AN ALGORITHM FOR COMPUTING THE RIEMANN ZETA FUNCTION BASED ON AN ANALYSIS OF BACKLUND'S REMAINDER ESTIMATE

By

Petra Margarete Menz

B. Sc., The University of Toronto, 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

MATHEMATICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1994

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Mathematics

The University of British Columbia
Vancouver, Canada

Date 06. Sept. 1994

Abstract

The Riemann zeta function, $\zeta(s)$ with complex argument s, is a widely used special function in mathematics. This thesis is motivated by the need of a cost reducing algorithm for the computation of $\zeta(s)$ using its Euler-Maclaurin series. The difficulty lies in finding small upper bounds, call them n and k, for the two sums in the Euler-Maclaurin series of $\zeta(s)$ which will compute $\zeta(s)$ to within any given accuracy ϵ for any complex argument s, and provide optimal computational cost in the use of the Euler-Maclaurin series.

This work is based on Backlund's remainder estimate for the Euler-Maclaurin remainder, since it provides a close enough relationship between n, k, s, and ϵ. We assumed that the cost of computing the Bernoulli numbers, which appear in the series, is fixed, and briefly discuss how this may influence high precision calculation. Based on our study of the behavior of Backlund's remainder estimate, we define the 'best' pair (n, k), and present a reliable method of computing the best pair. Furthermore, based on a computational analysis, we conjecture that there is a relationship between n and k which does not depend on s. We present two algorithms, one based on our method and the other on the conjecture, and compare their costs of finding n and k as well as computing the Euler-Maclaurin series with an algorithm presented by Cohen and Olivier. We conclude that our algorithm reduces the cost of computing $\zeta(s)$ drastically, and that good numerical techniques need to be applied to our method and conjecture for finding n and k in order to keep this computational cost low as well.

Table of Contents

List of Figures

Acknowledgement

I want to thank Dr. Bill Casselman, my thesis supervisor, for his acceptance of my interest in the presented problem, and his support and guidance throughout my work on this thesis when it seemed impossible to complete in such a short amount of time. I would like to thank Dr. Richard Froese for his willingness to read this thesis.

Most of all, I want to thank John Michael Stockie for his believe in my abilities and his readiness to listen. He has been an endless source of encouragement.

I would also like to thank Dr. Ragnar Buchweitz, who put me on the right track, and Kirby Kim, who has been there for me from the very beginning and without whom I would not be where I am today.

Finally, I would like to dedicate this thesis to my parents, Gerlinde and Peter Menz, and my sister, Manuela Schmidt, whose love from so far away pulled me through difficult times.

Chapter 1

Introduction

One of the most important special functions in mathematics is the Riemann zeta function, $\zeta(s)$, which takes complex argument s. Because of its widespread importance, it is essential to have an accurate and efficient technique of computing the zeta function. One method of approximating $\zeta(s)$ is via the Euler-Maclaurin (E-M) summation formula. Given $s \in \mathbb{C}$ and absolute error tolerance $\epsilon > 0$, we will present an algorithm for the computation of $\zeta(s)$ to within ϵ, which is based on the E-M series of $\zeta(s)$ and a simplified version of Backlund's remainder estimate (BRE) for the E-M remainder.

Our objective is to choose that number of terms in the E-M series which minimizes computational cost. We take into consideration that the terms of the two sums in the E-M series have different computational cost. This is due in part to the appearance of the Bernoulli numbers in one of the sums. Throughout this paper, we assume that there is a fixed cost for obtaining the Bernoulli numbers, since they can always be tabulated. However, a Bernoulli number table takes an excessive amount of storage space for high precision calculation. Therefore, we employ an estimate of the Bernoulli numbers. Then, by a recursive calculation of the sum involving the Bernoulli numbers, we can assume fixed computational cost for the terms of that sum.

Besides s and ϵ, BRE involves two unknowns, which we will call n and k, and which determine the number of terms used in the E-M series of $\zeta(s)$. So, our goal becomes finding that pair (n, k) which gives the least number of terms needed, i. e. which will minimize the computational cost. Based on a thorough computational analysis of BRE,

we conjecture that there is a relationship between the optimal n and k, which does not depend on s and ϵ, but only on the cost of computing terms from the E-M series of $\zeta(s)$. This relationship allows us to eliminate n, giving us a simplified version of BRE. Thereby, given s and ϵ, we make it possible to solve BRE for the single unknown k, and compute the E-M series of $\zeta(s)$ using the least number of terms.

The following is an outline of how the material of this thesis is presented. Chapter 2 covers in detail the proof of the E-M summation formula and an estimate of its remainder; introduces the Riemann zeta function, $\zeta(s)$, and some of its properties required for the algorithm; shows the application of the E-M summation formula to $\zeta(s)$; derives the growth ratio of consecutive terms in the E-M series of $\zeta(s)$; and gives a proof of BRE. At the end of Chapter 2, we also work through practical examples demonstrating the behaviour of the E-M series of $\zeta(s)$ and BRE for the real and complex cases. Thereby, we point out the difficulties involved in choosing optimal values for n and k. For those familiar with the E-M series of $\zeta(s)$ and BRE, we suggest that Chapter 2 be skipped.

Chapter 3 is an in-depth discussion of the behaviour of BRE. After a thorough investigation of the optimal n and k, we will present a method which allows us to compute the optimal n and k and give the aforementioned conjecture. Lastly, we will use the growth ratio, together with BRE, to approximate the size of the smallest term in the E-M series for any given s. We conclude Chapter 3 by considering the existence of a relationship between the n and k from the smallest term and the optimal n and k.

In Chapter 4 we present two algorithms for the computation of $\zeta(s)$. One which utilizes our method of computing the optimal n and k, and one which is based on our conjecture. Furthermore, we will compare these two algorithms as well as Cohen and Olivier's recent algorithm [2], and discuss their differences in distributing computational cost.

The Appendices comprise brief discussions of a validation of BRE, the Bernoulli polynomials, and the Gamma function, as well as Mathematica [6] programs which facilitate the above analyses and generate the graphs displayed throughout this work.

Chapter 2

Euler-Maclaurin Series of $\zeta(s)$

This chapter provides a detailed treatise of the E-M summation formula, its remainder, and its application to the Riemann zeta function. Furthermore, BRE is introduced and its practical use is demonstrated for real and complex cases. Through these examples, it will become evident that it is no easy task to solve for the two unknowns in BRE given $s \in \mathbb{C}$ and $\epsilon > 0$, and that we need a reliable method for choosing values for the two unknowns.

The E-M summation formula, its proof, and its application to $\zeta(s)$ all involve the Bernoulli polynomials and the Gamma function as well as some of their properties and estimates. The reader can find all relevant information regarding the Bernoulli polynomials and the Gamma function in Appendices B and C, respectively.

2.1 Statement and Proof of the E-M Summation Formula

Proposition 1 *Let f be a function defined on the interval $[a, b]$, where $a < b$ and $a, b \in \mathbb{Z}$. Suppose that f has continuous derivatives up to order $2k$ or $2k+1$. Then*

$$\begin{aligned} f(a) + f(a+1) + \ldots + f(b-1) &= \int_a^b f(x)\,dx - \frac{1}{2}(f(b) - f(a)) \\ &\quad + \sum_{j=1}^{k} \frac{\beta_{2j}}{(2j)!}(f^{(2j-1)}(b) - f^{(2j-1)}(a)) \\ &\quad + R_k, \end{aligned}$$

where

$$R_k = \begin{cases} -\frac{1}{(2k)!}\int_b^a f^{(2k)}(x)\,\psi_{2k}(x)\,dx\,, \\ \textit{or} \\ \frac{1}{(2k+1)!}\int_b^a f^{(2k+1)}(x)\,\psi_{2k+1}(x)\,dx\,, \end{cases} \tag{2.1}$$

depending on how smooth f is, and $\psi_m(x)$ is as defined by equation (B.22).

Proof: Let f be a function that satisfies the conditions of Proposition 1 on the interval $[a,b]$, where $a, b \in \boldsymbol{Z}$ and $a < b$.

Let $r \in \boldsymbol{Z}$ and consider the integral $\int_r^{r+1} f(x)\,\psi_1'(x)\,dx$ over the unit interval $[r, r+1]$. Using integration by parts and the periodicity 1 of ψ_1, we get:

$$\begin{aligned} & \int_r^{r+1} f(x)\psi_1'(x)\,dx \\ = \;& f(x)\psi_1(x)\big|_r^{r+1} - \int_r^{r+1} f'(x)\psi_1(x)\,dx \\ = \;& f(r+1)\lim_{R\to r+1^-}\psi_1(R) - f(r)\lim_{R\to r^+}\psi_1(R) - \int_r^{r+1} f'(x)\psi_1(x)\,dx \\ = \;& f(r+1)\frac{1}{2} - f(r)(-\frac{1}{2}) - \int_r^{r+1} f'(x)\psi_1(x)\,dx \\ = \;& \frac{1}{2}(f(r)+f(r+1)) - \int_r^{r+1} f'(x)\psi_1(x)\,dx\,. \end{aligned}$$

Using the fact that $\psi_m(r) = B_m(r) = B_m(0) = \beta_m$ for any integer r and $m > 1$ as well as property (B.23) of $\psi_m(x)$ for $m > 0$, we can continue with integration by parts successively and obtain:

$$\begin{aligned} & \int_r^{r+1} f(x)\psi_1'(x)\,dx \\ = \;& \frac{1}{2}(f(r)+f(r+1)) - \int_r^{r+1} f'(x)\psi_1(x)\,dx \\ = \;& \frac{1}{2}(f(r)+f(r+1)) - \frac{1}{2}f'(x)\psi_2(x)\Big|_r^{r+1} + \frac{1}{2}\int_r^{r+1} f''(x)\psi_2(x)\,dx \\ = \;& \frac{1}{2}(f(r)+f(r+1)) - \frac{\beta_2}{2}(f'(r+1)-f'(r)) + \frac{1}{2}\int_r^{r+1} f''(x)\psi_2(x)dx \\ \vdots \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}(f(r)+f(r+1)) + \sum_{j=2}^{k}(-1)^{j-1}\frac{\beta_j}{j!}\left(f^{(j-1)}(r+1) - f^{(j-1)}(r)\right) \\
&\quad + \frac{(-1)^k}{k!}\int_r^{r+1} f^{(k)}(x)\psi_k(x)\,dx \\
&= \frac{1}{2}(f(r)+f(r+1)) - \sum_{j=1}^{k}\frac{\beta_{2j}}{(2j)!}\left(f^{(2j-1)}(r+1) - f^{(2j-1)}(r)\right) - \overline{R}_k,
\end{aligned}
$$

since $\beta_m = 0$ for m odd, and $\overline{R}_k$ is R_k as given in Proposition 1 but restricted to the interval $[r, r+1]$. Next, subdivide the interval $[a, b]$ into its unit subintervals. Again, using the periodicity 1 of ψ_m, we can sum the above equation over the unit subintervals of $[a, b]$ and get:

$$
\begin{aligned}
\int_b^a f(x)\psi_1'(x)\,dx &= \frac{1}{2}f(a) + f(a+1) + \cdots + f(b-1) + \frac{1}{2}f(b) \\
&\quad - \sum_{j=1}^{k}\frac{\beta_{2j}}{(2j)!}\left(f^{(2j-1)}(b) - f^{(2j-1)}(a)\right) - R_k.
\end{aligned}
$$

Now, keeping in mind that by definition of $\psi_1(x)$ we have $\psi_1'(x) = 1$ for all x, we can rewrite this equation and Proposition 1 follows. □

We need to have an estimate for the remainder R_k of Proposition 1. Applying the absolute value to equation (2.1) and using the result of equation (B.22) we get

$$
\begin{aligned}
|R_k| &= \begin{cases} \frac{1}{(2k)!}\left|\int_b^a f^{(2k)}(x)\,\psi_{2k}(x)\,dx\right|, \\ or \\ \frac{1}{(2k+1)!}\left|\int_b^a f^{(2k+1)}(x)\,\psi_{2k+1}(x)\,dx\right|, \end{cases} \\
&\leq \begin{cases} \frac{1}{(2k)!}\int_b^a \left|f^{(2k)}(x)\right|\,|\psi_{2k}(x)|\,dx, \\ or \\ \frac{1}{(2k+1)!}\int_b^a \left|f^{(2k+1)}(x)\right|\,|\psi_{2k+1}(x)|\,dx, \end{cases}
\end{aligned}
$$

$$\leq \begin{cases} \frac{1}{(2k)!}\int_b^a \left|f^{(2k)}(x)\right| \left|\frac{2(2k)!}{(2\pi)^{2k}}\sum_{l=1}^{\infty}\frac{\cos 2\pi lx}{l^{2k}}\right| dx\,, \\ or \\ \frac{1}{(2k+1)!}\int_b^a \left|f^{(2k+1)}(x)\right| \left|\frac{2(2k+1)!}{(2\pi)^{2k+1}}\sum_{l=1}^{\infty}\frac{\sin 2\pi lx}{l^{2k+1}}\right| dx\,, \end{cases}$$

$$\leq \begin{cases} \frac{2\zeta(2k)}{(2\pi)^{2k}}\int_b^a \left|f^{(2k)}(x)\right| dx\,, \\ or \\ \frac{2\zeta(2k+1)}{(2\pi)^{2k+1}}\int_b^a \left|f^{(2k+1)}(x)\right| dx\,, \end{cases}$$

depending on how smooth f is. This is the standard way of estimating the E-M remainder. As we will see in Section 2.4 we need to take a similar but modified approach in estimating the E-M remainder of the zeta function.

2.2 E-M Summation Formula Applied to $\zeta(s)$

Next, we introduce Riemann's definition of the zeta function and some of its properties necessary for a computational algorithm, but our main concern is to derive the E-M series of $\zeta(s)$.

Definition 1 *The Riemann Zeta function $\zeta(s)$ is defined by*

$$\zeta(s) = \sum_{r=1}^{\infty}\frac{1}{r^s}, \tag{2.2}$$

for s in the set $\mathcal{S} := \{s \in \mathbb{C} \mid \Re(s) > 1\}$.

Proposition 2 *$\zeta(s)$ is analytic on $\mathcal{S}$.*

Proof: Let $f_r(s) = \frac{1}{r^s}$ for $r \in \mathbb{N}$. Let $s = \sigma + it \in \mathcal{S}$. Let M be any positive number such that $\Re(s) = \sigma \geq M > 1$. Then

$$|f_r(s)| = \left|\frac{1}{r^s}\right| = \left|\frac{1}{e^{s\ln r}}\right| = \frac{1}{e^{\sigma\ln r}} \leq \frac{1}{e^{M\ln r}} = \frac{1}{r^M},$$

and $\sum_{r=1}^{\infty} \frac{1}{r^M}$ converges since $M > 1$. Hence, by the Weierstrass Test $\sum_{r=1}^{\infty} f_r(s)$ converges uniformly on $\{s \in \mathbb{C} \mid \Re(s) \geq M\}$. Since further, $\{f_r(s)\}_{r=1}^{\infty}$ are analytic for all $s \in \mathbb{C}$, it follows that $\zeta(s)$ is analytic in $\mathcal{S}$. □

Proposition 3 (Functional Equation) *For all* $s \in \mathbb{C}\backslash\{0,1\}$

$$\zeta(s) = 2(2\pi)^{s-1} \sin\left(\frac{s\pi}{2}\right) \Gamma(1-s)\zeta(1-s). \tag{2.3}$$

Proof: By a change of variables in the definition of $\Gamma(s)$ (see Appendix C) we obtain

$$\begin{aligned} \Gamma(s) &= k^s \int_0^{\infty} t^s e^{-kt} \frac{dt}{t} \\ \Rightarrow \quad \pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \frac{1}{k^s} &= \int_0^{\infty} t^{\frac{s}{2}} e^{-\pi k^2 t} \frac{dt}{t}. \end{aligned}$$

Summing over k, we get

$$\begin{aligned} \pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) &= \sum_{k=1}^{\infty} \int_0^{\infty} t^{\frac{s}{2}} e^{-\pi k^2 t} \frac{dt}{t} \\ &= \int_0^{\infty} \left(\sum_{k=1}^{\infty} e^{-\pi k^2 t}\right) t^{\frac{s}{2}} \frac{dt}{t}. \end{aligned} \tag{2.4}$$

Now, let $\vartheta(t) = \sum_{-\infty}^{\infty} e^{-\pi k^2 t}$. By the Poisson summation formula [3, p.209] it follows that $\vartheta(t) = \frac{1}{\sqrt{t}} \vartheta(\frac{1}{t})$. Let $\psi(t) = \sum_{k=1}^{\infty} e^{-\pi k^s t}$. Then $\psi(t) = \frac{\vartheta(t)-1}{2}$, and together with the functional equation for $\vartheta(t)$ this implies that $\psi(\frac{1}{t}) = \sqrt{t}\psi(t) + \frac{\sqrt{t}-1}{2}$. So, equation (2.4) becomes

$$\begin{aligned} \pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) &= \int_0^{\infty} \psi(t) t^{\frac{s}{2}} \frac{dt}{t} \\ &= \int_0^1 \psi(t) t^{\frac{s}{2}} \frac{dt}{t} + \int_1^{\infty} \psi(t) t^{\frac{s}{2}} \frac{dt}{t} \\ &= \int_1^{\infty} \psi\left(\frac{1}{t}\right) t^{-\frac{s}{2}} \frac{dt}{t} + \int_1^{\infty} \psi(t) t^{\frac{s}{2}} \frac{dt}{t} \\ &= \int_1^{\infty} \psi(t) t^{-\frac{s}{2}+\frac{1}{2}} \frac{dt}{t} + \frac{1}{2} \int_1^{\infty} \left(t^{\frac{1}{2}} - 1\right) t^{-\frac{s}{2}} \frac{dt}{t} + \int_1^{\infty} \psi(t) t^{\frac{s}{2}} \frac{dt}{t} \\ &= \int_1^{\infty} \psi(t) \left(t^{\frac{s}{2}} + t^{\frac{1-s}{2}}\right) \frac{dt}{t} + \frac{1}{2} \int_1^{\infty} \left(t^{\frac{1-s}{2}} - t^{-\frac{s}{2}}\right) \frac{dt}{t} \\ &= \int_1^{\infty} \psi(t) \left(t^{\frac{s}{2}} + t^{\frac{1-s}{2}}\right) \frac{dt}{t} - \left(\frac{1}{s} + \frac{1}{1-s}\right). \end{aligned}$$

The right hand side of the above equation converges for all $s \in \mathbb{C}\backslash\{0,1\}$, and it is symmetrical with respect to s and $1-s$. It follows that

$$\pi^{-\frac{s}{2}}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \pi^{-\frac{1-s}{2}}\Gamma\left(\frac{1-s}{2}\right)\zeta(1-s)$$

$$\begin{aligned}\Rightarrow \zeta(s) &= \pi^{s-\frac{1}{2}}\frac{\Gamma(\frac{1-s}{2})}{\Gamma(\frac{s}{2})}\zeta(1-s) \\ &= (2\pi)^{s-1}2\sin\left(\frac{s\pi}{2}\right)\Gamma(1-s)\zeta(1-s)\end{aligned}$$

by equation (C.26) in Appendix C. □

Note: Throughout the rest of this paper, we write $s = \sigma + it$ for $\sigma, t \in \mathbb{R}$, unless otherwise stated.

Definition 2 *Let $n \in \mathbb{N}$. define $\zeta_n(s)$ to be the partial sum of $\zeta(s)$,*

$$\zeta_n(s) = \sum_{r=1}^{n-1}\frac{1}{r^s}.$$

Then we can write $\zeta(s)$ as $\zeta(s) = lim_{n\to\infty}\zeta_n(s)$. The following proposition uses the notation of Hutchinson [4] in a slightly modified form for easier cross references with the functions in the programs of Appendix D.

Proposition 4 *Let $n, k \in \mathbb{N}$. Then for $s \in \mathcal{S}$ the Riemann Zeta function satisfies*

$$\zeta(s) = \zeta_n(s) + \frac{1}{s-1}n^{1-s} + \frac{1}{2}n^{-s} + \sum_{j=1}^{k}T_j(n,s) + R(n,k,s), \tag{2.5}$$

where

$$\begin{aligned}T_j(n,s) &= \frac{\beta_{2j}}{(2j)!}\frac{s(s+1)...(s+2j-2)}{n^{s+2j-1}} \\ &= \frac{\beta_{2j}}{(2j)!}\frac{1}{n^{s+2j-1}}\prod_{m=0}^{2j-2}(s+m)\end{aligned} \tag{2.6}$$

and the remainder term is

$$\begin{aligned} R(n,k,s) &= -\frac{s(s+1)...(s+2k-1)}{(2k)!}\int_n^\infty \frac{\psi_{2k}(x)}{x^{s+2k}}\,dx, \\ &= -\frac{s(s+1)...(s+2k)}{(2k+1)!}\int_n^\infty \frac{\psi_{2k+1}(x)}{x^{s+2k+1}}\,dx. \end{aligned} \tag{2.7}$$

Proof: Let $n \in \mathbb{N}$ with $n > 1$. Let $s \in \mathbb{C}\backslash\{1\}$. Define $f(x) = \frac{1}{x^s}$. Then f is a function on the interval $[0,1]$ with continuous derivatives:

$$f^{(d)}(x) = (-1)^d \left(\frac{s(s+1)\dots(s+d-1)}{x^{s+d}}\right).$$

Hence, we can apply Proposition 1 to f on $[0,1]$ with derivative order 1 and get

$$\begin{aligned} f(1)+f(2)+\cdots+f(n-1) = \int_1^n f(x)\,dx &- \frac{1}{2}(f(n)-f(1)) \\ &+ \int_1^n f'(x)\psi_1(x)\,dx\,. \end{aligned}$$

This implies

$$\begin{aligned} \zeta_n(s) &= \int_1^n \frac{1}{x^s}\,dx - \frac{1}{2}\frac{1}{n^s} + \frac{1}{2} - s\int_1^n \frac{\psi_1(x)}{x^{s+1}}\,dx \\ &= \frac{-1}{(s-1)x^{s-1}}\Big|_1^n + \frac{1}{2} - \frac{1}{2}\frac{1}{n^s} - s\int_1^n \frac{\psi_1(x)}{x^{s+1}}\,dx \\ &= \frac{1}{2} + \frac{1}{s-1} - \frac{1}{s-1}\frac{1}{n^{s-1}} - \frac{1}{2}\frac{1}{n^s} - s\int_1^n \frac{\psi_1(x)}{x^{s+1}}\,dx\,. \end{aligned} \tag{2.8}$$

In other words,

$$-s\int_1^n \frac{\psi_1(x)}{x^{s+1}}\,dx = \frac{1}{2} + \frac{1}{s-1} - \frac{1}{s-1}\frac{1}{n^{s-1}} - \frac{1}{2}\frac{1}{n^s} + \zeta_n(s).$$

Now, let $n \to \infty$ in equation (2.8), so that

$$\begin{aligned} \zeta(s) &= \frac{1}{2} + \frac{1}{s-1} - s\int_1^\infty \frac{\psi_1(x)}{x^{s+1}}\,dx \\ &= \frac{1}{2} + \frac{1}{s-1} - s\int_n^\infty \frac{\psi_1(x)}{x^{s+1}}\,dx - s\int_1^n \frac{\psi_1(x)}{x^{s+1}}\,dx \\ &= \zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} - s\int_n^\infty \frac{\psi_1(x)}{x^{s+1}}\,dx\,. \end{aligned}$$

Continuing with integration by parts just as was done in detail in the proof of Proposition 1, we get

$$\begin{aligned}
\zeta(s) &= \zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} \\
&\quad + \frac{\beta_2}{2!}\frac{s}{n^{s+1}} + \cdots + \frac{\beta_{2n}}{(2n)!}\frac{s(s+1)\ldots(s+2n-2)}{n^{s+2n-1}} \\
&\quad - \frac{s(s+1)\ldots(s+2n-1)}{(2n)!}\int_n^\infty \frac{\psi_{2n}(x)}{x^{s+2n}}\,dx \\
&= \zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} \\
&\quad + \frac{\beta_2}{2!}\frac{s}{n^{s+1}} + \cdots + \frac{\beta_{2n}}{(2n)!}\frac{s(s+1)\ldots(s+2n-2)}{n^{s+2n-1}} \\
&\quad - \frac{s(s+1)\ldots(s+2n-1)}{(2n)!} \times \\
&\quad \left(\lim_{b\to\infty} \frac{\psi^{2n+1}(x)}{(2n+1)x^{s+2n}}\bigg|_n^b + \frac{s+2n}{2n+1}\int_n^\infty \frac{\psi_{2n+1}(x)}{x^{s+2n+1}}\,dx \right) \\
&= \zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} \\
&\quad + \frac{\beta_2}{2!}\frac{s}{n^{s+1}} + \cdots + \frac{\beta_{2n}}{(2n)!}\frac{s(s+1)\ldots(s+2n-2)}{n^{s+2n-1}} \\
&\quad - \frac{s(s+1)\ldots(s+2n)}{(2n+1)!}\int_n^\infty \frac{\psi_{2n+1}(x)}{x^{s+2n+1}}\,dx\,,
\end{aligned}$$

which proves Proposition 4. □

Remark 1: Actually, equation (2.5) makes sense for all $s \in \mathbb{C}$ with $\Re(s) > -2k$, because the integral for $R(n,k,s)$ converges throughout the half-plane $\Re(s+2k+1) > 1$. Hence, $\zeta(s)$ can be extended by analytic continuation to other values of s. Furthermore, equation (2.5) also shows that $\zeta(s)$ can be extended to a meromorphic function of s, since $\zeta(s)$ has as its only singularity a simple pole at $s = 1$.

Remark 2: Now, remember that the functional equation for $\zeta(s)$ is not valid for s equal to 0 or 1. The E-M series of $\zeta(s)$ (2.5) gives $\zeta(0) = -\frac{1}{2}$ and $\zeta(1) = \infty$.

Note: Throughout the rest of this paper, the letters n and k are solely reserved to represent the bounds for the two series respectively in the E-M summation formula for $\zeta(s)$ as given in equation (2.5).

2.3 Growth Ratio of the E-M Series for $\zeta(s)$

The terms in the E-M series of $\zeta(s)$ grow large in a graded way. This implies that if we ask, when do the terms of that series get large, it is not too important if we are off by a few terms. For the estimate of the growth ratio of the E-M series of $\zeta(s)$ we need the following three limits:

$$\lim_{k\to\infty} \sqrt{\frac{k+1}{k}} = 1,$$

$$\lim_{k\to\infty} \left(\frac{k+1}{k}\right)^{2k} = e^2, \text{ and}$$

$$\lim_{k\to\infty} \frac{(k+1)^2}{(2k+1)(2k+2)} = \frac{1}{2^2}.$$

Now, the ratio of two successive terms can be estimated for large k:

$$\begin{aligned}
\left|\frac{T_{k+1}(n,s)}{T_k(n,s)}\right| &= \left|\frac{\beta_{2k+2}(2k)!n^{1-s-2k-2}\prod_{m=0}^{2k}(s+m)}{\beta_{2k}(2k+2)!n^{1-s-2k}\prod_{m=0}^{2k-2}(s+m)}\right| \\
&= \left|\frac{\beta_{2k+2}}{\beta_{2k}}\right|\frac{1}{(2k+1)(2k+2)}\frac{1}{n^2}(s+2k-1)(s+2k) \\
&\sim \frac{\sqrt{2\pi(k+1)}(\frac{k+1}{\pi e})^{2k+2}}{\sqrt{2\pi k}(\frac{k}{\pi e})^{2k}}\frac{1}{(2k+1)(2k+2)}\frac{1}{n^2}(s+2k)^2 \\
&\sim \sqrt{\frac{k+1}{k}}\frac{1}{(\pi e)^2}\left(\frac{k+1}{k}\right)^{2k}\frac{(k+1)^2}{(2k+1)(2k+2)}\frac{1}{n^2}(s+2k)^2 \\
&\sim \left(\frac{s+2k}{2\pi n}\right)^2.
\end{aligned}$$

This implies that the terms in the E-M series (2.5) will start to grow when

$$|s+2k| \simeq 2\pi n. \tag{2.9}$$

2.4 Backlund's Remainder Estimate

A more practical estimate of the E-M remainder of $\zeta(s)$ than the one previously derived in Section 2.1 is due to Backlund [1]. It states that the magnitude of the remainder in the E-M series (2.5) is at most $\left|\frac{s+2k-1}{\sigma+2k-1}\right|$ times the magnitude of the first term omitted:

$$|R(n,k-1,s)| \leq \left|\frac{s+2k-1}{\sigma+2k-1}\right||T_k(n,s)|. \tag{2.10}$$

This shows that the error may get worse as the imaginary part of s gets larger. The practical effect of this is that in a given band $|\Re(s)| \leq b$, where $b \in I\!R^+$, we need roughly $\Im(s)$ many terms to calculate $\zeta(s)$. When s is real, the remainder is less than or equal to the first term omitted, i.e. the actual value of $\zeta(s)$ always lies between two consecutive partial sums of the E-M series (2.5).

Now, Backlund's estimate of the E-M remainder of $\zeta(s)$ is derived as follows:

$$
\begin{aligned}
&|R(n,k-1,s)| \\
&= \left|\frac{s(s+1)\dots(s+2k-2)}{(2k-1)!}\int_n^\infty \frac{\psi_{2k-1}(x)}{x^{s+2k-1}}\,dx\right| \\
&= \left|\frac{s(s+1)\dots(s+2k-2)}{(2k-1)!}\left(\lim_{b\to\infty}\left(\frac{\psi_{2k}(x)}{2kx^{s+2k-1}}\Big|_n^b\right) - \frac{-s-2k+1}{2k}\int_n^\infty \frac{\psi_{2k}(x)}{x^{s+2k}}\,dx\right)\right| \\
&= \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\left(\frac{\beta_{2k}\,n^{1-s-2k}}{s+2k-1} - \int_n^\infty \frac{\psi_{2k}(x)}{x^{s+2k}}\,dx\right)\right| \\
&= \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\int_n^\infty \frac{\beta_{2k}-\psi_{2k}(x)}{x^{s+2k}}\,dx\right| \\
&\le \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\right|\int_n^\infty \frac{|\beta_{2k}-\psi_{2k}(x)|}{x^{\sigma+2k}}\,dx \\
&= \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\right|\int_n^\infty \frac{(-1)^{k+1}\left(\beta_{2k}-\psi_{2k}(x)\right)}{x^{\sigma+2k}}\,dx \\
&= \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\right|\left(|\beta_{2k}|\int_n^\infty \frac{dx}{x^{\sigma+2k}} - \frac{\sigma+2k}{2k+1}\int_n^\infty (-1)^{k+1}\frac{\psi_{2k+1}(x)}{x^{\sigma+2k+1}}\,dx\right) \\
&\le \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\right||\beta_{2k}|\int_n^\infty \frac{dx}{x^{\sigma+2k}} \\
&= \left|\frac{s(s+1)\dots(s+2k-1)}{(2k)!}\right||\beta_{2k}|\,\frac{1}{(\sigma+2k-1)n^{\sigma+2k-1}} \\
&= \left|\frac{s+2k-1}{\sigma+2k-1}\right||T_k(n,s)|\,.
\end{aligned}
$$

For a validation of BRE see Appendix A, where a brief computational discussion is given on BRE compared with the absolute error.

2.5 The E-M Series of $\zeta(s)$ in Practice

We want to demonstrate how the E-M series of $\zeta(s)$ (2.5) works in practice. This will show that choosing n and k requires careful consideration of the calculations involved in the series. We will also find out that there is more than one pair $(n,\ k)$ that will yield a desired accuracy, and that we have to ask ourselves what the "best" pair $(n,\ k)$ should be. As suggested by Edwards [3, p.115], we need to compute the first several numbers in the sequence $\frac{B_{2j}}{(2j)!}s(s+1)\ldots(s+2j-2)$ and to see how large n must be in order to make the terms of the series decrease rapidly in size.

First, consider the case when s is real. Then $|R(n,k-1,s)| \leq |T_k(n,s)|$ by BRE (2.10). Let $s = 3$, and say, we want to compute $\zeta(3)$ to within six digits of accuracy. For $j = 1,2,3,4$ in the above sequence we get:

$$\frac{B_2}{2!}3 = \frac{1}{4} = 0.25$$

$$\frac{B_4}{4!}3\cdot 4\cdot 5 = -\frac{1}{12} = -0.08\bar{3}$$

$$\frac{B_6}{6!}3\cdot 4\cdot \ldots \cdot 7 = \frac{1}{12} = 0.08\bar{3}$$

$$\frac{B_8}{8!}3\cdot 4\cdot \ldots \cdot 9 = -\frac{3}{20} = -0.15$$

Table 2.1 shows many possible values for n and k which give $\zeta(3)$ to within six digits of accuracy, e. g. $n = 3, k = 4$ or $n = 6, k = 2$. However, we choose $n = 5, k = 2$ since the

$n =$ \ $k =$	1	2	3	4	5
2	$1.3\ 10^{-3}$	$3.3\ 10^{-4}$	$1.5\ 10^{-4}$	$1.0\ 10^{-4}$	$1.0\ 10^{-4}$
3	$1.1\ 10^{-4}$	$1.3\ 10^{-5}$	$2.5\ 10^{-6}$	$7.8\ 10^{-7}$	$3.4\ 10^{-7}$
4	$2.0\ 10^{-5}$	$1.3\ 10^{-6}$	$1.4\ 10^{-7}$	$2.5\ 10^{-8}$	$6.1\ 10^{-9}$
5	$5.3\ 10^{-6}$	$2.1\ 10^{-7}$	$1.5\ 10^{-8}$	$1.7\ 10^{-9}$	$2.7\ 10^{-10}$
6	$1.8\ 10^{-6}$	$5.0\ 10^{-8}$	$2.5\ 10^{-9}$	$1.9\ 10^{-10}$	$2.1\ 10^{-11}$

Table 2.1: Values of BRE for $s = 3$ as k ranges from 1 to 5 and n ranges from 2 to 6. Only two digits of accuracy are given!

terms of the series (2.5) decrease more rapidly in size than for the choice $n = 3, k = 4$, and we do not compute superfluous terms as with the choice $n = 6, k = 2$.

The following is the evaluation of $\zeta(3)$ using $n = 5, k = 2$:

$$\begin{aligned}
\tfrac{1}{1^3} &= 1.000\,000\,0 \\
\tfrac{1}{2^3} &= 0.125\,000\,0 \\
\tfrac{1}{3^3} &= 0.037\,037\,0 \\
\tfrac{1}{4^3} &= 0.015\,625\,0 \\
\tfrac{1}{5^3} &= 0.008\,000\,0 \\
\tfrac{1}{3-1}5^{1-3} &= 0.080\,000\,0 \\
\tfrac{1}{2}5^{-3} &= 0.004\,000\,0 \\
T_1(5,3) &= 0.000\,400\,0 \\
T_2(5,3) &= -0.000\,005\,3 \\
\hline
\zeta(3) &\sim 1.265\,067\,3
\end{aligned}$$

Now, consider the case when s is not real. Let $s = \frac{1}{2} + 10i$, and say, we want to

compute $\zeta\left(\frac{1}{2}+10i\right)$ to within six digits of accuracy. Here, $|s|$ is considerably larger than in the previous case. This implies that we need a much larger value for n in order to achieve comparable accuracy. However, the terms in $\zeta_n(s)$ are much more difficult to compute when s is complex. This is because $n^{-s} = n^{-\sigma}(\cos(t \ln n) - i \sin(t \ln n))$, so, computation involves a root, a logarithm and two trigonometric functions. Therefore, our goal is to keep n still as small as possible, but larger than k, because $\zeta_n(s)$ will then converge much faster. Table 2.2 shows many possible values for n and k which give

$n =$ \ $k =$	2	3	4	5	6
5	$1.7\ 10^{-3}$	$1.6\ 10^{-4}$	$2.2\ 10^{-5}$	$3.8\ 10^{-6}$	$8.6\ 10^{-7}$
6	$6.4\ 10^{-4}$	$4.1\ 10^{-5}$	$3.8\ 10^{-6}$	$4.7\ 10^{-7}$	$7.3\ 10^{-8}$
7	$2.7\ 10^{-4}$	$1.3\ 10^{-5}$	$8.8\ 10^{-7}$	$7.9\ 10^{-8}$	$9.1\ 10^{-9}$
8	$1.3\ 10^{-4}$	$4.8\ 10^{-6}$	$2.5\ 10^{-7}$	$1.7\ 10^{-8}$	$1.5\ 10^{-9}$
9	$6.9\ 10^{-5}$	$2.0\ 10^{-6}$	$8.1\ 10^{-8}$	$4.4\ 10^{-9}$	$3.1\ 10^{-10}$
10	$3.8\ 10^{-5}$	$9.0\ 10^{-7}$	$3.0\ 10^{-8}$	$1.3\ 10^{-9}$	$7.4\ 10^{-11}$
11	$2.3\ 10^{-5}$	$4.4\ 10^{-7}$	$1.2\ 10^{-8}$	$4.4\ 10^{-10}$	$2.0\ 10^{-11}$
12	$1.4\ 10^{-5}$	$2.3\ 10^{-7}$	$5.3\ 10^{-9}$	$1.6\ 10^{-10}$	$6.3\ 10^{-12}$

Table 2.2: Values of BRE for $s = \frac{1}{2} + 10i$ as k ranges from 2 to 6 and n ranges from 5 to 12. Only two digits of accuracy are given!

$\zeta\left(\frac{1}{2}+10i\right)$ to within six digits of accuracy, e. g. $n = 6, k = 5$ or $n = 12, k = 3$. However, we choose $n = 10, k = 3$ since the terms of the series (2.5) decrease more rapidly in size than for the choice $n = 6, k = 5$, and we do not compute superfluous terms as with the choice $n = 12, k = 3$. The following is the evaluation of $\zeta\left(\frac{1}{2}+10i\right)$ using $n = 10, k = 3$:

$$
\begin{array}{rcrcl}
\zeta_{10}\left(\frac{1}{2}+10i\right) & = & 1.370\,209\,21 & - & 0.386\,468\,41i \\
\frac{1}{-\frac{1}{2}+10i}10^{\frac{1}{2}-10i} & = & 0.279\,242\,35 & + & 0.147\,561\,28i \\
\frac{1}{2}10^{-\frac{1}{2}-10i} & = & \text{-}0.080\,761\,70 & + & 0.135\,932\,14i \\
T_1\left(10,\frac{1}{2}+10i\right) & = & \text{-}0.023\,328\,37 & - & 0.012\,327\,51i \\
T_2\left(10,\frac{1}{2}+10i\right) & = & \text{-}0.000\,456\,40 & - & 0.000\,042\,23i \\
T_3\left(10,\frac{1}{2}+10i\right) & = & \text{-}0.000\,009\,96 & + & 0.000\,007\,85i \\
\hline
\zeta\left(\frac{1}{2}+10i\right) & \sim & 1.544\,895\,12 & - & 0.115\,336\,89i
\end{array}
$$

The above two examples show that in order to compute $\zeta(s)$ to within a certain accuracy, we need some method that tells us how to choose n and k. This method has to be reliable and efficient, since it is too cumbersome to calculate a table of remainder values for each s, and decide by comparison what values of n and k to choose. Chapter 3 discusses the influence of n and k on the behavior of BRE, and explains how we can choose n and k in order to minimize the computations involved. This way we can develop an efficient algorithm for the calculation of $\zeta(s)$ based on the E-M series (2.5).

Chapter 3

Analysis of Backlund's Remainder Estimate

The main objective of this chapter is to find those n and k such that $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(n,s)| \leq \epsilon$ and such that the computational cost is minimized by the least number of terms in the E-M series of $\zeta(s)$. We want to do so in an efficient and accurate manner. To this end, we require an analysis of both the behavior of BRE, and the computational cost, as well as a reliable method of finding the optimal n and k.

3.1 Defining the Best Pair (n,k)

Throughout this section, let s be fixed, and write $\epsilon = e^{-D} > 0$, where $D \in \mathbb{R}^+$. First, we analyze the behavior of BRE for fixed k and n, and then we discuss the computational cost involved with the E-M series of $\zeta(s)$. This will lead to a definition of that pair (n,k) we require to solve our problem.

Case I: Let k be fixed. Then Backlund's estimate of $|R(n,k-1,s)|$ is of the form $C_1(s,k)\, n^{C_2(s,k)}$, where $C_1 := C_1(s,k)$, and $C_2 := C_2(s,k)$ are positive, non-zero constants depending on k and s. This means that if we want to find $\zeta(s)$ to within any accuracy e^{-D}, we need n to satisfy $C_2 \ln n - \ln C_1 \geq D$. Now, $f(n) := C_2 \ln n - \ln C_1$ as a function of n, for $n \in \mathbb{R}^+ \backslash \{0\}$, is monotone increasing and concave downward. Figure 3.1 shows the graphs of f for different s and k. As expected, the general shape of the graph of f is independent of the choices for s and k, since a change in C_1 results in a vertical translation of the curve of f,

and a change in C_2 results in an expansion or a contraction of the curve of f. Furthermore, f is invertible; so, given any $D \in \mathbb{R}^+$ we can always find a positive integer n such that $f(n) \geq D$.

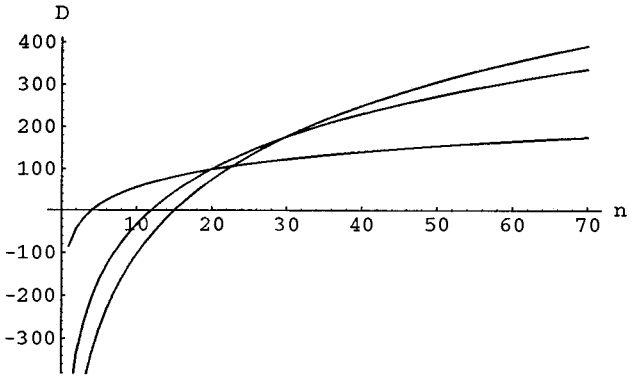

Figure 3.1: The graphs of $-\ln$ applied to Backlund's remainder estimate as a function of $n\ (= f(n))$ for three different values of s and k, where D denotes the number of accurate digits to the base e. The steepest curve has $s = \frac{1}{2} + 10i, k = 94$, the middle curve has $s = 3, k = 125$, and the curve with the most gradual slope has $s = 4 + 6i, k = 28$.

The above discussion implies that for any fixed k, we can always find an n such that $|R(n, k-1, s)| \leq \epsilon$. However, we want to keep computations at a minimum, and by randomly choosing k we have no influence on the number of terms used in the E-M series of $\zeta(s)$. In particular, we will not know, if the choice for k minimizes the number of terms used.

Case II: Let n be fixed. Then every term in Backlund's estimate of the remainder $|R(n, k-1, s)|$ depends on k. In order to show how BRE behaves for fixed s and n, let us simplify the estimate by applying estimate (B.24), Stirling's formula for factorials, and equation (C.25) to it. We can do this, because the asymptotic estimate $\frac{|\beta_{2k}|}{(2k)!} \sim \frac{2}{(2\pi)^{2k}}$ is sufficiently accurate for values of k as small as 3. We obtain

$$|R(n, k-1, s)| \leq \left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(n, s)|$$

$$\begin{aligned} &\simeq \left|\frac{s+2k-1}{\sigma+2k-1}\right| \frac{2}{(2\pi)^{2k}} \frac{1}{n^{\sigma+2k-1}} \left|\frac{\Gamma(s+2k-1)}{\Gamma(s)}\right| \\ &= C(s,n)\left|\frac{s+2k-1}{\sigma+2k-1}\right| \frac{|\Gamma(s+2k-1)|}{(2\pi n)^{2k}}, \end{aligned}$$

where $C := C(s,n) > 0$ is a constant depending only on fixed s and n. Let us define $g(k) := 2k \ln 2\pi n - \left(\ln\left|\frac{s+2k-1}{\sigma+2k-1}\right| + \ln|\Gamma(s+2k-1)| + \ln C\right)$. So, if we want to find $\zeta(s)$ to within accuracy e^{-D}, we need k to satisfy $g(k) \geq D$. Is this possible? In order to answer this question, we need to analyze $g(k)$, for $k \in \mathbb{R}^+ \backslash \{0\}$.

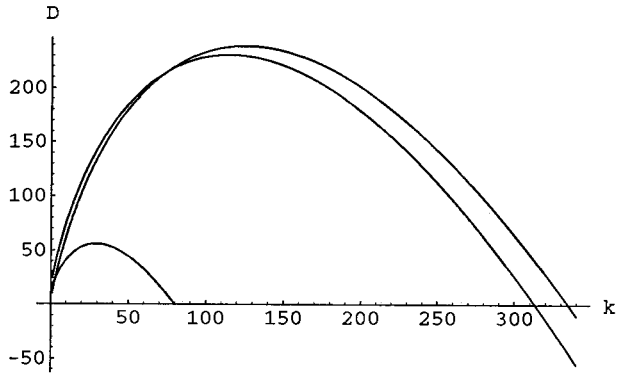

Figure 3.2: The graphs of $-\ln$ applied to Backlund's remainder estimate as a function of k $(= g(k))$ for three different values of s and n, where D denotes the number of accurate digits to the base e. The curve with the highest maximum has $s = \frac{1}{2} + 10i, n = 40$, the curve with the lowest maximum has $s = 4 + 6i, n = 10$, and the intermediate curve has $s = 3, n = 37$.

We have that $\ln\left|\frac{s+2k-1}{\sigma+2k-1}\right|$ increases for increasing k. Therefore, the behavior of g depends largely on the behavior of $\ln\frac{|\Gamma(s+2k-1)|}{(2\pi n)^{2k}}$. By definition of the Gamma function (see Appendix C), $|\Gamma(s+2k-1)|$ grows factorially in k, and $(2\pi n)^{2k}$ grows exponentially in k. So, there must exist a $k_o \in \mathbb{R}^+$ such that for $k < k_o$ we have $|\Gamma(s+2k-1)| < (2\pi n)^{2k}$, and for $k > k_o$ we have $|\Gamma(s+2k-1)| > (2\pi n)^{2k}$. In other words, g will increase to some best number D_o of accurate digits to the base e and then decrease again, where D_o depends on k_o. Finding k_o just means looking for the smallest term in the E-M series of $\zeta(s)$, but this term can be found by the

growth ratio (2.9). So, we are able to deduce from this growth ratio that k_o is that number, which satisfies $|s + 2k_o| \simeq 2\pi n$, for our fixed s and n. Figure 3.2 shows the graphs of g for various values of s and n. As expected, the general shape of g is independent of the choices for s and n, since modifying s or n shifts the maximum of the graph of g according to the growth ratio.

Now, we can answer our previous question if, given $D \in \mathbb{R}^+$, we can find a k that will satisfy $g(k) \geq D$. For fixed s and n, given any error $\epsilon = e^{-D}$, we can only find a k that satisfies the previous inequality if $D \leq D_o$, where D_o is as defined above. In other words, for fixed n, we cannot necessarily find a k such that $|R(n, k-1, s)| \leq \epsilon$ holds.

The behavior of Backlund's remainder estimate for fixed n and for fixed k shows that the best possible n and k, which satisfy $|R(n, k-1, s)| \leq \epsilon$, depend on k. This is because for any k we can always find an n that satisfies the previous inequality, but not vice versa. There is one further discussion we need to have regarding n and k, before we will define what exactly is meant by "best" pair (n, k): We have to analyze how the choice of n and k influence the cost of computing $\zeta(s)$ using the E-M series (2.5).

Essentially, the E-M series of $\zeta(s)$ consists of two sums, namely $\zeta_n(s) = \sum_{j=1}^{n-1} \frac{1}{j^s}$ and $\sum_{j=1}^{k} T_j(n, s)$. Let S_n denote the series $\zeta_n(s)$, and S_k denote the sum $\sum_{j=1}^{k} T_j(n, s)$. Then, n gives the number of terms to be used from the series S_n, and k gives the number of terms to be used from the series S_k. Since our goal is to write an efficient algorithm for the computation of $\zeta(s)$ using its E-M series, we need to analyze the work involved in computing S_n and S_k. This will determine how to choose the ideal integers n and k.

There are two important issues to keep in mind. First of all, if n is too large, we are not making efficient use of the E-M series for $\zeta(s)$, because the behavior of $\frac{1}{j^s}$ forces the addition of large and small numbers in the sum S_n. This causes problems in machine

computation, as valuable digits will be lost. On the other hand, if k is too large, we need many relatively expensive calculations of the terms $T_j(n,s)$ for $1 \leq j \leq k$. Lehmer [5] suggests a method in which the real and imaginary parts of $T_j(n,s)$ are computed recursively. However, this uses a lot of storage space and is again a question of machine economics. It is therefore important to find that pair (n,k) which will make efficient use of the behavior of the E-M series of $\zeta(s)$, and will be economical for numerical computation of $\zeta(s)$.

In machine arithmetic, an addition (subtraction) has negligible cost in comparison to a multiplication (division). Other operations such as a logarithm (exponential) or a table look-up cost a fixed amount. Assume we have a table of Bernoulli numbers. Then there is no cost associated with computing β_j. What is the cost of calculating S_k, or rather, of $\sum_{j=1}^{k} \frac{\beta_{2j}}{(2j)!} \frac{s(s+1)\ldots(s+2j-2)}{n^{2j}}$? The cost of computing the first term, $\frac{s\beta_2}{2!n^2}$, is 5 multiplications. Every other term of S_k can be obtained from its previous term by an additional 6 multiplicative operations:

$$\left(\frac{s(s+10)\ldots(s+2j-2)}{(2j)!n^{2j}}\right)\left(\frac{(s+2j-1)(s+2j)}{(2j+1)(2j+2)n^2}\right) = \frac{s(s+1)\ldots(s+2j)}{(2j+2)!n^{2j+2}}$$

Hence, S_k costs roughly $6k$ multiplicative operations to compute. There are some ways to speed up the computation (see Lehmer [5] and Hutchinson [4]) but, ultimately, the cost depends on k. Let us say, therefore, that the cost to compute S_k is pk multiplications, for some $p \geq 1$. The computation of S_n, on the other hand, costs roughly qn multiplications, since each term in the sum costs a fixed number q of multiplications, and there are $n-1$ terms in total.

The result of the above discussion is that the total cost of computing $\zeta(s)$ using its E-M series is roughly $qn+pk$, where q and p are as defined above. Therefore, we need to choose n and k such that $qn+pk$ is minimized. Together with the result of the discussion about the behavior of Backlund's remainder estimate for fixed n and for fixed k, we can

finally give the following definition for the best pair (n, k):

Definition 3 *Let $s \in \mathbb{C}\backslash\{1\}$ and $\epsilon > 0$. For any $k \in \mathbb{N}\backslash\{0\}$, we can find an $n \in \mathbb{N}$ such that $|R(n, k-1, s)| \leq \epsilon$ using Backlund's remainder estimate. The "best" pair (n, k) for the E-M series of $\zeta(s)$ such that $|R(n, k-1, s)| \leq \epsilon$ is defined by that positive integer n, for which $qn + pk$ is the smallest. The positive numbers q and p depend on the cost of computing a single term of S_n and S_k, respectively, and satisfy $p, q \geq 1$.*

Note: In the best pair (n, k), we will refer to n and k as the "best" n and "best" k, respectively.

3.2 Analyzing the Best Pair (n, k)

Let us look at BRE (2.10). We notice that, given s and ϵ, we can solve for n in terms of k, and then we can substitute this expression for n in $qn + pk$ to obtain an expression in k only. Furthermore, minimizing $qn + pk$ is equivalent to minimizing $\alpha n + k$ with $\alpha = \frac{q}{p}$ and $\alpha > 0$, since $p, q \geq 1$. Because the procedure of finding the smallest $\alpha n + k$ does not depend on α, we study the simpler case $\alpha = 1$ at first, and then look at the general case. Since BRE simplifies for $s \in \mathbb{R}$, we consider the real and complex cases separately.

3.2.1 Case I: s is Real, $\alpha = 1$

Throughout this section let $s \in \mathbb{R}$, i. e. $s = \sigma$, and let $\epsilon > 0$. Then Backlund's remainder estimate becomes

$$|R(n, k-1, s)| \leq |T_k(n, s)|\,.$$

We want to simplify $|T_k(n, s)|$ as much as possible without significant loss of accuracy. As in Case I of Section 3.1, we apply estimate (B.24), Stirling's formula for factorials, and

equation (C.25) to Backlund's estimate, and obtain an estimate which holds for $k \geq 3$, since the asymptotic estimate $\frac{|\beta_{2k}|}{(2k)!} \sim \frac{2}{(2\pi)^{2k}}$ is nearly an identity for $k \geq 3$:

$$|T_k(n,s)| \simeq \frac{2}{(2\pi)^{2k}} \frac{1}{n^{s+2k-1}} \frac{\Gamma(s+2k-1)}{\Gamma(s)}.$$

Given ϵ such that $|T_k(n,s)| = \epsilon$, we can solve the above estimate for n and get:

$$n = \left(\frac{2}{\epsilon(2\pi)^{2k}} \frac{\Gamma(s+2k-1)}{\Gamma(s)} \right)^{\frac{1}{s+2k-1}}. \tag{3.11}$$

Note that for now, we must consider $n \in \mathbb{R}^+ \backslash \{0\}$ in order to solve this equation, but shortly, we will return to $n \in \mathbb{N}$. Now, we can substitute the above expression for n in $n+k$, and obtain an expression involving only the variable k. Let us use this result and define the following function:

$$H_r(k) := \left(\frac{2}{\epsilon\Gamma(s)} \frac{\Gamma(s+2k-1)}{(2\pi)^{2k}} \right)^{\frac{1}{s+2k-1}} + k, \tag{3.12}$$

with $k \in \mathbb{R}^+ \backslash \{0\}$, and s and ϵ are parameters. We also require $k \geq 1$, since this function is developed from $R(n, k-1, s)$. In order to identify the best k, we need to minimize H_r. First, let us determine if H_r does in fact have a minimum.

Since s and ϵ are fixed, the term that embodies the dependence of H_r on k is $\frac{\Gamma(s+2k-1)}{(2\pi)^{2k}}$. Now, $\Gamma(s+2k-1)$ grows factorially in k by the definition of the Gamma function (see Appendix C), and $(2\pi)^{2k}$ grows exponentially in k. As a result we can show that there exists a $k_o \in \mathbb{R}$ such that for $k < k_o$ we have $\Gamma(s+2k-1) < (2\pi)^{2k}$, and for $k > k_o$ we have $\Gamma(s+2k-1) > (2\pi)^{2k}$. This implies that for increasing k, H_r first decreases to a minimum, and increases monotonically thereafter. Figures 3.3, 3.4, and 3.5 depict H_r for different s and ϵ. The graphs show that the general shape of H_r is independent of the choice for s and ϵ.

Since H_r has a minimum, we are able to find the best pair (n,k). Let k_{min} denote that k at which H_r has its minimum, i. e. k_{min} is the solution to $\frac{dH_r(k)}{dk} = 0$. Then the

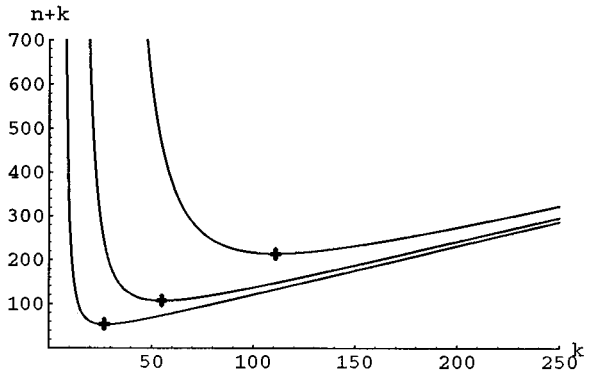

Figure 3.3: $H_r(k)$ $(= n + k)$ for $s = 3$ and $\epsilon_1 = 10^{-50}$, $\epsilon_2 = 10^{-100}$, and $\epsilon_3 = 10^{-200}$. The respective minima are at $(27, 53.8291)$, $(55, 107.272)$, and $(111, 214.346)$.

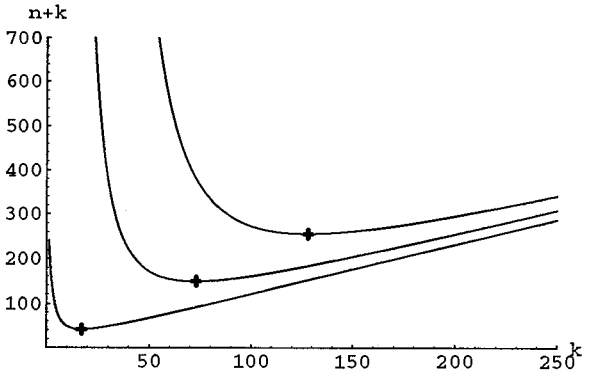

Figure 3.4: $H_r(k)$ $(n + k)$ for $s = 20$ and $\epsilon_1 = 10^{-50}$, $\epsilon_2 = 10^{-150}$, and $\epsilon_3 = 10^{-250}$. The respective minima are at $(17, 41.8848)$, $(73, 148.856)$, and $(128, 255.962)$.

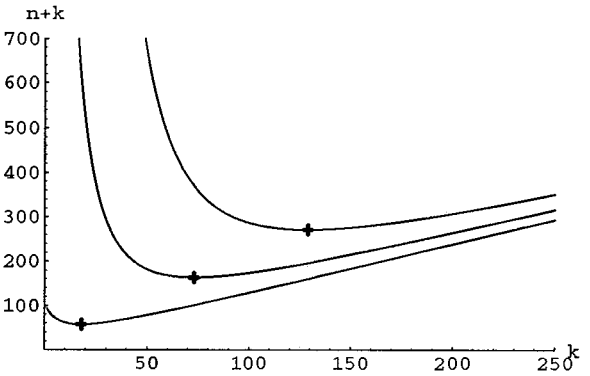

Figure 3.5: $H_r(k)$ $(= n + k)$ for $s = 50$ and $\epsilon_1 = 10^{-100}$, $\epsilon_2 = 10^{-200}$, and $\epsilon_3 = 10^{-300}$. The respective minima are at $(18, 57.0659)$, $(73, 148.1641)$, and $(129, 271.227)$.

best n, denote it by n_{min}, is given by $H_r(k_{min}) - k_{min}$. However, we require $n, k \in \mathbb{N}$, so let the best pair (n, k) be given by $(\lceil n_{min} \rceil, \lceil k_{min} \rceil)$. The remaining issue in finding k_{min} is solving $\frac{dH_r(k)}{dk} = 0$. Because of the complicated dependence of H on k in equation (3.12), an analytic expression for the minimum of H_r will be difficult to find (if it is even possible). Hence, we will instead utilize a numerical technique such as Newton's method; but considering the complicated forms of the derivatives of H_r, the secant method may be more appropriate.

Now that we can actually find the best pair (n, k), let us take a closer look at the sum of the best n and the best k for different s and ϵ. Examining the minima of Figures 3.3, 3.4, and 3.5, we notice a curious phenomenon: The value of $n + k$ is approximately twice the size of the best k; in other words, the best n is almost the same size as the best k. We computed the best n and the best k for different s as ϵ decreases, and depicted their graphs in Figure 3.6. These graphs show that the best n and the best k are almost identical. If this were true in general, it would allow us to replace n by k in $|T_k(n, s)|$. Then, BRE involves only one variable, namely k. This means that we can solve BRE for k given s and ϵ without having to find the best n and the best k. Therefore, we will investigate the relationship between the best pair (n, k) and the k found by taking $n = k$ in BRE.

Let s be fixed. We will compute the values for the best pair (n, k) as a function of ϵ, as well as those k which satisfy $|R(k, k-1, s)| \leq \epsilon$ according to BRE, and then compare the two results. To that effect, we define the following two functions:

$$f(\epsilon) := H_r(k_{min})|_\epsilon, \tag{3.13}$$

$$g(\epsilon) := k|_{|T_k(k,s)| \leq \epsilon}, \tag{3.14}$$

where $\epsilon > 0$ and k_{min} is that number which solves $\frac{dH_r(k)}{dk} = 0|_\epsilon$. In other words, $f(\epsilon)$ is the sum of the best n and the best k as a function of ϵ, and $g(\epsilon)$ is the smallest k for

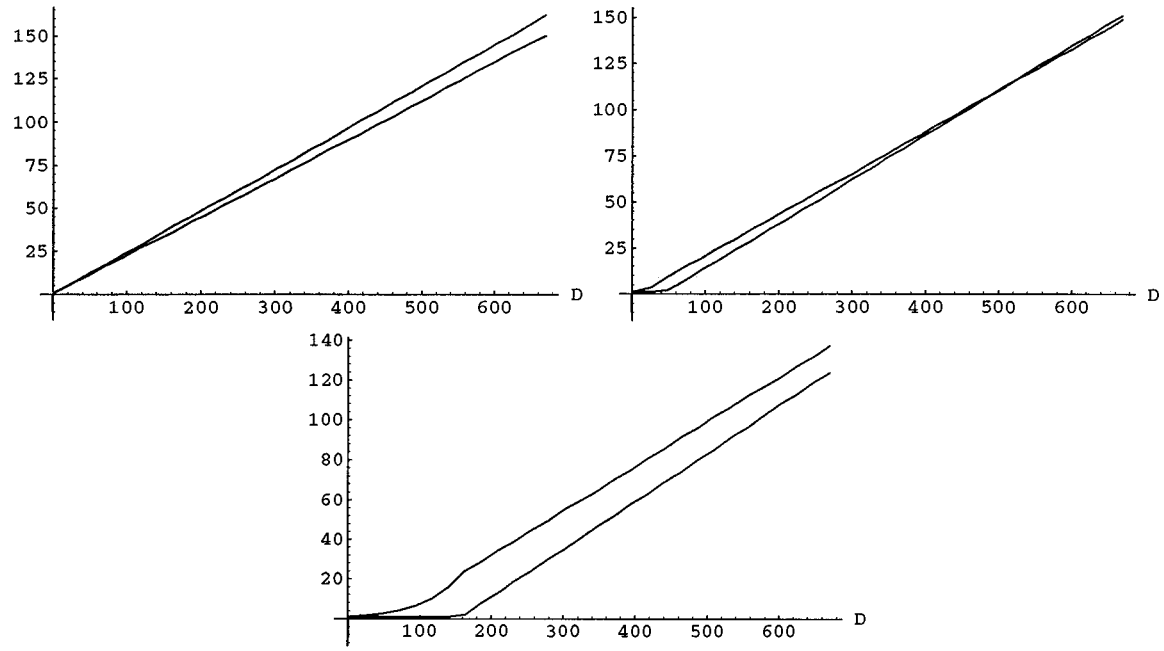

Figure 3.6: Plots of best n together with best k for $s = 3, 20$, and 50, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

which $|R(k, k-1, s)| \leq \epsilon$ using Backlund's remainder estimate. Since we want to show that we can choose $n = k$ in the Backlund's remainder estimate, we will compare $f(\epsilon)$ to $2g(\epsilon)$, i. e. $n + k$ to $2k$. Figure 3.7 depicts the graphs of $f(\epsilon)$ and $2g(\epsilon)$ for different s.

We observe that for decreasing ϵ, i. e. higher precision of $\zeta(s)$, the curves of f and $2g$ are nearly identical. This further supports the choice $n = k$ in Backlund's remainder estimate. In order to get a clearer picture of how f and $2g$ differ, we compute their ratio. We suspect that as $\epsilon \to 0$ we have $\frac{f(\epsilon)}{2g(\epsilon)} \to 1$. This conjecture is supported by Figure 3.8, which depicts the ratio $\frac{f}{2g}$ for the same values of s used in Figure 3.7. Consequently, it seems reasonable to let n equal to k in BRE, obtaining the same accuracy as if using the best n and the best k. In this way, we are able to reduce the number of variables in our problem from two to one, thereby significantly reducing the amount of computational effort required.

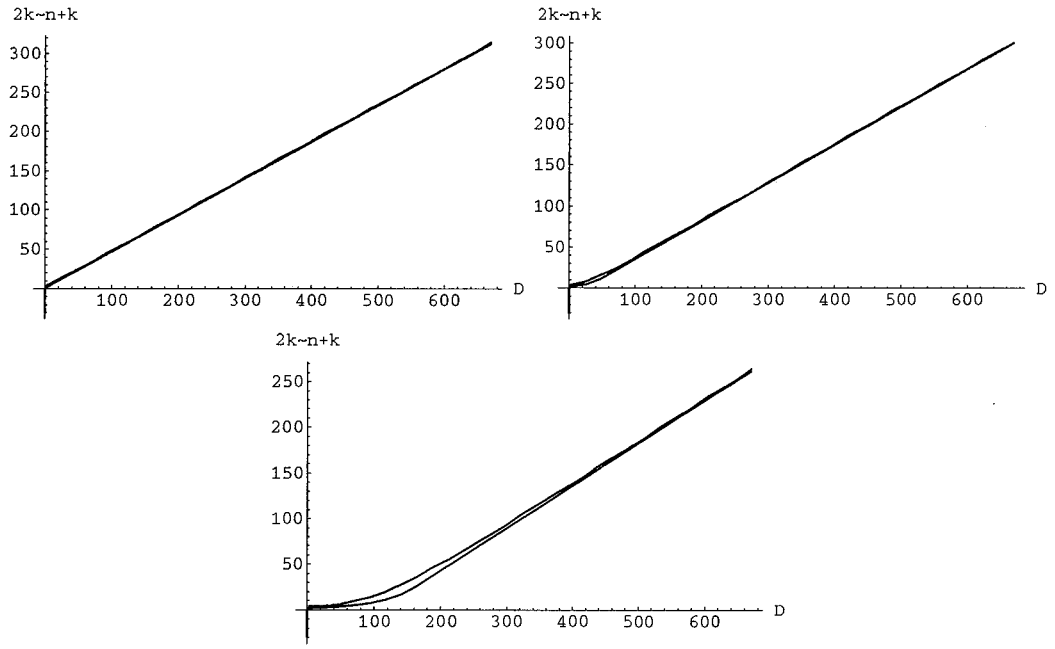

Figure 3.7: The graphs of $f(\epsilon)$ and $2g(\epsilon)$ for $s = 3, 20$, and 50, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

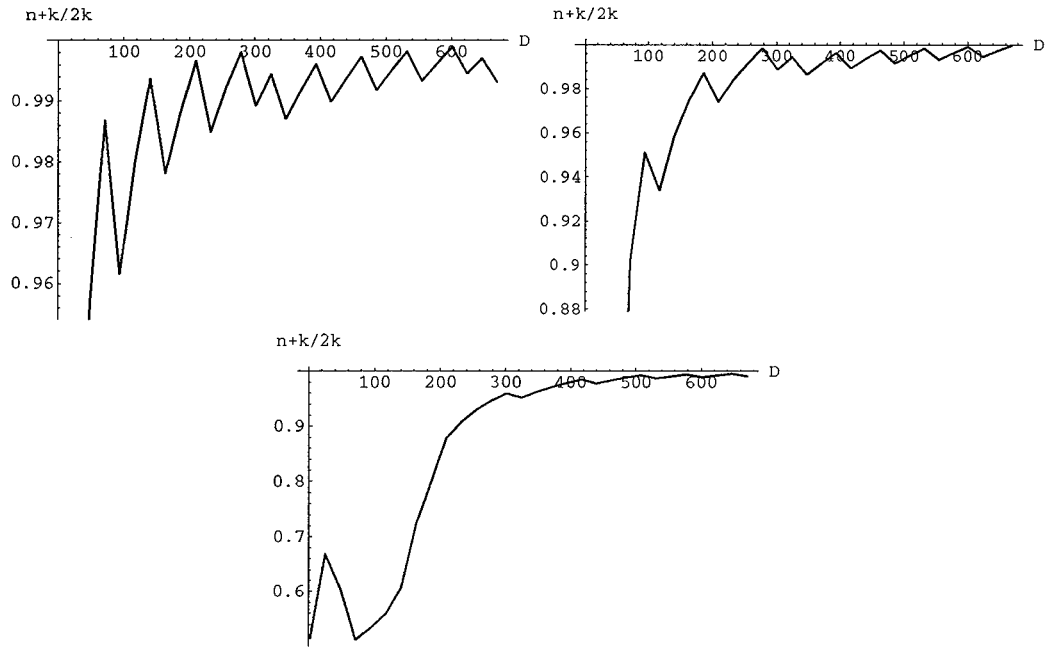

Figure 3.8: The graphs of the ratio $\frac{f(\epsilon)}{2g(\epsilon)}$ for $s = 3, 20$, and 50, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

Before we summarize the important result of this section, we will consider the case when s is complex.

3.2.2 Case II: s is Complex, $\alpha = 1$

Throughout this section let $s = \sigma + it \in \mathbb{C}$, and let $\epsilon > 0$. Remember, we want to solve BRE for n, and substitute the resulting expression in $n + k$ in order to get an expression in k only. Again, we need to simplify $|T_k(n, s)|$ as much as possible without significant loss of accuracy. Unlike the real case, $\frac{\Gamma(s+2k-1)}{\Gamma(s)} \in \mathbb{C}$, so we need to take the absolute value of this term. For computational reasons, we also need to estimate $|\Gamma(s+2k-1)|$. This is because, when trying to minimize $n + k$, we need to take the derivative of $|\Gamma(s+2k-1)|$ with respect to k, and Mathematica [6] attempts to take the derivative of the absolute value of this term, which does not exist except at zero for complex arguments. However, the derivative of this function with respect to the real variable k does exist everywhere. Therefore, we will apply estimate (B.24), Stirling's formula for factorials, equation (C.25) as well as Stirling's formula for the Gamma function (see Appendix C) to $|T_k(n, s)|$. We get:

$$\begin{aligned} |T_k(n,s)| &\sim \frac{2}{(2\pi)^{2k}} \frac{1}{n^{\sigma+2k-1}} \frac{1}{|\Gamma(s)|} \left| \sqrt{\frac{2\pi}{s+2k-1}} \left(\frac{s+2k-1}{e} \right)^{s+2k-1} \right| \\ &= \frac{2\sqrt{2\pi}}{(2\pi)^{2k}} \frac{1}{n^{\sigma+2k-1}} \frac{|s+2k-1|^{\sigma+2k-\frac{3}{2}} e^{-\Im s \arg s}}{|\Gamma(s)| e^{\sigma+2k-1}}. \end{aligned}$$

Then, given ϵ we can solve this estimate of BRE for n. Unlike the real case, we have the additional term $\left|\frac{s+2k-1}{\sigma+2k-1}\right|$. We get for n:

$$n = \left(\frac{2\sqrt{2\pi}}{\epsilon |\Gamma(s)| e^{\Im(s)\arg(s)+s-1}} \frac{|s+2k-1|^{\sigma+2k-\frac{1}{2}}}{(2\pi e)^{2k}} \right)^{\frac{1}{\sigma+2k-1}}, \tag{3.15}$$

with $n \in \mathbb{R}^+ \backslash \{0\}$. Substituting the above expression for n in $n + k$, we can define the following function:

$$H_c(k) := \left(\frac{2\sqrt{2\pi}}{\epsilon \, |\Gamma(s)| \, e^{\Im(s)\arg(s)+s-1}} \frac{|s+2k-1|^{\sigma+2k-\frac{1}{2}}}{(2\pi e)^{2k}} \right)^{\frac{1}{\sigma+2k-1}} + k \qquad (3.16)$$

$$= \frac{|s+2k-1|}{e} \left(\frac{2\sqrt{2\pi}}{\epsilon \, |\Gamma(s)| \, e^{\Im(s)\arg(s)}} \frac{\sqrt{|s+2k-1|}}{(2\pi e)^{2k}} \right)^{\frac{1}{\sigma+2k-1}} + k,$$

with $k > 0$, and s and ϵ are parameters. Now, we need to minimize H_c in order to identify the best k. So again, we need to determine if H_c does in fact have a minimum.

Since s and ϵ are parameters, the dependence of H_c on k is embodied by the fraction $\frac{|s+2k-1|^{\sigma+2k-1}}{(2\pi e)^{2k}}$. The term $|s + 2k - 1|^{\sigma+2k-1}$ grows factorially in k, and $(2\pi e)^{2k}$ grows exponentially in k. Similar to the real case, this implies that for increasing k, H_c first decreases to a minimum, and increases monotonically thereafter. Figures 3.9, and 3.10 depict H_c for different s and ϵ, and show that the general shape of H_c is independent of the choice for s and ϵ.

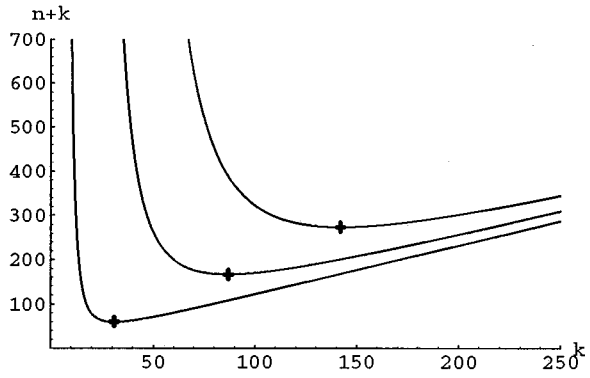

Figure 3.9: $H_c(k)$ $(= n + k)$ for $s = \frac{1}{2} + 10i$ and $\epsilon_1 = 10^{-50}$, $\epsilon_2 = 10^{-150}$, and $\epsilon_3 = 10^{-250}$. The respective minima are at $(31, 59.7432)$, $(87, 166.976)$, and $(142, 274.14)$.

As with the real case, H_c has a minimum and we let the best pair (n, k) be given by $(\lceil n_{min} \rceil, \lceil k_{min} \rceil)$, where k_{min} is the solution to $\frac{dH_c(k)}{dk} = 0$ and $n_{min} = H_c(k_{min}) - k_{min}$.

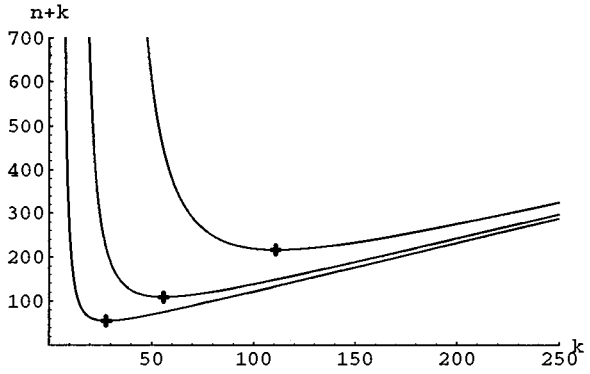

Figure 3.10: $H_c(k)$ $(= n + k)$ for $s = 5 + 9i$ and $\epsilon_1 = 10^{-50}$, $\epsilon_2 = 10^{-100}$, and $\epsilon_3 = 10^{-200}$. The respective minima are at $(28, 55.997)$, $(56, 109.609)$, and $(111, 216.763)$.

Since equation (3.16) is even more complicated than equation (3.12), we will again use a numerical method to find an analytic expression for the minimum of H_c.

Our task now is to check if the best pair (n, k) behaves the same way as in the real case. The minimum of Figures 3.9, and 3.10 seem to indicate that the best n is again about the same size as the best k. Figure 3.11 depict the graphs of the best n and the best k for different s as ϵ decreases, and show that the graphs in each of these Figures are almost identical. Therefore, we continue investigating the relationship between the best pair (n, k) and the k found by taking $n = k$ in BRE in the same manner as the real case.

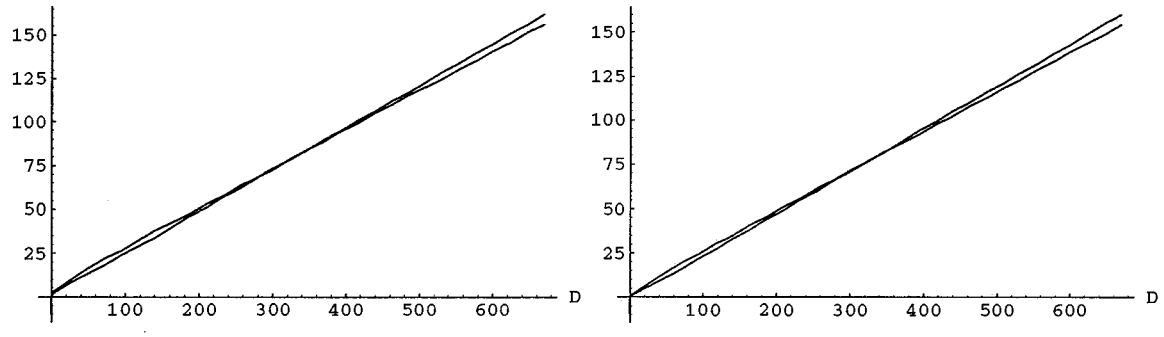

Figure 3.11: Plots of best n together with best k for $s = \frac{1}{2} + 10i$ and $s = 5 + 9i$, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

Let s be fixed. Similar to equation (3.14) we define the following function

$$g(\epsilon) \;:=\; k\Big|_{\left|\frac{s+2k-1}{\sigma+2k-1}\right||T_k(k,s)|\leq\epsilon},$$

and we let f be the same function as defined by equation (3.13) except we replace H_r by H_c. Figure 3.12 shows the graphs of $f(\epsilon)$ together with $2g(\epsilon)$ for different s. This way, we are comparing $n+k$ to $2k$, and notice that for decreasing ϵ the curves of f and $2g$ are nearly identical. Furthermore, the ratio $\frac{f}{2g}$, which is depicted in Figure 3.13 for the same values of s as in Figure 3.12, supports our previous conjecture that $\frac{f(\epsilon)}{2g(\epsilon)} \to 1$ as $\epsilon \to 0$. Therefore, we can draw the same conclusion as in the real case, namely, that it seems reasonable to let n equal to k in BRE.

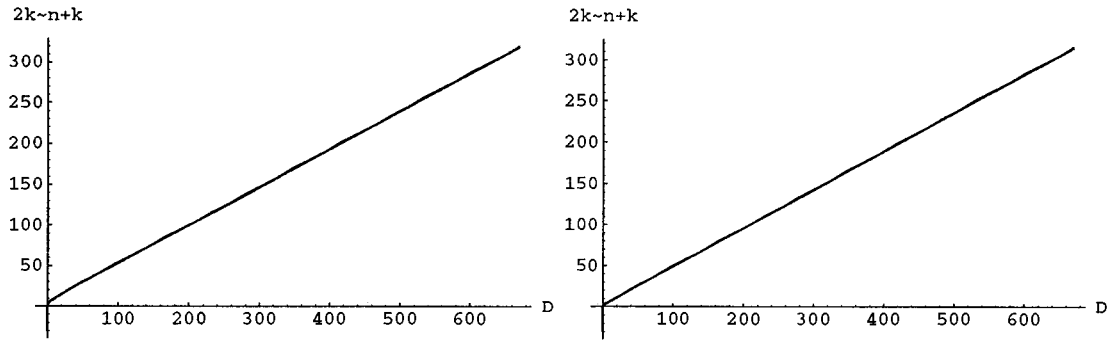

Figure 3.12: The graphs of $f(\epsilon)$ and $2g(\epsilon)$ for $s = \frac{1}{2} + 10i$ and $s = 5 + 9i$, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

Now, let us finally summarize the results from Sections 3.2.1 and 3.2.2.

3.2.3 Method for Finding the Best Pair (n, k) for $\alpha > 0$

Our original question was, given s and ϵ, what is the smallest $\alpha n + k$ that will give us $\zeta(s)$ to within ϵ using the E-M series of $\zeta(s)$? In Sections 3.2.1, and 3.2.2 we discussed the case $\alpha = 1$ for real and complex arguments s, respectively. As a result of that discussion, we found a method of finding the smallest $n+k$, which works equally well for

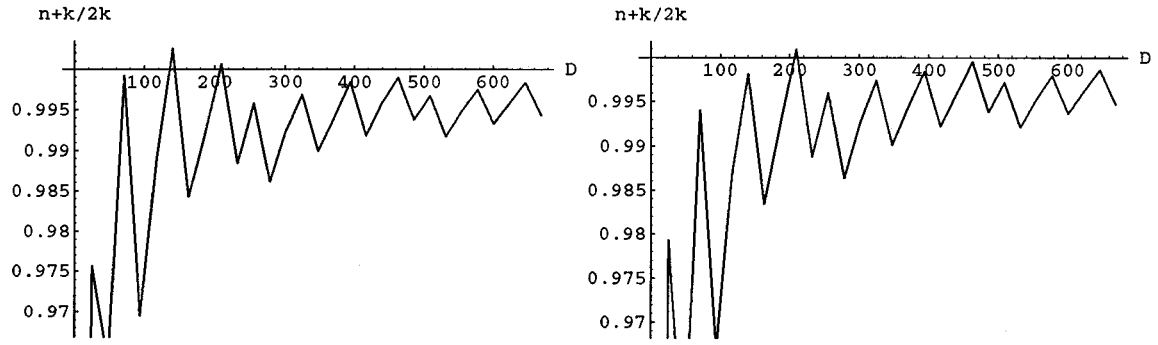

Figure 3.13: The graphs of the ratio $\frac{f(\epsilon)}{2g(\epsilon)}$ for $s = \frac{1}{2} + 10i$ and $s = 5 + 9i$, respectively, where $\epsilon = e^{-D}$ with $1 \leq D \leq 700$.

the real and complex cases, and can be summarized as follows. In essence, we simplified BRE by estimating it further applying estimate (B.24), Stirling's formula for factorials, equation (C.25), and in the complex case also applying Stirling's formula for the Gamma function (see Appendix C); solved the new estimate for n given ϵ; and substituted the resulting expression in $n + k$. Thereby, we obtained an expression in k only, and were able to employ a numerical technique for finding the minimum. In order to answer our original question, we generalize this method in the following way. Based on equations (3.11), and (3.15) write

$$n(k) := \begin{cases} \left(\frac{2}{\epsilon\Gamma(s)} \frac{\Gamma(s+2k-1)}{(2\pi)^{2k}}\right)^{\frac{1}{s+2k-1}}, & s \in \mathbb{R}, \\ \left(\frac{2\sqrt{2\pi}}{\epsilon|\Gamma(s)|e^{\Im(s)\arg(s)+s-1}} \frac{|s+2k-1|^{\sigma+2k-\frac{1}{2}}}{(2\pi e)^{2k}}\right)^{\frac{1}{\sigma+2k-1}}, & s \in \mathbb{C}, \end{cases}$$

and define $H_\alpha(k) := \alpha n(k) + k$ with $\alpha > 0$, so that

$$H_1(k) = \begin{cases} H_r(k), & s \in \mathbb{R}, \\ H_c(k), & s \in \mathbb{C}, \end{cases}$$

where H_r and H_c are as defined in Sections 3.2.1, and 3.2.2, respectively. By solving $\frac{dH_\alpha(k)}{dk} = 0$ using a numerical technique in the same way as before, we obtain the best k, which in turn gives us the best n by evaluating $n(bestk)$. So, we have demonstrated a practical method of finding the best pair (n, k) for arbitrary $\alpha > 0$.

3.2.4 Conjecture 1

Now that we have a working method for finding the best n and k, let us pick up from the intriguing result of Sections 3.2.1, and 3.2.2 about the best pair (n, k) for $\alpha = 1$. For various real and complex arguments s, we demonstrated via careful analysis and detailed graphing that best $n \simeq$ best k. This led us to substitute k for n in BRE. In order to verify the validity of this substitution, we looked at the ratio $\frac{best\ n + best\ k}{2k}$, where k is that number which gives $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(k,s)| \leq \epsilon$. We noticed that this ratio tends to 1 as $\epsilon \to 0$, and therefore were sufficiently convinced that taking $n = k$ in BRE and solving for k given s and ϵ will yield the best k as $\epsilon \to 0$.

We want to know if this result can be generalized to all $\alpha > 0$. To this end, we compute the ratio $\frac{best\ n}{best\ k}$ for different α at a number of s values as ϵ decreases. The scatter plot of Figure 3.14 shows this ratio. As ϵ decreases, this plot strongly suggests that the ratio converges for each α. This leads us to believe that there is a relationship between best n and best k, and we are prepared to make the following conjecture.

Conjecture 1 *Let $s \in \mathbb{C}\backslash\{1\}$. Let p, q be the cost of computing the terms $T_j(n,s)$ and $\frac{1}{j^s}$, respectively, for any $j \in \mathbb{N}$. Given $\epsilon > 0$, let (n_ϵ, k_ϵ) be the best pair (n,k) which solves $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(n,s)| \leq \epsilon$. Then there exists a real constant c, $0 < c < 1$, such that*

$$\lim_{\epsilon \to 0} n_\epsilon = \mu(p,q)\ k_\epsilon,$$

where

$$\mu(p,q) = (1-c)\frac{p}{q} + c.$$

Conjecture 1 has the following implications for BRE: It allows us to reduce the number of unknowns in BRE from two to one, so that we get the new E-M remainder estimate

$$|R(n, k-1, s)| \leq \left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(\mu(p,q)k, s)|,$$

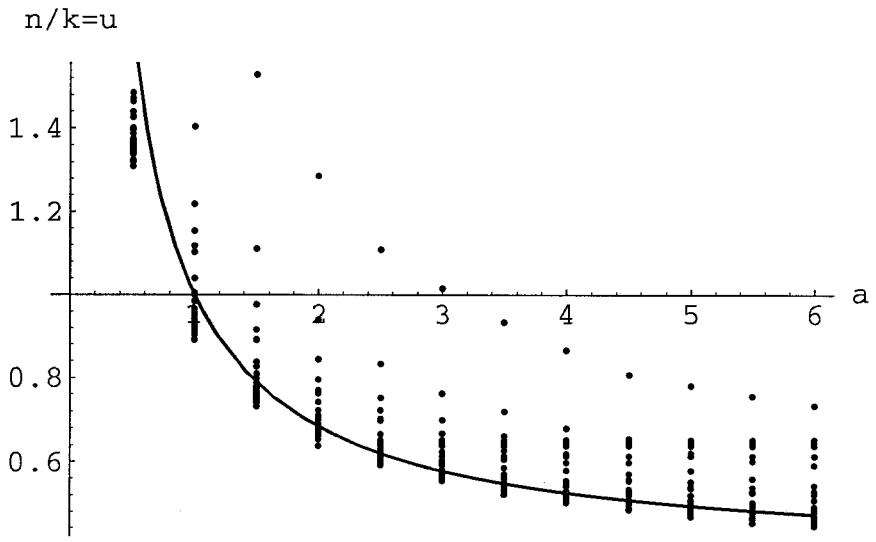

Figure 3.14: Scatter plot of the ratio $\frac{best\ n}{best\ k}$ as a function of α. At $\alpha = \frac{1}{2}, 1, \frac{3}{2}, \ldots, 6$ the ratio has been computed for the following s and ϵ values: $s = 3, 20, 50, \frac{1}{2} + 2i, \frac{1}{2} + 10i, \frac{1}{2} + 40i, 2 + 3i, 5 + 9i$, and $10 + 15i$; and $\epsilon = e^{-D}$ with $D = 230, 350, 460, 575$, and 700. The smooth curve is the graph of $\mu(\alpha) = (1 - \frac{1}{e})\frac{1}{\alpha} + \frac{1}{e}$.

which depends only on k. Furthermore, solving for k given s and ϵ, which can be done employing a numerical technique, ensures that the solution is the best possible k.

Consider the choice $c = \frac{1}{e}$, and remember that $\alpha = \frac{q}{p}$. Let us write $\mu(\alpha) = \mu(p, q)$. From Figure 3.14 we can see how well the graph of $\mu(\alpha)$ fits onto the scatter plot at the points of convergence for each α. How accurate is this choice of c? Figure 3.15 shows the graphs of best n together with $\mu(\alpha)$ *best* k for some arbitrarily chosen values of s and different α; and Figure 3.16 shows the ratio $\frac{\alpha\ best\ n + best\ k}{(\mu(\alpha)\alpha+1)k}$, for the previous values of s and α, and where k is again that number which gives $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(\mu(\alpha)k, s)| \leq \epsilon$. Since the plots of Figure 3.16 indicate that the ratio tends to a value near 1 as $\epsilon \to 0$, the chosen value, $\frac{1}{e}$, for c must be close to the actual value of c.

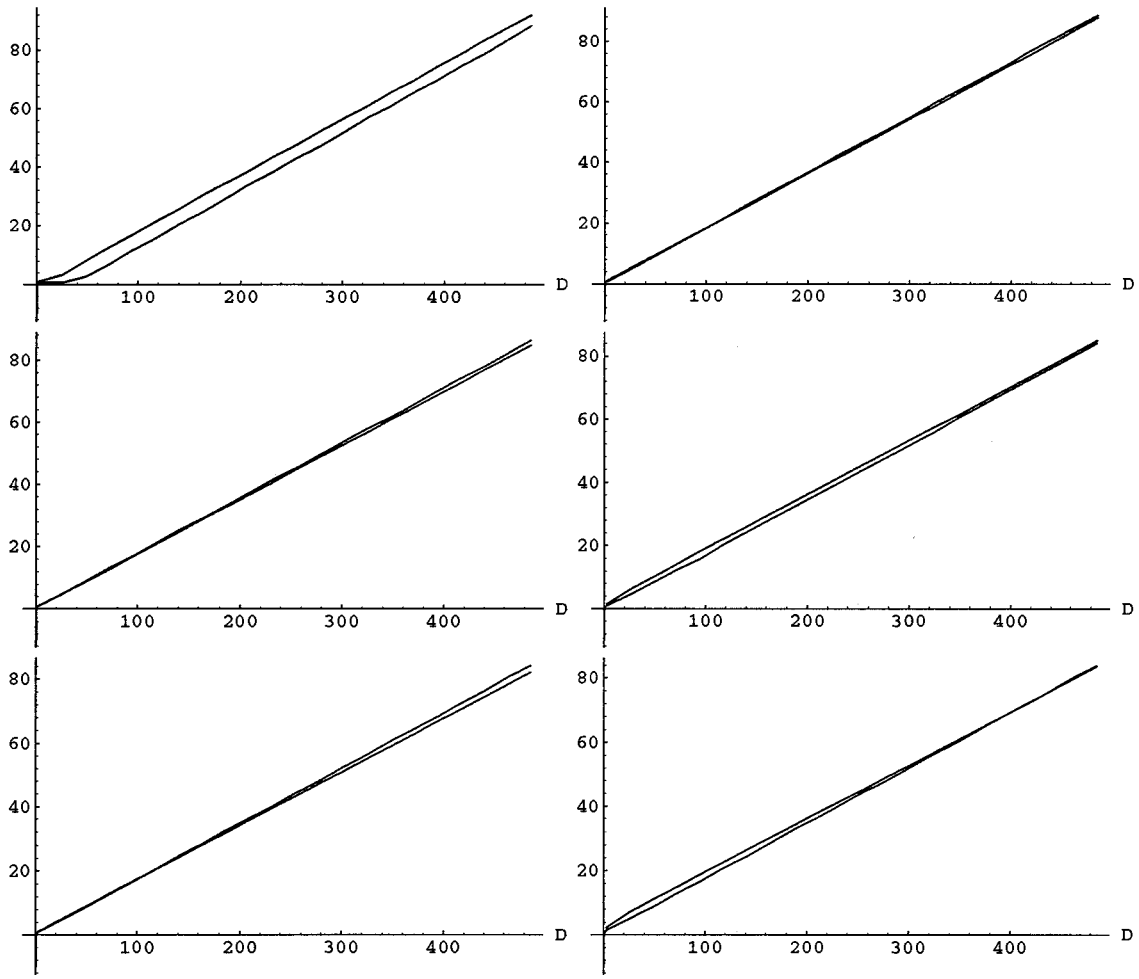

Figure 3.15: Plots of best n together with $\left(\left(1-\frac{1}{e}\right)\frac{1}{\alpha}+\frac{1}{e}\right)$ best k for $(s,\alpha)=(20,2)$, $(\frac{1}{2}+2i,3)$, $(3,4)$, $(5+9i,5)$, $(2+3i,6)$, and $(\frac{1}{2}+10i,7)$, where $\epsilon=e^{-D}$ with $1\leq D\leq 500$.

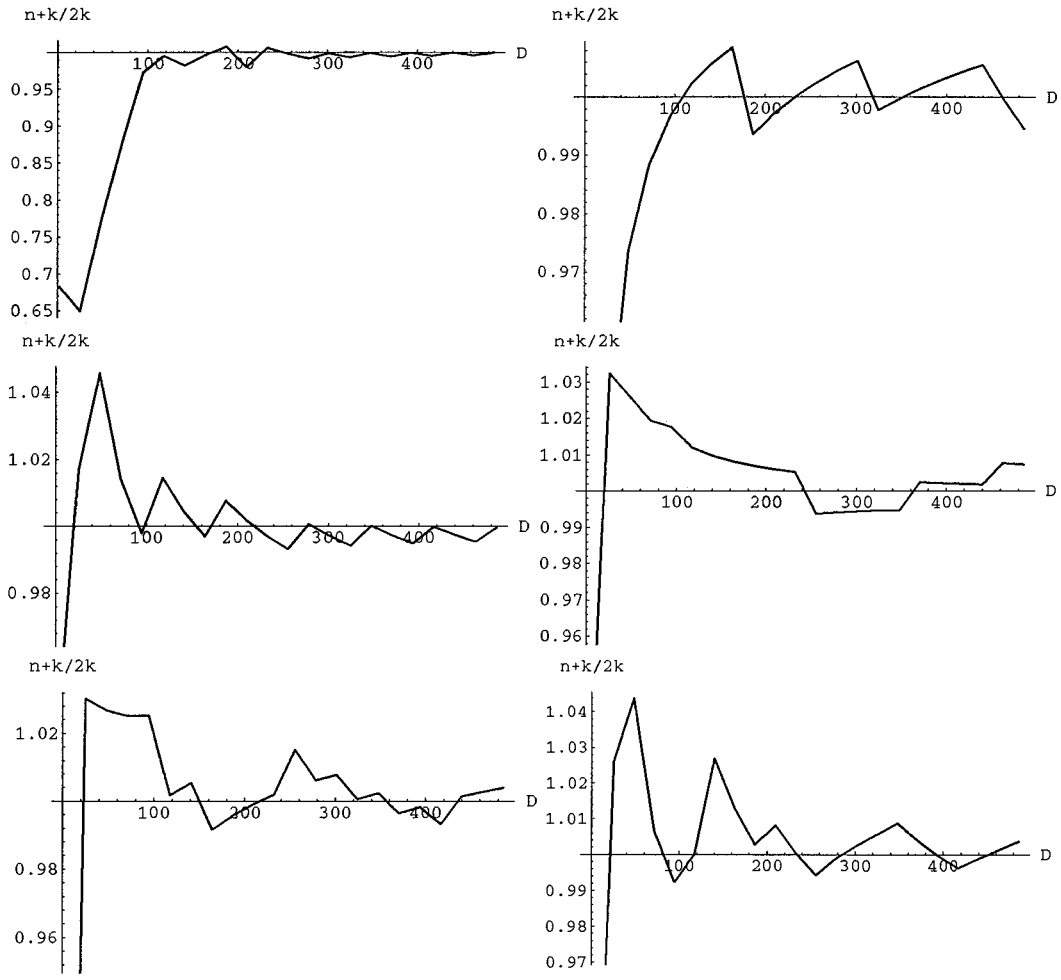

Figure 3.16: The graphs of the ratio $\frac{\alpha\ best\ n+best\ k}{(\mu(\alpha)\alpha+1)k}$ for $(s,\alpha) = (20,2)$, $(\frac{1}{2}+2i,3)$, $(3,4)$, $(5+9i,5)$, $(2+3i,6)$, and $(\frac{1}{2}+10i,7)$, where $\epsilon = e^{-D}$ with $1 \leq D \leq 500$.

3.3 The Smallest Pair (n, k)

In Section 3.1 we have defined that pair (n, k), which gives us the minimum number of terms to be used in the E-M series (2.5) for the computation of $\zeta(s)$, and we denoted this the best pair. Now, we want to introduce a different pair (n, k), which we shall call "smallest" pair, and discuss a possible relationship between the best and smallest pairs.

By the behavior of BRE for fixed k and s from Section 3.1, we know that the terms $T_k(n, s)$ first decrease to some best error, and then grow large again. The turning point of growth is given by the growth ratio of Section 2.3, which tells us that the size of the smallest term can be established with the relationship $|s + 2k| \simeq 2\pi n$. Let us find the magnitude of the smallest term by applying the growth ratio to BRE. As before, we also utilize estimate (B.24), Stirling's formula for factorials, equation (C.25), as well as Stirling's formula for the Gamma function (see Appendix C). We get

$$
\begin{aligned}
|R(n, k-1, s)| &\leq \left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(n, s)| \\
&\sim \frac{|s+2k-1|}{\sigma+2k-1} \frac{2}{(2\pi)^{2k}} \frac{|\Gamma(s+2k-1)|}{|\Gamma(s)|} \frac{1}{n^{\sigma+2k-1}} \\
&\sim \frac{|s+2k-1|}{\sigma+2k-1} \frac{2}{(2\pi)^{2k}} \frac{1}{|\Gamma(s)|} \left|\sqrt{\frac{2\pi}{s+2k-1}} \left(\frac{s+2k-1}{e}\right)^{s+2k-1}\right| \frac{1}{n^{\sigma+2k-1}} \\
&= \frac{\sqrt{|s+2k-1|}}{\sigma+2k-1} \frac{2}{(2\pi)^{2k-\frac{1}{2}}} \frac{1}{|\Gamma(s)|} \left(\frac{|s+2k-1|}{e}\right)^{\sigma+2k-1} \frac{e^{-t\arg(s+2k-1)}}{n^{\sigma+2k-1}} \\
&\sim \frac{\sqrt{|s+2k-1|}}{\sigma+2k-1} \frac{2}{(2\pi)^{2k-\frac{1}{2}}} \frac{1}{|\Gamma(s)|} \left(\frac{2\pi n}{e}\right)^{\sigma+2k-1} \frac{e^{-t\arg(s+2k-1)}}{n^{\sigma+2k-1}} \\
&= \frac{\sqrt{|s+2k-1|}}{\sigma+2k-1} \frac{2(2\pi)^{\sigma-\frac{1}{2}}}{|\Gamma(s)|} \frac{1}{e^{(\sigma+2k-1)+t\arg(s+2k-1)}}.
\end{aligned}
$$

Amazingly, this estimate depends only on k, so given s and ϵ we can solve for k employing a numerical technique, and n is given by $\frac{|s+2k|}{2\pi}$ according to the growth ratio. Based on this result, let us make the following definition.

Definition 4 *Let $s \in \mathbb{C}\backslash\{1\}$, and $\epsilon > 0$. The "smallest" pair (n,k) for the E-M series of $\zeta(s)$ such that $|R(n,k-1,s)|| \leq \epsilon$ is defined by the smallest positive integer k that satisfies*

$$\frac{\sqrt{|s+2k-1|}}{\sigma+2k-1}\frac{2(2\pi)^{\sigma-\frac{1}{2}}}{|\Gamma(s)|}\frac{1}{e^{(\sigma+2k-1)+t\arg(s+2k-1)}} \leq \epsilon, \tag{3.17}$$

and then

$$n := \left\lceil \frac{|s+2k|}{2\pi} \right\rceil.$$

Note: In the smallest pair (n,k), we will refer to n and k as the "smallest" n and "smallest" k, respectively.

We are interested, if there might be a relationship between the best and smallest pairs that will help us to reduce computational cost. A possible relationship does not ease computations any further if we compare solving equation (3.17) with solving equation $\left\lceil \left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(\mu(p,q)k,s)| \right\rceil = \epsilon$; however, both of these equations are less demanding to solve numerically than $\frac{dH_\alpha(k)}{dk} = 0$ from our working method for finding best n and best k. This is because the derivative $\frac{dH_\alpha(k)}{dk}$ is more complicated and involves the Polygamma function. So, just as with Conjecture 1 we can simplify the computations involved in finding best n and best k if their is a relationship between the best and smallest pairs. Therefore, we compute the ratio $\frac{smallest\ k}{best\ k}$ for different $\alpha > 0$ at a number of s values as ϵ decreases. Figure 3.17 depicts this ratio in a scatter plot. This plot provides strong evidence that the ratio converges for each α as ϵ decreases. Just as in Section 3.2.4, this leads us to believe that there is a relationship between smallest k and best k, and so we make the following conjecture.

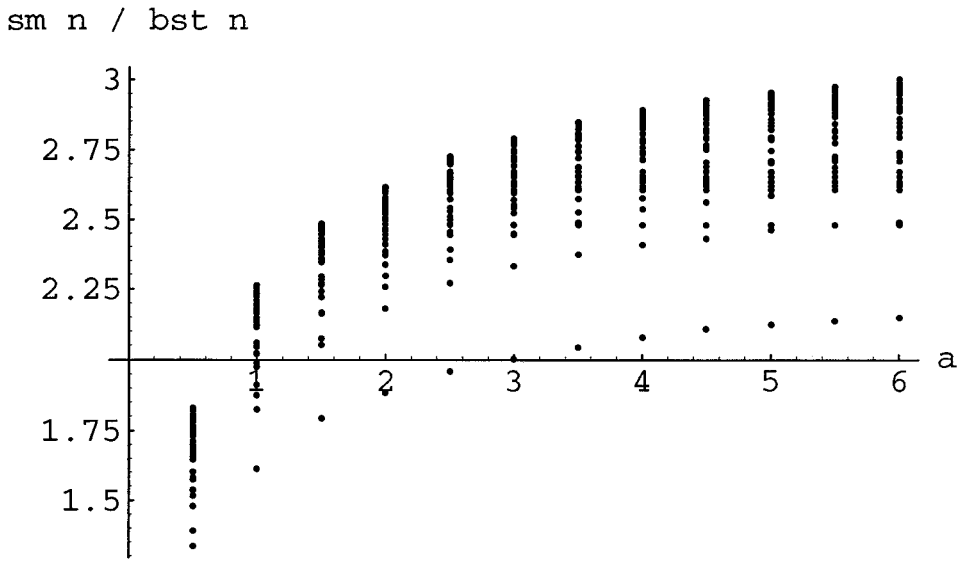

Figure 3.17: Scatter plot of the ratio $\frac{\textit{smallest } k}{\textit{best } k}$ as a function of α. At $\alpha = \frac{1}{2}, 1, \frac{3}{2}, \ldots, 6$ the ratio has been computed for the following s and ϵ values: $s = 3, 20, 50, \frac{1}{2} + 2i, \frac{1}{2} + 10i, \frac{1}{2} + 40i, 2 + 3i, 5 + 9i$, and $10 + 15i$; and $\epsilon = e^{-D}$ with $D = 230, 350, 460, 575$, and 700. The smooth curve is the graph of $\mu(\alpha) = (1 - \frac{1}{e})\frac{1}{\alpha} + \frac{1}{e}$.

Conjecture 2 *Let $s \in \mathbb{C}\backslash\{1\}$, and let p, q be the cost of computing the terms $T_j(n, s)$ and $\frac{1}{j^s}$, respectively, for any $j \in \mathbb{N}$. Given $\epsilon > 0$, let k_ϵ^b denote the best k, and let k_ϵ^s denote the smallest k. Then there exists a real function ν dependent on p, and q only such that*

$$\lim_{\epsilon \to 0} k_\epsilon^s = \nu(p, q) k_\epsilon^b.$$

Note: Conjectures 1 and 2 also imply a relationship between best n and smallest n:

Given $\epsilon > 0$, let n_ϵ^b denote the best n, and let n_ϵ^s denote the smallest n. Then

$$\lim_{\epsilon \to 0} n_\epsilon^s = \frac{\left| s + 2\,\frac{\nu(p,q)}{\mu(p,q)}\, n_\epsilon^b \right|}{2\pi}.$$

The result of this section has not given us any other insight than that of reducing the computational cost involved in finding best n and best k over the method given in

Section 3.2.3. However, Conjecture 2 involves the same amount of work in finding best n and best k as Conjecture 1; and since we justified Conjecture 1 in more detail, we will only employ the result of Conjecture 1 in our algorithm.

Chapter 4

Algorithm and Conclusion

4.1 Presentation and Discussion of Algorithm for Computing $\zeta(s)$

We finally outline an algorithm for the computation of the Riemann zeta function using the E-M series (2.5), which is based on the method described in Section 3.2.3:

1. **begin Zeta**(s, ϵ)
   (a) **initialize** p, q respectively
   (b) **if** $s = 0$ **then**
      i. **return** $-\frac{1}{2}$
   (c) **if** $s = 1$ **then**
      i. **return** ∞
   (d) **if** $\Re(s) < 0$ **then**
      i. **return** $2(2\pi)^{s-1} \sin\left(\frac{s\pi}{2}\right) \Gamma(1-s)$ **Zeta**$(1-s, \epsilon)$
   (e) **else**
      i. **let** $\alpha = \frac{q}{p}$
      ii. **let** k be that positive integer closest to the solution of $\frac{dH_\alpha(k)}{dk} = 0$
      iii. **let** $n = \lceil n(k) \rceil$
      iv. **return** $\zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} + \sum_{j=1}^{k} T_j(n, s)$
   (f) **end if**
2. **end Zeta**

Notes: The following is a list of explanations about variables and functions involved in the above algorithm:

- $s \in \mathbb{C}$ is the argument for the Riemann zeta function.
- $\epsilon > 0$ is the accuracy with which we want to compute $\zeta(s)$.
- p, q denote the cost of computing $T_j(n, s)$ and $\frac{1}{j^s}$, respectively.
- The Gamma function, Γ, on line 1.d.i. is built into most of the common computer languages, but if necessary, we can compute $\Gamma(s)$ using the Stirling series for $\log \Gamma(s)$.
- The functions $H_\alpha(k)$ and $n(k)$ on lines 1.f.ii. and 1.f.iii., respectively, are as defined in Section 3.2.3.
- On line 1.f.ii., any numerical root finding technique can be employed to solve $\frac{dH_\alpha(k)}{dk} = 0$ for k.
- The work required to compute the terms $T_j(n, s), 1 \leq j \leq k$, on line 1.f.iv. can be reduced by applying a recursive method such as suggested by Lehmer [5]. We have to take into account that such methods lower the computational cost and have to decrease p accordingly.

As the algorithm, stands it can be converted into any suitable programming language and applied for the computation of $\zeta(s)$ to within any number of accurate digits for all $s \in \mathbb{C}$. Before we discuss the complexity of this algorithm, we show how Conjecture 1 can be employed to yield a slightly different algorithm. Let c be the constant of Conjecture 1. As was shown in Section 3.2.4, $c = \frac{1}{e}$ is a reasonable choice; so, we will use it. The new algorithm is obtained by replacing lines 1.f.i. through 1.f.iii. with the following three lines:

let $\mu = (1 - c)\frac{p}{q} + c$

let k be the smallest positive integer that satisfies $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(\mu k, s)| \leq \epsilon$

let $n = \mu k$

We now have two operative algorithms for computing the values of the Riemann zeta function to any precision for any complex argument. Furthermore, these algorithms are optimal in that they have been derived so that the least number of terms are used from the E-M series of $\zeta(s)$;i. e. these algorithms achieve minimum cost in the computation of the series. However, these algorithms differ in their computational cost of finding n and k. We discuss their differences, and also draw another recent algorithm into the discussion, which has been developed by Cohen and Olivier [2]. For easy reference, we denote the algorithm based on the method of Section 3.2.3 by A, the algorithm using Conjecture 1 by B, and Cohen and Olivier's algorithm by C.

Cohen and Olivier's objective is to compute the value of the Riemann zeta function for any $s \in \mathbb{C}$ to within any given accuracy. Their algorithm C utilizes the relationship between n and k as given by the growth ratio, and an estimation of $\Gamma(s+2k)$ to find the values of n and k: For details see reference [2].

First, we look at the values of n and k given in Table 4.3 as computed by the three different algorithms A, B, and C; and compare them. Cohen and Olivier did not take into account the different computational cost of terms from the two sums in the E-M series of $\zeta(s)$. So for comparison reason, we assume that the cost is the same, and therefore, take $p = q = 1 \Rightarrow \alpha = 1$ in the computation of n and k. The subscripts A, B, and C of n and k indicate with what algorithm n and k have been computed. As was shown in Sections 3.2.1 and 3.2.2 it is immediately apparent from Table 4.3 that $k_A + n_A \simeq k_B + n_B$. More importantly, Table 4.3 also shows that

$$n_C + k_C \simeq \begin{cases} \frac{3}{2}(n_A + k_A), \\ 3k_B. \end{cases}$$

This means that algorithm C uses $1\frac{1}{2}$ times as many terms from the E-M series for the computation of $\zeta(s)$ as algorithms A or B. Further, the values of k_C compared to k_A or k_B are almost twice their size; however, k controls the number of terms used in the sum $\sum_{j=1}^{k} T_j(n,s)$, and it takes more work to compute $T_j(n,s)$ than $\frac{1}{j^s}$. This means that the cost of computing the E-M series is far less in either algorithm A or B than in algorithm C.

s	d	k_C	n_C	k_A	n_A	$k_B = n_B$
3	50	59	20	27	27	27
3	200	232	75	111	114	108
20	50	47	18	17	25	22
20	250	277	92	128	128	128
50	100	64	29	18	40	32
50	300	294	102	129	143	136
$\frac{1}{2}+10i$	50	67	22	31	29	30
$\frac{1}{2}+10i$	250	298	95	142	133	138
$5+9i$	50	62	21	28	28	28
$5+9i$	200	235	76	111	106	109

Table 4.3: Values of k and n as computed with algorithms A, B, and C, where s is the complex argument of the zeta function and $\epsilon = 10^{-d}$ is the given accuracy.

The question that needs to be answered next is, how does the computational cost of finding n and k compare in the three algorithms. As pointed out earlier, algorithm A's main work lies in solving $\frac{dH_\alpha(k)}{dk} = 0$ for k, for all $s \in \mathbb{C}\backslash\{1\}$. Regardless of the numerical technique employed, we need at least the first derivative of $\frac{dH_\alpha(k)}{dk}$. However, $\frac{dH_\alpha(k)}{dk}$ already involves the Log, Gamma, and Polygamma functions and is, therefore,

quite complex in structure. On the other hand, algorithm B's core work lies in solving $\left|\frac{s+2k-1}{\sigma+2k-1}\right| |T_k(\mu k, s)| = \epsilon$ for k, for all $s \in \mathbb{C}$, which only involves the Log and Gamma functions after the usual simplifications. Now, Newton's method converges quadratically, so for a good enough initial guess very few iterations should be required to solve for k with a tolerance of 1, which is all the accuracy we ask for since we take the ceiling of that k. In fact, with the approximations involved, being within one or two of the best k is reasonable. Algorithm C's main work comes from Newton's method applied to a function whose most intricate term is the Arctan; moreover, Cohen and Olivier provide an initial guess. Furthermore, this work is only required for $s \in \mathbb{C}$, since for $s \in \mathbb{R}$ algorithm C solves for n and k by given identities.

To sum up, we have that algorithm C requires very little cost in finding n and k and algorithm A the most. Algorithm B is less expensive than algorithm A, but somewhat more expensive than algorithm C in computing the values for n and k. Of course, much depends on the numerical root finding technique employed, but a numerical analysis is beyond the scope of this paper.

What we have, now, are two algorithms, namely B and C, with opposite qualities. Algorithm B takes some work in finding n and k, but uses the least computational cost in the E-M series, while algorithm C requires very little work in finding n and k, but uses a large number of terms from the E-M series. This implies that for lower precision calculation it is entirely possible that Cohen and Olivier's algorithm is less expensive than our algorithm, even though they use more terms in the E-M series of $\zeta(s)$. However, for high precision calculation, when the cost is dominated by the computation of the E-M series, our algorithm will excel due to the above points.

4.2 Conclusion and Future Work

We have met the objective of this paper in that we presented two algorithms for the computation of the Riemann zeta function, which are based on the E-M series of $\zeta(s)$ and minimize the cost of computing the E-M series with respect to BRE. One of the algorithms utilizes a method for computing the optimal values of the upper bounds, n and k, for the two sums in the E-M series. This method is derived and stated in Section 3.2. The other algorithm is based on Conjecture 1 in Section 3.2.4 and requires less computational cost in finding n and k than the previous algorithm.

There are two issues that have been mentioned, which deserve further investigation. First, in high precision calculation the Bernoulli numbers need to be dealt with in another way than tabulating them, as storage space becomes expensive. They can be estimated, but this causes a loss in accuracy of n and k. Secondly, for both algorithms an efficient numerical root finding technique has to be implemented. Neither of these issues have been addressed here, since they are beyond the scope of this paper.

Our main result is the conjectured relationship between n and k. In the case when $n = \mu k$, we have not performed an exhaustive analysis of BRE. Based on the graphs of Figure 3.15, we believe that there is a much simpler relationship between k, s, and ϵ than is given through BRE. We suggest that a relationship between k and s in the above case merits further investigation.

We also want to point out that considering a relationship between the upper bounds of the two sums in a series is worthwhile to look into in the general case. For example, using the identity (C.25) and the Stirling's series for the Gamma function, we get a series for $\log\Gamma(s)$ that has two sums. Since the Gamma function is related to the Riemann zeta function, e. g. by the Functional Equation (2.3), the technique employed in this paper should be applicable to the Gamma series as well.

Bibliography

[1] R. Backlund, *Über die Nullstellen der Riemannschen Zetafunktion*, Dissertation, Helsingfors, 1916.

[2] H.Cohen and M. Olivier, Calcul des valeurs de la fonction zêta de Riemann en multiprécision, *C.R. Acad. Sci. Paris*, t. 314, Série I, 427-430, 1992

[3] H. M. Edwards, *Riemann's Zeta Function*, Academic Press, New York and London, 1974.

[4] J. I. Hutchinson, On the Roots of the Riemann Zeta Function, *Trans Amer Math Soc*, 27:49-60 (1925).

[5] D. H. Lehmer, Extended Computation of the Riemann Zeta-Function, *Mathematica*, 3:102-108 (1956).

[6] S. Wolfram, *Mathematica:* A System for Doing Mathematics by Computer, Addison-Wesley Publishing Company, Inc., 1991.

Appendix A

Validation of Backlund's Remainder Estimate

Backlund's estimate of the E-M remainder is derived by leaving out a negative integral (see Section 2.4). We have been unable to evaluate or estimate this integral. However, we need to validate BRE as a good estimate of the actual error to justify its use. Therefore, we compute the ratio $\frac{AE}{BRE}$ as a function of n for fixed large k and different $s \in \mathbb{C}$, where AE denotes the absolute error. Remember that the absolute error is defined as the absolute value of the difference between the actual value and the computed value; in our case we have

$$AE = \left| \zeta(s) - \left(\zeta_n(s) + \frac{1}{s-1}\frac{1}{n^{s-1}} + \frac{1}{2}\frac{1}{n^s} + \sum_{j=1}^{k} T_j(n,s) \right) \right|.$$

Figure A.18 depicts the graphs of the ratio for nine values of s. It is evident from these graphs that in all cases considered, BRE increases the number of accurate digits asked for by about 20% when $n \geq k$. This means that n and k are slightly larger than required for a given accuracy. Since the simplification of the remainder introduced by BRE is needed to obtain a relationship between n and k, this small increase in cost is justified.

Furthermore, by taking $n = k$ in BRE we obtain an estimate for the error which is even sharper than in the general case, as is shown in Figure A.19. We used $n = k$ when comparing Cohen and Olivier's [2] algorithm to our algorithm in Section 4.1.

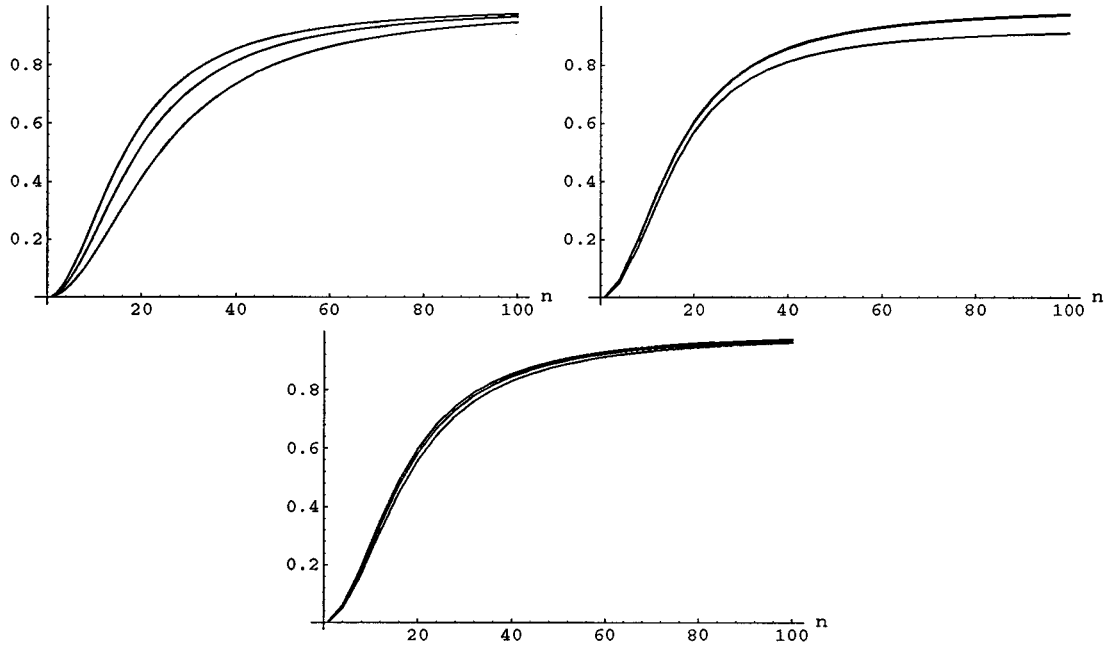

Figure A.18: The ratio $\frac{absolute\ error}{BRE}$, for $k = 40$ and $1 \leq n \leq 100$. The first plot corresponds to $s = 3, 20$, and 50; the second plot to $s = \frac{1}{2} + 2i$, $\frac{1}{2} + 10i$, and $\frac{1}{2} + 40i$; and the third plot to $s = 2 + 3i, 5 + 9i$, and $10 + 15i$. Note that for the complex values the ratios are very close to each other, so that their graphs overlap.

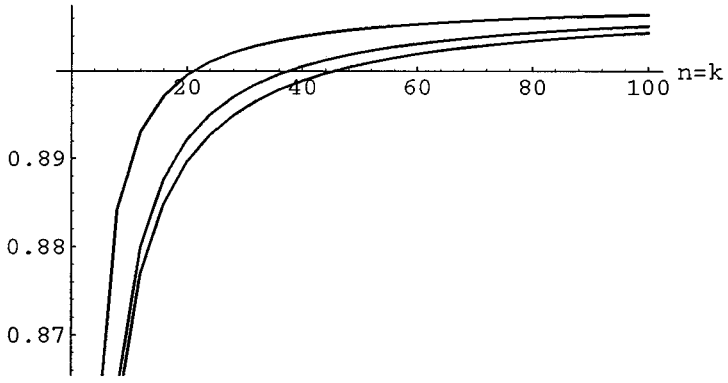

Figure A.19: The ratio $\frac{absolute\ error}{BRE}$, for $n = k$ with $1 \leq k \leq 100$, and $s = 3, \frac{1}{2} + 2i$, and $s = 2 + 3i$, respectively.

Appendix B

Bernoulli Polynomials

Definition 5 *Let $B_m(x)$ denote the m-th Bernoulli polynomial, where $x \in \mathbb{R}$. $B_m(x)$ is inductively defined by $B_0(x) = 1$ and for $m > 0$ by the two conditions*

$$B'_{m+1}(x) = (m+1)B_m(x) \quad \text{and} \tag{B.18}$$

$$\int_0^1 B_{m+1}(x)\,dx = 0. \tag{B.19}$$

For example,

$$\begin{aligned} B_1(x) &= x - \frac{1}{2}, \\ B_2(x) &= x^2 - x + \frac{1}{6}, \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x. \end{aligned}$$

Definition 6 *Let $\psi_m(x)$ be the periodic extension of the restriction of $B_m(x)$ to $x \in [0,1]$.*

One of its properties is that it is a continuous function for $m > 1$, since $B_m(x)$ restricted to $x \in [0,1]$ is continuous and $B_m(0) = B_m(1)$ for $m > 1$.

We want to express $\psi_m(x)$ in terms of its Fourier series. In general, the constant term of the Fourier series of $\psi_m(x)$ is zero because of condition (B.19) in the definition of the Bernoulli polynomial. The k-th Fourier coefficient of $\psi_1(x)$, $k \neq 0$, is given by

$$\begin{aligned} \int_0^1 B_1(x)\,e^{-2\pi kix}\,dx &= \int_0^1 (x - \frac{1}{2})\,e^{-2\pi kix}\,dx \\ &= -x\frac{e^{-2\pi kix}}{2\pi ki}\Big|_0^1 + \left(\frac{1}{2\pi ki} - \frac{1}{2}\right)\int_0^1 e^{-2\pi kix}\,dx \\ &= -\frac{1}{2\pi ki}. \end{aligned} \tag{B.20}$$

Assume that $m > 0$. Given the definition of the Fourier series for a continuous function f with period 1, i.e. $f(x) = \sum_{k=-\infty}^{\infty} \hat{f}(k)e^{2\pi kix}$, the k-th Fourier coefficient of f' is $2\pi ki\hat{f}(k)$. Together with condition (B.18) in the definition of the Bernoulli polynomial and the result of equation (B.20), this implies that the k-th Fourier coefficient of $\psi_m(x)$ must be $-\frac{m!}{(2\pi ki)^m}$. Hence, the Fourier series of $\psi_m(x)$ is given by

$$\psi_m(x) = -\sum_{k=-\infty, k\neq 0}^{\infty} \frac{m!}{(2\pi ki)^m} e^{2\pi kix} \tag{B.21}$$

for $m > 0$, or, alternately

$$\begin{aligned} \psi_{2m} &= (-1)^{m-1} \frac{2(2m)!}{(2\pi)^{2m}} \sum_{k=1}^{\infty} \frac{\cos 2\pi kx}{k^{2m}}, \\ \psi_{2m+1} &= (-1)^{m-1} \frac{2(2m+1)!}{(2\pi)^{2m+1}} \sum_{k=1}^{\infty} \frac{\sin 2\pi kx}{k^{2m+1}} \end{aligned} \tag{B.22}$$

for $m \geq 0$ and $2m \neq 0$. Furthermore, from equation (B.21) it can be easily seen that $\psi_m(x)$ has the following property for $m > 0$:

$$\psi_m'(x) = m\psi_{m-1}(x). \tag{B.23}$$

Definition 7 *Let β_m denote the m-th Bernoulli number, which is defined by the constant term of the m-th Bernoulli polynomial:*

$$\beta_m = B_m(0).$$

Since $B_m(0) = 0$ for m odd, it follows that for all m

$$\begin{aligned} \beta_{2m+1} &= 0 \quad \text{and} \\ \beta_{2m} &= B_{2m}(0). \end{aligned}$$

The following is an estimate of the magnitude of β_m for large m. Even though the first few Bernoulli numbers are quite small in size, their magnitude grows rapidly as can be seen from the following estimate:

$$\begin{aligned}
\beta_{2m} &= B_{2m}(0) \\
&= \psi_{2m}(0) \\
&= (-1)^{m-1}\frac{2(2m)!}{(2\pi)^{2m}}\sum_{k=1}^{\infty}\frac{1}{k^{2m}} \\
&\sim (-1)^{m-1}\frac{2(2m)!}{(2\pi)^{2m}},
\end{aligned}$$

since

$$\lim_{m\to\infty}\sum_{k=1}^{\infty}\frac{1}{k^{2m}} = \lim_{m\to\infty}\zeta(2m) = 1.$$

Applying Stirling's formula for factorials, we finally obtain the estimate

$$|\beta_{2m}| \sim 4\sqrt{\pi m}\left(\frac{m}{e\pi}\right)^{2m} \tag{B.24}$$

for large m.

Appendix C

Gamma Function

Definition 8 *The Gamma function* $\Gamma(s)$ *is defined by*

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t}\, dt$$

for $s \in \mathbb{C}$ *with* $\Re(s) > 0$.

Integration by parts shows that $\Gamma(s+1) = s\Gamma(s)$. More explicitly we have

$$\Gamma(s) = \frac{\Gamma(s+n)}{s(s+1)\dots(s+n-1)}. \tag{C.25}$$

This shows that $\Gamma(s)$ can be extended to all $s \in \mathbb{C}$ with $\Re(s) > -n$, except at the negative integers, where it will have simple poles. The following is a short collection of special values and properties of $\Gamma(s)$ used in this paper.

Special value $s = \frac{1}{2}$: By a change of variables in the definition of $\Gamma(s)$ we can calculate Γ at $s = \frac{1}{2}$:

$$\Gamma\left(\frac{1}{2}\right) = \int_{-\infty}^{\infty} e^{-t^2}\, dt = \sqrt{\pi}.$$

Euler's Formula: For all $s \in \mathbb{C}$

$$\Gamma(s) = \lim_{n\to\infty} \frac{n! n^s}{s(s+1)\dots(s+n)}.$$

Legendre's Duplication Formula: For all $s \in \mathbb{C}$

$$\Gamma\left(\frac{1}{2}\right)\Gamma(s) = 2^{s-1}\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{s+1}{2}\right).$$

Reflection Formula: For all $s \in \mathbb{C}$ with $0 < \Re(s) < 1$

$$\Gamma(s)\Gamma(1-s) = \frac{\pi}{\sin \pi s}.$$

Stirling's Formula: For large $s \in \mathbb{C}$ with $|\arg s| < \pi$

$$\Gamma(s) \sim \sqrt{\frac{2\pi}{s}}\left(\frac{s}{e}\right)^s.$$

Using Legendre's Duplication Formula and the Reflection Formula we can also get the following relationship:

$$\begin{aligned}
&\Gamma\left(\frac{1}{2}\right)\Gamma(1-s) = 2^{-s}\Gamma\left(\frac{1-s}{2}\right)\Gamma\left(\frac{2-s}{2}\right) \\
\Rightarrow\ &2^s\sqrt{\pi}\Gamma(1-s) = \Gamma\left(\frac{1-s}{2}\right)\Gamma\left(\frac{2-s}{2}\right) \\
\Rightarrow\ &2^s\pi^{-\frac{1}{2}}\sin\left(\frac{2\pi}{2}\right)\Gamma(1-s)\Gamma\left(\frac{s}{2}\right) = \Gamma\left(\frac{1-s}{2}\right) \\
\Rightarrow\ &(2\pi)^{s-1}2\sin\left(\frac{s\pi}{2}\right)\Gamma(1-s) = \pi^{s-\frac{1}{2}}\frac{\Gamma\left(\frac{1-s}{2}\right)}{\Gamma\left(\frac{s}{2}\right)}. \qquad \text{(C.26)}
\end{aligned}$$

Appendix D

Program

The following is the Mathematica code that produces all the graphs and tables in the text. All routines are thoroughly commented.

```
(* Z(s) denotes the Riemann Zeta function.
        R(n, k, s) is as defined in Section 2.2.
        The variable a in this program denotes alpha,
        which is defined in Section 3.3.2. *)

(* Zn[n, s] takes a complex number s and a positive
        integer n; and returns the partial sum of the
        Riemann Zeta function. *)

Zn[n_Integer, s_] := Block[{i},
        Sum[i^(-s), {i, n-1}]] /;
(n>0 && NumberQ[s]);

(* T[k, n, s] takes a complex number s and positive
        integers j and n; and returns T as defined in
the Euler-Maclaurin series for Z(s) of
Section 2.2. *)
```

```
T[k_Integer, n_Integer, s_] := Block[{a, b, c, d, i},
        a = (2k)!;
        b = n^(s+2k-1);
        c = BernoulliB[2k];
        d = Product[s + i, {i, 0, (2 k) - 2}];
        (c*d)/(a*b)] /;
(k>0 && n>0 && NumberQ[s]);

(* BRE[n, k, s] is Backlund's estimate for the remainder
        of the E-M series for (s). *)

BRE[n_Integer, k_Integer, s_] := Block[{a, b},
        a = Abs[(s+2k+1)/(Re[s]+2k+1)];
        b = Abs[T[k+1, n, s]];
        a*b] /;
(n>0 && k>0 && NumberQ[s]);

(* GErrVBRE[s, k, nmax] takes a complex number s,
and positive non-zero integers k and nmax;
and returns the graph of
(absolute error)/(Backlund's remainder estimate)
as a function of n, where 1 <= n <= nmax. *)

GErrVBRE[s_, k_Integer, nmax_Integer] :=
Block[{n, t1, t2, t3, t4, t5, t6, t7, t8, t9},
```

```
t1 = Table[Zs[n, k, s], {n, 4, nmax, 4}];
t2 = Zeta[s] - Prepend[t1, Zs[1, k, s]];
t3 = Map[Abs[#] &, t2];
t4 = Table[BRE[n, k, s], {n, 4, nmax, 4}];
t5 = Prepend[t4, BRE[1, k, s]];
t6 = N[Transpose[{t3, t5}], 200];
t7 = Apply[Divide, t6, 1];
t8 = Prepend[Table[n, {n, 4, nmax, 4}], 1];
t9 = Transpose[{t8, t7}];
ListPlot[t9,
PlotJoined -> True,
AxesLabel -> {"n", ""}]] /;
(NumberQ[s] && k>0 && nmax>0);

(* GErrVC[s, kmax] takes a complex number s,
and a positive non-zero integer kmax;
and returns the graph of
(absolute error)/(Backlund's remainder estimate)
as a function of k, where 1 <= k <= kmax
and n=k. *)

GErrVC[s_, kmax_Integer] :=
Block[{k, t1, t2, t3, t4, t5, t6, t7, t8, t9},
t1 = Table[Zs[k, k, s], {k, 4, kmax, 4}];
t2 = Zeta[s] - Prepend[t1, Zs[1, 1, s]];
t3 = Map[Abs[#] &, t2];
```

```
t4 = Table[BRE[k, k, s], {k, 4, kmax, 4}];
t5 = Prepend[t4, BRE[1, 1, s]];
t6 = N[Transpose[{t3, t5}], 200];
t7 = Apply[Divide, t6, 1];
t8 = Prepend[Table[k, {k, 4, kmax, 4}], 1];
t9 = Transpose[{t8, t7}];
ListPlot[t9,
PlotJoined -> True,
AxesLabel -> {"n+k", ""}]] /;
(NumberQ[s] && kmax>0);

(* GRn[s, n, kmax] takes a complex number s and positive
        integers n and kmax, and plots the graph of Backlund's
        remainder estimate for |R(n, k, s)| for fixed n,
        where kmax is the largest value for k. *)

GRn[s_, n_Integer, kmax_Integer:340] := Block[{b, t, k},
b = E;
        t = Table[{k, -Log[b, N[BRE[n, k, s], 300]]},
{k, kmax}];
        ListPlot[t,
                PlotJoined -> True,
                AxesLabel -> {"k", "D"}]] /;
(NumberQ[s] && n>0 && kmax>0);

(* GRk[s, k, nmax] takes a complex number s and positive
```

```
        integers k and nmax, and plots the graph of Backlund's
        remainder estimate for R(n, k, s) for fixed  k,
        where nmax is the largest value for n. *)

GRk[s_, k_Integer, nmax_Integer:70] := Block[{b, t, n},
b = E;
        t = Table[{n, -Log[b, N[BRE[n, k, s], 300]]},
{n, nmax}];
        ListPlot[t,
                PlotJoined -> True,
                AxesLabel -> {"n", "D"}]] /;
(NumberQ[s] && k>0 && nmax>0);

(* LogRSk[s, k] takes a complex number s and a positive
        integer k; and returns the logarithm to the
        base b of the size of the smallest term in the
        E-M sum for Z(s). Default value for b is e. *)

LogRSk[s_, k_Integer] := Block[{b, a1, a2, a3, a4, a5, a6},
        b = E;
        a1 = (1/2) Log[b, Abs[s+2k-1]];
        a2 = -Log[b, Re[s]+2k-1];
        a3 = Log[b, 2];
        a4 = (Re[s]-1) Log[b, 2 Pi];
        a5 = -(Re[s]-1) Log[b, Abs[s]];
        a6 = Im[s] (Arg[s]-Arg[s+2k-1])-(2k-1);
```

```
        N[a1+a2+a3+a4+a5+a6]] /;
(NumberQ[s] && k>1);

(* SkGR[s, d] takes a complex number s and an
        integer d, where d is the number of precision
        digits required in the calculation of Z(s), and
        returns that number k for which the estimate
        for the size of the smallest term in the
        E-M sum for Z(s) is less than or equal to 10^(-d). *)

SkGR[s_, d_Integer] := Block[{k, b, e},
        k = 2;
        b = E;
        e = N[- d Log[b, 10]];
        While[LogRSk[s, k] > e && k < 600, k++];
        k] /;
(NumberQ[s] && d>0);

(* Guessu[a] takes a positive number a greater than one;
        and returns u(a) as chosen in Section 3.2.4. *)

Guessu[a_] := N[(1-(1/E))(1/a)+(1/E)] /;
(NumberQ[a] && a>0);

(* H[k, s, d, a] takes a positive real number k, a
        complex number s, and a positive number d, and
```

```
        returns n+k as defined by the function H in
        Chapter 3.2.3. *)

H[k_, s_, d_, a_:1] := Block[{c1, c2, e},
        c1 = 2 10^d / (2 Pi)^(2k);
        c2 = Gamma[s+2k-1]/Gamma[s];
        e = 1/(s+2k-1);
        N[a ((c1 c2)^e) + k]] /;
(NumberQ[s] && Im[s]==0 && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

H[k_, s_, d_, a_:1] := Block[{c1, c2, c3, e},
        c1 = 2 10^d / (2 Pi)^(2k);
        c2 = Sqrt[(Re[s]+2k-1)^2 + (Im[s])^2] / (Re[s]+2k-1);
        c3 = Abs[Gamma[s+2k-1]/Gamma[s]];
        e = 1/(Re[s]+2k-1);
        N[a (c1 c2 c3)^e + k]] /;
(NumberQ[s] && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

(* HC[k, s, d, a] takes a positive real number k, a
        complex number s, and a positive number d, and
        returns n+k as defined by the function H in
        Chapter 3.2.3. *)

HC[k_, s_, d_, a_:1] := Block[{c1, c2, c3, e},
```

```
        c1 = ((Re[s]+2k-1)^2+Im[s]^2)^(1/2) / E;
        c2 = (2 Sqrt[2 Pi] 10^d) /
                (Abs[Gamma[s]] E^(Im[s] Arg[s]));
        c3 = ((Re[s]+2k-1)^2+Im[s]^2)^(1/4) /
                ((Re[s]+2k-1) * (2 Pi)^(2k));
        e = 1/(Re[s]+2k-1);
        N[a (c1 (c2 c3)^e) + k]] /;
(NumberQ[s] && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

(* BnH[k, s, d, a] takes positive numbers k, and d, and
        a complex number s; and returns the best n. *)

BnH[k_, s_, d_, a_:1] := (H[k, s, d, a]-k)/a /;
(NumberQ[k] && k>=1 && NumberQ[s] && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

(* GBESTnk[s, d, kmin, kmax] takes a complex number s,
        a positive number d, and positive integers kmin
        and kmax; and returns the graph of best (n, k)
        using H as defined in Section 3.2.3, where
        kmin <= k <= kmax. *)

GBESTnk[s_, d_, kmin_Integer, kmax_Integer, a_:1] :=
        Block[{t, k, m, g},
        t = Table[{k, H[k, s, d, a]}, {k, kmin, kmax}];
```

```
        m = Min[Map[#[[2]] &, N[t]]];
        c = Append[
                Position[Map[#[[2]]&, N[t]], m][[1]]+kmin-1,m];
        g = ListPlot[t,
                PlotJoined -> True,
                PlotRange -> {{kmin, kmax}, {1, 700}},
                AxesLabel -> {"k", "n+k"}];
        Show[{Graphics[{AbsoluteThickness[2], Cross[c]}],
                g}]] /;
(NumberQ[s] && NumberQ[d] && d>0 && kmin>0 &&
        kmax>kmin && NumberQ[a] && a>0);

(* BkdH[s, d, a] takes a complex number s and a
        positive number d; and returns that k at which
        H(k, s, d) has a minimum, using its derivative. *)

BkdH[s_, d_, a_:1] := Block[{F, x, y, k},
        F[y_] := D[H[x, s, d, a], x] /. x -> y;
        k = 1;
        While[F[k] < 0 && k < 400, k++];
        k] /;
(NumberQ[s] && Im[s]==0 && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

BkdH[s_, d_, a_:1] := Block[{F, x, y, k},
        F[y_] := D[HC[x, s, d, a], x] /. x -> y;
```

```
        k = 1;
        While[F[k] < 0 && k < 400, k++];
        k] /;
(NumberQ[s] && NumberQ[d] && d>0 && NumberQ[a] && a>0);

(* BkBRE[s, d, a] takes a complex number s and a positive
        number d; and returns the best k for which
        |R(k, k-1, s)| <= 10^(-d), using Backlund's remainder
        estimate. *)

BkBRE[s_, d_, a_:1] := Block[{b, e, u, k},
        b = E;
        e = N[- d Log[b, 10]];
        u = Guessu[a];
        k = 2;
        While[Log[b, N[BRE[Ceiling[u k], k-1, s], 300]] > e &&
k < 600, k++];
        k] /;
(NumberQ[s] && NumberQ[d] && d>0 &&
        NumberQ[a] && a>0);

(* GBkVBn[s, dmax, a] takes a complex number s and a
        positive integer dmax; and returns the graph of the
        best n versus the best k. *)

GBkVBn[s_, dmax_, a_:1] :=
```

```
        Block[{dm, u, d, t1, t2, t3, t4, t5, t6, t7},
        dm = Ceiling[dmax];
        u = Guessu[a];
        t1 = Table[{s, d, a}, {d, 1, dm, 10}];
        t2 = Table[d Log[10], {d, 1, dm, 10}];
        t3 = Apply[BkdH, t1, 1];
        t4 = Transpose[{t2, u t3}];
        t5 = Partition[Flatten[Transpose[{t3, t1}], 2], 4];
        t6 = Apply[BnH, t5, 1];
        t7 = Transpose[{t2, t6}];
        g1 = ListPlot[Join[{{0, -10}, {0, 0}}, t4],
                PlotJoined -> True,
                AxesLabel -> {"D", "k"}];
        g2 = ListPlot[t7,
                PlotJoined -> True,
                AxesLabel -> {"D", "n"}];
        Show[{g1, g2}, AxesLabel -> {"D", ""}]] /;
(NumberQ[s] && NumberQ[dmax] && dmax>0 &&
        NumberQ[a] && a>0);

(* Ratiok[s, dmax, a] takes a complex number s and a
        positive integer dmax; and returns the graph of
        H(kmin) = best n+k |d, the graph of two times
        those k for which |R(k, k-1, s)| <= 10^(-d)
        using Backlund's remainder estimate, and the graph
        of the ratio of 2k to (best n + best k), where
```

```
        the number of accurate digits d runs from 1 to
        10^(-dmax) in steps of 5. *)

Ratiok[s_, dmax_, a_:1] :=
        Block[{dm, c, d, t1, t2, t3, t4, t5, t6, t7, t8, g1, g2},
        dm = Ceiling[dmax];
        c = a Guessu[a] + 1;
        t1 = Table[{s, d, a}, {d, 1, dm, 5}];
        t2 = Table[d Log[10], {d, 1, dm, 5}];
        t3 = Apply[BkBRE, t1, 1];
        t4 = Transpose[{t2, c t3}];
        t5 = Apply[BkdH, t1, 1];
        t6 = Partition[Flatten[Transpose[{t5, t1}]], 4];
        t7 = If[Im[s]==0, Apply[H, t6, 1], Apply[HC, t6, 1]];
        t8 = Transpose[{t2, t7}];
        t9 = Apply[Divide, Transpose[{t7, c t3}], 1];
        t10 = Transpose[{t2, t9}];
        g1 = ListPlot[t4,
                PlotJoined -> True,
                AxesLabel -> {"D", "k"}];
        g2 = ListPlot[Join[{{0, -30}, {0, 0}}, t8],
                PlotJoined -> True,
                AxesLabel -> {"D", "an+k"}];
        Show[{g1, g2}, AxesLabel -> {"D", ""}];
        ListPlot[t10,
                PlotJoined -> True,
```

```
                AxesLabel -> {"D", ""}]] /;
(NumberQ[s] && dmax>0 && NumberQ[a] && a>0);

(* Gu[slist, dlist, amax, ainc] takes a list of
        complex s values, a list of positive d values, where
        d is the number of accurate digits, the maximum
        number amax of a, and the increase ainc of a from 1
        to amax. Gu returns a statistical plot of n/k
        values based on slist, dlist, and a values,
        together with the function u(a) as chosen in
        Section 3.2.4. *)

Gu[slist_List, dlist_List, amax_, ainc_:1] :=
        Block[{sl, dl, al, i, j, sdt, at, Hl, sdat, kt, ksdat1,
                ksdat2, nt, nkt1, nkt2, nkt3, g1, g2},
        sl = Length[slist];
        dl = Length[dlist];
        al = sl * dl;
        sdt = Flatten[Table[Transpose[{Table[slist[[i]], {dl}],
                dlist}], {i, sl}], 1];
        at = Table[a, {a, .5, amax, ainc}];
        Hl = Length[at];
        sdat = Table[Table[Append[sdt[[j]], at[[i]]], {j, al}],
                {i, Hl}];
        kt = Apply[BkdH, sdat, {2}];
        ksdat1 = Table[Table[Prepend[sdat[[i]][[j]],
```

```
                    kt[[i]][[j]]], {j, al}], {i, Hl}];
        ksdat2 = Partition[Flatten[ksdat1], 4];
        nt = Apply[BnH, ksdat2, {1}];
        nkt1 = Transpose[{nt, Flatten[kt]}];
        nkt2 = Partition[Apply[Divide, nkt1, {1}], al];
        nkt3 = Flatten[Table[Transpose[{Table[at[[i]], {al}],
                nkt2[[i]]}], {i, Hl}], 1];
        g1 = ListPlot[nkt3,
                AxesLabel -> {"a", "n/k"}];
        g2 = Plot[Guessu[a], {a, .5, amax},
                AxesLabel -> {"a", "u"}];
        Show[{g1, g2}, AxesLabel -> {"a", "n/k=u"}]] /;
(NumberQ[amax] && amax>=.5 && NumberQ[ainc]);

(* Gv[slist, dlist, amax, ainc] takes a list of
        complex s values, a list of positive d values, where
        d is the number of accurate digits, the maximum
        number amax of a, and the increase ainc of a from 1
        to amax. Gu returns a statistical plot of
        (smallest n)/(best n). *)

Gv[slist_List, dlist_List, amax_, ainc_:1] :=
        Block[{(*sl, dl, al, i, j, sdt, at, Hl, t1, t2, sdat,
                r1, r2, r3, r4, s1, s2, s3*)},
        sl = Length[slist];
        dl = Length[dlist];
```

```
        al = sl * dl;
        sdt = Flatten[Table[Transpose[{Table[slist[[i]], {dl}],
                dlist}], {i, sl}], 1];
        at = Table[a, {a, .5, amax, ainc}];
        Hl = Length[at];
        t1 = Apply[SkGR, sdt, 1];
        t2 = Flatten[Table[t1, {j, Hl}]];
        sdat = Table[Table[Append[sdt[[j]], at[[i]]], {j, al}],
                {i, Hl}];
        r1 = Apply[BkdH, sdat, {2}];
        r2 = Table[Table[Prepend[sdat[[i]][[j]],
                        r1[[i]][[j]]], {j, al}], {i, Hl}];
        r3 = Partition[Flatten[r2], 4];
        r4 = Apply[BnH, r3, {1}];
        s1 = Transpose[{t2, r4}];
        s2 = Partition[Apply[Divide, s1, {1}], al];
        s3 = Flatten[Table[Transpose[{Table[at[[i]], {al}],
                s2[[i]]}], {i, Hl}], 1];
        ListPlot[s3,
                AxesLabel -> {"a", "sm n / bst n"}]] /;
(NumberQ[amax] && amax>=.5 && NumberQ[ainc]);
```