analysis has played an important role in diverse areas of modern mathematics. For example, W. Tucker won the Moore prize for his recent work using interval analysis to prove that the Lorenz equations really do have a strange attractor; this solved one of Steven Smale's Problems for the 21st Century (see [Tuc02]); Hales and Ferguson [FH98] used interval methods in their resolution of the Kepler conjecture on sphere-packing, and Lanford [Lan82] used intervals to prove the existence of a universal limit—the Feigenbaum constant—in certain sequences of bifurcations.

The most important points of the interval approach to Problem 4 are:

- The algorithm is very general and will find the lowest critical point in the rectangle (provided interval arithmetic is available in a form that applies to the objective function and its partial derivatives).
- The results are verifiably correct if one uses intervals throughout.
- Getting the results to very high precision is not a problem.
- Having an interval root-finding method can lead to improvements in the optimization algorithm.
- And most important: The interval algorithm is a reasonable way to solve the problem whether or not one wants proved results. In short, interval thinking yields both a good algorithm and proved results.
- The basic ideas apply to functions of more variables, but life becomes more difficult in higher dimensions. See [Han92, Kea96] for discussions of various enhancements that can be used to improve the basic algorithm (one example: using the Hessian to eliminate n-dimensional intervals on which the function is not concave up).

4.6 A Validation Method for Roots

The pessimistic quote at the start of this chapter leads to the question, How can we be certain that the collection of 2720 critical points found in §4.4 is correct and complete? There are several ways to do this. One can use intervals to design a root-finding algorithm and check that it finds the same set of critical points. That can be done by a simple subdivision process where we keep only rectangles that have a chance of containing a zero, and constantly check and use the Krawczyk condition, $K(X) \subset X$. Here is a formal description.

Algorithm 4.4. Using Intervals to Find the Zeros of F: $\mathbb{R}^n \to \mathbb{R}^n$ in a Box.
Assumptions: Interval arithmetic is available for F.

Inputs: F, a continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^n$;
R, the box in which we want all zeros of F;
ϵ, an upper bound on the absolute error of the zero in terms of Euclidean distance;
`mtol`, a tolerance for combining boxes;
$i_{\max}$, a bound on the number of subdivision steps.
Output: A set of intervals with each one trapping a zero, and a set of intervals that are unresolved (these might contain none or one or more zeros).

Notation: For an n-dimensional box X, let m be the center of X and let M be a small box containing m. Then $K(X) = M - YF[M] + (I - YJ)(X - M)$, where J is the Jacobian interval matrix $F'[X]$ and Y is the inverse of the matrix obtained from J by replacing each box by its center. "Subdividing a box" here means dividing it into 2^n pieces by bisecting each side. "Combining" a set of boxes means replacing any two that have nonempty intersection with

the smallest box that contains both of them and repeating the process until no intersections remain. The Krawczyk condition for a box X is $K(X) \subset X$.

Step 1: Initialize: Let $\mathcal{R} = \{R\}$, $i = 0$, $a = \emptyset$.
Step 2: The main loop:
While $\mathcal{R} \neq \emptyset$ and $i < i_{\max}$:
Let $i = i + 1$;
Let $\mathcal{R}$ be the set of all boxes that arise by uniformly dividing each box in $\mathcal{R}$ into 2^n boxes;
If all boxes in $\mathcal{R}$ have their maximum side-dimension less than `mtol`, let $\mathcal{R}$ be the result of combining boxes in $\mathcal{R}$ until no further combining is possible;
For each $X \in \mathcal{R}$ compute $K(X)$, the Krawczyk image of X:
if $K(X) \cap X = \emptyset$, delete X from $\mathcal{R}$;
if $K(X) \subset X$,
iterate K starting from $K(X)$ until the tolerance (the size of the box, or the size of its F-image) is as desired;
add the resulting box to a and delete X from $\mathcal{R}$.
Step 3: Return a and $\mathcal{R}$, the latter being the unresolved boxes.

The combining step using `mtol` in the main loop requires some explanation. If a zero is near the edge of a box, then it can happen that the subdivision and Krawczyk contraction process will never succeed in isolating it. This can happen even in one dimension if the initial interval is $[-1, 1]$ and the zero is at 0, for then the subdivision process will produce two intervals with a very small overlap, thus placing the zero very near the edge. The combining step checks whether all remaining boxes are so small that we ought to have isolated the zeros (the merging tolerance is provided by the user); the fact that we have not done so ($\mathcal{R}$ is not empty) means that it would be wise to combine small boxes into larger ones. This will likely place the zero nearer the center of a box (but not at the exact center!), and the iterative process then succeeds. Of course, there is the chance of an infinite loop here. Thus it would be reasonable to run this algorithm first with `mtol` $= 0$ so that no combining takes place. If it fails to validate, it can be tried with a setting of, say, `mtol` $= 10^{-6}$.

For zeros x of F at which $F'(x)$ is singular, the Krawczyk condition is never satisfied, and the algorithm does not find those zeros. However, they are not lost, as they will be trapped within the unresolved intervals that are returned, and the user could investigate those further to try to determine if a zero lives in them.

Algorithm 4.4 succeeds in obtaining all 2720 zeros (in eight minutes). It is an important technique, especially in higher dimensions, where we might not have another way to get at the roots (see [Kea96]). But in cases where we think we already have the zeros, we should use an interval algorithm for *validation*; this is a central theme in the field of interval analysis: it is often more appropriate to use traditional numerical algorithms and heuristics to get results that are believed to be complete, and then use interval analysis to validate the results.

Here is a second approach, where we use interval analysis as a validation tool. It is a two-step approach based on the theorem about the Krawczyk operator in §4.5, and it works well for the problem at hand. Suppose the set of approximate zeros is r and has size m.

First we check that each zero in r is roughly correct: for each zero, let X be a small box centered at the zero and check that the Krawczyk condition $K(X) \subset X$ holds; this technique is called *ϵ-inflation* (originally due to Rump; see [Rum98] and references therein). Once this is done, we know that r contains approximations to a set of m true zeros. Then move to verify completeness by carrying out a subdivision process on the given domain, doing two things only with each box that shows up: if interval arithmetic or the Krawczyk operator shows that the enclosure for the F-values on the box does not contain the zero vector, it is discarded; and if the Krawczyk containment holds, so that the box does contain a zero, then it is again discarded, but a running count is incremented by 1. Boxes that remain are subdivided. When no boxes remain, we know that the count equals the number of zeros. If this count equals m, we know the set r was indeed a complete set of approximations to the zeros. Here is a formal description.

Algorithm 4.5. Using Intervals to Validate a Set of Zeros.
Assumptions: Interval arithmetic is available for F and its partial derivatives.

Inputs: F, a continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^n$;
R, the box that is the validation domain;
r, a list, believed to be complete and correct, of the set of zeros of F in R;
ϵ, an upper bound of the absolute error on the zero in terms of Euclidean distance;
`mtol`, a tolerance for combining boxes.

Output: True, if the true zeros in R coincide with the points in r, with the absolute error in each case less than ϵ; False otherwise.

Notation: As in Algorithm 4.4.

Step 1: ϵ-inflation:
For each point in r,
let X be the surrounding box of side-length $2\epsilon/\sqrt{n}$;
check that the Krawczyk condition holds for each X.
If there is any failure, stop and return False.
Step 2: Initialize the subdivision process. Let $\mathcal{R} = \{R\}$; $c = 0$.
Step 3: The main loop:
While $\mathcal{R} \neq \emptyset$:
Let $\mathcal{R}$ be the set of boxes obtained by subdividing each box in $\mathcal{R}$;
Delete from $\mathcal{R}$ any box for which the F-interval does not straddle 0;
Delete from $\mathcal{R}$ any box X for which the Krawczyk condition holds,
increase c by 1 for every such box;
Delete from $\mathcal{R}$ any box X for which $K(X) \cap X = \emptyset$;
If all boxes in $\mathcal{R}$ have their maximum side-dimension less than `mtol`,
let $\mathcal{R}$ be the result of combining boxes in $\mathcal{R}$
until no further combining is possible.
Step 4: If $c =$ the number of points in r, return True, otherwise False.

For the gradient of the function f of Problem 4, the contour method locates the 2720 zeros in 30 seconds. The validation method then takes 30 seconds to check that the zeros

are correct and another 5.5 minutes to verify that the set is complete. While the speedup over the method of using intervals from the beginning to find all the zeros is modest, the validation method is a more elegant algorithm and illustrates the important idea that, when possible, one should put off the interval work to the last stage of a computational project.

Algorithms 4.4 and 4.5 can be combined into a single algorithm that finds and validates the zeros. The key observation is that there is no need to iterate the Krawczyk operator when one can just use the traditional Newton method.

Algorithm 4.6. Using Intervals to Find and Validate a Set of Zeros.
Assumptions: Interval arithmetic is available for F and its partial derivatives.

Inputs: F, a continuously differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^n$;
R, the box that is the validation domain;
ϵ, an upper bound on the absolute error of the zeros in terms of Euclidean distance;
`mtol`, a tolerance for combining boxes.

Output: Validated approximations to the zeros, with the absolute error in each case less than ϵ.

Notation: As before, with s being a set of rough approximations to the zeros, and r the set of final approximations.

Step 1: Let $s = \emptyset$; follow Steps 1 and 2 of Algorithm 4.4, except that when $K(X) \subset X$ add the center of X to s.
Step 2: Apply Newton's method to each point in s, putting the result in r.
Step 3: Use ϵ-inflation as in Step 1 of Algorithm 4.5 to verify that r is a complete set of zeros to the desired tolerance.
Step 4: Return r.

Algorithm 4.6 finds and validates all the critical points to the challenge problem in 5.5 minutes, a 9% time savings over the combination of Algorithms 4.4 and 4.5. The material on the web page for this book includes the *Mathematica* programs `ValidateRoots` and `FindAndValidateRoots`.

4.7 Harder Problems

What sort of problem could have been asked instead of Problem 4 that would have been a little, or a lot, harder? There certainly would have been additional difficulties if there were more dimensions or if the objective function did not have interval arithmetic easily available. Indeed, the optimization needed to solve Problem 5 is exactly of this sort (four dimensions, complicated objective), and that is quite a difficult problem for general-purpose minimization algorithms. Yet another sort of problem would arise if the objective function had not a single minimum, but a continuous set, such as a line or a plane (see [Han92]; there are some examples there). But let us look only at the dimension issue for a moment. Here

is a slight variation of Problem 4, but in three dimensions: What is the global minimum of

$$g(x, y, z) = e^{\sin(50x)} + \sin(60e^y)\sin(60z) + \sin(70\sin x)\cos(10z) + \sin\sin(80y) - \sin(10(x+z)) + (x^2+y^2+z^2)/4\ ?$$

The methods of this chapter work well on this problem, except for the techniques that tried to find all the critical points in a box: there are probably over 100,000 such points. An approach that combines various methods to advantage would proceed as follows:

1. Use the genetic minimization routine with 200 points per generation and a scale factor of 0.9 to discover that g gets as small as -3.327. Differential evolution works too.

2. Use traditional root-finding (Newton) on the gradient to get the more accurate value of -3.32834 (or hundreds of digits if desired).

3. Use interval arithmetic as in §4.3 to then prove that the global minimum is inside the cube $[-0.77, 0.77]^3$.

4. Use the basic interval algorithm of §4.3, with upper bound -3.328 and Krawczyk iteration taking over after 11 rounds, to prove that the minimum is -3.32834 and occurs near $(-0.15, 0.29, -0.28)$. This takes almost a minute, and the total number of boxes examined is 9408.

A *Mathematica* Session

```
g[x_, y_, z_] := e^Sin[50 x] + Sin[60 e^y] Sin[60 z] + Sin[70 Sin[x]] Cos[10 z] +
    Sin[Sin[80 y]] - Sin[10 (x + z)] + 1/4 (x^2 + y^2 + z^2);

IntervalMinimize[g[x, y, z], {x, -0.77, 0.77}, {y, -0.77, 0.77},
  {z, -0.77, 0.77}, -3.328]
```

```
{Interval[{-3.328338345663281, -3.328338345663262}],
  {{x → Interval[{-0.1580368204689058, -0.1580368204689057}],
     y → Interval[{0.2910230486091526, 0.2910230486091528}],
     z → Interval[{-0.2892977987325703, -0.2892977987325701}]}}}
```

So for at least one complicated example, the algorithm works well in three dimensions. If the dimension is increased in such a way that the number of local minima grows with the space (that is, exponentially), one would expect a slowdown in the algorithm, though it will be sensitive to the location of the minima and still might work well.

As a final example, we mention that the Krawczyk validation method given at the end of §4.6 can solve the related problem of finding the critical points of $g(x, y, z)$. Because there are many critical points we work in a small box only ($[0, 0.1] \times [0, 0.05] \times [0, 0.1]$) and show some of them, together with the three surfaces $g_x = 0$, $g_y = 0$, $g_z = 0$, in Figure 4.8. There are six zeros in this box.

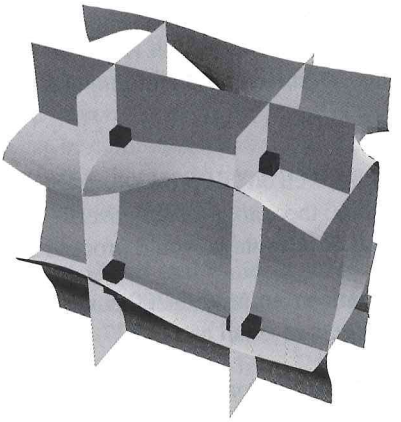

Figure 4.8. *A very small portion—within* $[0, 0.1] \times [0, 0.05] \times [0, 0.1]$*—of the three surfaces* $g_x = 0$, $g_y = 0$, $g_z = 0$ *whose intersections form the critical points of* g*, with the roots obtained by the interval method shown as small cubes.*

4.8 Summary

Interval arithmetic is powerful and can obtain certifiably correct answers to optimization and root-finding problems. The technique is useful not only as a validation tool, but also as a complete algorithm. For many problems it makes sense to use other techniques first, and then use interval arithmetic to validate the results, if that is desired. Random search methods, especially evolutionary algorithms, can be very efficient at getting approximate results. The contour approach is somewhat specialized, but it too is very efficient at solving this, and other, two-dimensional problems.

Chapter 5

A Complex Optimization

Dirk Laurie

Die Lekkerland se pad is 'n lang, lang pad wat in die rondte loop.
Hy slinger deur die bosse en hy kronkel om die rante tot daar waar sy end moet wees.
En dáár begin hy weer!
Hy dwaal deur die vleie en hy boggel oor die bulte en op een plek raak hy weg.
Maar onder die braambos begin hy weer en hy drentel al om en om en om die diep kuil vol soet water wat nooit opdroog nie.
O, dit is 'n lang, lang pad![32]

—W. O. Kühne [Küh82]

If ζ is a real function of a real variable z, then the relation between ζ and z, which may be written $\zeta = f(z)$, can be visualized by a curve in the plane, namely the locus of a point whose coordinates referred to rectangular axes in the plane are (z, ζ). No such simple and convenient geometrical method can be found for visualizing an equation $\zeta = f(z)$, considered as defining the dependence of one complex number $\zeta = \xi + i\eta$ on another complex number $z = x + iy$.

—E. T. Whittaker and G. N. Watson [WW96, p. 41]

Problem 5

Let $f(z) = 1/\Gamma(z)$, where $\Gamma(z)$ is the gamma function, and let $p(z)$ be the cubic polynomial that best approximates $f(z)$ on the unit disk in the supremum norm $\|\cdot\|_\infty$. What is $\|f - p\|_\infty$?

[32]The Lekkerland road is a long, long road that runs in a circle. It winds through the bushes and it coils round the ridges to there where its end should be.// And there it starts again!// It wanders through the wetlands and it hunches over the hillocks and in one place it dwindles away.// But under the brambles it starts again and it saunters all round and round and round the deep pool of sweet water that never dries up.// Oh, it is a long, long road! —W. O. Kühne, translated by Dirk Laurie

5.1 A First Look

We change Trefethen's notation so that p is a generic cubic polynomial

$$p(z) = az^3 + bz^2 + cz + d,$$

and the optimal polynomial[33] is

$$p_{\text{opt}}(z) = a_{\text{opt}}z^3 + b_{\text{opt}}z^2 + c_{\text{opt}}z + d_{\text{opt}}.$$

We need to compute

$$\epsilon_{\text{opt}} = \|f - p_{\text{opt}}\|_\infty.$$

Since $f - p$ is an entire function, by the maximum principle the maximum of $|f - p|$ over the unit disk occurs on the unit circle. So we can revise the definition of the supremum norm in this case to read

$$\|f - p\|_\infty = \max_{\theta\in[0,2\pi]} |f(e^{i\theta}) - p(e^{i\theta})|.$$

Minimax problems are notoriously hard for general-purpose optimization algorithms, so we will try to get as far as we can before invoking such an algorithm; see §5.2.

Any p trivially gives us an upper bound

$$\epsilon_{\text{opt}} \leqslant \|f - p\|_\infty.$$

One candidate for p that is easy to find comes from the first few terms of the Maclaurin series [AS84, form. (6.1.34)] of f, which gives, to four decimals,

$$p_0(z) = -0.6559z^3 + 0.5772z^2 + z.$$

The usual objections to a Maclaurin polynomial, when seen as a way of approximating a real function on a finite interval, do not apply in the case of approximation on the unit circle. On the contrary, the Maclaurin polynomial is optimal in the L_2 norm, defined by

$$\|g\|_2 = \left(\frac{1}{2\pi}\int_0^{2\pi} |g(e^{i\theta})|^2 \, d\theta\right)^{1/2}.$$

The reason for this optimality is that by setting $z = e^{i\theta}$ the Maclaurin series of $g(z)$ passes into the Fourier series of $g(e^{i\theta})$ on the interval $0 < \theta \leqslant 2\pi$.

Another easy way of getting a first approximation would be to discretize the unit circle using n equidistant points θ_j and then to solve the discrete least squares problem, that is, to minimize

$$E(p, n) = \frac{1}{n}\sum_{j=1}^{n} |f(e^{i\theta_j}) - p(e^{i\theta_j})|^2$$

over all cubic polynomials p. Clearly, $E(p, n)$ is merely a discretization of $\|f - p\|_2^2$, and the polynomial thus found therefore tends rapidly to the Maclaurin polynomial as n increases. By calculating $E(p_0, n)$ for increasing values of n, we find that $\|f - p_0\|_2 \doteq 0.177$.

[33] Existence and uniqueness of the optimal polynomial follows from general results on complex Chebyshev approximation; for references, see §5.8.

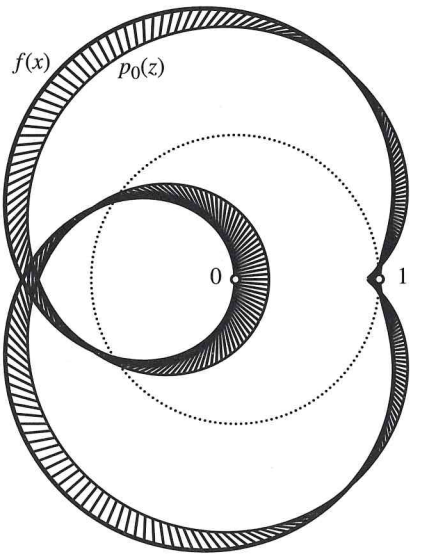

Figure 5.1. *The plots of z, $f(z)$, and $p_0(z)$ in the complex plane when $|z| = 1$.*

To get a visual impression of what we are trying to do, see Figure 5.1, which is a graph of $f(e^{i\theta})$ and of $p_0(e^{i\theta})$ when $\theta = 0°, 1°, 2°, \ldots, 360°$, superimposed on the unit circle. Corresponding points have been connected to make it clear that the quantity $f(z) - p_0(z)$ keeps changing its angle all the time and that $p_0(z)$ is almost never the closest point to $f(z)$ on the graph of p_0.

When θ is restricted to this set of points, we find that $\max |f(\theta) - p_0(\theta)| \doteq 0.235$, so, allowing for a small discretization error, we assert that

$$0.17 < \epsilon_{\text{opt}} < 0.24$$

since

$$\|f - p_0\|_2 \leqslant \|f - p_{\text{opt}}\|_2 \leqslant \|f - p_{\text{opt}}\|_\infty \leqslant \|f - p_0\|_\infty.$$

In a sense, we have one digit: both bounds round to 0.2.

The coefficients of p_0 are real. Can we assume that the coefficients of p_{opt} are also real? The cubic polynomial

$$\hat{p}(z) = \frac{1}{2}(p_{\text{opt}}(z) + \overline{p_{\text{opt}}(\overline{z})})$$

has real coefficients, and for all z, by the symmetry $f(\overline{z}) = \overline{f(z)}$ [AS84, form. (6.1.23)],

$$\begin{aligned} |f(z) - \hat{p}(z)| &\leqslant \frac{1}{2}(|f(z) - p_{\text{opt}}(z)| + |f(z) - \overline{p_{\text{opt}}(\overline{z})}|) \\ &= \frac{1}{2}(|f(z) - p_{\text{opt}}(z)| + |\overline{f(\overline{z})} - \overline{p_{\text{opt}}(\overline{z})}|) \end{aligned}$$

$$= \frac{1}{2}(|f(z) - p_{\text{opt}}(z)| + |f(\overline{z}) - p_{\text{opt}}(\overline{z})|) \leqslant \frac{1}{2}(\epsilon_{\text{opt}} + \epsilon_{\text{opt}}).$$

Thus, $\hat{p}$ is at least as good as p_{opt}, and therefore *it is sufficient to search only among polynomials with real coefficients.*

5.2 Optimization by General-Purpose Methods

General-purpose optimization software does not fare too well on Problem 5. Typically it will halt well before reaching the answer. Yet there is a relatively new technique that does remarkably well here and with a variety of other optimization problems in $\mathbb{R}^n$; it can get a dozen digits in only 40 seconds (*Mathematica* on a 1.25 GHz Macintosh). The method is a type of evolutionary algorithm called *differential evolution*, and it succeeds on many problems where traditional random search methods such as simple evolutionary methods or simulated annealing fail. For an example of success with a simple evolutionary algorithm see §4.2. While no search method can match the analytic methods presented elsewhere in this chapter, the fact that they work at all on this problem is noteworthy and indicates that differential evolution would be a good method to try in situations where there is no known analytic approach.

The idea underlying the algorithm is similar to basic evolutionary methods, with the important twist that the children at each generation are formed by taking a linear combination of parents (as opposed to being simply mutations of a single parent). Each member of the next generation (the new child) has the form $p_1 + r(p_2 - p_3)$, where the p_i are members of the current generation and r is an amplification factor. The formula for children relies heavily on the difference between two of the parents, and this is where the method's name comes from. Here is a formal description.

Algorithm 5.1. Minimization by Differential Evolution.
Input: F, a continuously differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$;
R, a box that serves as the starting seed for the first generation;
N, the population size (number of d-vectors in each generation);
$n_{\max}$, a bound on the number of generations;
r, the amplification factor.

Output: An approximation to the global minimum of F.

Step 1: Initialize the first generation to be a list of N d-vectors by choosing randomly within R. Set up a vector to consist of the F-values at these vectors.

Step 2: The main loop:
For n from 1 to $n_{\max} - 1$, form the $(n+1)$th generation as follows:
For each parent p_i in the nth generation,
construct a child c_i
by randomly choosing indices j, k, and m and letting
$c_i = p_j + r(p_k - p_m)$.

If $F(c_i) < F(p_i)$
let c_i be the new ith member of the $(n+1)$st generation;
else p_i becomes this member.
After all the members of the next generation are determined,
update the value array to contain their F-values.

Step 3: Return the vector in the final generation having the smallest F-value.

Note that a generation is not updated until the entire generation is formed. That is, if a child is to count, it is placed in temporary storage. After the loop is complete, the current generation is replaced by the new generation. Algorithm 5.1 is the barest-possible outline of differential evolution, but it is good enough to solve the problem at hand. One usually selects the indices j, k, and m in such a way that they are distinct from each other and also from i, but ignoring this point speeds up the random integer generation and has no serious consequences. It is also customary to add a crossover step, where some of the entries in the child c_i are replaced by the corresponding entries in the parent p_i, using a probabilistic rule of some sort. But the problem at hand runs faster and more efficiently with no crossover. An amplification factor of 0.4 seems to work.

We need an objective function, of course. Given a cubic $az^3 + bz^2 + cz + d$, where $z = e^{i\theta}$, we can, by symmetry, focus on the domain $\theta \in [0, \pi]$ and use a standard optimization technique starting from a number of seeds, such as is used by *Mathematica*'s `FindMaximum` function, for example. Experiments show that it is sufficient to take seeds near 1.40 and 2.26, and also to throw in the local extremum that always arises (by symmetry) at $\theta = \pi$; that is simply $|a - b + c - d|$. The largest of the three values returned is taken as the value of the objective function.

A *Mathematica* Session

```
maxerror[{a_, b_, c_, d_}] := Max[Abs[a - b + c - d],
   FindMaximum[t = e^(i θ); Abs[a t^3 + b t^2 + c t + d - 1 / Gamma[t]],
       {θ, # - 0.03, # + 0.03}, PrecisionGoal → 12][[1]] & /@ {1.40, 2.26}];
```

Differential evolution is used by *Mathematica*'s `NMinimize` function, but we obtained greater control and understanding by writing our own code. Now, to solve Problem 5 one can let the population size be 60, the amplification factor be 0.4, and the iteration bound be 300. It makes sense to check for convergence, but comparing the best values at successive generations will lead to premature stoppage, since the best might well not change over just one generation. Checking for agreement every 10 generations is more reasonable. When we do that we see convergence to within 10^{-12} in about 180 generations, and therefore about 10,000 evaluations of the objective function. Here is a bare-bones *Mathematica* implementation.[34] One enhancement would be to gradually increase the precision in the objective function as the generations evolve.

[34]MATLAB code can be found on the web page for this book.

A *Mathematica* Session

```
DifferentialEvolution[n_, seeds_] :=
  Module[{best, current, children, vals, childval},
   current = Table[Random[Real, #] & /@ seeds, {n}];
   vals = maxerror /@ current;
   oldbest = Min[vals];
   Do[children = Table[{1, 0.4, -0.4}.
       current[[Table[Random[Integer, {1, n}], {3}]]], {n}];
    Do[If[(childval = maxerror[children[[j]]]) < vals[[j]],
      {current[[j]], vals[[j]]} = {children[[j]], childval}], {j, n}];
    If[Mod[i, 10] == 0, best = Min[vals];
     If[Abs[best - oldbest] < 10^-14, Break[], oldbest = best]], {i, 300}];
   First[Sort[Transpose[{vals, current}]]]];
```

And here is a typical run:

```
SeedRandom[1];
DifferentialEvolution[60, {{-2, 2}, {-2, 2}, {-2, 2}, {-2, 2}}]
```

```
{0.2143352345904104, {-0.6033432200279722,
  0.6252119165808774, 1.019761853131395, 0.005541951112956156}}
```

The numbers in the bottom line are, respectively, ϵ_{opt}, a_{opt}, b_{opt}, c_{opt}, and d_{opt}.

Because of randomness, not every run with these parameters will converge to the right answer, but with $N = 60$ it generally does (353 successes in 400 trials). Figure 5.2 shows the convergence. A comparison of the results of several random runs gives evidence that 12 digits are correct:

$$\epsilon_{\text{opt}} \doteq 0.21433\,52345\,90.$$

For a more comprehensive treatment of the method and further applications, see [CDG99].

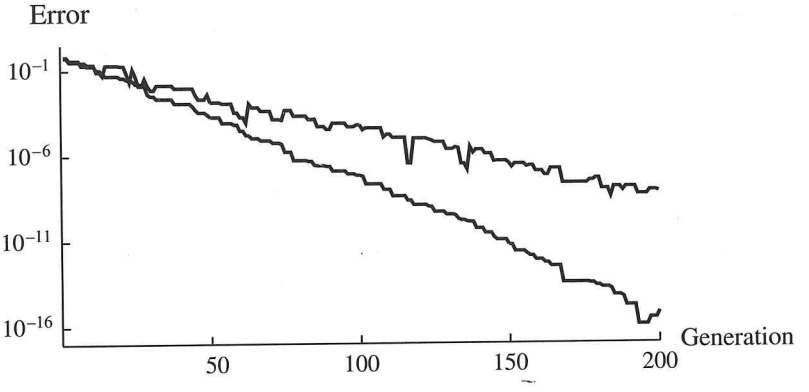

Figure 5.2. *The lower data set shows the convergence to the true answer for a run of 200 generations; the convergence is remarkably steady at about one new digit every 13 generations. The upper data set is the distance from the best cubic at each generation to the truly best cubic (in the Euclidean norm in 4-space).*

5.3 Improving the Lower Bound

Having seen what a sophisticated search technique can do, we return to the basics. Consider the related problems where we minimize the absolute value not of $f - p$, but of its real and imaginary parts separately; that is, we find

$$\epsilon_R = \min_p \epsilon_R(p), \qquad \epsilon_R(p) = \max_{0 \leqslant \theta \leqslant 2\pi} |\mathrm{Re}(f(e^{i\theta}) - p(e^{i\theta}))|;$$
$$\epsilon_I = \min_p \epsilon_I(p), \qquad \epsilon_I(p) = \max_{0 \leqslant \theta \leqslant 2\pi} |\mathrm{Im}(f(e^{i\theta}) - p(e^{i\theta}))|;$$

where in both cases p runs over all cubic polynomials with real coefficients. Clearly ϵ_R and ϵ_I are lower bounds for ϵ_{opt}. These problems involve the Chebyshev approximation of a real-valued function and are therefore easier to solve than the original problem. Let us see what they bring us.

A standard theorem on real Chebyshev approximation (a proof in the case of approximation by polynomials is given in [Rut90, Thms. 7.4 and 7.5] is as follows.

Theorem 5.1 (Alternation Theorem). *Let the continuous functions $f_1, f_2, \ldots, f_n$ be such that the interpolation problem*

$$y_j = \sum_{k=1}^{n} a_k f_k(x_j), \qquad j = 1, 2, \ldots, n,$$

always has a unique solution $a_1, a_2, \ldots, a_n$ when the x_j are distinct points in an interval $[a, b]$. Then the function

$$q(x) = \sum_{k=1}^{n} c_k f_k(x)$$

is the best approximation on $[a, b]$ in the supremum norm to a given continuous function $y(x)$ if and only if $n+1$ points $a \leqslant x_1 < x_2 < \cdots < x_{n+1} \leqslant b$ can be found where $|y(x) - q(x)|$ assumes its maximum value, such that the signs of $y(x_j) - q(x_j)$, $j = 1, 2, \ldots, n+1$, alternate.

The theorem immediately suggests an iterative method, known as the multiple exchange algorithm or the second Remes[35] algorithm [Rem34b, Rem34a].

Algorithm R

0. Find a starting value of q good enough so that step 1, below, is possible. Set ϵ_{old} to some impossibly large value.

1. Find $n+1$ points $x_1 < x_2 < \cdots < x_{n+1}$, where $y - q$ has a local extremum, such that the signs of $y(x_j) - q(x_j)$ alternate.

[35] This name is sometimes transliterated as "Remez," based on the current standard for transliteration from Russian into English. However, the papers for which the author is best known were published in French, and presumably he knew how to spell his own name.

2. Solve the linear equations

$$y(x_j) = \sum_{k=1}^{n} c_k f_k(x_j) + \operatorname{sgn}(y(x_j) - q(x_j))\epsilon, \qquad j = 1, 2, \ldots, n+1,$$

for the unknowns $c_1, c_2, \ldots, c_n, \epsilon$.

3. If $|\epsilon| \geqslant \epsilon_{\text{old}}$, exit. Otherwise replace q by $\sum_{k=1}^{n} c_k f_k$, ϵ_{old} by $|\epsilon|$, and return to step 1.

The algorithm becomes slightly more complicated when there are several ways to choose the alternating extrema of $y - q$; fortunately that is not the case here. We can trivially guarantee that the algorithm terminates in a finite number of steps since, on a computer, there are only finitely many values that ϵ can take, and we stop as soon as no progress is made. Of course, if the initial values are bad, we may stop at a point which is not close to a solution to the problem.

If the functions f_k are smooth enough, then the multiple exchange algorithm can be expected to be quadratically convergent; that is, each iteration approximately doubles the number of correct digits. Moreover, as is more fully explained in Chapter 9, when we consider ϵ as a function of $c_1, c_2, \ldots, c_n$ around the optimum, then the value of ϵ will be correct to twice as many digits as the coefficients are.

Let us try this on the real part of f. The approximating functions are $\operatorname{Re}(e^{ik\theta}) = \cos k\theta$, $k = 0, 1, 2, 3$. These functions satisfy the unique interpolation condition (also called the Haar condition) over $[0, \pi]$. Our initial approximation is the real part of p_0, namely

$$q_0(\theta) = -0.6559 \cos 3\theta + 0.5772 \cos 2\theta + \cos \theta.$$

The five extrema occur at approximately 37°, 73°, 108°, 142°, and 180°. Algorithm R gives

$$q_1(\theta) = -0.589085 \cos 3\theta + 0.657586 \cos 2\theta + 1.067908 \cos \theta + 0.032616$$

with $\epsilon_R(q_1) \doteq 0.211379$. We now have

$$0.211 < \epsilon_{\text{opt}} < 0.236;$$

the polynomial p_1 corresponding to q_1 does not improve on the upper bound obtained from p_0.

In the case of the imaginary part, the approximating functions are $\operatorname{Im}(e^{ik\theta}) = \sin k\theta$, $k = 1, 2, 3$, which satisfy the Haar condition over $[0, \pi]$. There are only three parameters, and the optimum is at

$$q_2(\theta) = -0.596767 \sin 3\theta + 0.636408 \sin 2\theta + 1.028102 \sin \theta,$$

with $\epsilon_I(q_2) \doteq 0.212559$. Somewhat surprisingly, this trigonometric polynomial not only gives a better lower bound than q_1 did (which had one more parameter available) but it turns out that the corresponding polynomial

$$p_2(z) = -0.596767z^3 + 0.636408z^2 + 1.028102z$$

does surprisingly well on the real part, so much so that the least-squares upper bound is also improved, and we have

$$0.2125 < \epsilon_{\text{opt}} < \|f - p_2\|_\infty \doteq 0.2319.$$

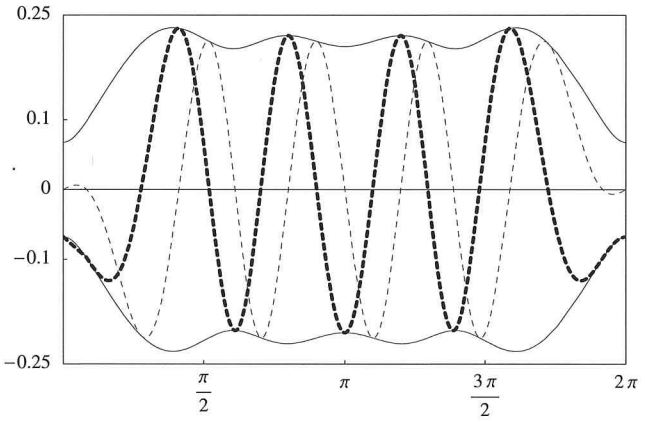

Figure 5.3. *The error in p_2. The solid line gives $|f - p_2|$ and $-|f - p_2|$; the heavy and light dashed lines give the real and imaginary parts of $f - p_2$, respectively.*

(Figure 5.3 shows a graph of the error in p_2.) In retrospect this is perhaps not so surprising, since in principle if we know either the real part or the imaginary part of an analytic function, the other can be found up to an additive constant by applying the Cauchy–Riemann equations.

This picture suggests an easy way to improve the approximation: the constant term does not affect the imaginary part, but it does affect the real part. Note that $\max \mathrm{Re}(f - p_2) > \max \mathrm{Re}(p_2 - f)$. Therefore, for small values of $d > 0$, we should find that $p_2(z) + d$ is a better approximation to f than p_2.

The optimal d satisfies

$$\|f - p_2 - d\|_\infty = |f(-1) - p_2(-1) - d| = p_2(-1) + d,$$

which can be solved by fixed-point iteration from the equivalent equation

$$d = \tfrac{1}{2}(|f - p_2 - d|_\infty - p_2(-1) + d).$$

This gives us the improved approximation

$$p_3(z) = p_2(z) + d = -0.596767z^3 + 0.636408z^2 + 1.028102z + 0.014142.$$

The upper bound is reduced quite a bit by doing this, giving

$$0.2125 < \epsilon_{\mathrm{opt}} < \|f - p_3\|_\infty < 0.2193.$$

We now have two digits: $\epsilon_{\mathrm{opt}} \doteq 0.21$.

The graph of $|f - p_3|$ in Figure 5.4 shows five peaks, one more than the number of parameters, but the two inner peaks are slightly lower than the central and outer peaks. We suspect that the optimal error curve will look like this one, but with all five peaks at the same height.

5.4 Discrete Complex Approximation

There is another way of finding lower bounds: suppose that we somehow are able to solve the minimax problem on a subset S of the unit circle; that is, we can find p_S such that $\epsilon_S(p_S)$

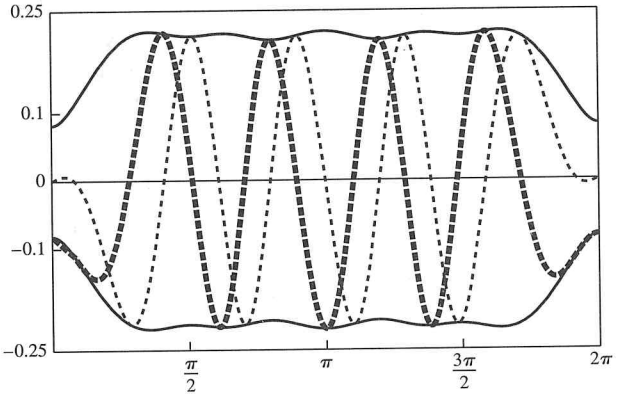

Figure 5.4. *The error in p_3. The solid line gives $|f - p_3|$ and $-|f - p_3|$; the heavy and light dashed lines give the real and imaginary parts of $f - p_3$, respectively.*

is a minimum, where $\epsilon_S(q) = \max_{z \in S} |f(z) - q(z)|$. Then $\epsilon_S(p_S)$ is a lower bound for ϵ_{opt}. This is because

$$\epsilon_S(p_S) \leqslant \epsilon_S(p_{\text{opt}}) \leqslant \max_{|z|=1} |f(z) - p_{\text{opt}}(z)| = \epsilon_{\text{opt}}.$$

Now suppose that S consists only of points where $|f(z) - p_S(z)| = \|f(z) - p_S(z)\|_\infty$; that is, there is no point z outside of S where $|f(z) - p_S(z)|$ is greater than its maximum on S. Then $\epsilon_S(p_S)$ is also an upper bound for ϵ_{opt}, which means that $\epsilon_S(p_S) = \epsilon_{\text{opt}}$. Since we expect the equation $|f(z) - p_S(z)| = \epsilon_S(p_S)$ to hold for only a finite number of values of z, the case of particular interest occurs when S is finite.

We are thus led to the following analogue of Algorithm R for minimizing $\|f - p\|_\infty$ when f is a given complex-valued function and p comes from some n-dimensional linear space $\mathcal{P}$.

Algorithm C

0. Find a starting value of p good enough so that step 1, below, is possible. Set ϵ_{old} to some impossibly large value.

1. Find the set $S = \{z_1, z_2, \ldots, z_m\}$ of points where $|f - p|$ has a local maximum.

2. Solve the restricted problem of finding p that minimizes $\epsilon = \max_j |f(z_j) - p(z_j)|$.

3. If $\epsilon \geqslant \epsilon_{\text{old}}$, exit. Otherwise replace ϵ_{old} by ϵ, and return to step 1.

If this algorithm converges, the answer is the solution we want; of course, it is possible that it might diverge.

The crunch is in step 2. Let $p(z) = \sum_{k=1}^{n} x_k f_k(z)$, where the functions $f_1, f_2, \ldots, f_n$ form a basis for the space $\mathcal{P}$. Then step 2 is the special case where $a_{j,k} = f_k(z_j)$, and $b_j = f(z_j)$ of the discrete linear complex Chebyshev approximation problem:

Given an $m \times n$ matrix A and an m-vector b with complex entries, find a complex n-vector x such that $\|Ax - b\|_\infty$ is a minimum.

Although we happen to know that the x_k are real, there is no theoretical advantage in making that assumption. Watson [Wat88] gave the following characterization theorem.

Theorem 5.2 *The vector x minimizes $\|r\|_\infty$, where $r = b - Ax$, if and only if there exists a set $\mathcal{I}$ containing d indices, where $d \leqslant 2n + 1$, and a real m-vector w, such that:*

(a) $|r_j| = \|r\|_\infty$, $j \in \mathcal{I}$;

(b) $w_j > 0$ *if* $j \in \mathcal{I}$, *and* $w_j = 0$ *for* $j \notin \mathcal{I}$;

(c) $A^*Wr = 0$ *with* $W = \operatorname{diag}(w_j)$, *where* A^* *is the Hermitian transpose of* A.

The set $\mathcal{I}$ is called an active set *and the vector w a* dual solution. *In addition, if all $n \times n$ submatrices of A are nonsingular (the discrete Haar condition), the Chebyshev solution x (but not necessarily the active set $\mathcal{I}$ or the dual solution w) is unique. In that case, $d \geqslant n + 1$.*

The practical implications of this theorem are that, once the set $\mathcal{I}$ is fixed, one obtains d real equations of the form $|r_j| = \epsilon$, $j \in I$, from (a) and n complex equations from (c) for the $d + 1$ real unknowns $\epsilon, w_1, w_2, \ldots, w_d$ and the n complex unknowns x_k. However, the system of equations is not underdetermined, since it is homogeneous in the w_j; we could pick any of the w_j, set it to 1, and solve for the others. If any w_j is then nonpositive, we would know that $\mathcal{I}$ is wrong.

It is in general a combinatorial problem, not at all easy when m is large, to find the active set $\mathcal{I}$ (see, for example, [LV94]). In the present case we are lucky: the error graph for p_3, currently our best available approximation, gives reason to think that S contains only five points, so $m = 5 = n + 1$; and since the Haar condition holds in the space of cubic polynomials, all five points are active.

By symmetry, we can write S in terms of two unknowns as

$$S = \{e^{i\theta_1}, e^{i\theta_2}, -1, e^{-i\theta_2}, e^{-i\theta_1}\},$$

where θ_1 is near 65° and θ_2 is near 113°. The characterization equations then reduce to

$$|f(z_j) - (az_j^3 + bz_j^2 + cz_j + d)|^2 = \epsilon^2, \qquad j = 1, 2, 3, 4, 5;$$

$$\sum_{j=1}^{5} \overline{z_j}^k w_j (f(z_j) - (az_j^3 + bz_j^2 + cz_j + d)) = 0, \qquad k = 0, 1, 2, 3.$$

We have squared the first set of equations in order to obtain expressions that are analytic in terms of the coefficients. Now let $w_3 = 1$; note that, by symmetry, we can discard the equations with $j = 4, 5$, and also take $w_5 = \overline{w_1}$, $w_4 = \overline{w_2}$; then we are left with seven equations in seven unknowns $a, b, c, d, w_1, w_2, \epsilon$, since for the inner iteration the z_j values are fixed. So Algorithm C involves the solution of a system of nonlinear equations *at every iteration*. Also, the question of initial values for the w_j arises.

Surely one can simplify things a little more? Indeed one can, and there are two ways. One of these does away with the outer iteration by differentiating the square of the residual (treated in §5.5), so that the inner iteration converges to the correct answer; the other does away with the inner iteration by explicitly solving the nonlinear system (treated in §5.6). These two sections are independent of each other.

5.5 A Necessary Condition for Optimality

Consider the second set of equations in Theorem 5.2:

$$\sum_{j=1}^{5} \overline{z_j}^k w_j r_j = 0, \qquad k = 0, 1, 2, 3.$$

If we knew the r_j's, these equations (taking $w_3 = 1$) would form a system of four equations in four unknowns w_1, w_2, w_4, w_5. By symmetry, $r_5 = \overline{r_1}$, $r_4 = \overline{r_2}$, $w_5 = \overline{w_1}$, $w_4 = \overline{w_2}$; moreover, we make the assumption that $r_3 < 0$, leading to $r_3 = -\epsilon$. This assumption is based on what we know of p_3, which is already a very good approximant. What is left of the characterization constraints? Well, nothing so far forces the w_j values to be real; this yields the two equations $\operatorname{Im} w_1(z_1, z_2) = 0$ and $\operatorname{Im} w_2(z_1, z_2) = 0$, where the functions w_1 and w_2 are implicitly defined by the system of linear equations. Analytical expressions for $r_1 w_1$ and $r_2 w_2$ are easy to obtain using Cramer's rule and Vandermonde determinants: let

$$V(a, b, c, d) = \begin{vmatrix} 1 & 1 & 1 & 1 \\ a & b & c & d \\ a^2 & b^2 & c^2 & d^2 \\ a^3 & b^3 & c^3 & d^3 \end{vmatrix} = (a-b)(a-c)(a-d)(b-c)(b-d)(c-d),$$

then

$$r_1 w_1 = \frac{V(-1, \overline{z_2}, z_2, z_1)}{V(\overline{z_1}, \overline{z_2}, z_2, z_1)}\,\epsilon, \qquad r_2 w_2 = \frac{V(\overline{z_1}, -1, z_2, z_1)}{V(\overline{z_1}, \overline{z_2}, z_2, z_1)}\,\epsilon. \tag{5.1}$$

Some further simplifications are possible; for example, the denominators are clearly real and therefore irrelevant when testing whether w_1 and w_2 are real.

If Algorithm C converges, the optimal points z_j are such that

$$\left(\frac{d}{d\theta}|f(e^{i\theta}) - p(e^{i\theta})|^2\right)_{\theta=\theta_j} = 0.$$

Since for f and p, differentiation with respect to $z = e^{i\theta}$ is natural, we use $z'(\theta) = iz$ and the chain rule, giving

$$\begin{aligned} 0 &= \left(\frac{d}{d\theta}|f(e^{i\theta}) - p(e^{i\theta})|^2\right)_{\theta=\theta_j} \\ &= \overline{(f(e^{i\theta_j}) - p(e^{i\theta_j}))}\, iz_j(f'(e^{i\theta_j}) - p'(e^{i\theta_j})) \\ &\qquad + (f(e^{i\theta_j}) - p(e^{i\theta_j}))\, \overline{iz_j(f'(e^{i\theta_j}) - p'(e^{i\theta_j}))} \\ &= -2\operatorname{Im}\left(\overline{r}_j z_j (f'(e^{i\theta_j}) - p'(e^{i\theta_j}))\right). \end{aligned}$$

Collecting all our information, we see that the following six equations in the six unknowns $a, b, c, d, \theta_1, \theta_2$ are necessary optimality conditions:

$$|r_j| - \epsilon = 0, \quad \operatorname{Im} w_j(z_1, z_2) = 0, \quad \operatorname{Im}\left(\overline{r_j} z_j (f'(z_j) - p'(z_j))\right) = 0, \quad j = 1, 2, \tag{5.2}$$

where the functions w_1 and w_2 are defined in (5.1) and the auxiliary quantities are

$$\epsilon = d - c + b - a, \qquad z_j = e^{i\theta_j}, \qquad r_j = f(z_j) - p(z_j).$$

The equation for ϵ comes from $1/\Gamma(-1) = 0$, plus the assumption that $r_3 = -\epsilon$. The derivative of f is given by $f'(z) = -\psi(z) f(z)$, where $\psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function.[36]

Instead of the approach with which the previous section ends, whereby a system of seven nonlinear equations must be solved at each step of an iteration involving the z_j's, we now have the alternative of solving the system (5.2) of only six nonlinear equations and obtaining p_{opt} and the critical values z_j without any outer iteration. It is true that our derivation does not guarantee that the solution of (5.2) is unique. To check that the solution thus found is indeed optimal, one could simply plot $|f(z) - p(z)|$ over the whole interval of interest. But it is easier simply to check that $w_1(z_1, z_2)$ and $w_2(z_1, z_2)$ are positive: in that case, a result due to Videnskiĭ (see [Sin70b, Thm. 1.4, p. 182] or [SL68, Lemma 1, p. 450]) guarantees optimality.

Six equations in six unknowns, with no outer iteration: that is surely a viable approach, given the good initial values that we already have. Newton's method,

$$u_{m+1} = u_m - J(F, u_m)^{-1} F(u_m),$$

for improving the approximate solution u_m of a nonlinear system $F(u) = 0$ when the Jacobian $J(F, u)$ can be calculated is likely to work.

An Octave Session

The functions `p5func` (which evaluates the six functions in (5.2) given the six unknowns as a vector) and `jac` (which calculates a numerical Jacobian of a given function) have been predefined and can be found on the web page for this book. I have deleted some padding from the actual output.

```
>> u=[0.014142; 1.028102; 0.636408; -0.596767; [65;113]*pi/180];
>> while true,
>>    f=p5func(u); [u(1)-u(2)+u(3)-u(4), max(abs(f))]
>>    if abs(f)<5e-15, break, end;
>>    u=u-jac('p5func',u)\p5func(u);
>> end
>> u

   ans =
      2.19215000000000e-01   4.38761683759301e+00
```

[36] The complex gamma and digamma functions are available for Octave and MATLAB as contributed implementations by Paul Godfrey:
`http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=978`.

```
    2.09343333405806e-01   7.48097169131265e-01
    2.12244608568497e-01   3.48454464754578e-01
    2.14102287307571e-01   4.63264212148370e-03
    2.14335804172636e-01   4.31572906936413e-05
    2.14335234577871e-01   2.74815371456513e-09
    2.14335234590463e-01   5.82867087928207e-15
    2.14335234590459e-01   2.27453094590858e-15
u =
    5.54195073311451e-03
    1.01976185298384e+00
    6.25211916433389e-01
   -6.03343220407797e-01
    1.40319917081600e+00
    2.26237744961289e+00
```

The convergence criterion has not been plucked out of thin air: since $\epsilon = d-c+b-a \doteq 0.2$ but $|d|+|c|+|b|+|a| \doteq 2$, we expect to lose one digit to roundoff in the calculation of ϵ, and therefore in IEEE double precision (almost 16 digits)[37] one expects that the result will be off by a few units in the 15th digit. We obtain

$$p_4(z) = -0.60334322040780z^3 + 0.62521191643339z^2 + 1.01976185298384z + .00554195073311$$

(the error in p_4 is shown in Figure 5.5) and

$$\theta_1 \doteq 1.4031992, \qquad \theta_2 \doteq 2.2623774, \qquad \epsilon_{\text{opt}} \doteq 0.21433523459046.$$

The angles are near 80.4° and 129.6°, quite some distance from the starting angles; if our initial values were less good, Newton's method might well have failed to converge.

Note that although ϵ_{opt} agrees to 13 digits with the result found in §5.2, the coefficients of p_4 agree to only 8 digits with those given there. As a general rule, in an optimization problem one can find the value of a smooth function at the optimum much more accurately than one can find the location of the optimum. The reason for this is explained more fully in §9.3 in a univariate setting. Thus, when working to 16 digits, we can expect only about 8 correct digits in the coefficients even though the function value itself is correct to nearly the full working precision.

This does not mean that we can afford to display the coefficients of p_4 to less than full accuracy, only that there are certain perturbations of the coefficients (e.g., to change (a, b, c, d) by a multiple of $(1, 1, 1, 1)$) to which $\|f-p\|_\infty$ is very insensitive when p is near p_4. There are other perturbations (e.g., to change (a, b, c, d) by a multiple of $(1, -1, 1, -1)$) to which $\|f-p\|_\infty$ is not insensitive. Thus, the quantity $\|f-p\|_\infty$, when seen as a function of four variables, is smooth in some directions and nonsmooth in others. This property may account for the difficulty that general-purpose optimization methods have on this problem.

It is astonishing just how flat the whole graph is, barely discernible from a horizontal line in the range $72° < \theta < 288°$, which represents three-fifths of the whole. In fact,

[37] The methods of this and the next section also work for 10,000 digits; see Appendix B.

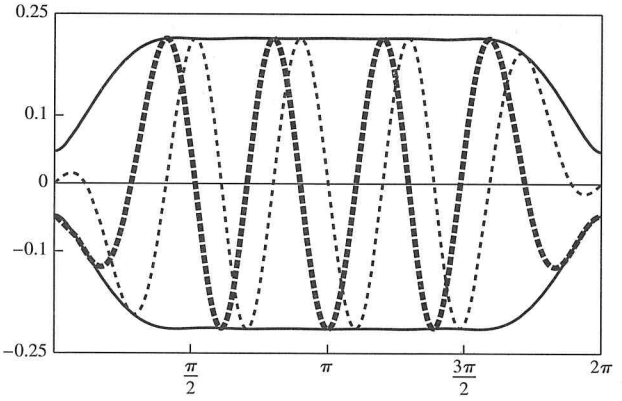

Figure 5.5. *The error in p_4. The solid line gives $|f - p_4|$ and $-|f - p_4|$; the heavy and light dashed lines give the real and imaginary parts of $f - p_4$, respectively.*

$0.21280 < |f(z) - p_4(z)| < 0.21434$ for every $z = e^{i\theta}$ in that region. This property[38] is also one that could cause difficulties for general-purpose optimization methods.

5.6 Implementation of Algorithm C

The discrete linear complex Chebyshev approximation problem has an explicit solution [LV94] when $m = n + 1$:

> Note that the left nullspace of A is one-dimensional. There is thus a unique vector $y = (y_1, y_2, y_3, y_4, y_5)$ in that space satisfying $y_3 = -1$; therefore $\epsilon y = Wr$, so that $w_j = |y_j|$ and $r_j = \epsilon \operatorname{sgn} y_j$. Find that vector y, and then solve the $m \times m$ linear system

$$\begin{pmatrix} 1 & z_1 & z_1^2 & z_1^3 \\ 1 & z_2 & z_2^2 & z_2^3 \\ 1 & z_3 & z_3^2 & z_3^3 \\ 1 & z_4 & z_4^2 & z_4^3 \\ 1 & z_5 & z_5^2 & z_5^3 \end{pmatrix} \begin{pmatrix} d \\ c \\ b \\ a \end{pmatrix} + \epsilon \begin{pmatrix} \operatorname{sgn} y_1 \\ \operatorname{sgn} y_2 \\ \operatorname{sgn} y_3 \\ \operatorname{sgn} y_4 \\ \operatorname{sgn} y_5 \end{pmatrix} = \begin{pmatrix} f(z_1) \\ f(z_2) \\ f(z_3) \\ f(z_4) \\ f(z_5) \end{pmatrix} \tag{5.3}$$

> for the unknowns a, b, c, d, and ϵ.

The system can be simplified to take advantage of conjugate symmetry so that only real quantities appear. So each iteration of Algorithm C involves finding two extrema of $|f(e^{i\theta}) - p(e^{i\theta})|$ near their previous values, followed by the solution of a 5×5 linear system to update p.

It is also possible to take a purely dual view of the problem: all the quantities in (5.3) depend only on the two critical angles θ_1 and θ_2. We can think of (5.3) as defining a function $\epsilon(\theta_1, \theta_2)$. This value of ϵ, as we have seen, is a lower bound for ϵ_{opt}. The dual point of view is simply this: find a local *maximum* of ϵ as a function of θ_1 and θ_2.

[38] The flatness of the error in the best approximation is a well-known feature (called "near-circularity of the error curve") of complex Chebyshev approximation, and in fact is less pronounced for $1/\Gamma$ than for more well-behaved functions [Tre81].

An Octave Session *(Solving the dual problem by Newton's method)*

The routines `hessian` and `cjac` are prewritten, giving numerical approximations to, respectively, the Hessian matrix and the Jacobian of a function of several variables. Their code can be found on the web page for this book.

```
>> function [c,err]=p5solve(z)
>>   for p=1:4, A(:,p)=z.^(p-1); end
>>   b=isgamma(z); d=sign(b-A*(A\b)); c=[A d]\b; err=c(5); c(5)=[];
>> endfunction
>> function y=epsilon(th)
>>   z=z=exp(i*th); z=[z;-1;conj(z)]; [c,e]=p5solve(z); y=real(e);
>> endfunction
>> th=[80;130]*pi/180; f0=epsilon(th);
>> while true,
>>    dth=hessian('epsilon',th)\cjac('epsilon',th)';
>>    new=th-dth; f1=epsilon(new);
>>    [new' f1], if abs(f1-f0)<1e-15, break; end
>>    f0=f1; th=new;
>> end

   ans =
     1.403354225245182   2.262356557886901   0.214335228788773
     1.403199207182680   2.262377441623623   0.214335234590459
     1.403199170789188   2.262377449611380   0.214335234590460
```

For the dual algorithm, the initial values from p_3 are not good enough for Newton's method to converge to the correct values, which is why the very good initial values of 80° and 130° were used here for the demo version of the code. The version of the code on the web page for this book uses a slightly more sophisticated variation of Newton's method in which monotonic convergence of ϵ is enforced, which does converge from the p_3 initial values.

5.7 Evaluation of the Gamma Function

Some people feel that one can trust developers who produce software to evaluate a mathematical function: surely they will have made such a deep study of the function in question that the casual user need not be concerned with the details. This is probably true for the elementary functions as implemented in hardware on standard IEEE-compliant processors, but in the case of the gamma function, there are at least two reasons why one must know something about the implementation. One reason is that only a handful of languages provide the gamma function over the whole complex plane; another is that the gamma function takes much longer to evaluate than do the elementary functions. We may have to write our own routine, and we certainly need to understand the behavior of other people's routines.

In the Octave programs above, use was made of a handwritten routine `isgamma`, since neither Octave nor MATLAB comes with a built-in routine that can evaluate `Gamma(z)` when z is not real. This routine is a straightforward implementation of [AS84, form. (6.1.34)], which gives $1/\Gamma(z)$ as a power series around $z = 0$, taking into account all the coefficients that are significant when working to 16 decimal places. The series seems to be Taylor-made for our application, where $1/\Gamma(z)$ is required in IEEE double precision for $|z| = 1$; it is not suitable when $|z|$ is much larger than that.

The same series was given to 20 decimal places by Luke [Luk75, pp. 1,2] who additionally warmed the hearts of do-it-yourself enthusiasts by supplying a recursion formula for the coefficients in the series. That formula requires the values of Euler's constant γ and of $\zeta(k)$, $k = 2, 3, 4, \ldots$, where ζ is the Riemann zeta function, and is therefore not yet the ultimate answer.

A popular way of evaluating $\Gamma(z)$ is via Stirling's formula [AS84, form. (6.1.42)]

$$\log\Gamma(z) = (z - \tfrac{1}{2})\log z - z + \tfrac{1}{2}\log(2\pi) + \sum_{m=1}^{n} \frac{B_{2m} z^{-2m+1}}{2m(2m-1)} + R_n, \tag{5.4}$$

$$|R_n| = \frac{|B_{2n+2} z^{-2n-1}| K_n}{(2n+1)(2n+2)}, \tag{5.5}$$

where B_n is the nth Bernoulli number, defined recursively by

$$B_n = \sum_{k=0}^{n} \binom{n}{k} B_k, \qquad n = 2, 3, \ldots;$$

starting from $B_0 = 1$ (note that the recursion formula says nothing about B_n, but can be thought of as defining B_{n-1} in terms of its predecessors). Several estimates for K_n are given in [Luk75, pp. 8, 9], but we will only need the simplest: $K_n \leqslant 1$ when $\arg z \leqslant \pi/4$. For those values of z, the error term in (5.5) is smaller in magnitude than the first neglected term.

For $n \geqslant 4$, the estimate (see [AS84, forms. (6.1.38) and (23.1.15)])

$$4\sqrt{\pi n}\left(\frac{n}{\pi e}\right)^{2n} \leqslant |B_{2n}| \leqslant 4.08\sqrt{\pi n}\left(\frac{n}{\pi e}\right)^{2n} \tag{5.6}$$

holds. Thus for $K_n \leqslant 1$ and $n \geqslant 4$,

$$|R_n| \leqslant \frac{2.04|z|\sqrt{\pi(n+1)}}{(n+1)(2n+1)}\left(\frac{n+1}{\pi e|z|}\right)^{2n+2}.$$

A reasonable (not the best) place to stop is when $n + 1 = \lfloor \frac{1}{2}\pi e|z| \rfloor$. In that case, for $\arg z \leqslant \pi/4$ and $|z| > 1$,

$$|R_n| < 2^{-\pi e|z|}. \tag{5.7}$$

We have no way, for fixed z, of substantially decreasing this bound, but we can increase z to make the bound small enough and then recurse back, using the relation

$$\log\Gamma(z+N) - \log\Gamma(z) = \log \prod_{k=0}^{N-1} (z+k).$$

This process will cause some cancellation of significant figures, but nothing catastrophic: for example, for z on the unit circle, if we want 10,000 significant digits, we will need to take N near 4000. Then $\log\Gamma(z+N)$ will be near 30,000, so we can expect to lose about five digits to cancellation.

This, in essence, is the way almost all multiprecision languages with a routine for Γ compute it. By making pessimistic decisions all along the way, a careful implementation can come close to delivering a guaranteed precision.

Why have we gone to such lengths to explain what every implementor knows? This is the reason: *Not only is the gamma function a very expensive function to compute, it also has some rather counterintuitive timing behavior.*

The normal paradigm when comparing optimization methods for speed is to base the comparison on how many function evaluations are required. Underlying this way of thinking is the tacit assumption that not only are the function values far and away the most expensive part of the computation, but also that all function values take approximately the same time to compute. *This assumption is not true in the case of the gamma function.*

A PARI/GP Session

```
? \p1000
?  realprecision = 1001 significant digits (1000 digits displayed)
? #
   timer = 1 (on)
? g=gamma(Pi/4);
   time = 11,121 ms.
? g2=gamma(Pi/4+0.1*I);
   time = 703 ms.
? z3=zeta(3);
   time = 78 ms.
? b2500=bernreal(2500);
   time = 0 ms.
```

Note that it took over 11 seconds to compute the first value of the gamma function, but less than 1 second to compute the next one, even though the second argument is complex. The clue lies in that Bernoulli number that we got instantly. In a fresh PARI/GP session, we get:

```
? b2500=bernreal(2500);
   time = 8,940 ms.
```

So almost all the time goes into the one-off computation of the necessary Bernoulli numbers. Once they are known, the second and later evaluations of the gamma function go quickly. Other multiprecision packages like Maple and *Mathematica* show similar behavior.

To summarize: When going for 10,000 digits on *this* problem (see Appendix B), there is little point in trying to optimize the solution method itself, since that first value of the gamma function totally swamps the computing time.

5.8 More Theory and Other Methods

We have no more than touched upon the very rich theory of approximation in the Chebyshev norm, and in particular we have barely scratched the surface of the results available for approximation by elements of subspaces of a complex vector space. Much more has been proved; for example, there is a general theorem due to Kolmogorov [Kol48] that gives existence and uniqueness of the solution, containing Theorems 5.1 and 5.2 as simple

corollaries. The interested reader is referred to Watson [Wat00] for a historical survey of theory and computational methods for approximation in real vector spaces, and to Singer [Sin70b] for a self-contained treatise in a functional analysis framework, which includes complex vector spaces but omits numerical methods. Both authors have a strong sense of historical responsibility and give numerous references to original sources. Another useful work is [SL68], also a self-contained treatise, but less dauntingly abstract than [Sin70b].

A Google™ search on January 22, 2004, for the exact phrase "complex Chebyshev approximation" returned 181 hits. One of these, an algorithm by Tang [Tan88], was successfully used by at least one winning team.

The special case of which this problem is an example, namely, Chebyshev approximation by a polynomial on the unit circle, has been the subject of some recent papers [Tse96, BT99].

Further, there is the MATLAB package COCA,[39] written by Fischer and Modersitzki, that calculates linear Chebyshev approximations in the complex plane based on techniques similar to those in §5.6. Using it, Problem 5 can successfully be solved with a few lines of code, which can be found on the web page for this book.

5.9 A Harder Problem

Problem 5 can be made substantially harder by introducing one tiny change in the wording:

> *Let* $f(z) = 1/\Gamma(z)$, *where* $\Gamma(z)$ *is the gamma function, and let* $p(z)$ *be the cubic polynomial that best approximates* $f(z)$ *on the unit disk in the* L_1 *norm* $\|\cdot\|_1$. *What is* $\|f - p\|_1$?

The definition of the L_1 norm is

$$\|g\|_1 = \frac{1}{\pi}\int_0^{2\pi}\int_0^1 |g(re^{i\theta})|\, r\, dr\, d\theta.$$

[39] http://www.math.mu-luebeck.de/workers/modersitzki/COCA/coca5.html