J. Aust. Math. Soc. **107** (2019), 319–337
doi:10.1017/S1446788718000393

ON THE ACCURACY OF ASYMPTOTIC APPROXIMATIONS TO THE LOG-GAMMA AND RIEMANN–SIEGEL THETA FUNCTIONS

RICHARD P. BRENT

(Received 12 July 2017; accepted 10 October 2018; first published online 21 December 2018)

Communicated by B. Sims

In memory of Jonathan Borwein 1951–2016

Abstract

We give bounds on the error in the asymptotic approximation of the log-Gamma function $\ln\Gamma(z)$ for complex z in the right half-plane. These improve on earlier bounds by Behnke and Sommer [*Theorie der analytischen Funktionen einer komplexen Veränderlichen*, 2nd edn (Springer, Berlin, 1962)], Spira [‘Calculation of the Gamma function by Stirling’s formula’, *Math. Comp.* **25** (1971), 317–322], and Hare [‘Computing the principal branch of log-Gamma’, *J. Algorithms* **25** (1997), 221–236]. We show that $|R_{k+1}(z)/T_k(z)| < \sqrt{\pi k}$ for nonzero z in the right half-plane, where $T_k(z)$ is the kth term in the asymptotic series, and $R_{k+1}(z)$ is the error incurred in truncating the series after k terms. We deduce similar bounds for asymptotic approximation of the Riemann–Siegel theta function $\vartheta(t)$. We show that the accuracy of a well-known approximation to $\vartheta(t)$ can be improved by including an exponentially small term in the approximation. This improves the attainable accuracy for real $t > 0$ from $O(\exp(-\pi t))$ to $O(\exp(-2\pi t))$. We discuss a similar example due to Olver [‘Error bounds for asymptotic expansions, with an application to cylinder functions of large argument’, in: *Asymptotic Solutions of Differential Equations and Their Applications* (ed. C. H. Wilcox) (Wiley, New York, 1964), 16–18], and a connection with the Stokes phenomenon.

2010 *Mathematics subject classification*: primary 11M06; secondary 33B15, 33B99, 65D20.

Keywords and phrases: asymptotics, Gamma function, log-Gamma function, Riemann zeta function, Riemann–Siegel theta function.

1. Introduction

The *Riemann–Siegel theta function* $\vartheta(t)$, which occurs in the theory of the Riemann zeta function [11, Section 6.5], is defined for real t by

$$\vartheta(t) := \arg\Gamma\left(\frac{it}{2}+\frac{1}{4}\right) - \frac{t}{2}\log\pi. \tag{1.1}$$

The argument is defined so that $\vartheta(t)$ is continuous on $\mathbb{R}$, and $\vartheta(0) = 0$. Clearly $\vartheta(t)$ is an odd function, that is, $\vartheta(-t) = -\vartheta(t)$ for all real t, so there is no essential loss of generality in assuming that t is positive.

The significance of $\vartheta(t)$ is the fact that $Z(t) := \exp(i\vartheta(t))\,\zeta(\frac{1}{2} + it)$ is a real-valued function. Thus, zeros of $\zeta(s)$ on the critical line $\mathrm{Re}(s) = \frac{1}{2}$ can be detected by sign changes of $Z(t)$. In a sense, $\vartheta(t)$ encodes half the information contained in $\zeta(\frac{1}{2} + it)$ (albeit the less interesting half), while $Z(t)$ encodes the other half.

The motivation for this paper was an attempt to give a straightforward proof for the well-known asymptotic expansion

$$\vartheta(t) \sim \frac{t}{2}\log\left(\frac{t}{2\pi e}\right) - \frac{\pi}{8} + \sum_{j=1}^{\infty} \frac{(1 - 2^{1-2j})\,|B_{2j}|}{4j(2j-1)\,t^{2j-1}}, \tag{1.2}$$

and to obtain a rigorous bound on the error incurred in truncating the sum after k terms. A bound

$$\frac{(2k)!}{(2\pi)^{2k+2}\,t^{2k+1}} + \exp(-\pi t) \tag{1.3}$$

was stated in [7, following Equation (2.3)], but no proof was given, and in fact the bound is incorrect[1]. For example, with $k = 3$ and $t = 9.5$, the error exceeds the bound by a factor of 1.011.

To obtain a satisfactory error bound to replace (1.3) we needed an error bound for Stirling's asymptotic approximation [1, (6.1.40)] to $\ln\Gamma(z)$ on the imaginary axis $\mathrm{Re}(z) = 0$. We found several such bounds in the literature, but they were not entirely satisfactory for our purposes (see Remarks 2.4–2.8). Hence, Theorems 2.1 and 2.10 and Corollary 2.2 give new error bounds on Stirling's approximation. These bounds are valid in the right half-plane ($\mathrm{Re}(z) \ge 0$, $z \ne 0$), and improve on previous bounds when z is on or sufficiently close to the imaginary axis.

Stirling's approximation leads, via the duplication formula for the Gamma function, to an asymptotic expansion

$$\ln\Gamma\left(z + \frac{1}{2}\right) \sim z\log z - z + \frac{1}{2}\log(2\pi) + \sum_{j=1}^{\infty} \frac{B_{2j}(\frac{1}{2})}{2j(2j-1)\,z^{2j-1}}$$

that is mentioned by Gauss [13, Equation [59] of Art. 29], and in some sense is due to Stirling; see [25]. It is the special case $a = \frac{1}{2}$ of an expansion for $\ln\Gamma(z + a)$ that was considered, for $a \in [0, 1]$ and real positive z, by Hermite [16]. See also Askey and Roy [2, 5.11.8], and Nemes [19, (1.6)]. Using our bounds on the error in Stirling's approximation to $\ln\Gamma(z)$, we deduce bounds on the error in Gauss's approximation to $\ln\Gamma(z + \frac{1}{2})$. The bounds are almost the same as those for Stirling's approximation, the only difference being that a factor $\eta_k = 1/(1 - 2^{1-2k})$ multiplies some of the bounds (see Theorems 3.2 and 3.5 and Corollary 3.3 in Section 3).

[1] We have taken into account a typographical error in Equation (2.3), where B_{2k} should be replaced by $|B_{2k}|$, as previously noted in [9, footnote on page 682].

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

These bounds, in the case where $z = it$ ($t \in \mathbb{R}$), are what is needed to give bounds on the approximation of $\vartheta(t)$. See Theorem 4.5 and Corollaries 4.7 and 4.9 in Section 4 for these bounds. One such result (see (4.9) below) is a bound

$$\eta_k\,(\pi k)^{1/2}\,\widetilde{T}_k(t) + \tfrac{1}{2}e^{-\pi t} \tag{1.4}$$

on the error if the sum in (1.2) is truncated after the kth term $\widetilde{T}_k(t)$.

Perhaps surprisingly, we obtain a smaller bound if an exponentially small term $\frac{1}{2}\arctan(\exp(-\pi t))$ is included in the approximation of $\vartheta(t)$. The term $\frac{1}{2}\exp(-\pi t)$ in (1.4) can then be omitted (see Theorem 4.5 and Corollary 4.7). This is discussed in Sections 4–5. In Section 5 we show that the attainable error, if the terms in the asymptotic series are summed until the smallest term is reached, is of order $\exp(-\pi t)$ if (as usual) the arctan term is omitted from the approximation, but is reduced to $O(\exp(-2\pi t))$ if the arctan term is included. This observation is perhaps implicit in the work of Berry [4, Section 4] and Gabcke [12, Satz 4.2.3], but our presentation makes it explicit[1].

2. Asymptotic approximation of $\ln\Gamma(z)$

Regarding notation: variables $s, z \in \mathbb{C}$; $c, r, t, u, x, y, \varepsilon, \eta, \theta, \psi \in \mathbb{R}$; and $j, k, m, n \in \mathbb{N}^*$ (the positive integers). 'log' denotes the principal branch of the natural logarithm on the cut plane $\mathbb{C}\backslash(-\infty, 0]$. The (closed) right half-plane is $\mathcal{H} := \{z \in \mathbb{C} : \mathrm{Re}(z) \geq 0\}$, and $\mathcal{H}^* := \mathcal{H}\backslash\{0\}$. We define constants η_k for $k \in \mathbb{N}^*$ by $\eta_k := 1/(1 - 2^{1-2k})$.

The proper domain for the log-Gamma function $\ln\Gamma$ is a Riemann surface. However, for our purposes it is sufficient to take the (principal branch of the) log-Gamma function to be an analytic function on the cut-plane $\mathbb{C}\backslash(-\infty, 0]$, such that $\ln\Gamma(x) = \log(\Gamma(x))$ is real for positive real x. In a software implementation of the function $\ln\Gamma(z)$, care has to be taken because $\ln\Gamma(z)$ and $\ln(\Gamma(z))$ may differ by a multiple of $2\pi i$; see Hare [15].

In this section we consider approximation of $\ln\Gamma(z)$ for $z \in \mathbb{C}\backslash(-\infty, 0]$. When computing $\Gamma(z)$ or $\ln\Gamma(z)$, we can use the reflection formula

$$\Gamma(z)\Gamma(-z) = -\frac{\pi}{z\sin(\pi z)}$$

if $\mathrm{Re}(z) < 0$, $z \notin \mathbb{Z}$. Thus, in the following we assume that $\mathrm{Re}(z) \geq 0$.

We recall Stirling's approximation, taking $k - 1$ terms in the asymptotic expansion with a remainder R_k:

$$\ln\Gamma(z) = \left(z - \frac{1}{2}\right)\log z - z + \frac{1}{2}\log(2\pi) + \sum_{j=1}^{k-1} T_j(z) + R_k(z), \tag{2.1}$$

[1] The fact that the error in the Riemann–Siegel approximation to $Z(t)$ is of order $\exp(-\pi t)$ was observed empirically by the author in 1977, when writing the review [6]. A detailed theoretical explanation was later given by Berry [4].

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

where

$$T_j(z) = \frac{B_{2j}}{2j(2j-1)z^{2j-1}} \tag{2.2}$$

and $R_k(z)$ is a 'remainder' or 'error' term that may be written as

$$R_k(z) = \int_0^\infty \frac{B_{2k} - B_{2k}(\{u\})}{2k\,(u+z)^{2k}}\,du. \tag{2.3}$$

Here $\{u\} := u - \lfloor u \rfloor$ denotes the fractional part of u, $B_{2k}(u)$ is a Bernoulli polynomial, and $B_{2k} = B_{2k}(0)$ is a Bernoulli number, so $B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, etc. See Olver [22, Sections 8.1, 8.4] for the definitions and a proof of (2.3).

A different representation of the remainder is often convenient. Using (2.3) and $R_k(z) = T_k(z) + R_{k+1}(z)$, we see that the error after taking k terms (instead of $k-1$) in the sum is[1]

$$R_{k+1}(z) = -\int_0^\infty \frac{B_{2k}(\{u\})}{2k\,(u+z)^{2k}}\,du. \tag{2.4}$$

If z is real and positive, then the asymptotic series (2.1) is strictly enveloping in the sense of Pólya and Szegö [23, Ch. 4], so $R_k(z)$ has the same sign as the first term omitted, which is $T_k(z)$. Also, $R_k(z)$ is smaller in magnitude than this term, that is, $|R_k(z)| < |T_k(z)|$ (in fact this inequality holds whenever $|\arg(z)| \le \pi/4$; see Remark 2.7).

In the case of complex z in the right half-plane, the error $R_k(z)$ may be larger in absolute value than the first omitted term. This case is covered by Theorem 2.1 and Corollary 2.2, which improve on earlier results by Spira [24] and Hare [15, Proposition 4.1].

THEOREM 2.1. *If $z \in \mathcal{H}^*$, $R_k(z)$ is defined by Equation (2.1), and $T_j(z)$ by (2.2), then*

$$|R_{k+1}(z)| \le \frac{\pi^{1/2}\,\Gamma(k+\frac{1}{2})}{\Gamma(k)}\,|T_k(z)| \tag{2.5}$$

and

$$|R_k(z)| \le \left(\frac{\pi^{1/2}\,\Gamma(k+\frac{1}{2})}{\Gamma(k)} + 1\right)|T_k(z)|. \tag{2.6}$$

PROOF. Let $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$. From (2.4),

$$|R_{k+1}(z)| = \left|\int_0^\infty \frac{B_{2k}(\{u\})}{2k(u+z)^{2k}}\,du\right| \le \frac{|B_{2k}|}{2k}\int_0^\infty |u+z|^{-2k}\,du. \tag{2.7}$$

Since $x \ge 0$, inside the integral we have that

$$|u+z|^2 = (u+x)^2 + y^2 \ge u^2 + x^2 + y^2 = u^2 + |z|^2.$$

[1] We have followed Olver's convention. Other authors may include k terms in the sum in (2.1). Thus, their R_k may correspond to our R_{k+1}, and care has to be taken when comparing bounds in the literature. See, for example, Abramowitz and Stegun [1, (6.1.42)].

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Making a change of variables $u \mapsto |z|\tan\psi$ gives

$$\begin{aligned}\int_0^\infty |u+z|^{-2k}\,du &\le \int_0^\infty (u^2+|z|^2)^{-k}\,du\\ &= |z|^{1-2k}\int_0^{\pi/2}\cos^{2k-2}\psi\,d\psi\\ &= \frac{\pi^{1/2}}{2}\,\frac{\Gamma(k-\frac{1}{2})}{\Gamma(k)}\,|z|^{1-2k},\end{aligned}$$

where the closed form for the integral is known as 'Wallis's formula'; see, for example, [1, (6.1.49)]. Thus, inequality (2.5) follows from (2.7).

Inequality (2.6) follows easily from (2.5) and the triangle inequality

$$|R_k(z)| = |T_k(z) + R_{k+1}(z)| \le |T_k(z)| + |R_{k+1}(z)|. \tag{2.8}$$

□

During a computation, we may wish to bound the error term as a multiple of either the last term included in the approximating sum, or the first term omitted. Hence, the following corollary of Theorem 2.1 is useful.

Corollary 2.2. *If $z \in \mathcal{H}^*$ and $R_k(z)$ is defined by Equation (2.1), then*

$$\left|\frac{R_{k+1}(z)}{T_k(z)}\right| < \sqrt{\pi k} \tag{2.9}$$

and

$$\left|\frac{R_k(z)}{T_k(z)}\right| < 1 + \sqrt{\pi k}. \tag{2.10}$$

Proof. From [8, Equation (21)],

$$\ln\Gamma\Big(x+\frac{1}{2}\Big) - \ln\Gamma(x) - \frac{1}{2}\log(x) \sim -\frac{1}{8x} + \cdots,$$

where the asymptotic series on the right is strictly enveloping for positive real x. Thus, we have $\log(\Gamma(x+\frac{1}{2})/\Gamma(x)) < \frac{1}{2}\log x$, which implies that $\Gamma(k+\frac{1}{2})/\Gamma(k) < \sqrt{k}$. Inequality (2.9) now follows from (2.5) of Theorem 2.1 and the definition of $T_k(z)$. Inequality (2.10) follows similarly, from (2.6) of Theorem 2.1, or directly from (2.8). □

Remark 2.3. The device of converting a bound on $R_{k+1}(z)$ into a bound on $R_k(z)$, of the same order in $|z|$, via the triangle inequality (2.8), also applies to the bounds given in Sections 3–4 below. For the sake of brevity we do not always give such bounds explicitly.

In Remarks 2.4–2.8 we comment briefly on some related bounds that may be found in the literature, allowing for different notations. Here and elsewhere, we define $\theta = \theta(z) := \arg z$ (not to be confused with $\vartheta(t)$ of (1.1)).

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Remark 2.4. Spira [24, Equation (4)] obtains a bound of the same form as our (2.5), but larger by a factor of approximately $4\sqrt{k/\pi}$. This is primarily because he uses a rather crude upper bound on the relevant integral instead of using Wallis's formula[1].

Remark 2.5. Hare [15, Proposition 4.1] obtains a bound of the form $c(k)/|\mathrm{Im}(z)|^{2k-1}$, assuming that $\mathrm{Im}(z) \neq 0$, but without the assumption that $\mathrm{Re}(z) \geq 0$. Here $c(k) = 4\pi^{1/2}\Gamma(k + \frac{1}{2})/\Gamma(k) \sim 4\sqrt{\pi k}$. When both bounds are applicable, our bound (2.6) is better by a factor of about $4/|\sin\theta|^{2k-1}$ (for large k). A problem with a bound such as Hare's, involving $|\mathrm{Im}(z)|$ rather than $|z|$, is that the bound can not be reduced by applying the recurrence $\Gamma(z+1) = z\Gamma(z)$.

Remark 2.6. In Behnke and Sommer [3, (18) page 304] we find a bound that (in our notation) is

$$\left|\frac{R_{k+1}(z)}{T_{k+1}(z)}\right| < 1 + \frac{2k+1}{2}\sqrt{\frac{\pi}{k}}, \tag{2.11}$$

valid for $k \geq 1$ and $\mathrm{Re}(z) \geq 0$, $z \neq 0$. It is interesting to note that this pre-dates the bounds of Spira [24] and Hare [15]. To compare with our bounds, make a change of variables $k \mapsto k+1$ in (2.10) to obtain

$$\left|\frac{R_{k+1}(z)}{T_{k+1}(z)}\right| < 1 + \sqrt{\pi(k+1)}. \tag{2.12}$$

Since $k + 1 < (k + \frac{1}{2})^2/k$, our bound (2.12) is always smaller than Behnke and Sommer's bound (2.11), although the ratio tends to 1 as $k \to \infty$. Note that our bound (2.10) gives a valid bound $1 + \sqrt{\pi}$ on $|R_1(z)/T_1(z)|$, whereas (2.11) requires $k \geq 1$ as the right-hand side is undefined if $k = 0$.

Remark 2.7. A bound due to Whittaker and Watson [27, page 252] (see also [1, (6.1.42)]), valid for $\mathrm{Re}(z) > 0$, is

$$|R_k(z)| \leq K(z)\,|T_k(z)|, \tag{2.13}$$

where $K(z) = \sup_{u\geq 0} |z^2/(u^2 + z^2)|$. It is easy to see that $K(z)$ depends only on $\theta(z)$. A geometric argument shows that

$$K(z) = \begin{cases} 1 & \text{if } |\theta| \leq \pi/4, \\ \dfrac{1}{|\sin(2\theta)|} & \text{if } |\theta| \in (\pi/4, \pi/2). \end{cases}$$

Thus, the bound (2.13) is preferable to those mentioned in Remarks 2.4–2.6 (and to our bound (2.10)) if $|\theta| \leq \pi/4$, but it becomes poor as $|\theta|$ approaches $\pi/2$.

Remark 2.8. A bound due to Stieltjes (see Olver [22, (8.4.06)]) is

$$|R_k(z)| \leq |T_k(z)| \sec^{2k}(\theta/2), \tag{2.14}$$

[1] We note that the proof given by Spira [24, top of page 319] is incomplete: he only proves a bound of the form $c(k)/|\mathrm{Im}(z)|^{2k-1}$, not the claimed $c(k)/|z|^{2k-1}$.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

where $|\theta| < \pi$. This differs from our bound (2.6) by a factor of approximately $\sec^{2k}(\theta/2)/\sqrt{\pi k}$. If $\theta \approx \pi/2$ this factor is approximately $2^k/\sqrt{\pi k}$, which is greater than 1 for all $k \geq 1$. Thus, (2.14) is better than our bound only if $|\theta|$ is sufficiently small. However, if $|\theta| \leq \pi/4$ we should prefer the bound (2.13).

It is natural to ask if an upper bound of order $k^{1/2}$ for $|R_{k+1}(z)/T_k(z)|$, as in Corollary 2.2, is the best possible. Certainly, when $|\arg(z)| \leq \pi/4$, or when $|T_k(z)|$ is much larger than $|T_{k+1}(z)|$, the bound is not optimal. However, without imposing conditions on k and/or z, the bounds of Corollary 2.2 are the best possible, up to constant factors. We sketch a proof of this. Let n be a sufficiently large positive integer, and $z = iy$, where $y = n/\pi$. Thus, n is close to the index of the minimal term $|T_j(z)|$. Also, there is no cancellation in the sum $T_1(z) + T_2(z) + \cdots + T_n(z)$, since, using (2.2),

$$i\,T_j(iy) = \frac{i\,(-1)^{j-1}|B_{2j}|}{2j(2j-1)\,(iy)^{2j-1}} = \frac{|B_{2j}|}{2j(2j-1)\,y^{2j-1}}$$

is real and positive. Using Stirling's approximation to estimate $T_j(z)$ and $T_n(z)$, if $j = n - \delta$ and $\delta^2 \leq y$ then

$$\left|\frac{T_j(z)}{T_n(z)}\right| = 1 + O\left(\frac{\delta^2}{y}\right).$$

We can choose a positive integer $\delta = O(y^{1/2})$ so that $1/2 \leq |T_j(z)/T_n(z)| \leq 2$ for $n - \delta \leq j \leq n$. Hence $|T_{n-\delta}(z) + \cdots + T_{n-1}(z)| \geq \delta\,|T_n(z)|/2$. For some k in the interval $[n-\delta, n]$, we must have $|R_{k+1}(z)/T_n(z)| \geq \delta/4$, so $|R_{k+1}(z)/T_k(z)| \geq \delta/8$ is of order $y^{1/2}$.

Numerical evidence confirms this conclusion. Taking $n = 100$, $y = n/\pi$, and $k = 90$, we find that $|R_{k+1}(iy)/T_k(iy)| \approx 4.62$. If $n = 400$, $y = n/\pi$, $k = 383$, then $|R_{k+1}(iy)/T_k(iy)| \approx 10.15$. Thus, it appears that the constant $\sqrt{\pi}$ appearing in Corollary 2.2 cannot be reduced by a factor greater than 4 when z lies on, or sufficiently close to[1], the imaginary axis.

In Theorem 2.10, we obtain bounds that are better than the bounds given in Theorem 2.1 and Corollary 2.2, provided the condition $k \leq |z|$ is satisfied. If $|z|$ is too small, we can apply the recurrence $\ln\Gamma(z) = \ln\Gamma(z+1) - \log z$ as often as necessary and then apply Theorem 2.10.

Before stating Theorem 2.10, we define some constants c_k which enter into the proof of the theorem. Assuming that $T_k(z)$ is defined by (2.2), let

$$c_k := \sum_{j=1}^{2k}\left|\frac{T_{k+j}(k)}{T_k(k)}\right| + \sqrt{3k\pi}\left|\frac{T_{3k}(k)}{T_k(k)}\right|.$$

The following lemma is the reason for introducing the constants c_k.

[1] The proof that we have outlined can be modified to cover a region of the form $\mathrm{Re}(z) \geq 0$, $|\mathrm{Im}(z)| \geq c\mathrm{Re}(z)^2$, where c is a sufficiently large positive constant. On the other hand, by Whittaker and Watson's bound (2.13), it cannot be extended into the sector $|\theta| < \pi/2 - \varepsilon$ ($|z|$ sufficiently large), since in that region $|R_k(z)/T_k(z)|$ and $|R_{k+1}(z)/T_k(z)|$ are $O(1/\varepsilon)$.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Table 1. The constants c_k (rounded up to 6 decimal places).

k	c_k	k	c_k	k	c_k
1	0.072 096	6	0.107 384	15	0.110 498
2	0.103 961	7	0.108 089	20	0.111 050
3	0.104 294	8	0.108 634	25	0.111 384
4	0.105 304	9	0.109 067	30	0.111 609
5	0.106 460	10	0.109 419	50	0.112 060

Lemma 2.9. *If $z \in \mathcal{H}^*$, $R_k(z)$ is defined by Equation (2.1), and $k \le |z|$, then*

$$\left|\frac{R_{k+1}(z)}{T_k(z)}\right| \le c_k\,(k/|z|)^2. \tag{2.15}$$

Proof. For all $m \in \mathbb{N}$,

$$R_{k+1}(z) = \sum_{j=1}^{m} T_{k+j}(z) + R_{k+m+1}(z). \tag{2.16}$$

Now

$$|R_{k+m+1}(z)| \le \sqrt{(k+m)\pi}\,|T_{k+m}(z)|,$$

by Corollary 2.2 with k replaced by $k+m$. Taking norms in (2.16), choosing $m = 2k$, and dividing both sides by $|T_{k+1}(z)|$,

$$\left|\frac{R_{k+1}(z)}{T_{k+1}(z)}\right| \le \frac{1}{|T_{k+1}(z)|}\left(\sum_{j=1}^{2k} |T_{k+j}(z)| + \sqrt{3k\pi}\,|T_{3k}(z)|\right).$$

Since $|T_{k+j}(z)/T_{k+1}(z)|$ has the form $c/|z|^{2j-2}$, it is a nonincreasing function of $|z|$ (assuming $j \ge 1$), so its maximum occurs when $|z|$ is minimal, that is, when $|z| = k$. Thus

$$\left|\frac{R_{k+1}(z)}{T_{k+1}(z)}\right| \le \frac{1}{|T_{k+1}(k)|}\left(\sum_{j=1}^{2k} |T_{k+j}(k)| + \sqrt{3k\pi}\,|T_{3k}(k)|\right) = c_k\left|\frac{T_k(k)}{T_{k+1}(k)}\right|.$$

Since $T_{k+1}(z)/T_k(z)$ has the form c/z^2,

$$\left|\frac{T_{k+1}(z)}{T_k(z)}\right| = (k/|z|)^2\left|\frac{T_{k+1}(k)}{T_k(k)}\right|,$$

and (2.15) follows. □

Numerical values of c_k for various $k \le 50$ are given in Table 1. The c_k appear to increase monotonically to the limit $1/(\pi^2 - 1) \approx 0.112\,745$. We have verified monotonicity, and that $c_k < 1/(\pi^2 - 1)$, for $k \le 100$.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

THEOREM 2.10. *If $z \in \mathcal{H}^*$, $R_k(z)$ is defined by Equation (2.1), and $k \le |z|$, then*

$$\left|\frac{R_{k+1}(z)}{T_k(z)}\right| < \frac{(k/|z|)^2}{\pi^2 - 1} \le \frac{1}{\pi^2 - 1} < 0.113 \tag{2.17}$$

and

$$\left|\frac{R_k(z)}{T_k(z)}\right| < 1 + \frac{(k/|z|)^2}{\pi^2 - 1} \le \frac{\pi^2}{\pi^2 - 1} < 1.113. \tag{2.18}$$

PROOF. Let

$$\mu := \left(\frac{k}{\pi|z|}\right)^2 \le \frac{1}{\pi^2}$$

and $m := \lfloor k^{1/2} \rfloor$. For brevity, we write R_k for $R_k(z)$ and T_k for $T_k(z)$. Since $R_{k+1} = T_{k+1} + T_{k+2} + \cdots + T_{k+m} + R_{k+m+1}$, we have $|R_{k+1}/T_k| \le S + E$, where

$$S := \sum_{j=1}^{m} \left|\frac{T_{k+j}}{T_k}\right| \quad \text{and} \quad E := \left|\frac{R_{k+m+1}}{T_k}\right|.$$

Since $|B_{2k}| = 2(2k)!\,\zeta(2k)/(2\pi)^{2k}$,

$$\left|\frac{T_{k+j}}{T_k}\right| \le \frac{(2k + 2j - 2)!}{(2k - 2)!}\,|2\pi z|^{-2j}.$$

Using the assumption $k \le |z|$, it follows that

$$\left|\frac{T_{k+j}}{T_k}\right| \le \mu^j \prod_{n=1}^{2j} \left(1 + \frac{n-2}{2k}\right). \tag{2.19}$$

Now $1 + x \le \exp(x)$ for all $x \in \mathbb{R}$, so

$$\left|\frac{T_{k+j}}{T_k}\right| \le \mu^j \prod_{n=1}^{2j} \exp\left(\frac{n-2}{2k}\right) = \mu^j \exp\left(\frac{(2j-3)j}{2k}\right).$$

By convexity, $1 \le \exp(x) \le 1 + (e - 1)x$ for all $x \in [0, 1]$. It follows that, for $2 \le j \le m$,

$$\left|\frac{T_{k+j}}{T_k}\right| \le \mu^j \left(1 + (e-1)\frac{(2j-3)j}{2k}\right). \tag{2.20}$$

Also, for the special case $j = 1$, inequality (2.19) gives

$$\left|\frac{T_{k+1}}{T_k}\right| \le \mu\left(1 - \frac{1}{2k}\right). \tag{2.21}$$

From (2.20) to (2.21),

$$\begin{aligned} S &\le -\frac{\mu}{2k} + \sum_{j=1}^{m} \mu^j + \frac{e-1}{2k} \sum_{j=2}^{m} (2j-3)j\mu^j \\ &< -\frac{\mu}{2k} + \sum_{j=1}^{\infty} \mu^j + \frac{e-1}{2k} \sum_{j=2}^{\infty} (2j-3)j\mu^j \\ &= -\frac{\mu}{2k} + \frac{\mu}{1-\mu} + \left(\frac{e-1}{2k}\right)\frac{\mu^2(2 + 3\mu - \mu^2)}{(1-\mu)^3}. \end{aligned}$$

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Thus

$$\frac{\mu}{1-\mu} - S > \frac{\mu}{2k}\Big[1 - \frac{(e-1)\mu(2+3\mu-\mu^2)}{(1-\mu)^3}\Big].$$

Since $\mu(2+3\mu-\mu^2)/(1-\mu)^3 = \sum_{j=2}^{\infty}(2j-3)j\mu^{j-1}$ is monotonic increasing on $[0, 1/\pi^2]$, the factor in square brackets attains its minimum on $[0, 1/\pi^2]$ at $\mu = 1/\pi^2$, and a numerical computation shows that the minimum is greater than $\pi^2/22$. Thus,

$$\frac{\mu}{1-\mu} - S > \frac{\pi^2\mu}{44k}.$$

Now consider E. We have

$$E = \left|\frac{R_{k+m+1}}{T_k}\right| = \left|\frac{T_{k+m}}{T_k}\right| \cdot \left|\frac{R_{k+m+1}}{T_{k+m}}\right|.$$

The first factor on the right is at most $\mu^m e$, by (2.20) with $j = m$; the second factor is at most $\sqrt{\pi(k+m)}$, by an application of Corollary 2.2 with k replaced by $k+m$. This gives

$$E \le \mu^m e\sqrt{\pi(k+m)} \le \mu^{\sqrt{k}-1} e\sqrt{2\pi k}.$$

Thus $kE/\mu \le \mu^{\sqrt{k}-2} e\sqrt{2\pi k^3} \ll 1/k$, so there exists k_0 such that, for all $k \ge k_0$, $kE/\mu < \pi^2/44$, so $E < \pi^2\mu/(44k)$ and $\mu/(1-\mu) > S + E$. A computation shows that we can take $k_0 = 34$. Thus, for all $k \ge k_0$,

$$\left|\frac{R_{k+1}}{T_k}\right| < \frac{\mu}{1-\mu} = \frac{k^2}{\pi^2|z|^2 - k^2} \le \frac{(k/|z|)^2}{\pi^2 - 1}.$$

This proves the desired inequality (2.17) for $k \ge k_0$.

By a straightforward numerical computation, we can verify that (2.17) also holds for $1 \le k \le 33$ (see Lemma 2.9 and Table 1). This concludes the proof of (2.17). Finally, (2.18) follows from (2.17) and the triangle inequality. □

Remark 2.11. It is reasonable to conjecture the slightly stronger inequalities

$$\left|\frac{R_{k+1}(z)}{T_k(z)}\right| < \frac{k^2}{\pi^2|z|^2 - k^2}, \qquad \left|\frac{R_k(z)}{T_k(z)}\right| < \frac{\pi^2|z|^2}{\pi^2|z|^2 - k^2}, \tag{2.22}$$

for all (k, z) such that $|z| \ge k \ge 1$. This has been verified numerically, and the proof of Theorem 2.10 shows that (2.22) holds for $k \ge 34$. However, our proof of (2.17) for $k \le 33$, using Lemma 2.9 and the constants c_k, is insufficient to prove (2.22). Hence, we leave (2.22) as a conjecture.

3. Asymptotic approximation of $\ln\Gamma(z + \frac{1}{2})$

In this section we deduce, from the results of Section 2, an asymptotic series for $\ln\Gamma(z + \frac{1}{2})$ in descending odd powers of z. The series was given by Gauss [13, Art. 29]; by using the results of Section 2 we obtain new error bounds for $z \in \mathcal{H}^*$.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Replacing z by $2z$ in (2.1) and then subtracting (2.1) gives

$$\ln\Gamma(2z) - \ln\Gamma(z) = z\log z + (2\log 2 - 1)z - \frac{1}{2}\log 2 + \sum_{j=1}^{k-1}\widehat{T}_j(z) + \widehat{R}_k(z), \tag{3.1}$$

where $\widehat{T}_j(z) = T_j(2z) - T_j(z)$ and $\widehat{R}_k(z) = R_k(2z) - R_k(z)$. More explicitly, using [22, (8.1.12)] for $B_{2j}(\frac{1}{2})$,

$$\widehat{T}_j(z) = -(1 - 2^{1-2j})T_j(z) = -\frac{(1 - 2^{1-2j})B_{2j}}{2j(2j-1)z^{2j-1}} = \frac{B_{2j}(\frac{1}{2})}{2j(2j-1)z^{2j-1}}. \tag{3.2}$$

Also, $\widehat{R}_k(z) = \widehat{T}_k(z) + \widehat{R}_{k+1}(z)$, where

$$\widehat{R}_{k+1}(z) = -\int_0^\infty \frac{2^{1-2k}B_{2k}(\{2u\}) - B_{2k}(\{u\})}{2k(u+z)^{2k}}\,du. \tag{3.3}$$

Using the duplication formula $\Gamma(z+\frac{1}{2}) = 2^{1-2z}\pi^{1/2}\Gamma(2z)/\Gamma(z)$, Equation (3.1) immediately gives Gauss's asymptotic expansion of $\ln\Gamma(z+\frac{1}{2})$:

$$\ln\Gamma\Big(z+\frac{1}{2}\Big) = z\log z - z + \frac{1}{2}\log(2\pi) + \sum_{j=1}^{k-1}\widehat{T}_j(z) + \widehat{R}_k(z). \tag{3.4}$$

The following lemma enables us to simplify the 'kernel' function appearing in the integral (3.3).

Lemma 3.1. *For $k \ge 1$ and all real u,*

$$2^{1-2k}B_{2k}(\{2u\}) - B_{2k}(\{u\}) = B_{2k}(\{u+\tfrac{1}{2}\}).$$

Proof. This follows from the known identities [1, (23.1.8) and (23.1.10)]

$$B_{2k}(u) = B_{2k}(1-u)$$

and

$$2^{1-2k}B_{2k}(2u) - B_{2k}(u) = B_{2k}(u+\tfrac{1}{2}). \qquad \square$$

Using Lemma 3.1, we see from (3.3) that

$$\widehat{R}_{k+1}(z) = -\int_0^\infty \frac{B_{2k}(\{u+\frac{1}{2}\})}{2k(u+z)^{2k}}\,du. \tag{3.5}$$

We can now prove an analogue of Theorem 2.1. The upper bound on $|\widehat{R}_k(z)|$ is the same as the bound that we obtained for $|R_k(z)|$, but the bound on $|\widehat{R}_k(z)/\widehat{T}_k(z)|$ is larger than the bound on $|R_k(z)/T_k(z)|$ by a factor $\eta_k = 1/(1-2^{1-2k}) \le 2$.

Theorem 3.2. *If $z \in \mathcal{H}^*$ and $\widehat{R}_k(z)$ is defined by Equation (3.4), then*

$$\left|\frac{\widehat{R}_{k+1}(z)}{\widehat{T}_k(z)}\right| \le \eta_k \frac{\pi^{1/2}\Gamma(k+\frac{1}{2})}{\Gamma(k)}.$$

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Proof. This is almost identical to the proof of Theorem 2.1, the only difference being that we use (3.5) to bound $\widehat{R}_{k+1}(z)$ instead of (2.4) to bound $R_{k+1}(z)$. This increases the bound by a factor $\eta_k = |T_k(z)/\widehat{T}_k(z)|$. □

Corollary 3.3. *Under the conditions of Theorem 3.2,*

$$\left|\frac{\widehat{R}_{k+1}(z)}{\widehat{T}_k(z)}\right| < \eta_k \sqrt{\pi k}. \tag{3.6}$$

Remark 3.4. The factor η_k in (3.6) can be omitted if $k \geq 3$ or $|z| \geq 1$. A proof is given in an earlier version of this paper (arXiv:1609.03682v1, proof of Corollary 3).

Theorem 3.5. *If $z \in \mathcal{H}^*$, $\widehat{R}_k(z)$ is defined by Equation (3.4), and $k \leq |z|$, then*

$$\left|\frac{\widehat{R}_{k+1}(z)}{\widehat{T}_k(z)}\right| < \eta_k \frac{(k/|z|)^2}{\pi^2 - 1}$$

and

$$\left|\frac{\widehat{R}_k(z)}{\widehat{T}_k(z)}\right| < 1 + \eta_k \frac{(k/|z|)^2}{\pi^2 - 1}.$$

Proof. This is the same as the proof of Theorem 2.10, except that we have to allow for the additional factor η_k that arises because the errors are normalised by $\widehat{T}_k(z)$ instead of by $T_k(z)$. □

Remark 3.6. By a small modification of Lemma 2.9, if $k \leq |z|$ then

$$|\widetilde{R}_{k+1}(z)/\widetilde{T}_k(z)| \leq \eta_k c_k (k/|z|)^2.$$

4. The Riemann–Siegel theta function

In this section we consider the Riemann–Siegel theta function $\vartheta(t)$ defined by (1.1). Lemma 4.1 gives an equivalent expression for $\vartheta(t)$ that is better for our purposes than the definition.

Lemma 4.1. *For all $t \in \mathbb{R}$,*

$$\vartheta(t) = \frac{1}{2}\arg\Gamma\Big(it + \frac{1}{2}\Big) - \frac{1}{2}t\log(2\pi) - \frac{\pi}{8} + \frac{1}{2}\arctan(e^{-\pi t}).$$

Proof. The reflection formula $\Gamma(s)\Gamma(1-s) = \pi/\sin(\pi s)$ with $s = (it/2) + \frac{1}{4}$ gives

$$\Gamma\Big(\frac{it}{2} + \frac{1}{4}\Big)\Gamma\Big(-\frac{it}{2} + \frac{3}{4}\Big) = \frac{\pi}{\sin\pi(\frac{it}{2} + \frac{1}{4})}, \tag{4.1}$$

and the duplication formula $\Gamma(s)\Gamma(s + \frac{1}{2}) = 2^{1-2s}\pi^{1/2}\Gamma(2s)$ gives

$$\Gamma\Big(\frac{it}{2} + \frac{1}{4}\Big)\Gamma\Big(\frac{it}{2} + \frac{3}{4}\Big) = 2^{1/2-it}\pi^{1/2}\Gamma\Big(it + \frac{1}{2}\Big). \tag{4.2}$$

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Multiplying (4.1) and (4.2) gives

$$\Gamma\Big(\frac{it}{2}+\frac{1}{4}\Big)^2\Big|\Gamma\Big(\frac{it}{2}+\frac{3}{4}\Big)\Big|^2 = \frac{2^{1/2-it}\pi^{3/2}\Gamma(it+\frac{1}{2})}{\sin\pi(\frac{it}{2}+\frac{1}{4})}.$$

Taking the argument of each side and simplifying, using the fact that

$$\arctan\Big(\frac{1-e^{-\pi t}}{1+e^{-\pi t}}\Big) = \frac{\pi}{4} - \arctan(e^{-\pi t}),$$

proves the lemma. □

Using the representation of $\vartheta(t)$ given in Lemma 4.1, and the results of Section 3, we obtain an asymptotic approximation of $\vartheta(t)$ together with error bounds. This is summarised in Theorems 4.2 and 4.5. As far as we are aware, this is the first time that a rigorous error bound applicable for all $k \geq 1$ and all real $t > 0$ has been given. Most authors seem to restrict themselves to small k and sufficiently large t. For example, Edwards [11, (2) in Section 6.5] takes $k = 2$ and t 'large'; Gabcke [12, Satz 4.2.3(d)] takes $k = 4$ and $t \geq 10$.

THEOREM 4.2. *For all real $t > 0$,*

$$\vartheta(t) = \frac{t}{2}\log\Big(\frac{t}{2\pi e}\Big) - \frac{\pi}{8} + \frac{\arctan(e^{-\pi t})}{2} + \sum_{j=1}^{k-1}\widetilde{T}_j(t) + \widetilde{R}_k(t), \tag{4.3}$$

where

$$\widetilde{T}_j(t) := \frac{1}{2}|\widehat{T}_j(t)| = \frac{|B_{2j}(\frac{1}{2})|}{4j(2j-1)t^{2j-1}}$$

and

$$\widetilde{R}_k(t) := \mathrm{Im}(\tfrac{1}{2}\widehat{R}_k(it)). \tag{4.4}$$

PROOF. From Lemma 4.1,

$$2\vartheta(t) = \mathrm{Im}(\ln\Gamma(it+\tfrac{1}{2})) - t\log(2\pi) - \pi/4 + \arctan(e^{-\pi t}).$$

Using (3.4) with $z = it$ for the $\ln\Gamma(it+\frac{1}{2})$ term, we obtain

$$\begin{aligned}2\vartheta(t) = \mathrm{Im}\Big(it\log(it) - it + \sum_{j=1}^{k-1}\widehat{T}_j(it) + \widehat{R}_k(it)\Big) - t\log(2\pi)\\ - \pi/4 + \arctan(e^{-\pi t})\end{aligned}$$

Since $B_{2j} = (-1)^{j-1}|B_{2j}|$ and $B_{2j}(\frac{1}{2}) = -(1-2^{1-2j})B_{2j}$, we see from (3.2) that $\mathrm{Im}(\widehat{T}_j(it)) = |\widehat{T}_j(t)|$. Also, $\mathrm{Im}(it\log i) = \mathrm{Im}(it\cdot i\pi/2) = 0$. Thus,

$$\begin{aligned}2\vartheta(t) &= t\log t - t + \sum_{j=1}^{k-1}|\widehat{T}_j(t)| + \mathrm{Im}(\widehat{R}_k(it)) - t\log(2\pi) - \pi/4 + \arctan(e^{-\pi t})\\ &= t\log\Big(\frac{t}{2\pi e}\Big) - \frac{\pi}{4} + \arctan(e^{-\pi t}) + 2\sum_{j=1}^{k-1}\widetilde{T}_j(t) + 2\widetilde{R}_k(t).\end{aligned}$$

Thus, the result (4.3) follows. □

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Remark 4.3. The first few terms of the asymptotic expansion for $\vartheta(t)$ are derived in a different manner by Edwards [11, Section 6.5]; his method does not easily lead to an expression for the general term or to an error bound valid for all k.

Lemma 4.4. *For all real $t > 0$,*

$$\widetilde{R}_1(t) = \mathrm{Im}\biggl(\int_0^\infty \frac{B_2(\frac{1}{2}) - B_2(\{u + \frac{1}{2}\})}{4(u + it)^2}\, du\biggr) \tag{4.5}$$

and

$$\widetilde{R}_{k+1}(t) = \mathrm{Im}\biggl(-\int_0^\infty \frac{B_{2k}(\{u + \frac{1}{2}\})}{4k(u + it)^{2k}}\, du\biggr). \tag{4.6}$$

Proof. Equation (4.6) follows from (3.5) and the definition (4.4) of $\widetilde{R}_k(t)$. For (4.5) we use $\widetilde{R}_1(t) = \widetilde{T}_1(t) + \widetilde{R}_2(t)$, where $\widetilde{R}_2(t)$ is given by (4.6) with $k = 1$. □

Theorem 4.5. *If t and $\widetilde{R}_k(t)$ are as in Theorem 4.2, then*

$$|\widetilde{R}_{k+1}(t)| \le \frac{\pi^{1/2}\,\Gamma(k - \frac{1}{2})\,|B_{2k}|}{8\,k!\,t^{2k-1}}. \tag{4.7}$$

Proof. We use Theorem 3.2 and (3.2) to bound $\widetilde{R}_{k+1}(t) = \frac{1}{2}\mathrm{Im}(\widehat{R}_{k+1}(it))$. Note that the η_k factor in Theorem 3.2 cancels a factor in (3.2). □

Remark 4.6. From (3.4), using the fact that $\mathrm{Re}(\widehat{T}_j(it)) = 0$,

$$\begin{aligned}
\mathrm{Re}(\widehat{R}_k(it)) &= \mathrm{Re}\Bigl(\ln\Gamma\Bigl(it + \frac{1}{2}\Bigr) - it\log(it) + it - \frac{1}{2}\log(2\pi)\Bigr)\\
&= \log\Bigl|\Gamma\Bigl(it + \frac{1}{2}\Bigr)\Bigr| + \frac{\pi t}{2} - \frac{1}{2}\log(2\pi)\\
&= \frac{1}{2}\log\Bigl(\frac{\pi}{\cosh \pi t}\Bigr) + \frac{\pi t}{2} - \frac{1}{2}\log(2\pi) \ \ \text{(using [1, (6.1.30)])}\\
&= -\frac{1}{2}\log(1 + e^{-2\pi t}) = -\frac{1}{2}e^{-2\pi t} + O(e^{-4\pi t}),
\end{aligned}$$

so $\mathrm{Re}(\widehat{R}_k(it))$ is exponentially small, but nonzero. Thus $|\widetilde{R}_k(t)| < \frac{1}{2}|\widehat{R}_k(it)|$, and it follows that inequality (4.7) is strict.

Corollary 4.7. *If $t > 0$ then*

$$\left|\frac{\widetilde{R}_{k+1}(t)}{\widetilde{T}_k(t)}\right| < \eta_k\sqrt{\pi k}. \tag{4.8}$$

Proof. This follows from Corollary 3.3 with $z = it$. □

Remark 4.8. The factor η_k in Corollary 4.7 can be omitted if $k \ge 3$ or $t \ge 1$ (see Remark 3.4).

Corollary 4.9. *If $t \ge k > 0$, then*

$$\left|\frac{\widetilde{R}_{k+1}(t)}{\widetilde{T}_k(t)}\right| < \eta_k\frac{(k/t)^2}{\pi^2 - 1}.$$

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Proof. This follows from Theorem 3.5 with $z = it$. □

Remark 4.10. The factor η_k in Corollary 4.9 can be omitted if $k \geq 3$. This follows for sufficiently large k from a slight modification of the proof of Theorem 3.5, and for small k from the observation that $\eta_k c_k < 1/(\pi^2 - 1)$ for $k \geq 3$ (see Remark 3.6 and Table 1). If $1 \leq k \leq 2$ we can use the bound $\eta_k c_k (k/t)^2$ that follows from Remark 3.6.

In the literature, the asymptotic approximation (4.3) always seems to be stated without the exponentially small arctan term. See, for example, Edwards [11, (1) page 120], Gabcke [12, Satz 4.2.3(c)] and Lehmer [17, (5) page 104]. The arctan term appears in some related formulas, such as Gram [14, (7) page 300] and Gabcke [12, Satz 4.2.3(a)]. See also the discussion in Berry [4, Section 4].

It is valid to omit the arctan term if all we want is an asymptotic series in the sense of Poincaré (see Olver [22, Section 1.7.3]). However, it is not desirable if we want to minimise the error in the approximation. If we omit the arctan term, then the upper bounds on $|\widetilde{R}_k(t)|$ have to be increased accordingly. Since $\arctan(e^{-\pi t}) < e^{-\pi t}$ for $t \geq 0$, it is sufficient to add $\frac{1}{2}e^{-\pi t}$ to the bound on $|\widetilde{R}_{k+1}(t)|$ in (4.7). The bound of Corollary 4.7 can be replaced by

$$|\widetilde{R}_{k+1}(t)| < \eta_k \sqrt{\pi k}\, \widetilde{T}_k(t) + \tfrac{1}{2}e^{-\pi t}. \tag{4.9}$$

Of course, $\frac{1}{2}e^{-\pi t}$ is negligible if t is large, for example when searching for high zeros of $\zeta(s)$ on the critical line. When t is not so large, the arctan term may be significant. We discuss this in the next section.

Remark 4.11. Other situations where an exponentially small contribution is significant are mentioned by Watson [26, Sections 7.22–7.23], in connection with the Stokes phenomenon [18, 20] and the asymptotic expansions of the Bessel functions $J_\nu(z)$ and $I_\nu(z)$. An example that is similar to ours, but somewhat simpler, was given by Olver [21] and is discussed by Meyer [18, Appendix].

5. Attainable accuracy

In this section we consider the accuracy of the asymptotic expansion of $\vartheta(t)$ if t is fixed and we choose (close to) the optimal number of terms to sum.

Assume that t is fixed and positive. The terms $\widetilde{T}_k(t)$ initially decrease (unless $t \leq \sqrt{7/120} \approx 0.2415$), but eventually increase in value, so it is of interest to determine the index of a minimal term. Define

$$k_{\min} = k_{\min}(t) := \min\{k \geq 1 : \widetilde{T}_k(t) \leq \widetilde{T}_{k+1}(t)\}$$

and

$$\widetilde{T}_{\min}(t) := \widetilde{T}_{k_{\min}}(t).$$

Lemma 5.1 shows that, for all $t > 0$, the sequence of terms $(\widetilde{T}_k(t))_{k\geq 1}$ is unimodal, and that $\widetilde{T}_{\min}(t)$ is a minimal term.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

LEMMA 5.1. *Fix $t > 0$. Then*

(1) *for* $1 \le k < k_{\min}(t)$, $\widetilde{T}_k(t) > \widetilde{T}_{k+1}(t) > 0$;
(2) *for* $k = k_{\min}(t)$, $0 < \widetilde{T}_k(t) \le \widetilde{T}_{k+1}(t)$;
(3) *for* $k > k_{\min}(t)$, $0 < \widetilde{T}_k(t) < \widetilde{T}_{k+1}(t)$;
(4) $\widetilde{T}_{\min}(t) = \min_{k \ge 1} \widetilde{T}_k(t)$.

PROOF. We sketch the proof. Observe that, for all $k \in \mathbb{N}^*$,

$$R(k) := \frac{\widetilde{T}_{k+1}(t)/\widetilde{T}_{k+2}(t)}{\widetilde{T}_k(t)/\widetilde{T}_{k+1}(t)}$$

is independent of t, and can be shown to lie in the interval $(0, 1)$. (This is clear for large k, since

$$R(k) = \frac{k(2k-1)}{(k+1)(2k+1)}(1 + O(4^{-k})),$$

and can be verified by a numerical computation for small k.) Thus

$$\frac{\widetilde{T}_{k+1}(t)}{\widetilde{T}_{k+2}(t)} < \frac{\widetilde{T}_k(t)}{\widetilde{T}_{k+1}(t)}.$$

Inequalities (1)–(3) of the lemma now follow easily, and equality (4) follows from (1)–(3). □

LEMMA 5.2. *For large positive $t \in \mathbb{R}$,*

$$k_{\min}(t) = \pi t + O(1)$$

and, if $k = \pi t + O(1)$, then

$$\widetilde{T}_k(t) = \frac{e^{-2\pi t}}{2\pi\sqrt{t}}\left(1 + O\left(\frac{1}{t}\right)\right).$$

PROOF. We sketch the proof. Using $|B_{2k}| = 2(2k)!\,\zeta(2k)/(2\pi)^{2k}$,

$$\frac{\widetilde{T}_k(t)}{\widetilde{T}_{k+1}(t)} = \frac{2k(2k-1)}{4\pi^2 t^2}(1 + O(4^{-k})). \tag{5.1}$$

Thus, $k_{\min} = \pi t + O(1)$, where the $O(1)$ term covers the $1 + O(4^{-k})$ factor and the effect of rounding to the nearest integer.

The estimate of $\widetilde{T}_k(t)$ follows from Stirling's approximation. Write $k = \pi t/(1 + \varepsilon)$, so $\varepsilon = O(1/t)$. Then

$$\begin{aligned}
\widetilde{T}_k(t) &= \frac{(1 - 2^{1-2k})\,\zeta(2k)\,(2k)!}{2k(2k-1)\,(2\pi)^{2k}\,t^{2k-1}} \\
&= \frac{t}{4k^2}\left(\frac{2k}{e}\right)^{2k}\frac{\sqrt{4k\pi}}{(k(1+\varepsilon))^{2k}}\,(1 + O(\varepsilon)) \\
&= \frac{e^{-2k-2k\varepsilon}}{2\pi\sqrt{t}}\,(1 + O(\varepsilon)) \\
&= \frac{e^{-2\pi t}}{2\pi\sqrt{t}}\,(1 + O(\varepsilon))
\end{aligned}$$

which concludes the proof. □

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

Remark 5.3. If we minimise $(\pi k)^{1/2}\,\widetilde{T}_k(t)$ instead of $\widetilde{T}_k(t)$, the minimum is still at $k = \pi t + O(1)$. The difference between the indices of the two minima can be subsumed by the $O(1)$ term.

Corollary 5.4. *If $k = \pi t + O(1)$, then $|\widetilde{R}_{k+1}(t)| < \frac{1}{2}e^{-2\pi t}(1 + O(1/t))$.*

Proof. This follows from (4.8) and the second half of Lemma 5.2. □

From Lemma 5.2 and Corollary 5.4, we can guarantee an error that does not exceed $\frac{1}{2}e^{-2\pi t}(1 + O(1/t))$ by taking $k_{\min}(t) = \pi t + O(1)$ terms in the approximation

$$\vartheta(t) \approx \frac{t}{2}\log\Big(\frac{t}{2\pi e}\Big) - \frac{\pi}{8} + \frac{\arctan(e^{-\pi t})}{2} + \sum_{j=1}^{k_{\min}(t)} \widetilde{T}_j(t). \tag{5.2}$$

On the other hand, if we use the 'standard' approximation

$$\vartheta(t) \approx \frac{t}{2}\log\Big(\frac{t}{2\pi e}\Big) - \frac{\pi}{8} + \sum_{j=1}^{k_{\min}(t)} \widetilde{T}_j(t), \tag{5.3}$$

we can only guarantee an error not exceeding $\frac{1}{2}e^{-\pi t} + O(e^{-2\pi t})$. Thus, the arctan term is numerically significant, even though it is asymptotically smaller than any term $\widetilde{T}_j(t)$. This is illustrated by Table 2, where we give, for various $t \in [1, 100]$, $k_{\min}(t)$ and

A : the error in the standard approximation (5.3) after taking $k_{\min}(t)$ terms, normalised by the smallest term $\widetilde{T}_{\min}(t) \approx e^{-2\pi t}/(2\pi t^{1/2})$;

B : the error bound (4.8) (this is already normalised);

C : the error in the approximation (5.2), normalised by the smallest term, that is, $\widetilde{R}_{k+1}(t)/\widetilde{T}_k(t)$ for $k = k_{\min}(t)$;

D : the error in the empirically improved approximation

$$\begin{aligned}\vartheta(t) \approx{}& \frac{t}{2}\log\Big(\frac{t}{2\pi e}\Big) - \frac{\pi}{8} + \frac{\arctan(e^{-\pi t})}{2}\\ &+ \sum_{j=1}^{k_{\min}(t)} \widetilde{T}_j(t) + \Big(\pi t - k_{\min}(t) + \frac{1}{12}\Big)\widetilde{T}_{\min}(t),\end{aligned} \tag{5.4}$$

normalised by $\widetilde{T}_{\min}(t)$, as for columns A and C.

It can be seen that $k_{\min}(t)$ is usually $\lfloor \pi t + 5/4 \rfloor$. This is as expected from (5.1). The normalised value A is approximately $\pi t^{1/2}\exp(\pi t)$, which is large because $\widetilde{T}_{\min}(t)$, given by Lemma 5.2, is much smaller than the error, which is about $\frac{1}{2}\exp(-\pi t)$.

Column B gives upper bounds on the absolute values of the entries in column C; it is clear that the upper bounds are conservative (although necessarily so, by the discussion near the end of Section 2).

It can be observed that the entries in column C are negative. This suggests that we would be better off truncating the sum after $k_{\min} - 1$ terms instead of $k_{\min}$ terms (which

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

TABLE 2. Normalised errors—see text for A, B, C, D.

t	$k_{\min}$	A	B	C	D
1	4	7.2×10^{1}	3.57	−0.79	-1.1×10^{-2}
2	7	2.4×10^{3}	4.69	−0.63	$+2.4 \times 10^{-4}$
5	16	4.6×10^{7}	7.09	−0.21	$+2.8 \times 10^{-3}$
10	32	4.4×10^{14}	10.0	−0.50	$+8.3 \times 10^{-4}$
20	64	2.7×10^{28}	14.2	−1.08	$+8.3 \times 10^{-5}$
50	158	3.7×10^{69}	22.3	−0.84	-1.5×10^{-4}
100	315	8.6×10^{137}	31.5	−0.76	-5.2×10^{-5}

would have the effect of adding 1 to the entries in column C). However, a much better approximation is obtained by adding a 'correction term'

$$(\pi t - k_{\min}(t) + \tfrac{1}{12})\widetilde{T}_{\min}(t)$$

as in (5.4). The motivation for the correction term is to smooth out the sawtooth nature of approximation C, which has jumps at the values of t where $k_{\min}(t)$ changes. This explains the addition of $(\pi t - k_{\min}(t) + c)\,\widetilde{T}_{\min}(t)$, where c is an arbitrary constant. Column D gives numerical evidence for a constant close to $\frac{1}{12}$. We do not have a theoretical explanation for the value of this constant, although it is clearly related to the asymptotic location of the positive zero(s) of the function $\widetilde{R}_{k+1}(t)$ given by (4.6). It may be relevant that, for large k, $B_{2k}(u + \frac{1}{2})$ behaves like a scaled version of $\cos(2\pi u)$: see Dilcher [10, Theorem 1].

Acknowledgements

It is a pleasure to acknowledge discussions with Jon Borwein on the subject of this paper in the year preceding his untimely death. Jon had a great interest in algorithms for the accurate and reliable computation of special functions; see, for example, his paper [5] on computation of the Riemann zeta function. The author was supported in part by Australian Research Council grant DP140101417 (held by Jon Borwein, David Bailey and the author).

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).

[2] R. A. Askey and R. Roy, 'Gamma function', *NIST Digital Library of Mathematical Functions* Ch. 5, http://dlmf.nist.gov/.

[3] H. Behnke and F. Sommer, *Theorie der analytischen Funktionen einer komplexen Veränderlichen*, 2nd edn (Springer, Berlin, 1962).

[4] M. V. Berry, 'The Riemann–Siegel expansion for the zeta function: high orders and remainders', *Proc. R. Soc. Lond. A* **450** (1995), 439–462.

[5] J. M. Borwein, D. M. Bradley and R. E. Crandall, 'Computational strategies for the Riemann zeta function', *J. Comput. Appl. Math.* **121** (2000), 247–296.

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press

[6] R. P. Brent, 'F. D. Crary & J. B. Rosser, High precision coefficients related to the zeta function [review]', *Math. Comp.* **31** (1977), 803–804.

[7] R. P. Brent, 'On the zeros of the Riemann zeta function in the critical strip', *Math. Comp.* **33** (1979), 1361–1372.

[8] R. P. Brent, 'Asymptotic approximation of central binomial coefficients with rigorous error bounds', 2016. arXiv:1608.04834v1.

[9] R. P. Brent, J. van de Lune, H. J. J. te Riele and D. T. Winter, 'On the zeros of the Riemann zeta function in the critical strip, II', *Math. Comp.* **39** (1982), 681–688.

[10] K. Dilcher, 'Asymptotic behaviour of Bernoulli, Euler, and generalized Bernoulli polynomials', *J. Approx. Theory* **49** (1987), 321–330.

[11] H. M. Edwards, *Riemann's Zeta Function* (Academic Press, New York, 1974), reprinted by Dover Publications, 2001.

[12] W. Gabcke, 'Neue Herleitung und explizite Restabschätzung der Riemann–Siegel-Formel', Dissertation, Mathematisch-Naturwissenschaftlichen, Göttingen, 1979.

[13] C. F. Gauss, 'Disquisitiones generales circa seriem infinitam ...', *Comm. Soc. Reg. Sci. Göttingensis Rec.* **2** (1813), reprinted in *Carl Friedrich Gauss Werke*, Bd. 3, Göttingen, 1876, 123–162.

[14] J.-P. Gram, 'Note sur les zéros de la fonction $\zeta(s)$ de Riemann', *Acta Math.* **27** (1908), 289–304.

[15] D. E. G. Hare, 'Computing the principal branch of log-Gamma', *J. Algorithms* **25** (1997), 221–236.

[16] M. Ch. Hermite, 'Sur la fonction $\log\Gamma(a)$', *J. reine angew. Math.* **115** (1895), 201–208.

[17] D. H. Lehmer, 'Extended computation of the Riemann zeta function', *Mathematika* **3** (1956), 102–108.

[18] R. E. Meyer, 'A simple explanation of the Stokes phenomenon', *SIAM Rev.* **31** (1989), 435–445.

[19] G. Nemes, 'Generalization of Binet's Gamma function formulas', *Integral Transforms Spec. Funct.* **24** (2013), 597–606.

[20] A. B. Olde Daalhuis, S. J. Chapman, J. R. King, J. R. Ockendon and and R. H. Tew, 'Stokes phenomenon and matched asymptotic expansions', *SIAM J. Appl. Math.* **55** (1995), 1469–1483.

[21] F. W. J. Olver, 'Error bounds for asymptotic expansions, with an application to cylinder functions of large argument', in: *Asymptotic Solutions of Differential Equations and Their Applications* (ed. C. H. Wilcox) (Wiley, New York, 1964), 163–183.

[22] F. W. J. Olver, *Asymptotics and Special Functions* (Academic Press, New York, 1974).

[23] G. Pólya and G. Szegö, *Problems and Theorems in Analysis I*, Springer Classics in Mathematics (Springer, Berlin, 1972).

[24] R. Spira, 'Calculation of the Gamma function by Stirling's formula', *Math. Comp.* **25** (1971), 317–322.

[25] I. Tweddle, 'Approximating $n!$, historical origins and error analysis', *Amer. J. Phys.* **52** (1984), 487–488.

[26] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd edn (Cambridge University Press, Cambridge, 1941).

[27] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 3rd edn (Cambridge University Press, Cambridge, 1920).

RICHARD P. BRENT, Mathematical Sciences Institute,
Australian National University, Canberra,
ACT 2600, Australia
e-mail: richard.brent@anu.edu.au
and
CARMA, University of Newcastle, Callaghan,
NSW 2308, Australia

https://doi.org/10.1017/S1446788718000393 Published online by Cambridge University Press