Divergent series: taming the tails

M. V. Berry & C. J. Howls

1 H. H. Wills Physics Laboratory, Tyndall Avenue, Bristol BS8 1TL, UK
2 Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK.

1 Introduction

By the 17th century, in what became the theory of convergent series, it was beginning to be understood how a sum of infinitely many terms could be finite; this is now a fully developed and largely standard element of every mathematician’s education. Contrasting with it is the theory of series that do not converge, especially those in which the terms first get smaller but then increase factorially: this is the class of ‘asymptotic series’, encountered frequently in applications, with which this article is mainly concerned. Although now a vibrant area of research, the development of the theory of divergent series has been tortuous and often accompanied by controversy. As a pedagogical device to explain the subtle concepts involved, we will focus on the contributions of individuals and describe how the ideas developed during several (overlapping) historical epochs, often driven by applications ranging from wave physics to number theory. This article complements the accompanying Companion article by P D Miller on ‘Perturbation Theory (including Asymptotic Expansions)’.

2 The Classical Period

In 1747, the Reverend Thomas Bayes (better known for his theorem in probability theory) sent a letter to Mr John Canton, F.R.S; it was published posthumously in 1763. Bayes demonstrated that the series now known as Stirling’s expansion for log(z!), “asserted by some eminent mathematicians,” does not converge. Arguing from the recurrence relation relating successive terms of the series, he showed that the coefficients “increase at a greater rate than what can be compensated by an increase of the powers of z, though z represent a number ever so large.” As we would say now, this expansion of the factorial function is a factorially divergent asymptotic series. The explicit form of the series, written formally as an equality, is

\[
\log(z!) = (z + 1/2) \log z + \log \sqrt{2\pi} - z + \frac{1}{2\pi^2z} \sum_{r=0}^{\infty} (-1)^r \frac{a_r}{(2\pi z)^{2r}},
\]

where

\[
a_r = (2r)! \sum_{n=1}^{\infty} \frac{1}{n^{2r+2}}.
\]

Bayes claimed that Stirling’s series “can never properly express any quantity at all” and the methods used to obtain it “are not to be depended upon.”

Leonhard Euler, in extensive investigations of a wide variety of divergent series beginning several years after Bayes, took the opposite view. He argued that such series have a precise meaning, to be decoded by suitable resummation techniques (several of which he invented): “if we employ [the] definition … that … the sum of a series is that quantity which generates the series, all doubts with respect to divergent series vanish and no further controversy remains.”

With the development of rigorous analysis in the 19th century, Euler’s view, which as we will see is the modern one, was sidelined and even derided. As Niels Henrik Abel wrote in 1828: “Divergent series are the invention of the devil, and it is shameful to base on them any demonstration whatsoever.” Nevertheless, divergent series, especially factorially divergent ones, repeatedly arose in application. Towards the end of the century they were embraced by Oliver Heaviside, who used them in pioneering studies of radio wave propagation. He obtained reliable results using undisciplined semi-empirical arguments that were criticised by mathematicians, much to his disappointment: “It is not easy to get up any enthusiasm after it has been artificially cooled by the wet blanket of rigorists.”

3 The Neoclassical Period

In 1886, Henri Poincaré published a definition of asymptotic power series, involving a large parameter z, that was both a culmination of previous
work by analysts and the foundation of much of the rigorous mathematics that followed. A series of the form $\sum_{n=0}^{\infty} a_n/z^n$ is defined as asymptotic by Poincaré if the error resulting from truncation at the term $n = N$ vanishes as $K/z^{N+1}(K > 0)$ as $|z| \to \infty$ in a certain sector of the complex $z$ plane. In retrospect, Poincaré’s definition seems a retrograde step, because although it encompasses convergent as well as divergent series in one theory, it fails to address the distinctive features of divergent series that ultimately lead to the correct interpretation that can also cure their divergence.

It was George Stokes, in research inspired by physics nearly four decades before Poincaré, who laid the foundations of modern asymptotics. He tackled the problem of approximating an integral devised by George Airy to describe waves near caustics, the most familiar example being the rainbow. This is what we now call the Airy function $Ai(z)$, defined by the oscillatory integral

$$Ai(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp \left( \frac{i}{3} t^3 + izt \right) dt,$$

the rainbow-crossing variable being $\text{Re} z$ (Figure 1) and the light intensity being $\text{Ai}^2(z)$. Stokes derived the asymptotic expansion representing the Airy function for $z > 0$, and showed that it is factorially divergent. His innovation was to truncate this series not at a fixed order $N$ but at its smallest term (optimal truncation), corresponding to an order $N(z)$ that increases with $z$. By studying the remainder left after optimal truncation, he showed that it is possible to achieve exponential accuracy (Figure 1) far beyond the power-law accuracy envisaged in Poincaré’s definition. We will call such optimal truncation superasymptotics.

Superasymptotics enabled Stokes to understand a much deeper phenomenon, one that is fundamental to the understanding of divergent series. In $Ai(z)$, $z > 0$ corresponds to the dark side of the rainbow, where the function decays exponentially: physically, this represents an evanescent wave. On the bright side $z < 0$, the function oscillates trigonometrically, that is, as the sum of two complex exponential contributions, each representing a wave; the interference of these waves generates the ‘supernumerary rainbows’ (whose observation was one of the phenomena earlier adduced by Thomas Young in support of his view that light is a wave phenomenon). One of these complex exponentials is the continuation across $z = 0$ of the evanescent wave on the dark side. But where does the other originate?

Stokes’s great discovery was that this second exponential appears during continuation of $Ai(z)$ in the complex plane from positive to negative $z$, across what is now called a ‘Stokes line’, where the dark-side exponential reaches its maximum size. Alternatively stated, the small (subdominant) exponential appears when maximally hidden behind the large (dominant) one. For $Ai(z)$ the Stokes line is $\text{arg} z = 120^\circ$ (Figure 2).

Stokes thought that the least term in the asymptotic series representing the large exponential constitutes an irreducible vagueness in the description of $Ai(z)$ in his superasymptotic scheme. By quantitative analysis of the size of this least term, Stokes concluded that only at maximal dominance could this obscure the small ex-
natural, which could then appear without inconsistency. As we will explain later, Stokes was wrong to claim that supersymptotics—optimal truncation—represents the best approximation that can be achieved within asymptotics. But his identification of the Stokes line with the birth of the small exponential (Figure 3) was correct. Moreover, he also appreciated that the concept was not restricted to $\text{Ai}(z)$ but applies to a wide variety of functions arising from integrals, solutions of differential equations and recurrence equations, etc., for which the associated asymptotic series are factorially divergent.

This Stokes phenomenon, connecting different exponentials representing the same function, is central to our current understanding of such divergent series, and is the feature that distinguishes them most sharply from convergent ones. In view of this seminal contribution, it is ironic that George (‘G H’) Hardy makes no mention of the Stokes phenomenon in his textbook ‘Divergent series’. Nor does he exempt Stokes from his devastating assessment of 19th century English mathematics: “there [has been] no first-rate subject, except music, in which England has occupied so consistently humiliating a position. And what have been the peculiar characteristics of such English mathematics ... ? ... for the most part, amateurism, ignorance, incompetence, and triviality.”

4 The Modern Period

Late in the 19th century, Jean-Gaston Darboux showed that for a wide class of functions the high derivatives diverge factorially. This would become an important ingredient in later research, for the following reason. Asymptotic expansions (particularly those encountered in physics and applied mathematics) are often based on local approximations: the steepest-descent method for approximating integrals is based on local expansion about a saddle-point, the phase-integral method for solving differential equations (e.g. the Wentzel-Kramers-Brillouin (WKB) approximation to Schrödinger’s equation in quantum mechanics) is based on local expansions of the coefficients, etc. Therefore successive orders of approximation involve successive derivatives, and
the high orders, responsible for the divergence of the series, involve high derivatives.

Another major late 19th century ingredient of our modern understanding was Émile Borel’s development of a powerful summation method in which the factorials causing the high orders to diverge are tamed by replacing them by their integral representation. Often this enables the series to be summed ‘under the integral sign’. Underlying the method is the formal equality

$$
\sum_{r=0}^{\infty} \frac{a_r}{z^r} = \sum_{r=0}^{\infty} \frac{a_r r!}{z^r r!} = \int_{0}^{\infty} dt \, e^{-t} \sum_{r=0}^{\infty} \frac{a_r}{r!} \left( \frac{t}{z} \right)^r .
$$

Reading this from right to left is instructive. Interchanging summation and integration shows why the series on the left diverges if the $a_r$ increase factorally (as in the cases we are considering): the integral is over a semi-infinite range, yet the sum in the integrand converges only for $|t/z| < 1$. Borel summation effectively repairs an analytical transgression that may have caused the divergence of the series. The power of Borel summation is that, as was fully appreciated only later, it can be analytically continued across Stokes lines, where some other summation techniques fail (for example Padé approximants).

Now we come to the central development in modern asymptotics. In a seminal and visionary advance, motivated initially by mathematical difficulties in evaluating some integrals occurring in solid-state physics and developed in a series of papers culminating in a book published in 1973, Robert Dingle synthesized earlier ideas into a comprehensive theory of factorially divergent asymptotic series.

Dingle’s starting point was Euler’s insight that divergent series are obtained by a sequence of precisely specified mathematical operations on the integral or differential equation defining the function being approximated, so the resulting series must represent the function exactly, albeit in coded form, which is the task of asymptotics to decode. Next was the realization that Darboux’s expression of high derivatives in terms of factorials implies that the high orders of a wide class of asymptotic series diverge similarly. In turn this means that the terms beyond Stokes’s optimal truncation—representing the tails of such series beyond supersymptotics—can all be Borel-summed in the same way.

The next insight was Dingle’s most original contribution. Consider a function represented by several different formal asymptotic series (for example those corresponding to the two exponentials in $\text{Ai}(z)$), each representing the function differently in sectors of the complex plane separated by Stokes lines. Since each series is a formally complete representation of the function, each must contain, coded into its high orders, information about all the other series. Thus Darboux’s factorials are simply the first terms of asymptotic expansions of each of the late terms of the original series. Dingle appreciated that the natural variables implied by Darboux’s theory are the differences between the various exponents; usually these are proportional to the large asymptotic parameter. In the simplest case, where there are only two exponentials, there is one such variable, which Dingle called the singulant, denoted $F$. For the Airy function $\text{Ai}(z)$, $F = 4z^{3/2}/3$.

We exhibit Dingle’s expression for the high orders for an integral with two saddle-points $a$ and $b$, corresponding to exponentials $\exp(-F_a)$ and $\exp(-F_b)$ with $F_{ab} = F_b - F_a$ and series with terms $T_n^{(a)}$ and $T_n^{(b)}$: for $r \gg 1$, the terms of the $a$ series are related to those of the $b$ series by

$$
T_r^{(a)} = K \left( \frac{(r-1)!}{F_{ab}^r} \right) \left( T_0^{(b)} + \frac{F_{ab}}{(r-1)} T_1^{(b)} + \frac{F_{ab}^2}{(r-1)(r-2)} T_2^{(b)} + \ldots \right) ,
$$

in which $K$ is a constant. This shows that although the early terms $T_0^{(a)}, T_1^{(a)}, T_2^{(a)}, \ldots$ of an asymptotic series can rapidly get extremely complicated, the high orders display a miraculous functional simplicity.

With Borel’s as the chosen summation method, Dingle’s late terms formula enabled the divergent tails of series to be summed in terms of certain terminant integrals, and then re-expanded to generate new asymptotic series, exponentially small compared with the starting series. He envisaged that “these terminant expansions can themselves be closed with new terminants; and so on, stage after stage.” Such resummations, beyond supersymptotics, were later called hyperasymptotics.

Thus Dingle envisaged a universal technique
for repeated resummation of factorially divergent series, to obtain successively more accurate exponential improvements far beyond that achievable by Stokes’s optimal truncation of the original series. The meaning of universality is that although the early terms—the ones that get successively smaller—can be very different for different functions, the summation method for the tails is always the same, involving terminant integrals that are the same for a wide variety of functions. The method automatically incorporates the Stokes phenomenon. Although Dingle clearly envisaged the hyperasymptotic resummation scheme as described above, he applied it only to the first stage; this was sufficient to illustrate the high improvement in numerical accuracy as compared with optimal truncation.

Like Stokes before him, Dingle presented his new ideas not in the ‘lemma, theorem, proof’ style familiar to mathematicians, but in the discursive manner of a theoretical physicist. Perhaps this is why it has taken several decades for the originality of his approach to be widely appreciated and accepted. Meanwhile his explicit relation, connecting the early and late terms of different asymptotic series representing the same function, was rediscovered independently by several people. In particular, Jean Écalle coined the term resurgence for the phenomenon, and in a sophisticated and comprehensive framework applied it to a very wide class of functions.

5 The Postmodern Period

One of the first steps beyond superasymptotics into hyperasymptotics was an application of Dingle’s ideas to give a detailed description of the Stokes phenomenon. In 1988, one of us (Michael Berry) resummed the divergent tail of the dominant series of the expansion, near a Stokes line, of a wide class of functions including $\text{Ai}(z)$, to give birth to the change in the subdominant exponential, occurring not suddenly as in previous accounts of the phenomenon, but smoothly and in a universal manner. In terms of Dingle’s singulant $F$, now defined as the difference between the exponents of the dominant and subdominant exponentials, the Stokes line corresponds to the positive real axis in the complex $F$ plane, asymptotics corresponds to $\Re F \gg 1$, and the Stokes phenomenon corresponds to crossing the Stokes line, that is $\Im F$ passing through zero.

The result of the resummation is that the change in the coefficient of the small exponential—the Stokes multiplier—is universal for all factorially divergent series, and proportional to

$$\frac{1}{2} \left( 1 + \text{Erf} \left( \frac{\Im F}{\sqrt{2\Re F}} \right) \right).$$

In the limit $\Re F \to \infty$ this becomes the unit step. For $\Re F$ large but finite, the formula describes the smooth change in the multiplier (Figure 4), and makes precise the description given by Stokes in 1902 after thinking about divergent series for more than half a century: “...the inferior term enters as it were into a mist, is hidden for a little from view, and comes out with its coefficient changed. The range during which the inferior term remains in a mist decreases indefinitely as the [large parameter] increases indefinitely.” The smoothing shows that the ‘range’ referred to by Stokes, that is the effective thickness of the Stokes line, is of order $\sqrt{\Re F}$.

The full hyperasymptotic repeated resummation scheme envisaged by Dingle has been implemented in several ways. We (the present authors) investigated one-dimensional integrals with several saddle-points, each associated with an exponential and its corresponding asymptotic se-
Figure 5: Terms in the first four stages in the hyperasymptotic approximation to \( \text{Ai}(4.326\ldots) = 4.496\ldots \times 10^{-4} \), i.e. \( F = 4 z^{3/2}/3 = 12 \), normalized so that the lowest approximation is unity. For the lowest approximation, i.e. no correction terms, the fractional error is \( \varepsilon \approx 0.01 \); after stage 0 of hyperasymptotics, i.e. optimal truncation of the series (supersymptotics), \( \varepsilon \approx 3.6 \times 10^{-7} \); after stage 1, \( \varepsilon \approx 1.3 \times 10^{-11} \); after stage 2, \( \varepsilon \approx 4.4 \times 10^{-14} \); after stage 3, \( \varepsilon \approx 6.1 \times 10^{-15} \). At each stage, the error is of the same order as the first neglected term.

With each ‘hyperseries’ truncated at its least term, this incorporated all subdominant exponentials and all associated Stokes phenomena; and the accuracy obtained far exceeded supersymptotics (Figure 5) but was nevertheless limited.

It was clear from the start that in many cases unlimited accuracy could, in principle, be achieved with hyperasymptotics, by truncating the hyperseries not at the smallest term but beyond it (although this introduces numerical stability issues associated with the cancellation of larger terms). This version of the hypersymptotic programme was carried out by Adri Olde Daalhuis, who reworked the whole theory, introducing mathematical rigor and effective algorithms for computing Dingle’s terminant integrals and their multidimensional generalizations, and applied the theory to differential equations with arbitrary finite numbers of transition points.

There has been an explosion of further developments. Écalle’s rigorous formal theory of resurgence has been developed in several ways, based on the Borel (effectively inverse-Laplace) transform. This converts the factorially-divergent series into a convergent one, with radius of convergence determined by singularities on a Riemann sheet. These singularities are responsible for the divergence of the original series, and for integrals discussed above, correspond to the adjacent saddle points. In the Borel plane, complex and microlocal analysis allows the resurgence linkages between asymptotic contributions to be uncovered and exact remainder terms to be established. Notable results include exponentially accurate representations of quantum eigenvalues (R. Balian & C. Bloch, A. Voros, F. Pham, E. Delabaere); this inspired the work of the current authors on quantum eigenvalue counting functions, linking the divergence of the series expansion of smoothed spectral functions to oscillatory corrections involving the classical periodic orbits.

T. Kawai and Y. Takei in Kyoto have extended ‘formally exact’, exponentially accurate, WKB analysis to several areas, most notably to Painlevé equations. They have also developed a theory of ‘virtual turning points’ and ‘new Stokes curves’. In the familiar WKB situation, with only two wave-like asymptotic contributions, Stokes lines emerge from classical turning points, and never cross. With three or more asymptotic contributions, Stokes lines can cross in the complex plane at points where the WKB solutions are not singular. Local analysis shows that an extra, active, ‘new Stokes line’ sprouts from one side only of this regular point; this can be shown to emerge from a distant virtual turning point, where, unexpectedly, the WKB solutions are not singular. This discovery has been explained by C.J. Howls, P. Langman and A.B. Olde Daalhuis in terms of the Riemann sheet structure of the Borel plane and linked hyperasymptotic expansions, and independently by S.J. Chapman and D.B. Mortimer in terms of matched asymptotics.

Groups led by S.J. Chapman and J.R. King have developed and applied the work of M. Kruskal and H. Segur to a variety of nonlinear and PDE problems. This involves a local matched-asymptotic analysis near the distant Borel singularities that generate the factorially-divergent terms in the expansion, to identify the form of late terms, thereby allowing for an optimal truncation and exponentially accurate approach. Applications include: selection problems in viscous fluids, gravity-capillary solitary
waves, oscillating shock solutions in Kuramoto-Shivashinsky equations, elastic buckling, non-linear instabilities in pattern formation, ship wave modelling and the seeking of reflectionless hull profiles. Using a similar approach, O. Costin and S. Tanveer have identified and quantified the effect of ‘daughter singularities’ not present in initial data of PDE problems, but which are generated at infinitesimally short times. In so doing, they have also found a formally exact Borel representation for small-time solutions of 3D Navier-Stokes equations, offering a promising tool to explore the global existence problem.

Other applications include quantum transitions, quantum spectra, the Riemann zeta function high on the critical line, and even the philosophy of representing physical theories describing phenomena at different scales by singular relations.

Further Reading