COMPUTING THE GAMMA FUNCTION USING CONTOUR INTEGRALS AND RATIONAL APPROXIMATIONS

THOMAS SCHMELZER[†§] AND LLOYD N. TREFETHEN[‡]

Abstract. Some of the best methods for computing the gamma function are based on numerical evaluation of Hankel's contour integral. For example, Temme evaluates this integral based on steepest-decent contours by the trapezoid rule. Here we investigate a different approach to the integral: the application of the trapezoid rule on Talbot-type contours using optimal parameters recently derived by Weideman for computing inverse Laplace transforms. Relatedly, we also investigate quadrature formulas derived from best approximations to $\exp(z)$ on the negative real axis, following Cody, Meinardus and Varga. The two methods are closely related and both converge geometrically. We find that the new methods are competitive with existing ones, even though they are based on generic tools rather than on specific analysis of the gamma function.

Key words. gamma function, Hankel contour, numerical quadrature

AMS subject classification. 65D20, 33F05

1. The gamma function. In his childhood Gauss rediscovered that the sum of the first n positive integers is given by

$$\sum_{k=1}^{n} k = \frac{n\,(n+1)}{2},$$

a formula which can be considered as an interpolation valid even for non-integers. Starting in 1729 Euler discussed in a series of three letters to Goldbach, well known for the Goldbach conjecture, the problem of the product of the first n integers, which is today known as the factorial of n, $n!$. Davis [5] gives details about the history of the gamma function. We start here with the standard definition

$$\Gamma\,(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad \text{Re } z > 0, \tag{1.1}$$

where

$$t^{z-1} = e^{(z-1)\log t} \text{ and } \log t \in \mathbb{R}.$$

The gamma function is analytic in the open right half-plane. Partial integration yields

$$\Gamma\,(z+1) = z\Gamma\,(z)\,, \tag{1.2}$$

and since $\Gamma\,(1) = 1$, we have

$$\Gamma\,(n+1) = n!.$$

Any confusion caused by this identity dates back to Legendre. It is possible to continue the gamma function analytically into the left half-plane. This is often done by a representation of the reciprocal gamma function as an infinite product [1, Eq. 6.1.2]:

$$\frac{1}{\Gamma\,(z)} = \lim_{n\to\infty} \frac{n^{-z}}{n!} z\,(z+1)\dots(z+n)\,,$$

[†]Computing Laboratory, Oxford University, United Kingdom, thoms@comlab.ox.ac.uk
[‡]Computing Laboratory, Oxford University, United Kingdom, lnt@comlab.ox.ac.uk
[§]Thomas Schmelzer is supported by a Rhodes Scholarship.

valid for all z. This representation shows that $\Gamma(z)$ has poles for $z = 0, -1, -2, \ldots$. Of more practical use is the reflection formula [1, Eq. 6.1.17]

$$\Gamma(z)\,\Gamma(1-z) = \frac{\pi}{\sin \pi z}, \qquad z \notin \mathbb{Z}. \tag{1.3}$$

This identity implies $\Gamma(1/2) = \sqrt{\pi}$. It is standard to approximate the gamma function only for Re $z \geq 1/2$ and exploit (1.3) for Re $z < 1/2$.

2. Hankel's representation. An alternative representation for the reciprocal gamma function, which is an entire function, is due to Hankel [10]. Substituting $t = su$ in (1.1) yields

$$F(s) := \frac{\Gamma(z)}{s^z} = \int_0^\infty u^{z-1} e^{-su} du,$$

which can be regarded as the Laplace transform of u^{z-1}. Hence u^{z-1} can be interpreted as an inverse Laplace transform:

$$u^{z-1} = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi i} \int_{\mathcal{C}} e^{ku} F(k) dk.$$

The path $\mathcal{C}$ is any deformed Bromwich contour such that $\mathcal{C}$ winds around the negative real axis in the anti-clockwise sense. Now we substitute $s = ku$, which yields

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i} \int_{\mathcal{C}} s^{-z} e^s ds. \tag{2.1}$$

The numerical evaluation of integrals of the form

$$I = \frac{1}{2\pi i} \int_{\mathcal{C}} e^s f(s) ds \tag{2.2}$$

has been discussed by Trefethen, et al. [21]. The function s^{-z} has a branch cut on $\mathbb{R}^- = (-\infty, 0]$ but is analytic everywhere else. Hence (2.2) is independent of $\mathcal{C}$ under mild assumptions. The freedom to choose the path for inverse Laplace transforms has aroused a good deal of research interest. Recently Weideman [21, 22, 23] has optimized parameters for the cotangent contours introduced by Talbot [17] as well as for other contours in the form of parabolas and hyperbolas. Here we focus on different numerical methods for which (2.1) is the common basis. In particular we shall compare:

1. steepest descent contours,
2. Talbot-type contours,
3. rational approximation of e^s on $(-\infty, 0]$.

The first of these methods is an existing one and the other two are new. Methods we do not compare are those of Spouge, Lanczos and Stirling. Comments on these and on what is done in practice can be found in §7.

In addition we mention in §6 a generalization of (2.1) for matrices and introduce an idea for solving linear systems of the form $\Gamma(A)x = c$ without computing $\Gamma(A)$.

3. Saddle point method. Saddle point methods in general are extensively discussed in the book by Bender and Orszag [2, Section 6.6]. The reciprocal gamma function is a standard example for this technique presented in this and many other textbooks. We keep the details to a minimum and follow an approach of Temme [18],

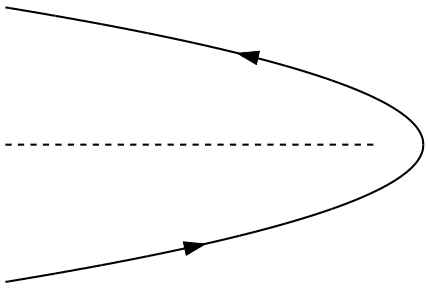


FIG. 2.1. *A typical Hankel contour, winding around the negative real axis (dashed) in the anti-clockwise sense.*

who advocates the numerical evaluation of the integral along a steepest descent contour. A zero of the first derivative of an analytic function f indicates a saddle point of $\left|e^f\right|$. Through this point runs a path $\mathcal{C}$ where f has constant imaginary part and decreasing real part. This is a very desirable property for asymptotic analysis and numerical quadrature schemes. In order to apply these ideas here we fix the movable saddle by a change of variable $s = zt$. We get

$$\frac{1}{\Gamma(z)} = \frac{e^z z^{1-z}}{2\pi i} \int_{\mathcal{C}} e^{z\phi(t)} dt \tag{3.1}$$

where $\phi(t) = t - 1 - \ln t$. If z is real and positive, then the integrand in (3.1) decreases exponentially as t moves away from 1 along the steepest descent contour. For complex z, on the other hand, the decrease becomes oscillatory, and in the limit of pure imaginary z, there is no decrease at all. Thus let us assume that z is a positive real number. Let $t = \rho e^{i\theta}$ be the steepest descent path parameterized by the radius ρ and the argument θ. The vanishing imaginary part at $t = 1$ induces the equation

$$0 = \text{Im } \phi(t) = \rho \sin\theta - \theta.$$

Hence the path is given by $\rho = \theta / \sin\theta$. Temme [18] gives the reparametrization

$$\frac{1}{\Gamma(z)} = \frac{e^z z^{1-z}}{2\pi} \int_{-\pi}^{\pi} e^{-z\Phi(\theta)} d\theta$$

where

$$\Phi(\theta) = 1 - \theta \cot\theta + \ln \frac{\theta}{\sin\theta}$$

with $\Phi(0) = 0$.

The integral can be approximated by the midpoint rule, which is exponentially accurate. See [20] for a review of this phenomenon of high accuracy. The approximated integral is

$$I_N(z) = \frac{e^z z^{1-z}}{N} \sum_{k=1}^{N} e^{-z\Phi(\theta_k)}, \tag{3.2}$$

where the nodes are

$$\theta_k = -\pi + \left(k - \frac{1}{2}\right)\frac{2\pi}{N}, \qquad 1 \le k \le N.$$

This set of nodes is exponentially accurate, but it is not optimal for large z, for the nodes closer to $-\pi$ and π contribute negligibly because of the fast decay along the path. We could delete some of these points to make the method even more efficient, truncating the interval to $[-\tau, \tau]$ instead of $[-\pi, \pi]$.

4. Direct Contour Integration. Instead of working with saddle points, another approach is to apply the trapezoidal rule directly to (2.1). This makes it easy to evaluate $\Gamma(z)$ for complex as well as real arguments. Let $\phi(\theta)$ be an analytic function that maps the real line $\mathbb{R}$ onto the contour $\mathcal{C}$. Then (2.1) can be written as

$$I = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \phi(\theta)^{-z} e^{\phi(\theta)} \phi'(\theta)\, d\theta. \tag{4.1}$$

Because of the term $e^{\phi(\theta)}$, the integrand decreases exponentially as $|\theta| \to \infty$, so that one commits an exponentially small error by truncating $\mathbb{R}$ to a finite interval. For simplicity we shall arbitrarily fix this interval as $[-\pi, \pi]$. In $[-\pi, \pi]$ we take N points θ_k spaced regularly at a distance $2\pi/N$, and our trapezoid approximation to (2.1) becomes

$$I_N = -iN^{-1} \sum_{k=1}^{N} e^{s_k} s_k^{-z} w_k, \tag{4.2}$$

where $s_k = \phi(\theta_k)$ and $w_k = \phi'(\theta_k)$. MATLAB codes are given in Fig. 4.1.

```
function I = ContourIntegral(z,contour,N,f)
  [s,w] = feval(contour,N);                    % contour is a function
  I = zeros(size(z));                          % the different sums
  for k = 1:N                                  % quadrature via
    I = I+w(k)*exp(s(k)).*feval(f,s(k),z);     % evaluating f at the nodes
  end

function [s,w] = contourCot(N)
  t = (-N+1:2:N-1)*pi/N;                       % angles theta
  a = 0.5017; b = 0.2645i; ct = 0.6407*t; d = 0.6122;
  s = N*(a*t.*cot(ct)-d+b*t).';                % poles
  w = -i*(a*cot(ct)-a*ct./sin(ct).^2+b).';     % weights

function f = IntGamma(s,z)
  % for the reciprocal gamma function
  f = s.^(-z);
```

FIG. 4.1. *MATLAB codes to evaluate* (2.2) *by* (4.2). *The function* $f(s) = s^{-z}$ *and the contour* $\mathcal{C}$ *are defined in separate M-files and addressed as handles.*

Note that there is still the freedom left to choose a particular path. In Program 31 of the textbook [19], a closed circle with center $c = -11$ and radius $r = 16$ is used with 70 equidistant nodes on it. Although this contour crosses the branch cut, it does so sufficiently far down the real axis that the error introduced thereby is less than 10^{-11}.

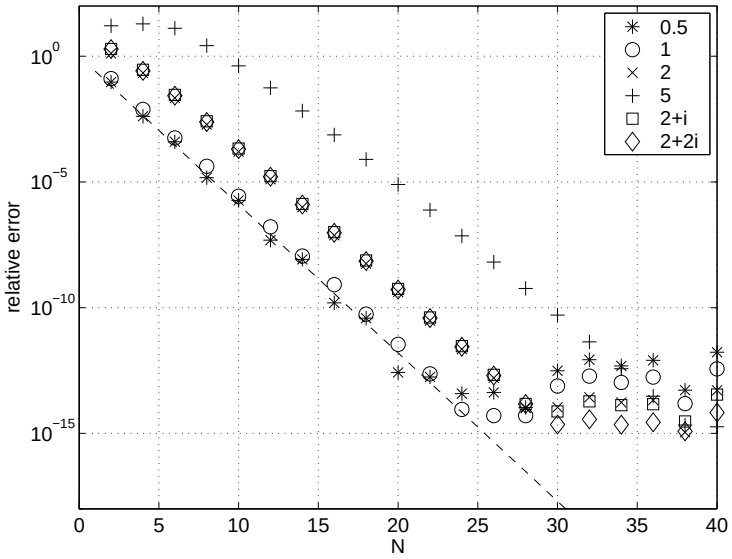


FIG. 4.2. *Convergence of I_N to $1/\Gamma(z)$ for the cotangent contour* (4.2), (4.5), *for six different values of z. The dashed line shows 3.89^{-N}, confirming Weideman's analysis.*

A more systematic approach has been pursued by Weideman [21, 22, 23], who has proposed, in particular, parameters for parabolic, hyperbolic and cotangent contours:

1. Parabolic contour

$$s(\theta) = N\left[0.1309 - 0.1194\theta^2 + 0.2500i\theta\right], \tag{4.3}$$

2. Hyperbolic contour

$$s(\theta) = 2.246N\left[1 - \sin(1.1721 - 0.3443i\theta)\right], \tag{4.4}$$

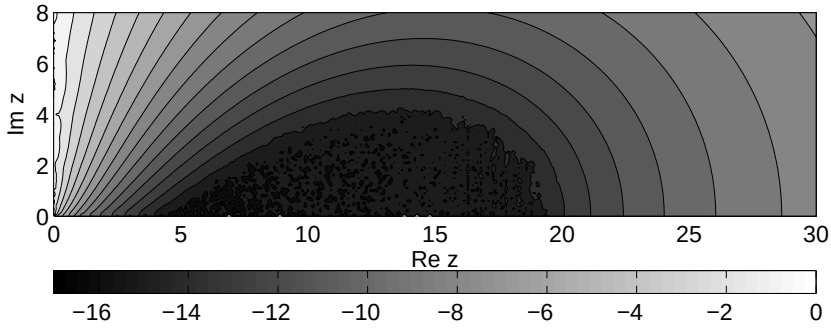
3. Cotangent contour

$$s(\theta) = N\left[0.5017\theta\cot(0.6407\theta) - 0.6122 + 0.2645i\theta\right]. \tag{4.5}$$

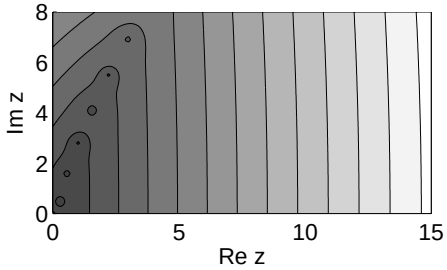
Using equidistant nodes with respect to θ, all of these contours show geometric convergence at rates approximately $O(3^{-N})$. Figure 4.2 illustrates this behaviour by showing convergence as $N \to \infty$ for six values of z. According to Weideman the convergence rate for the cotangent contour is $O(3.89^{-N})$, which is shown as a dashed line in the figure.

In Fig. 4.3, this behavior is compared in a region of the z-plane to the convergence for the parabolic and hyperbolic contours, the steepest descent contours, and the method of rational approximation to be introduced in the next section. All the methods are geometrically convergent (except steepest descents near the imaginary axis), and the cotangent contours and rational approximations are the best.

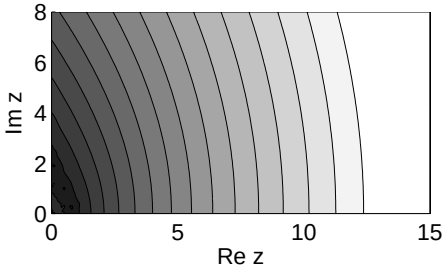
For all of these Talbot-type contours we encounter the same non-optimality effect as for the saddle point method: The decay of the integrand is so fast that the leftmost nodes make a negligible contribution. The source of this phenomenon is the fact that Weideman's analysis considers only the factor e^s in (2.1), treating the factor s^{-z} as of order 1, whereas in fact, when z has large real part, s^{-z} is very small. This



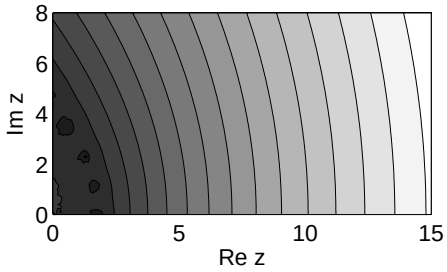
(a) *Saddle point method* (3.2), $N = 32$.



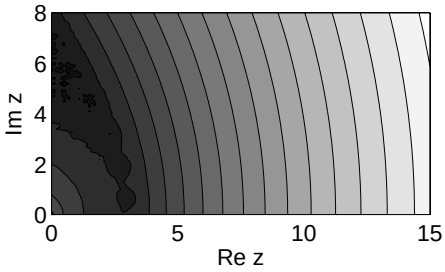
(b) *Circular contour from* [19], $N = 70$.



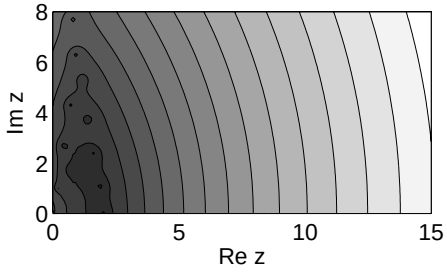
(c) *Parabolic contour* (4.3), $N = 32$.



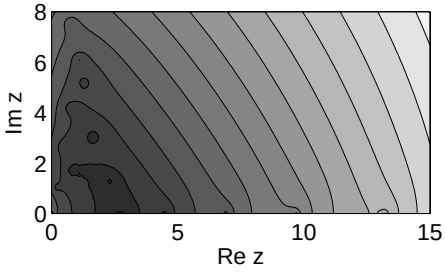
(d) *Hyperbolic contour* (4.4), $N = 32$.



(e) *Cotangent contour* (4.5), $N = 32$.



(f) *CMV approximation* (5.1) *with no shift*, $N = 16$.



(g) *CMV approximation* (5.1) *with shift* $b = 1$, $N = 16$.

FIG. 4.3. *Relative error in evaluating* $\Gamma(z)$ *in various points of the* z*-plane. The colorbar in (a) indicates the scale for all seven plots (logs base* 10*). In practice, one would improve accuracy by reducing values of* z *to a fundamental strip, as shown in Figs.* 4.4 *and* 5.3.

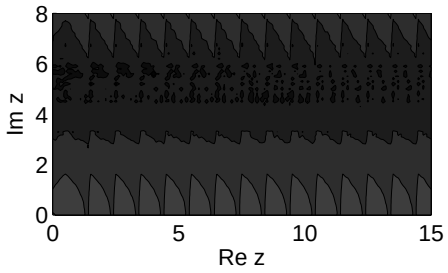


FIG. 4.4. *Relative error in evaluating* $\Gamma(z)$ *using a cotangent contour* (4.5)*,* $N = 32$ *in* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$ *and applying* (1.2) *and* (1.3) *for other points of the z-plane. The shading is the same as in Fig.* 4.3.

```
% gammatalbot - Thomas Schmezler & Nick Trefethen November 2005
%
% For real arguments this is around 20 times slower than Matlab's
% gamma, a factor roughly equal to the product of:
%   5 since this is an M-file rather than a .mex file
%   2 since it uses Talbot quadrature rather than best approximation
%   2 since the real symmetry is not exploited in the sum

function g = gammatalbot(z)                % complex Gamma function
  r = find(real(z)<0.5);                   % reflect to real(z)>=0.5
  z(r) = 1-z(r);
  shift = floor(real(z)-0.5);              % shift to fundamental strip
  zz = z-shift;
  g = 1./ContourIntegral(zz,@contourCot,32,@IntGamma);

  while any(shift)>0
    f = find(shift>0);
    g(f) = g(f).*zz(f);
    shift(f) = shift(f)-1;
    zz(f) = zz(f)+1;
  end
  g(r) = -pi./(g(r).*sin(pi*(z(r)-1)));    % reflect back
  j = find(imag(z)==0); g(j) = real(g(j)); % real inputs -> real outputs
```

FIG. 4.5. *A MATLAB routine for computing the the gamma function. The fundamental identities* (1.2) *and* (1.3) *are used to reduce all arguments to the strip* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$*. The code makes use of the functions listed in Fig.* 4.1.

effect is ubiquitous when computing with a fixed path and fixed nodes for all $z \in \mathbb{C}$. We could take advantage of it by fine-tuning Weideman's parameters in a manner specific to the gamma function, but we shall not do that here since our interest is in the application of generic methods for integrals of the form (2.2). Also, it is simpler and just as effective to use the fundamental identities (1.2) and (1.3) to reduce all arguments to the strip $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$. The effect of such reductions is illustrated for the cotangent contour in Fig. 4.4.

5. Rational Approximation. In a recent paper we interpreted the trapezoidal rule on a Hankel contour as a rational approximation of $\exp(z)$ on the negative real

axis [21]. The analysis of best Chebyshev approximations of this kind is a problem made famous by Cody, Meinardus and Varga [4]; the errors are known to decrease asymptotically at the rate $O(H^N)$, where $H = 1/9.28903\ldots$ is known as *Halphen's constant* [9]. As shown in [21], these approximations can be used directly to evaluate integrals (2.2), bypassing the consideration of Talbot contours and the trapezoid rule. Given N, we define the best type (N, N) approximation to $\exp(s)$ to be the unique real rational function r_N^* of type (N, N) such that

$$\sup_{s\in\mathbb{R}^-} |r_N^*(s) - \exp(s)| = \inf_{r\in R_N} \sup_{s\in\mathbb{R}^-} |r(s) - \exp(s)|$$

where R_N denotes the set of all rational functions of type (N, N). The coefficients of the polynomials in the numerator and denominator of r_N^* are given to very high accuracy in a paper by Carpenter et al. [3]. A practical way of determining these approximants on the fly is the Carathéodory-Fejér (CF) method. (In principle, the CF approximation is not best but near-best, but its difference from the true best approximation is negligible for $N \geq 2$ [21].) The function r_N^* can be represented in a partial fraction representation, that is, by N poles $p_1, \ldots, p_N$ and residues $c_1, \ldots, c_N$ such that

$$r_N^*(s) = \sum_{k=1}^{N} \frac{c_k}{s - p_k} + c_0.$$

We define $\tilde{r}_N(s)$ to be the portion of this expression in the sum, i.e., $\tilde{r}_N(s) = r_N^*(s) - r_N^*(\infty)$, a rational function of type $(N-1, N)$ whose deviation from $\exp(s)$ on $\mathbb{R}^-$ decreases at the same asymptotic rate as that of r_N^* as $N \to \infty$.

These rational approximants can be used as the basis of another method for evaluating $1/\Gamma(z)$. We simply replace e^s in (2.1) by $\tilde{r}_N$ to obtain, with the aid of residue calculus,

$$I_N = \frac{1}{2\pi i} \int_{\mathcal{C}} \tilde{r}_N(s) s^{-z} ds = -\sum_{k=1}^{N} c_k p_k^{-z}, \tag{5.1}$$

which converges for $\mathrm{Re}\ z > 0$ as the decay of the integrand at infinity is fast enough. For $\mathrm{Re}\ z > 1$ we also have

$$I_N = \frac{1}{2\pi i} \int_{\mathcal{C}} r_N^*(s) s^{-z} ds. \tag{5.2}$$

For even N the poles come in conjugate pairs and (5.1) simplifies for real z to

$$I_N = -\sum_{k=1}^{N/2} 2\mathrm{Re}\left(c_k p_k^{-z}\right)$$

provided the first $N/2$ poles are all in the upper half-plane or all in the lower half-plane.

For each z satisfying $\mathrm{Re}\ z > 0$ or $\mathrm{Re}\ z > 1$ as appropriate, I_N appears to converge to $1/\Gamma(z)$ at a geometric rate controlled by the same constant $H = 1/9.28903\ldots$. A proof of this claim would follow from the following result, which we believe is true but do not yet have a proof of.

CONJECTURE 5.1. *Let $\{r_N^*\}$ be the best approximations over $\mathbb{R}^-$ as defined above, let K be a compact set in $\mathbb{C}$, and let $\|\cdot\|_K$ denote the supremum norm over K. Then*

$$\limsup_{N\to\infty} \|\exp(s) - r_N^*(s)\|_K^{1/N} \le H = \frac{1}{9.28903\ldots}.$$

Here is the result that follows from the conjecture:

THEOREM 5.2. *Let $\{\tilde{r}_N\}$ and $\{r_N^*\}$ be the rational approximations defined above and let z be fixed with* Re $z > 0$. *Then the approximations* (5.1) and (5.2) (provided Re $z > 1$) satisfy

$$\limsup_{N\to\infty} \left|\frac{1}{\Gamma(z)} - I_N(z)\right|^{1/N} \le H = \frac{1}{9.28903\ldots}.$$

Partial proof, assuming the validity of Conjecture 5.1. We introduce a special Hankel contour $\mathcal{C}_\rho$. It consists of a circle of radius ρ enclosing the origin and two rays joining $\rho e^{-i\pi}$ and $\rho e^{+i\pi}$ with the point $-\infty$. An upper bound for the error is deduced on $\mathcal{C}_\rho$. For the case of r_N^*, for example, we get by using (2.1) and (5.2)

$$\left|\frac{1}{\Gamma(z)} - I_N(z)\right| \le \frac{1}{2\pi} \|r_N^*(s) - \exp(s)\|_{\mathcal{C}_\rho} \int_{\mathcal{C}_\rho} \left|s^{-z}\right| |ds|$$

and we note that for any s, $|s^{-z}| \le |s|^{-a} e^{\pi|b|}$ for $z = a + bi$ with $a > 1$. From here we readily obtain

$$\int_{\mathcal{C}_\rho} \left|s^{-z}\right| |ds| \le \left(2\pi + \frac{2}{a-1}\right) e^{|b|\pi} \rho^{1-a}.$$

The convergence of $r_N^*(s)$ to $\exp(s)$ on the circle of radius ρ can be estimated by Conjecture 5.1, and therefore

$$\limsup_{N\to\infty} \left|\frac{1}{\Gamma(z)} - I_N(z)\right|^{1/N} \le H.$$

It remains to show that the result just proved for r_N^* and Re $z > 1$ also holds for $\tilde{r}_N$ and Re $z > 0$. To do this split up the integral to obtain the estimate

$$\left|\frac{1}{\Gamma(z)} - I_N(z)\right| \le \frac{1}{2\pi} \|s\,(\tilde{r}_N(s) - \exp(s))\|_{\mathcal{C}_\rho} \int_{\mathcal{C}_\rho} \left|s^{-z-1}\right| |ds|\,.$$

The function $s\,(\tilde{r}_N - \exp(s))$ in the left-hand term of this estimate approaches a constant as $s \to -\infty$ for each N, since $\tilde{r}_N - \exp(s)$ decreases at the rate $O(s^{-1})$. The essential point in showing that these Nth roots approach H as required is to make sure that the leftmost extremum of $\tilde{r}_N(s) - \exp(s)$ does not occur at a value of s that is exponentially large, in which case the Nth root of this value of s might fail to converge to 1. In fact, the results of Aptekarev and Magnus appear to confirm numerical evidence that the location of this extremum grows just algebraically, but we will not attempt a rigorous proof here. □

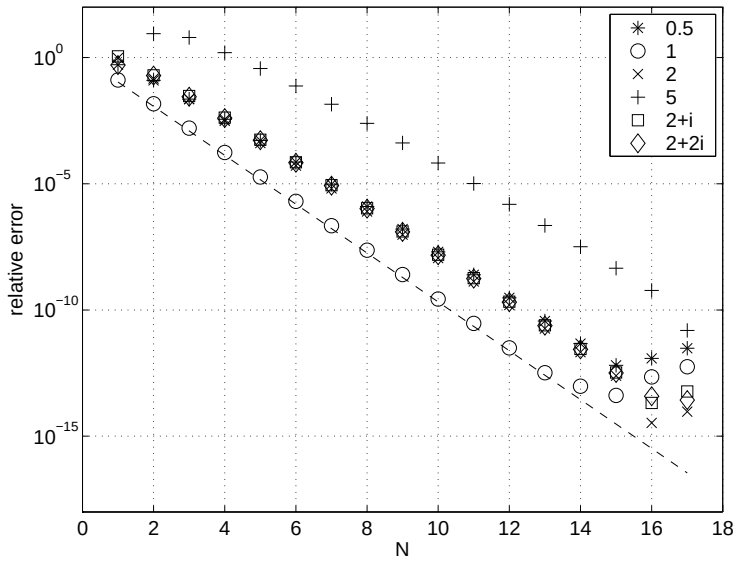


FIG. 5.1. *Convergence for the near-best rational approximation* (5.1) *of type* $(N-1, N)$ *with no shift. The convergence is about twice as fast as in Fig.* 4.2*, with fifteen integrand evaluations sufficing to produce near machine precision. The dashed line shows* 9.28903^{-N}*, confirming Theorem* 5.2.

The fundamental property $\exp(a+b) = \exp(a)\exp(b)$ for any two complex arguments can be exploited in our algorithm. Given a positive parameter b, the function $\tilde{r}_N^b(s) = \exp(b)\tilde{r}_N(s-b)$ can be regarded as an approximation of $\exp(s)$ in the interval $(-\infty, b]$. In particular, equation (5.1) is the special case of this approximation for $b = 0$:

$$I_N^b = \frac{1}{2\pi i}\int_{\mathcal{C}} \tilde{r}_N^b(s) s^{-z} ds = -\sum_{k=1}^{N} e^b c_k (p_k + b)^{-z}. \tag{5.3}$$

It is easily proved that the *shifted* rational approximation $\tilde{r}_N^b(s)$ still converges with the same asymptotic rate H^N. In experiments we have observed that a shift of $O(1)$ gives better results especially for real arguments, as illustrated in Fig. 5.2 and Fig. 4.3(g), where we used a shift of $b = 1$. The results for $b = 0$ are given in Fig. 5.1.

6. Matrix arguments. Hankel's contour integral (2.1) can be generalized to square matrices A, and one can apply the methods introduced here to compute $\Gamma(A)^{-1}$ or to compute the solution vector x in a linear system $\Gamma(A)x = c$ without computing $\Gamma(A)$. We have confirmed this by numerical experiments not reported here. A drawback of such methods is that it is expensive to compute $s_k^{-A}c$ for every node; methods based on the algorithms of Spouge and Lanczos might be more efficient. We are currently not aware of applications where $\Gamma(A)$ is used for matrix arguments.

7. Other methods and existing software. There are a variety of existing methods for computing the gamma function. Are our methods competitive with these? As far as we can tell, the answer seems to be yes, they are "in the ballpark" in the sense of coming within a factor of 1–10 of the best methods, notably

- the method of Lanczos [11],
- the method of Spouge [16],

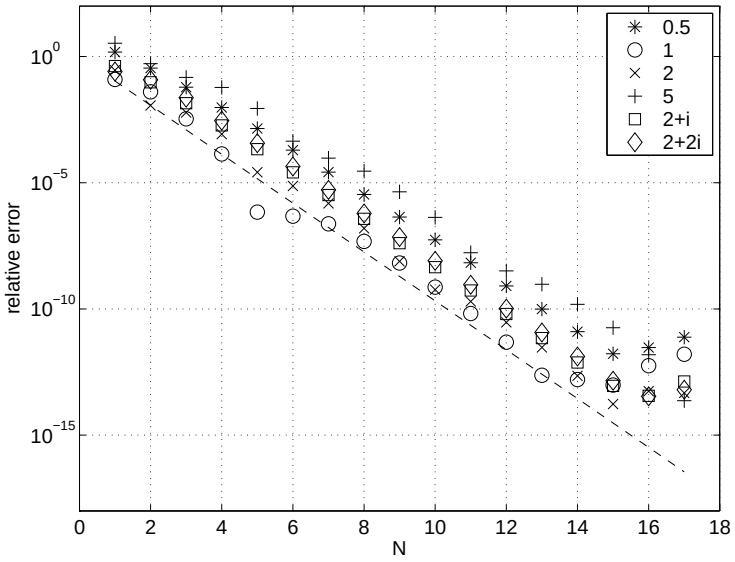


FIG. 5.2. *Convergence for the near-best rational approximation* (5.1) *of type* $(N-1, N)$ *with shift* $b = 1$. *Though the asymptotic behaviour is the same, the constants are better than in Fig.* 5.1, *and the use of such a shift might be a good idea in practice.*

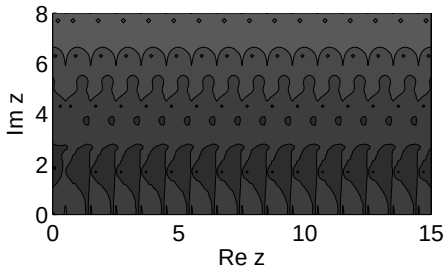


FIG. 5.3. *Relative error in evaluating* $\Gamma(z)$ *using a CMV approximation,* $N = 16$ *with no shift solely in* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$ *and applying* (1.2) *and* (1.3) *for other points of the* z*-plane. The shading is the same as in Fig.* 4.3.

- the asymptotic Stirling series [1, Eq. 6.1.37].

We emphasize that these methods are specialised algorithms designed for computing the gamma function and its close relatives, whereas our ideas are applicable in a much larger framework.

7.1. The method of Spouge. The method of Spouge is attractive because of its simplicity and precise error estimates. Spouge introduced the approximation

$$\Gamma(z+1) \approx (z+\gamma)^{z+1/2} e^{-(z+\gamma)} \sqrt{2\pi} \left[c_0 + \sum_{k=1}^{N} \frac{c_k(\gamma)}{z+k} \right],$$

which is valid for $\mathrm{Re}\,(z+\gamma) > 0$ and dependent on a positive real parameter γ with $N = \lceil \gamma \rceil - 1$, which converges to an equality as $\gamma \to \infty$. Here $c_0 = 1$, and the other

coefficients are given by

$$c_k(\gamma) = \frac{1}{\sqrt{2\pi}} \frac{(-1)^{k-1}}{(k-1)!} (-k+\gamma)^{k-1/2} e^{-k+\gamma}, \qquad 1 \le k \le N.$$

The absolute error for this approximation can be bounded [16, Theorem 1.3.1] by

$$E_N(z) \le \left| \gamma(z) \frac{1}{\sqrt{N+1}(2\pi)^{N+3/2}} \right|.$$

Note that the relative error does not depend on z, making Spouge's method especially attractive for uniform approximations in the right half-plane. The above inequality implies that the method converges at least as fast as (6.28^{-N}), a rate lying midway between (3.89^{-N}) for Talbot contours and (9.29^{-N}) for best rational approximations. Actually, experiments suggest a better convergence rate, closer to $O(10^{-N})$.

7.2. The method of Lanczos. The method of Lanczos is closely related to that of Spouge. Lanczos's method is based on the fast evaluation of the integral

$$F_\gamma(z) = \int_0^e [v(1-\log v)]^z \, v^\gamma dv$$

where γ is a positive free parameter. The integral is approximated by a rational function

$$F_{N,\gamma}(z) = a_0 + \sum_{k=1}^{N} a_k/(z+k).$$

A variety of methods for computing the coefficients are discussed in a recent thesis by Pugh [14]. Their rate of decay depends strongly on a good choice for γ. However, it is unclear if it makes sense to ask about the asymptotic behaviour for $N \to \infty$. Little is known about the decay of the error $|F_\gamma(z) - F_{N,\gamma}(z)|$ [14, Chapter 11]. Lanczos claimed that the higher γ becomes, the smaller is the value of the coefficients at which the convergence begins to slow down. At the same time, however, we have to wait longer before the asymptotic stage is reached. Pugh [14] calls this behaviour the *Lanczos shelf* and is interested in finding good pairs of γ and N in order to guarantee a certain precision in the right half-plane. Godfrey [8] gives a 15-term expansion that provides an accuracy of about 15 significant digits along the real axis and about 13 digits in the rest of the complex plane. Because of the simple form of $F_{N,\gamma}(z)$, Lanczos's method is particularly suitable for matrix arguments.

7.3. Stirling's method. The asymptotic series that generalizes Stirling's formula[1] is still a standard and powerful method for evaluating the gamma function. There is a great deal of literature discussing efficient strategies and error estimates for these series (see the references in [12]). The goal here is to minimize the number of terms used to achieve the desired accuracy. This can be done in two ways, either by shifting the argument to the right or by enforcing a faster asymptotic decay of the relative error using more terms in the series. (For fixed z and $N \to \infty$ the series does not converge.) The method is especially attractive for arguments with large real part working in an arbitrary precision environment. Using an asymptotic series for $\log\Gamma(z)$, the error is bounded for $\mathrm{Re}\, z \ge 0$ by $|B_{2N}/(2N-1)|\,|z|^{1-2N}$ where B_{2N} denotes a Bernoulli number. This simple error estimate is due to Spira [15].

[1] Stirling was a student at the same Oxford college we both belong to, Balliol.

7.4. Software. Software libraries and programming environments for scientific computing all have routines to compute the gamma function, although quite a few do not deal with complex arguments. For our small survey we explored online documentations for various products and yet it often remains unclear exactly which methods are used. For real arguments, a popular trick is to work with a rational Chebyshev approximation on the interval $[1, 2]$ and map this interval by the recurrence relation (1.2) to larger regions of the real line. The routine in the NAG library[2] seems to map this interval to the whole real line, whereas MATLAB[3] uses a Stirling approximation for arguments larger than 12. On the fundamental interval, MATLAB uses a rational Chebyshev approximation of type $(8, 8)$. As the MATLAB routine was originally designed for Fortran[4] we imagine that many Fortran providers compilers use essentially the same method.

None of the above products provides a function for complex arguments. For Fortran the IMSL Library[5] has a routine of this kind. As there are no references to the work of Lanczos and Spouge in the IMSL documentation, we presume that it is based on asymptotic series.

Mathematica[6] uses the asymptotic Binet formula, which is another name for Stirling series. We presume Maple uses the same method, since the Maple documentation gives a reference to the classic book on special functions [6], which appeared before the methods of Lanczos and Spouge were introduced. Somewhat more interesting are the comments in [13]:

> There are a variety of methods in use for calculating the function $\Gamma(z)$ numerically, but none is quite as neat as the approximation derived by Lanczos. This scheme is entirely specific to the gamma function, seemingly plucked from thin air.

8. Conclusions. We have shown that $\Gamma(z)$ can be evaluated with geometric accuracy by two types of generic related methods:

- Applying the trapezoidal rule on Talbot contours
- Using best rational approximations on the negative real axis.

Typically the second method is about twice as fast as the first. However, the first is much simpler to implement as the construction of the best rational approximation is not trivial.

Amongst the Talbot contours, the cotangent contour has the best results. Using a shift from $(-\infty, 0]$ to $(-\infty, 1]$, one can improve the the results for the best rational approximation a bit. For smaller values of z in the right half-plane the approximations are excellent, and using the fundamental recurrence relation for the gamma function one can extend the region of accuracy.

Even though the methods we have introduced are based on generic tools rather than on specific analysis of the gamma function, they are competitive with existing ones. The gamma function is just one of many special functions that have integral representations which can be evaluated efficiently by Talbot-type contours and rational approximations (see [7] for further examples). We believe that these methods can

[2] `http://www.nag.co.uk/numeric/FN/manual/pdf/c03/c03m02_gamma_fun_fn03.pdf`

[3] In MATLAB 7.0 the command `type gamma` gives the source code of the corresponding `mex`-file. Previous versions do not offer this possibility.

[4] `http://csit1cwe.fsu.edu/extra_link/xlhpf/xlflrm03.htm`

[5] `http://www.vni.com/books/dod/pdf/SFun.pdf`

[6] `http://documents.wolfram.com/v5/TheMathematicaBook/MathematicaReferenceGuide/SomeNotesOnInternalImplementation/A.9.4.html`

be useful in many areas of scientific computing.

Acknowledgments. We are grateful to André Weideman for discussions throughout this project, to Nico Temme for expert advice on the gamma function, and to Alphonse Magnus and Alexander Aptekarev for advice about Conjecture 5.1 and the challenges involved in proving it.

REFERENCES

[1] Milton Abramowitz and Irene A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Dover, New York, 1965.
[2] Carl M. Bender and Steven A. Orszag, *Advanced mathematical methods for scientists and engineers*, McGraw-Hill, New York, 1978.
[3] A. J. Carpenter, A. Ruttan, and R. S. Varga, *Extended numerical computations on the "1/9" conjecture in rational approximation theory*, in Rational approximation and interpolation, P.R. Graves-Morris, E.B. Saff, and R.S. Varga, eds., Lecture Notes in Math. 1105, Springer, 1984, pp. 383–411.
[4] W. J. Cody, G. Meinardus, and R. S. Varga, *Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems*, J. Approximation Theory, 2 (1969), pp. 50–65.
[5] Philip J. Davis, *Leonhard Euler's integral: A historical profile of the gamma function*, Amer. Math. Monthly, 66 (1959), pp. 849–869.
[6] Arthur Erdélyi, Wilhelm Magnus, Fritz Oberhettinger, and Francesco G. Tricomi, *Higher transcendental functions Vols. I, II*, McGraw-Hill, New York, 1953.
[7] Amparo Gil, Javier Segura, and Nico M. Temme, *Computing special functions by using quadrature rules*, Numer. Algorithms, 33 (2003), pp. 265–275.
[8] Paul Godfrey, *A note on the computation of the convergent Lanczos complex gamma approximation*. http://my.fit.edu/~gabdo/gamma.txt, 2001.
[9] A. A. Gonchar and E. A. Rakhmanov, *Equilibrium distributions and degree of rational approximation of analytic functions*, Math. USSR Sbornik, 62 (1989), pp. 305–348.
[10] H. Hankel, *Die Euler'schen Integrale bei unbeschränkter Variabilität des Arguments*, Zeitschrift für Math. und Phys., 9 (1864), pp. 1–21.
[11] C. Lanczos, *A precision approximation of the gamma function*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 1 (1964), pp. 86–96.
[12] Edward W. Ng, *A comparison of computational methods and algorithms for the complex gamma function*, ACM Trans. Math. Software, 1 (1975), pp. 56–70.
[13] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical recipes in C*, Cambridge University Press, Cambridge, second ed., 1992.
[14] Glendon R. Pugh, *An analysis of the Lanczos gamma approximation*, PhD thesis, University of British Columbia, 2004.
[15] Robert Spira, *Calculation of the gamma function by Stirling's formula*, Math. Comp., 25 (1971), pp. 317–322.
[16] John L. Spouge, *Computation of the gamma, digamma, and trigamma functions*, SIAM J. Numer. Anal., 31 (1994), pp. 931–944.
[17] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.
[18] Nico M. Temme, *Special functions: An introduction to the classical functions of mathematical physics*, Wiley, New York, 1996.
[19] Lloyd N. Trefethen, *Spectral methods in MATLAB*, SIAM, Philadelphia, PA, 2000.
[20] Lloyd N. Trefethen and J. A. C. Weideman, *The fast trapezoid rule in scientific computing*, manuscript in preparation (2005).
[21] Lloyd N. Trefethen, J. A. C. Weideman, and Thomas Schmelzer, *Talbot quadratures and rational approximations*, BIT, accepted (2005).
[22] J. A. C. Weideman, *Optimizing Talbot's contours for the inversion of the Laplace transform*, Tech. Report NA 05/05, Oxford University Computing Laboratory, 2005.
[23] J. A. C. Weideman and Lloyd N. Trefethen, *Parabolic and hyperbolic contours for computing the Bromwich integral*, submitted to Math. Comp.