Mellin transforms and asymptotics: the mergesort recurrence

Philippe Flajolet, Mordecai Golin

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

1 9 9 2

25ème anniversaire

N° 1612

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

MELLIN TRANSFORMS AND ASYMPTOTICS : THE MERGESORT RECURRENCE

Philippe FLAJOLET
Mordecai GOLIN

Février 1992

RR-1612

Mellin Transforms and Asymptotics: The Mergesort Recurrence

Philippe FLAJOLET and Mordecai GOLIN
Algorithms Project
INRIA Rocquencourt
F-78153 Le Chesnay, France

Abstract. *Mellin transforms and Dirichlet series are useful in quantifying periodicity phenomena present in recursive divide-and-conquer algorithms. This note illustrates the techniques by providing a precise analysis of the standard top-down recursive mergesort algorithm, in the average-case as well as in the worst-case. It also derives the variance and shows that the cost of mergesort has a Gaussian limiting distribution. The approach is applicable to a number of divide-and-conquer recurrences.*

Transformation de Mellin et asymptotique: La recurrence du tri-fusion

Résumé. La transformation de Mellin et les séries de Dirichlet sont utiles pour quantifier les phénomènes de périodicité qui se présentent dans les algorithmes récursifs de type "diviser-pour-régner". Cette note illustre ces techniques en produisant une analyse précise de l'algorithme de tri-fusion récursif descendant, ce dans le cas moyen et dans le pire cas. On y déduit aussi une estimation de variance et l'on montre que le coût du tri-fusion admet une loi limite Gaussienne. L'approche suivie est applicable à nombre de récurrences diviser-pour-régner.

Mellin Transforms and Asymptotics: The Mergesort Recurrence

Philippe FLAJOLET and Mordecai GOLIN
Algorithms Project
INRIA Rocquencourt
F-78153 Le Chesnay, France

February 7, 1992

Abstract. Mellin transforms and Dirichlet series are useful in quantifying periodicity phenomena present in recursive divide-and-conquer algorithms. This note illustrates the techniques by providing a precise analysis of the standard top-down recursive mergesort algorithm, in the average-case as well as in the worst-case. It also derives the variance and shows that the cost of mergesort has a Gaussian limiting distribution. The approach is applicable to a number of divide-and-conquer recurrences.

Many algorithms are based on a recursive *divide-and-conquer* strategy. Accordingly, their complexity is expressed by recurrences of the usual divide-and-conquer form [5]. Typical examples are heapsort, mergesort, Karatsuba's multiprecision multiplication, discrete Fourier transforms, binomial queues, sorting networks, etc. It is relatively easy to determine general orders of growth for solutions to these recurrences as explained in standard texts, see the "master theorem" of [5]. However, a precise asymptotic analysis is often appreciably more delicate.

At a more detailed level, divide-and-conquer recurrences tend to have solutions that involve periodicities, many of which are of a fractal nature. It is our purpose here to discuss the analysis of such periodicity phenomena while focussing on the analysis of the standard top-down recursive mergesort algorithm.

The methods employed — Mellin transforms, Dirichlet series, Perron's formula — borrow from classical analytic number theory [3]. Related problems with emphasis on digit sums and exact summatory formulæ are discussed in [9].

1 Mergesort

Let $T(n)$ denote the worst time cost measured in the number of comparisons that are required for sorting n elements by the `MergeSort` procedure of Fig. 1, and let $U(n)$ be the corresponding average cost. We have

$$\begin{aligned} T(n) &= T(\lfloor \tfrac{n}{2} \rfloor) + T(\lceil \tfrac{n}{2} \rceil) + n - 1 \\ U(n) &= U(\lfloor \tfrac{n}{2} \rfloor) + U(\lceil \tfrac{n}{2} \rceil) + n - \gamma_n \end{aligned} \qquad (1)$$

for $n \geq 2$, with $T(1) = U(1) = 0$, and

$$\gamma_n = \frac{\lfloor \frac{n}{2} \rfloor}{\lceil \frac{n}{2} \rceil + 1} + \frac{\lceil \frac{n}{2} \rceil}{\lfloor \frac{n}{2} \rfloor + 1}.$$

This results from the cost of merging two files of size a and b which is

$$a + b - 1 \quad \text{and} \quad a + b - \frac{a}{b+1} - \frac{b}{a+1},$$

in the worst case and average cases respectively (see [13, p. 165] for a fuller description of recursive mergesort and [12, ex. 5.2.4-2] for a derivation of the average case cost of merging).

Algorithm `MergeSort`$(a[1..n])$;
- `MergeSort`$(a[1..\lfloor n/2 \rfloor])$;
- `MergeSort`$(a[\lfloor n/2 \rfloor + 1..n])$;
- `Merge`$(a[1..\lfloor n/2 \rfloor], a[\lfloor n/2 \rfloor + 1..n])$;

Figure 1: Top-Down Recursive Mergesort.

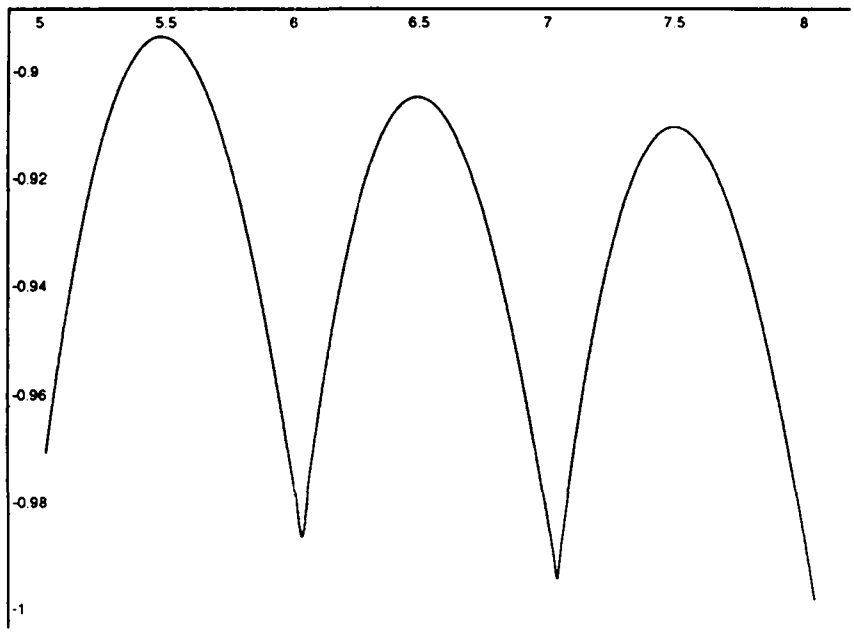

Figure 2: The fluctuation in the worst case behavior of Mergesort, in the form of the coefficient of the linear term $\frac{1}{n}[T(n) - n \lg n]$ as a function of $\lg n \equiv \log_2 n$ for $n = 32..256$. From Theorems 1 and 2, the periodic function involved, $A(u)$, fluctuates in $[-1, -0.91392,]$ with mean value $a_0 = -0.94269$.

The precise behavior of $T(n)$ is essentially known. The main term is $n \lg n$ and $T(n)$ also contains a simple periodic function in $\lg n \equiv \log_2 n$. (Recall the usual notation for fractional parts, $\{u\} = u - \lfloor u \rfloor$.) The periodicities are apparent from Fig. 2 with "cusps" whenever $\lg n$ is an integer.

Theorem 1 *The worst case cost $T(n)$ satisfies*

$$T(n) = n\lg n + nA(\lg n) + 1,$$

where $A(u)$ is the periodic function

$$A(u) = 1 - \{u\} - 2^{1-\{u\}}.$$

Proof. It is easy to check that

$$\begin{aligned} T(n) &= \sum_{k=1}^{n} \lceil \lg n \rceil \\ &= n\lceil \lg n \rceil - 2^{\lceil \lg n \rceil} + 1. \end{aligned}$$

(See [11, p. 400], where a closely related function is discussed.) The statement then follows from writing

$$\lceil \lg n \rceil = \lg n + 1 - \{\lg n\},$$

for any n not a power of 2. □

Knuth analyzes a bottom up version of Mergesort in the average case (Algorithm L, see [12, 5.2.4 and 5.2.4-13]), when n is power of 2. Knuth's analysis is also valid for top down recursive Mergesort in this special case. When $n = 2^k$, the recurrence for $U(n)$ can be unfolded to derive

$$U(2^k) = n\lg n + \beta n + o(n)$$

where

$$\beta = -\sum_{j \geq 0} \frac{1}{2^j + 1} = -1.26449\,97803.$$

For general n, no such formula is known. (See however Equation (11) at the end of Section 4 for some related analyses.) In what follows we will outline an approach that permits the analysis of mergesort type recurrences and demonstrate it by analyzing $U(n)$.

2 The Mergesort Recurrences

We approach the analysis of $T(n)$ and $U(n)$ via the computation of some associated Dirichlet series.

Let $\{w_n\}$ be a sequence of numbers. The Dirichlet generating function of w_n is defined to be

$$W(s) = \sum_{n=1}^{\infty} \frac{w_n}{n^s}.$$

The coefficients of Dirichlet series can be recovered by an inversion formula known as the Mellin–Perron formula which belongs to the galaxy of methods relating to Mellin transform analysis.

Lemma 1 (Mellin–Perron) *Assume the Dirichlet series $W(s)$ converges absolutely for $\Re(s) > 2$. Then,*

$$\frac{n}{2i\pi} \int_{3-i\infty}^{3+i\infty} W(s) n^s \frac{ds}{s(s+1)} = \sum_{k=1}^{n-1} (n-k) w_k. \quad (2)$$

Proof. For completeness, we sketch the proof of this classical result. See [3] for a closely related result. For the more general version and its relation to Mellin inversion, see [9]. Take $x > 0$ and consider the integral

$$J(x) = \frac{1}{2i\pi} \int_{3-i\infty}^{3+i\infty} x^s \frac{ds}{s(s+1)}.$$

By closing the line of integration by a large semicircle to the left (when $x \geq 1$) or to the right (when $x \leq 1$), and taking residues into account, we find that

$$J(x) = \begin{cases} 0 & \text{if } x \leq 1 \\ 1 - x^{-1} & \text{if } x \geq 1. \end{cases}$$

The left hand side of Equation (2) is therefore equal to

$$n \sum_{k=1}^{\infty} J\left(\frac{n}{k}\right) w_k = \sum_{k=1}^{n-1} (n-k) w_k$$

and the proof of the lemma follows. □

An iterated sum

$$\sum_{k=1}^{n-1} (n-k) w_k = \sum_{k=1}^{n-1} \sum_{l=1}^{k} w_l$$

of coefficients of a Dirichlet series is thus expressible by an integral applied to the series itself.

In order to recover the mergesort quantities $T(n)$ and $U(n)$, we will determine the Dirichlet series of their second differences. Then we will use the Mellin-Perron formula to derive an integral representation of the given quantity. We conclude by evaluating the integral via the residue theorem. As in other Mellin type analyses, this provides an asymptotic expansion for the quantities of interest.

This technique, which is familiar from analytic number theory, is analogous to a common technique in combinatorial counting. In the latter case, generating functions are ordinary, their singularities play a crucial rôle, and the asymptotic behavior of the coefficients of the power series is found by utilizing the Cauchy integral formula.

Consider the general divide-and-conquer recurrence scheme

$$f_n = f_{\lfloor n/2 \rfloor} + f_{\lceil n/2 \rceil} + e_n, \quad (3)$$

for $n \geq 2$, where e_n is a known sequence and f_n is the sequence to be analyzed. An initial condition fixing the value f_1 is also assumed. In order to make the notation unambiguous we formally set $e_0 = f_0 = e_1 = 0$. The functions $T(n)$ and $U(n)$ both satisfy this scheme: for $T(n)$, $e_n = n - 1$ and for $U(n)$, $e_n = n - \gamma_n$.

Distinguishing between odd and even cases, we find that for $m > 0$

$$\begin{cases} f_{2m} &= 2f_m + e_{2m} \\ f_{2m+1} &= f_m + f_{m+1} + e_{2m+1} \end{cases} \quad (4)$$

Taking backward differences with $\nabla f_n = f_n - f_{n-1}$ and $\nabla e_n = e_n - e_{n-1}$ yields

$$\begin{cases} \nabla f_{2m} &= \nabla f_m + \nabla e_{2m} \\ \nabla f_{2m+1} &= \nabla f_{m+1} + \nabla e_{2m+1} \end{cases} \quad (5)$$

for $m > 0$. Taking forward differences of the preceeding quantities, $\Delta\nabla f_n = \nabla f_{n+1} - \nabla f_n$ and $\Delta\nabla e_n = \nabla e_{n+1} - \nabla e_n$, we arrive at

$$\begin{cases} \Delta\nabla f_{2m} &= \Delta\nabla f_m + \Delta\nabla e_{2m} \\ \Delta\nabla f_{2m+1} &= \Delta\nabla e_{2m+1}, \end{cases} \quad (6)$$

for $m \geq 1$, with $\Delta\nabla f_1 = f_2 - 2f_1 = e_2$.

Define the Dirichlet generating function corresponding to $w_n = \Delta\nabla f_n$,

$$W(s) = \sum_{n=1}^{\infty} \frac{\Delta\nabla f_n}{n^s}.$$

Then, from (6), multiplying w_n by n^{-s} and summing over n, we find

$$W(s) = \frac{W(s)}{2^s} + \Delta\nabla f_1 + \sum_{n=2}^{\infty} \frac{\Delta\nabla e_n}{n^s}.$$

Solving for $W(s)$, we attain an explicit form for $W(s)$. Since $\sum_{k=1}^{n-1}(n-k)\Delta\nabla f_k = f_n - nf_1$ the Mellin-Perron formula yields a direct integral representation of f_n :

Lemma 2 *Consider the recurrence*

$$f_n = f_{\lfloor n/2 \rfloor} + f_{\lceil n/2 \rceil} + e_n,$$

for $n \geq 2$, with f_1 given and $e_n = O(n)$. The solution satisfies

$$f_n = nf_1 + \frac{n}{2i\pi} \int_{3-i\infty}^{3+i\infty} \frac{\Xi(s) n^s}{1-2^{-s}} \frac{ds}{s(s+1)},$$

where

$$\Xi(s) = \Delta\nabla f_1 + \sum_{n=2}^{\infty} \frac{\Delta\nabla e_n}{n^s}.$$

(The growth condition on e_n ensures existence of associated Dirichlet series when $\Re(s) > 2$, in accordance with the conditions of Lemma 1.)

3 Worst Case Analysis

As an easy application of Lemma 2 we quickly sketch how it can be used to derive an alternate expression involving a Fourier series for the value $T(n)$, the *worst case* number of comparisons performed by mergesort.

Theorem 2 *The worst case cost $T(n)$ satisfies*

$$T(n) = n\lg n + nA(\lg n) + 1$$

where $A(u)$ is a periodic function with mean value

$$a_0 = \frac{1}{2} - \frac{1}{\log 2} = -0.94269\,50408,$$

and $A(u)$ has the explicit Fourier expansion,

$$A(u) = \sum_{k \in \mathbf{Z}} a_k e^{2ik\pi u},$$

where, for $k \in \mathbf{Z} \setminus \{0\}$,

$$a_k = \frac{1}{\log 2} \frac{1}{\chi_k(\chi_k+1)} \qquad \textit{with} \quad \chi_k = \frac{2ik\pi}{\log 2}.$$

The extreme values of $A(u)$ are

$$-\frac{1+\log\log 2}{\log 2} = -0.91392, \textit{and} \quad -1.$$

Proof. We apply Lemma 2 with $f_n = T(n)$. For this case we have $e_n = n-1$ and $f_1 = 0$ so $\Delta\nabla f_1 = e_2 = 1$ and $\Delta\nabla e_n = 0$ for all n. Thus $\Xi(s) = 1$ and

$$\frac{f_n}{n} = \frac{1}{2i\pi} \int_{3-i\infty}^{3+i\infty} \frac{n^s}{1-2^{-s}} \frac{ds}{s(s+1)}. \qquad (7)$$

We can evaluate this integral using standard methods. Fix $\alpha < -1$. Let $R > 0$ and Γ be the counterclockwise contour around $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ where[1]

$$\begin{aligned} \Gamma_1 &= \{3+iy : |y| \leq R\} \\ \Gamma_2 &= \{x+iR : \alpha \leq x \leq 3\} \\ \Gamma_3 &= \{\alpha+iy : |y| \leq R\} \\ \Gamma_4 &= \{x-iR : \alpha \leq x \leq 3\}. \end{aligned}$$

See Figure 3. Set $I(s) = \frac{n^s}{1-2^{-s}} \frac{1}{s(s+1)}$ to be the kernel of the integral in (7). Letting $R \uparrow \infty$ we find that $\frac{1}{2i\pi}\int_{\Gamma_1} I(s)\,ds$ becomes the integral in (7), $|\int_{\Gamma_2} I(s)ds|$ and $|\int_{\Gamma_4} I(s)ds|$ are both $O(1/R^2)$ and

$$\left|\int_{\Gamma_3} I(s)ds\right| \to \left|\int_{\alpha+i\infty}^{\alpha-i\infty} I(s)ds\right| \leq 4n^{\alpha}.$$

The residue theorem therefore yields that f_n/n equals $O(n^{\alpha})$ plus the sum of the residues of $I(s)$ inside Γ.

We can actually do better. Since $I(s)$ is analytic for *all* s with $\Re(s) < -1$ we may let α go to $-\infty$ getting progressively smaller and smaller error terms. This shows that f_n/n is *exactly* equal to the sum of the residues of $I(s)$ inside Γ. The singularities of $I(s)$ are

[1] We further assume that R is of the form $(2j+1)\pi/\log 2$ for integer j, so that the contour passes halfway between poles of the integrand.

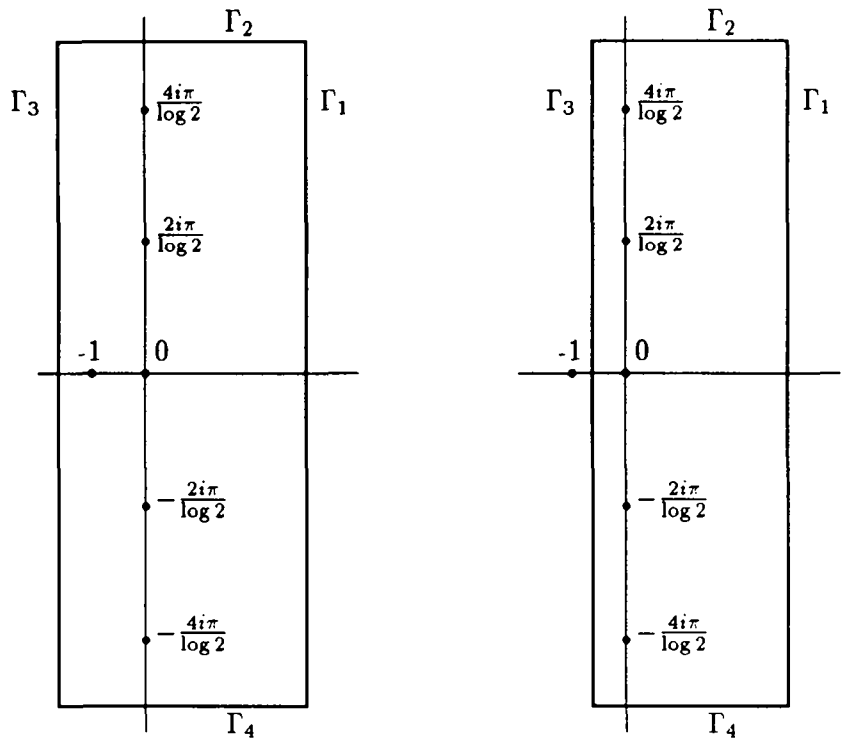

Figure 3: The two contours employed in the proofs of Theorem 2 (left) and Theorem 3 (right). Singularities are represented by dots. Note that the contour on the left contains the singularity at -1 while the contour on the right does not.

1. A double pole at $s = 0$ with residue $\lg n + \frac{1}{2} - \frac{1}{\log 2}$.
2. A simple pole at $s = -1$ with residue $\frac{1}{n}$.
3. Simple poles at $s = 2ki\pi/\log 2$, $k \in \mathbf{Z} \setminus \{0\}$ with residues $a_k e^{2ik\pi \lg n}$.

Thus, as promised, we have shown that

$$T(n) = n\lg n + nA(\lg n) + 1$$

where $A(u)$ is defined by the designated Fourier series. This Fourier series is uniformly convergent because $a_k = O(1/k^2)$.
The extreme values of $A(u)$ are calculated using standard techniques. □

We note that a computation of the Fourier series of $A(u)$ directly from Theorem 1 is also feasible and in fact yields the Fourier series derived in the last theorem (providing a convenient check on the validity of the theorem). However, the calculations performed above are needed in the analysis of the average case behavior in the next section.

4 Average Case Analysis

We now proceed with the real purpose of this paper, the analysis of the *average* number of comparisons performed by mergesort, $U(n)$.

Theorem 3 *(i). Let $\epsilon > 0$. The average case cost $U(n)$ of Mergesort satisfies*

$$U(n) = n\lg n + nB(\lg n) + O\left(n^{\epsilon}\right),$$

where $B(u)$ is a continuous non differentiable periodic function with period 1 that has an explicit Fourier expansion.

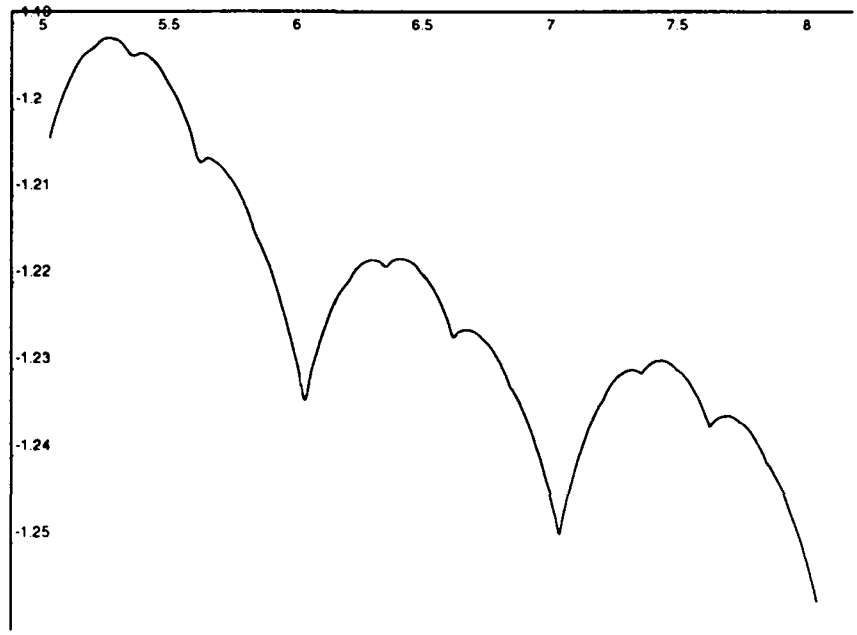

Figure 4: The fluctuation in the average case behavior of Mergesort, graphing the coefficient of the linear term $\frac{1}{n}[U(n) - n\lg n]$ using a logarithmic scale for $n = 32..256$. From Theorem 3, the periodic function involved, $B(u)$, fluctuates in $[-1.26449, -1.24075]$ with mean value $b_0 = -1.24815$.

(ii). The mean value b_0 of $B(u)$ equals

$$\frac{1}{2} - \frac{1}{\log 2} - \frac{1}{\log 2} \sum_{m=1}^{\infty} \frac{2}{(m+1)(m+2)} \log\left(\frac{2m+1}{2m}\right).$$

Numerically,

$$b_0 = -1.24815\,20420\,99653\,88489\ldots.$$

(iii). $B(u) = \sum_{k\in\mathbf{Z}} b_k e^{2ik\pi u}$ where b_0 is as above and the other Fourier coefficients of $B(u)$ are, for $k \in \mathbf{Z} \setminus \{0\}$,

$$b_k = \frac{1}{\log 2} \frac{1 + \Psi(\chi_k)}{\chi_k(\chi_k + 1)} \qquad \textit{where } \chi_k = \frac{2ik\pi}{\log 2},$$

and

$$\Psi(s) = \sum_{m=1}^{\infty} \frac{2}{(m+1)(m+2)} \left[\frac{-1}{(2m)^s} + \frac{1}{(2m+1)^s}\right].$$

This Fourier series is uniformly convergent to $B(u)$.

(iv). The extreme values of $B(u)$ are

$$\beta = -1.26449\,97803\ldots \textit{ and } -1.24075\,0572 \pm 10^{-9}.$$

Proof. The proof follows the paradigm laid down by Theorem 3. We first use Lemma 2 to derive an integral form for $f_n = U(n)$ and then use residue analysis to evaluate the integral.

For $f_n = U(n)$ we are given $f_1 = 0$ and $\Delta\nabla f_1 = e_2 = 1$. We are also given that for all $m > 0$

$$\begin{cases} e_{2m} &= 2m - 2 + \frac{2}{m+1} \\ e_{2m+1} &= 2m - 1 + \frac{2}{m+2}, \end{cases} \tag{8}$$

and thus

$$-\Delta\nabla e_{2m} = \frac{2}{(m+1)(m+2)} = \Delta\nabla e_{2m+1}.$$

Summing over all m we may write

$$\Xi(s) = \Delta\nabla f_1 + \sum_{n=2}^{\infty} \frac{\Delta\nabla e_n}{n^s} = 1 + \Psi(s)$$

where

$$\Psi(s) = \sum_{m=1}^{\infty} \frac{2}{(m+1)(m+2)} \left[\frac{-1}{(2m)^s} + \frac{1}{(2m+1)^s}\right]$$

converges absolutely and is $O(1)$ on any imaginary line $\Re(s) = \alpha \geq -1 + \epsilon$. Lemma 2 therefore

tells us that

$$\frac{f_n}{n} = \frac{1}{2i\pi}\int_{3-i\infty}^{3+i\infty} \frac{n^s}{1-2^{-s}}\frac{ds}{s(s+1).}$$

$$+\frac{1}{2i\pi}\int_{3-i\infty}^{3+i\infty} \frac{n^s\Psi(s)}{1-2^{-s}}\frac{ds}{s(s+1).} \quad (9)$$

The first integral on the right-hand side was already evaluated during the proof of Theorem 2 and shown to be equal to $\lg n + A(\lg n) + 1$ where $A(u) = \sum_k a_k 2^{ik\pi u}$.

The second integral can be evaluated using similar techniques. Fix $\alpha = -1 + \epsilon$. Let $R > 0$ and Γ be the counterclockwise contour around $\Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ where

$$\begin{aligned}\Gamma_1 &= \{3 + iy : |y| \le R\}\\ \Gamma_2 &= \{x + iR : \alpha \le x \le 3\}\\ \Gamma_3 &= \{\alpha + iy : |y| \le R\}\\ \Gamma_4 &= \{x - iR : \alpha \le x \le 3\}.\end{aligned}$$

See Figure 3. Set $I(s) = \frac{n^s\Psi(s)}{1-2^{-s}}\frac{1}{s(s+1)}$ to be the kernel of the second integral in (9). Letting $R \uparrow \infty$ we find that $\frac{1}{2i\pi}\int_{\Gamma_1} I(s)\,ds$ becomes the second integral in (9), $|\int_{\Gamma_2} I(s)ds|$ and $|\int_{\Gamma_4} I(s)ds|$ are both $O(1/R^2)$ and

$$\left|\int_{\Gamma_3} I(s)ds\right| \to \left|\int_{\alpha+i\infty}^{\alpha-i\infty} I(s)ds\right| = O(n^{-1+\epsilon}).$$

The constants implicit in the $O()$ notation are dependent upon ϵ.

Thus f_n/n equals $O(n^{-1+\epsilon})$ plus the sum of the residues of $I(s)$ inside Γ. The singularities of $I(s)$ *inside* Γ are

1. A simple pole at $s = 0$ with residue $\frac{\Psi'(0)}{\log 2}$.
2. Simple poles at $s = \chi_k = 2ki\pi/\log 2$, $k \in \mathbf{Z} \setminus \{0\}$ with residues $b_k e^{2ik\pi\lg n}$.

Summing these residues, multiplying by n and then adding the previously calculated contribution from the first integral yields the required result with the stated b_k's.

Note that $I(s)$ does have a simple pole at $s = -1$ but we do not count its residue because it is outside Γ. There are technical reasons (the behavior of $\Psi(s)$ towards $\pm i\infty$ when $\Re(s) = \alpha \le -1$) which stop us from setting $\alpha < -1$ and surrounding this last pole by Γ.

When $\Re(s) = 0$ then $|\Psi(s)| < 2$, so that $b_k = O(1/k^2)$; thus the Fourier series is uniformly convergent and the function $B(u)$ is continuous. Differentiability properties and numerical estimates are discussed below. □

Non Differentiability. There is an interesting decomposition of the periodic part of the average case behavior $B(u)$ in terms of the periodic part of the worst case $A(u)$. Define first

$$A^*(u) = A(u) - a_0, \quad B^*(u) = B(u) - b_0,$$

both functions having mean value 0. We have

$$B^*(u) - A^*(u) = \sum_{m=1}^{\infty} \psi_m A^*(u - \lg m), \quad (10)$$

where the ψ_m are the coefficients of the Dirichlet series $\Psi(s) = \sum_{m\ge 2} \frac{\psi_m}{m^s}$:

$$-\psi_{2m} = \frac{2}{(m+1)(m+2)} = \psi_{2m+1}.$$

To derive (10), take the Fourier expansion of $B^*(u) - A^*(u)$, expand the Fourier coefficients as sums since they are special values of a Dirichlet series, and exchange summations:

$$\begin{aligned}B^*(u) - A^*(u) &= \frac{1}{\log 2}\sum_k \frac{e^{2ik\pi u}}{\chi_k(\chi_k+1)}\Psi(\chi_k)\\ &= \frac{1}{\log 2}\sum_k \frac{e^{2ik\pi u}}{\chi_k(\chi_k+1)}\sum_{m=2}^{\infty}\psi_m e^{-2ik\pi\lg m}\\ &= \frac{1}{\log 2}\sum_{m=2}^{\infty}\psi_m\left[\sum_k \frac{e^{2ik\pi(u-\lg m)}}{\chi_k(\chi_k+1)}\right]\\ &= \sum_{m=2}^{\infty}\psi_m A^*(u-\lg m),\end{aligned}$$

the summations on k being for $k \in \mathbf{Z} \setminus \{0\}$.

This unusual decomposition (10) explains the behavior of $U(n)$ in Fig. 4. First, $A(u)$ and $A^*(u)$ have a cusp at $u = 0$, where the derivative has a finite jump. The function $B^*(u)$ is $A^*(u)$ to which is added a sum of pseudo-harmonics $A^*(u - \lg m)$ with decreasing amplitudes ψ_m. The harmonics corresponding to $m = 2, 4, 8$ are the same as those

of $A^*(m)$ up to scaling, and their presence explains the cusp of $B^*(u)$ at $u = 0$ which is visible on the graph of Fig. 4. We also have two less pronounced cusps at $\{\lg 3\} = 0.58$ and at $\{\lg 5\} = 0.32$ induced by the harmonics corresponding to $m = 3$ and $m = 5$. More generally, this decomposition allows us to prove the following property: *The function $B(u)$ is non differentiable (cusp-like) at any point of the form $u = \lg(p/2^r)$.* Stated differently, $B(\lg v)$ has a cusp at any dyadic rational $v = p/2^r$.

Numerical Computations. These have been carried out with the help of the Maple system. The computation of the *mean value* b_0 to great accuracy can be achieved simply by appealing to a general purpose series acceleration method discussed by Vardi in his entertaining book [15]. We have $\Psi'(0) = \sum_{m=1}^{\infty} \theta(1/m)$, where

$$\theta(y) = \frac{2y^2}{(1+y)(1+2y)} \log(1 + y/2).$$

The function $\theta(y)$ is analytic near $y = 0$ with a singularity at $y = -1/2$. Thus

$$\theta(y) = y^3 - \frac{13}{4}y^4 + \frac{67}{6}y^5 - \cdots = \sum_{j=3}^{\infty} c_j y^j,$$

where the $|c_j|$ grow essentially like 2^j. Select some small number m_0 (for instance $m_0 = 10$), and rewrite $\Psi'(0)$ as

$$\Psi'(0) = \sum_{m=1}^{m_0} \theta(\frac{1}{m}) + \sum_{j=3}^{\infty} c_j[\zeta(j) - \sum_{m=1}^{m_0} m^{-j}].$$

This form is obtained by separating the first m_0 terms, expanding each $\theta(1/m)$, and interchanging summations, which introduces the Riemann zeta function, $\zeta(s) = \sum_{n \geq 1} n^{-s}$. Standard facts about the zeta function tell us that the infinite series converges like $(2/m_0)^j$. In this way, with 80 terms and $m_0 = 10$, we evaluate $\Psi'(0)$ to 50 digits in a matter of one minute of computation time.

Regarding the computation of *extreme values* of $B(u)$ accurately, the approach via the Fourier series does not seem to be practicable, since the Fourier coefficients only decrease as $O(k^{-2})$.

Consider instead the sequence $U(a2^k)$ for some fixed integer a. By unwinding the recurrence, we find

$$U(a2^k) = ak2^k + 2^k U(a) - a2^k \sum_{j=0}^{k-1} \frac{1}{a2^j + 1}.$$

Rewriting $U(a2^k)$ in terms of $n = a2^k$, and taking care of the error terms yields for these particular values of n,

$$U(n) = n\lg n + \beta(a)n + o(n),$$

where

$$\beta(a) = \frac{U(a)}{a} - \lg a - \sum_{j=0}^{\infty} \frac{1}{a2^j + 1}. \qquad (11)$$

This formula generalizes the one given by Knuth for the average case, when $n = 2^k$. Comparing with Theorem 3, we find that

$$\beta(a) = B(\lg a).$$

The computation of $\beta(a)$ for all values a in an integer interval like $[2^{15}..2^{16}]$ (again in a matter of minutes) then furnishes the values of B with the required accuracy.

One final note. From these estimates, Mergesort has been found to have an average case complexity about

$$n\lg n - (1.25 \pm 0.01)n + o(n).$$

This appears to be not far from the information theoretic lower bound which is

$$\lg n! = n\lg n - n\lg e + o(n) = n\lg n - 1.44n + o(n).$$

5 Distribution

The cost of Mergesort is the sum of the costs of the individual merges, which are independent random variables with a known distribution. Merging two files of size m and n costs $m + n - S$, where the random variable S has distribution [12, p. 620]

$$\Pr\{S \geq s\} = \frac{\binom{m+n-s}{m} + \binom{m+n-s}{n}}{\binom{m+n}{n}}. \qquad (12)$$

Then, the variance $V(n)$ of Mergesort applied to random data of size n is a solution to the recurrence

$$f_n = f_{\lfloor n/2 \rfloor} + f_{\lceil n/2 \rceil} + e_n,$$

where e_n is equal to the variance of the cost of a single merge of type $(\lceil n/2 \rceil, \lfloor n/2 \rfloor)$,

$$e_{2m+1} = e_{2m+2} = \frac{2m(m+1)^2}{(m+2)^2(m+3)}.$$

The analysis unwinds exactly as in the average case. Applying Lemma 2 we find

Theorem 4 *The variance of the MergeSort algorithm applied to data of size n satisfies*

$$V(n) = n \cdot C(\log_2 n) + o(n),$$

where $C(u)$ is a continuous periodic function with period 1 and mean value

$$c_0 = \frac{1}{\log 2} \sum_{m=1}^{\infty} \frac{2m(5m^2+10m+1)}{(m+1)(m+2)^2(m+3)^2} \log \frac{2m+1}{2m}$$

which evaluates to $c_0 \approx 0.3454995688$. $C(u) = \sum_{k \in \mathbf{Z}} c_k e^{2ik\pi u}$, where for $k \in \mathbf{Z} \setminus \{0\}$,

$$c_k = \frac{1}{\log 2} \frac{\Psi(\chi_k)}{\chi_k(\chi_k+1)} \qquad \text{where } \chi_k = \frac{2ik\pi}{\log 2}$$

and

$$\Psi(s) = \sum_{m=1}^{\infty} \frac{2m(5m^2+10m+1)}{(m+1)(m+2)^2(m+3)^2} \left[\frac{-1}{(2m)^s} + \frac{1}{(2m+1)^s} \right].$$

Like the function $B(u)$ that describes the fluctuation of the average cost the function $C(u)$ is continuous but non-differentiable with cusps at the logarithms of dyadic rationals, a dense set of points. Numerically, its range of fluctuation is found to lie in the interval $[0.30, 0.36]$.

Finally, using standard extensions of the central limit theorem to sums of independent—but not necessarily identically distributed—random variables, we have:

Theorem 5 *The cost X_n of Mergesort applied to random data of size n converges in distribution to a normal variable,*

$$\Pr\left\{ \frac{X_n - U(n)}{\sqrt{V(n)}} \le \mu \right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu} e^{-t^2/2}\, dt.$$

Proof.(Sketch) From (12), each individual merge cost is found to have a third moment bounded by an absolute constant; from Theorem 4 the variance is $\Omega(n)$. The proof then directly follows from Lyapounov's generalization of the central limit theorem [4, p. 371]. □

In particular the cost is very close to its average estimate with high probability. This is illustrated by Figure 5. Notice that we even verify our theorems by *using samples of size 1.* The (fractal) periodic functions are thus far from being an artifact of our analysis but closely mirror the reality of the algorithm's behavior.

6 Conclusion

Divide-and-conquer recurrences are naturally associated with Dirichlet series that satisfy various sorts of functional relations [1, 2] and that can be proven to have meromorphic continuations in the whole of the complex plane. As we have seen here and as in [9], the Mellin–Perron formula then normally allows us to recover asymptotic properties of the original sequence. Several complications may however occur, and we offer a brief comment. First, the intervening Dirichlet series are often not explicit, and one has to operate with infinite functional relations. One such example is the Thue–Morse sequence that appears in [10] in connection with a probabilistic estimation algorithm. The Thue–Morse sequence is defined as $\epsilon_n = (-1)^{\nu(n)}$, where $\nu(n)$ designates the sum of digits of the binary representation of n. Sequences such as these lead to infinite functional equations and integral representations and are typical of the forms which have to be dealt with in more general cases.

Another problem is that each sequence has a certain degree of "smoothness" that dictates a certain level of summations. For mergesort, we were able to operate with the Mellin–Perron formula relative to double sums and the integrals we had to evaluate were nicely convergent. In general, this need not be the case. Take for example the cost of Karatsuba multiplication,

$$K(n) = 3K(\lceil \frac{n}{2} \rceil) + n.$$

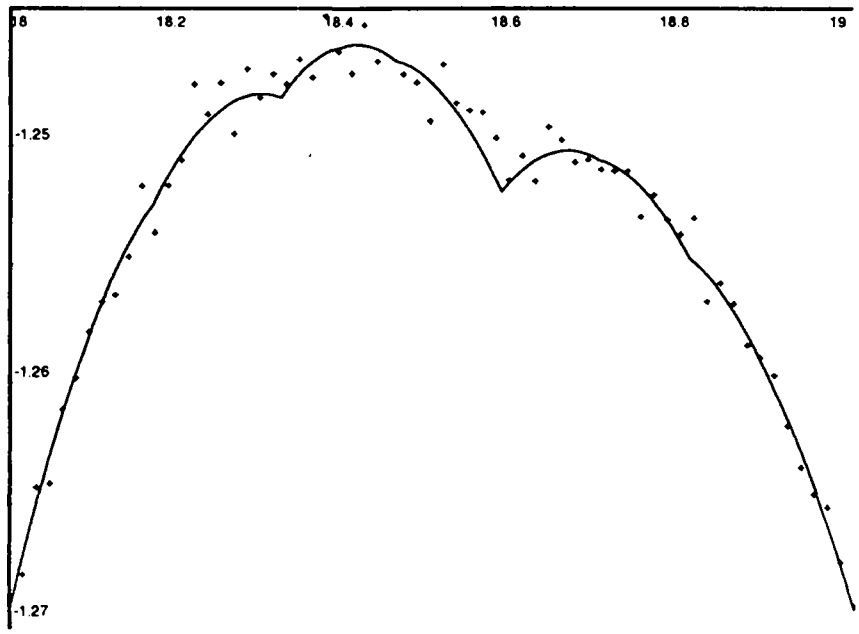

Figure 5: Display of the limit periodic curve $B(u)$ for $u \in [18, 19]$. Each cross represents one simulation of the cost X_n for $n \in [2^{18}, 2^{19}]$, with a logarithmic scale for n and the usual normalization $(X_n - n\lg n)/n$; X_n was simulated by running top-down mergesort on a random file of n-elements and counting the comparisons performed.

From the defining equation, the Dirichlet series of *first* differences has an explicit form,

$$\sum_{n=1}^{\infty} \frac{\Delta K_n}{n^s} = \frac{\zeta(s)}{1 - 3 \cdot 2^{-s}}.$$

In this case the suitable form of the Mellin–Perron formula involves a different kernel with a denominator of the form $1/s$ instead of the $1/(s(s+1))$ that we have encountered so far (and was discussed in Lemma 1), i.e.

$$\frac{K(n) + K(n+1)}{2} = \frac{1}{2i\pi} \int_{3-i\infty}^{3+i\infty} \frac{\zeta(s) n^s}{1 - 3 \cdot 2^{-s}} \frac{ds}{s}.$$

This poses specific convergence problems. Accordingly, the sequence exhibits a discontinuous behavior, for instance $K(2^n + 1)/K(2^n) \to K(2)/K(1) = 5/2$. In that case, it is the sum $\sum_{n=1}^{N} K(n)$ that appears to be amenable to our treatment: see the closely related example of "triadic binary numbers" in [9].

We propose to return to a more thorough analysis of divide–and–conquer recurrences by means of analytic techniques in a companion paper.

Acknowledgements: The work of both authors was supported in part by the Basic Research Action of the E.C. under contract No. 3075 (Project ALCOM).

References

[1] ALLOUCHE, J.-P. Automates finis en théorie des nombres. *Expositiones Mathematicae 5* (1987), 239–266.

[2] ALLOUCHE, J.-P., AND COHEN, H. Dirichlet series and curious infinite products. *Bulletin of the London Mathematical Society 17* (1985), 531–538.

[3] APOSTOL, T. M. *Introduction to Analytic Number Theory.* Springer-Verlag, 1976.

[4] BILLINGSLEY, P. *Probability and Measure*, 2nd ed. John Wiley & Sons, 1986.

[5] CORMEN, T. H., LEISERSON, C. E., AND RIVEST, R. L. *Introduction to Algorithms.* MIT Press, New York, 1990.

[6] DELANGE, H. Sur la fonction sommatoire de la fonction somme des chiffres. *L'enseignement Mathématique XXI*, 1 (1975), 31–47.

[7] DUMAS, P. *Récurrences Mahlériennes, suites automatiques, et études asymptotiques.* Doctorat de mathématiques, Université de Bordeaux I, 1992. In preparation.

[8] DUMONT, J.-M., AND THOMAS, A. Systèmes de numération et fonctions fractales relatifs aux substitutions. *Theoretical Computer Science 65* (1989), 153–169.

[9] FLAJOLET, P., GRABNER, P., KIRSCHENHOFER, P., PRODINGER, H., AND TICHY, R. Mellin transforms and asymptotics: Digital sums, July 1991. 23 pages. INRIA Research Report. Submitted to *Theoretical Computer Science.*

[10] FLAJOLET, P., AND MARTIN, G. N. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci. 31*, 2 (Oct. 1985), 182–209.

[11] KNUTH, D. E. *The Art of Computer Programming*, vol. 1: Fundamental Algorithms. Addison-Wesley, 1968. Second edition, 1973.

[12] KNUTH, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, 1973.

[13] SEDGEWICK, R. *Algorithms*, second ed. Addison–Wesley, Reading, Mass., 1988.

[14] STOLARSKY, K. B. Power and exponential sums of digital sums related to binomial coefficients. *SIAM Journal on Applied Mathematics 32*, 4 (1977), 717–730.

[15] VARDI, I. *Computational Recreations in Mathematica.* Addison Wesley, 1991.

ISSN 0249-6399