Cambridge University Press & Assessment
978-1-108-83662-3 — Analytic Combinatorics in Several Variables
Robin Pemantle , Mark C. Wilson , Stephen Melczer
Excerpt
More Information

PART I

COMBINATORIAL ENUMERATION

Cambridge University Press & Assessment
978-1-108-83662-3 — Analytic Combinatorics in Several Variables
Robin Pemantle , Mark C. Wilson , Stephen Melczer
Excerpt
More Information

1

Introduction

Consider an array[1] of complex numbers

$$\left\{a_{\boldsymbol{r}} : \boldsymbol{r} \in \mathbb{N}^d\right\} := \{a_{r_1,\dots,r_d} : r_1, \dots, r_d \in \mathbb{N}\}$$

where, as in the rest of this book, we include zero in the set $\mathbb{N} = \{0, 1, 2, \dots\}$. The numbers $a_{\boldsymbol{r}}$ usually come with a story – a reason they are interesting. Often, they count a class of objects parametrized by $\boldsymbol{r}$. For example, it could be that $a_{\boldsymbol{r}}$ is the multinomial coefficient $a_{\boldsymbol{r}} = \binom{|\boldsymbol{r}|}{r_1 \cdots r_d}$, in which case $a_{\boldsymbol{r}}$ counts sequences of elements in $\{1, \dots, d\}$ with r_1 occurrences of 1, r_2 occurrences of 2, and so forth up to r_d occurrences of the symbol d. Another frequent source of these arrays is probability theory, where the numbers $a_{\boldsymbol{r}} \in [0, 1]$ are probabilities of events parametrized by $\boldsymbol{r}$. For example, a_{rs} might be the probability that a simple random walk of r steps in $\{-1, 1\}$ ends at the integer point s.

Definition 1.1 (running notation)**.** Throughout this text we use d to denote the dimension of an arbitrary array, and often employ r, s, and t as synonyms for r_1, r_2, and r_3, respectively, so as to avoid subscripts in low-dimensional examples. We also use the notation $|\boldsymbol{r}| := \sum_{j=1}^{d} |r_j|$ for any vector $\boldsymbol{r}$, which helps us normalize in a way convenient for combinatorial examples.

How might one *understand* an array of numbers? In some cases there may be a simple explicit formula, for instance the multinomial coefficients are given by a ratio of factorials. When a formula of such brevity exists, we don't need fancy techniques to describe the array. Unfortunately, this rarely happens. Often, if a formula exists at all, it will not be in closed form but will include indefinite summation. As Stanley [Sta97, Ex.1.1.4] notes in his foundational text on enumeration, "There are actually formulas in the literature (nameless here

[1] To simplify our presentation in this introduction we consider arrays indexed by vectors of natural numbers, while later in the text we generalize to arrays indexed by integer vectors.

forevermore) for certain counting functions whose evaluation requires listing all of the objects being counted! Such a 'formula' is completely worthless." Less egregious are the formulae containing functions that are rare or complicated and whose properties are not immediately familiar to us. It is not clear how much good comes from this kind of formula.

Another way of describing arrays of numbers is via recursions. The simplest examples are finite linear recurrences, such as the recurrence $a_{r,s} = a_{r-1,s} + a_{r,s-1}$ for the binomial coefficients $a_{r,s} = \binom{r+s}{r}$. A recursion for $a_{\boldsymbol{r}}$ in terms of values $\{a_{\boldsymbol{s}} : \boldsymbol{s} < \boldsymbol{r}\}$ whose indices precede $\boldsymbol{r}$ in the coordinatewise partial order may be unwieldy, perhaps requiring evaluation of a complicated function of all $a_{\boldsymbol{s}}$ with $\boldsymbol{s} < \boldsymbol{r}$, but if the recursion is of bounded complexity then it can give an efficient algorithm for computing $a_{\boldsymbol{r}}$. Still, we will see that even in the case of simple recursions the estimation of $a_{\boldsymbol{r}}$ may not be straightforward. Thus, while we look for recursions to help us understand number arrays, and for efficient methods of computation, they rarely provide definitive descriptions.

A third way of understanding an array of numbers is via an estimate. For instance, Stirling's formula, which approximates

$$n! \approx \frac{n^n}{e^n}\sqrt{2\pi n}$$

for large n, yields an approximation

$$\binom{r+s}{r} \approx \left(\frac{r+s}{r}\right)^r \left(\frac{r+s}{s}\right)^s \sqrt{\frac{r+s}{2\pi rs}} \tag{1.1}$$

for the binomial coefficients when r and s are large. If number-theoretic properties of the binomial coefficients are required then we are better off sticking with a ratio of factorials; when their approximate size is paramount, the estimate (1.1) is better.

A fourth way to understand an array of numbers is to encode it algebraically. The *generating function* (often abbreviated GF) of the array $\{a_{\boldsymbol{r}}\}$ is the formal series $F(\boldsymbol{z}) := \sum_{\boldsymbol{r}\in\mathbb{N}^d} a_{\boldsymbol{r}} \boldsymbol{z}^{\boldsymbol{r}}$. Here $\boldsymbol{z}$ is a d-dimensional vector of indeterminates $(z_1, \ldots, z_d)$ and we use the notation $\boldsymbol{z}^{\boldsymbol{r}} := z_1^{r_1} \cdots z_d^{r_d}$. In our running example of multinomial coefficients, we have the generating function

$$F(\boldsymbol{z}) = \sum_{\boldsymbol{r}\in\mathbb{N}^d} \binom{|\boldsymbol{r}|}{r_1 \; \cdots \; r_d} z_1^{r_1} \cdots z_d^{r_d} = \frac{1}{1 - z_1 - \cdots - z_d},$$

where the final expression can be viewed either as a multiplicative inverse in a formal power series ring, or as an analytic function over an appropriate domain of $\mathbb{C}^d$. Stanley calls the generating function "the most useful but the most difficult to understand" method for describing a sequence or array.

The algebraic form of a generating function is intimately related to recursions – and exact formulae – for its coefficient sequence $a_{\boldsymbol{r}}$, as well as combinatorial decompositions for the objects enumerated by $a_{\boldsymbol{r}}$. In a complementary manner, the analytic properties of a generating function correspond to estimates of $a_{\boldsymbol{r}}$.

1.1 Generating functions and asymptotics

In this text we are chiefly concerned with the asymptotic behavior of $a_{\boldsymbol{r}}$ as $\boldsymbol{r} \to \infty$ in certain *directions*. To discuss the behavior of sequences as their indices go off to infinity, we introduce some standard asymptotic notation.

Definition 1.2 (asymptotic notation)**.** If f and g are real-valued functions then we write

- $f = O(g)$ if and only if $\limsup_{x \to x_0} |f(x)/g(x)| < \infty$,
- $f = o(g)$ if and only if $\lim_{x \to x_0} f(x)/g(x) = 0$,
- $f \sim g$ if and only if $\lim_{x \to x_0} f(x)/g(x) = 1$,
- $f = \Omega(g)$ when $g = O(f)$, and
- $f = \Theta(g)$ when both $f = O(g)$ and $g = O(f)$,

for some value x_0 understood in context, typically 0 or $+\infty$.

As $n \to \infty$ the function $f(n)$ is said to be ***rapidly decreasing*** if $f(n) = O\left(n^{-K}\right)$ for every $K > 0$, ***exponentially decaying*** if $f(n) = O(e^{-cn})$ for some $c > 0$, and ***super-exponentially decaying*** if $f(n) = O(e^{-cn})$ for every $c > 0$.

Remark. An alternative definition is that $f = O(g)$ when there exists $C > 0$ and an open neighborhood N of x such that $f(x) \leq Cg(x)$ for all $x \in N$. In this case C is called an ***implied constant***. One may increase C and decrease N and still maintain the inequality, so implied constants are not unique, even if they are chosen to give a tight inequality.

Example 1.3. As $n \to \infty$ the function $f(n) = 1/n!$ decays super-exponentially, while 2^{-n} decays exponentially and $e^{-\sqrt{n}}$ approaches zero but does not decay exponentially. ◁

An ***asymptotic scale*** is a sequence $\{g_j : j \in \mathbb{N}\}$ of functions satisfying $g_{j+1} = o(g_j)$ for all $j \geq 0$. An ***asymptotic expansion*** (also called ***asymptotic series*** or ***asymptotic development***)

$$f \approx \sum_{j=0}^{\infty} c_j g_j$$

for a function f in terms of an asymptotic scale $\{g_j : j \in \mathbb{N}\}$ and constants $c_j \in \mathbb{C}$ is said to hold if

$$f - \sum_{j=0}^{M-1} c_j g_j = O(g_M) \tag{1.2}$$

for every $M \geq 1$.

Remark. It is possible that $c_j = 0$ for all j. For example, this will happen if $g_j(n) = n^{-j}$ and f is exponentially decaying. In this case there is no leading term in the expansion. Otherwise, the ***leading term of an asymptotic expansion*** is the first non-zero term $c_k g_k$ in the expansion.

Example 1.4. Stirling's famous approximation to the factorial can be refined to give an asymptotic series

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \sum_{\ell \geq 0} c_\ell n^{-\ell}$$

with coefficient sequence $\{c_\ell\}$ beginning $1, 1/12, 1/288, -139/51840, \ldots$. ◃

Example 1.5. Let $f \in C^\infty(\mathbb{R})$ be a smooth real function defined on a neighborhood of zero, so that $c_n = f^{(n)}(0)/n!$ is the n^{th} term in its Taylor expansion. If f is not analytic then this expansion may not converge to f, and may even diverge for all non-zero x, but Taylor's Theorem with remainder always implies

$$f(x) = \sum_{n=0}^{M-1} c_n x^n + c_M \xi^M$$

for some $\xi > 0$ bounded close to the origin. This proves that

$$f \approx \sum_{n \geq 0} c_n x^n$$

is always an asymptotic expansion for f near zero. ◃

Remark. Following Poincaré, many authors use the symbol $\sim$ to denote both asymptotic equivalence of functions and asymptotic series expansions. However, this overloading of notation can lead to inconsistencies. We thus follow texts such as [dBru81] in using $\approx$ for asymptotic expansions.

Exercise 1.1. Let $f(x) = e^x$. Prove that $f(x) \sim 1$ as $x \to 0$ but $f(x) \not\approx 1$ as an asymptotic expansion in powers of x at $x = 0$.

All these notations hold in the multivariate case as well, except that if the limit value $\boldsymbol{z}_0$ is infinity then a statement such as $f(\boldsymbol{z}) = O(g(\boldsymbol{z}))$ must also specify how $\boldsymbol{z}$ approaches the limit. A ***direction*** is a ray in $\mathbb{R}^d$ defined by all positive multiples of a fixed non-zero vector, which can also be viewed as an element of $(d-1)$-dimensional real projective space $\mathbb{RP}^{d-1}$. Often we will parametrize directions of interest by taking $\boldsymbol{r} \to \infty$ while fixing or bounding the normalized vector $\hat{\boldsymbol{r}} := \boldsymbol{r}/|\boldsymbol{r}|$, where, as introduced above,

$$|\boldsymbol{r}| = |r_1| + |r_2| + \cdots + |r_d|.$$

Sometimes we shall loosely refer to "the direction $\boldsymbol{r}$", by which we mean the direction parametrized by $\hat{\boldsymbol{r}}$, or the ray determined by $\boldsymbol{r}$.

Definition 1.6. A ***multivariate asymptotic expansion***

$$f_{\boldsymbol{r}} \approx \sum_{j=0}^{\infty} c_j g_j(\boldsymbol{r})$$

holds on a compact set of directions $D \subseteq \mathbb{RP}^{d-1}$ if each $c_j \in \mathbb{C}$, each $g_j = o(g_{j+1})$, and $f_{\boldsymbol{r}} - \sum_{j=0}^{M-1} g_j(\boldsymbol{r}) = O(g_M)$ for each M as $\boldsymbol{r} \to \infty$ with $\hat{\boldsymbol{r}} \in D$. This asymptotic expansion is a ***uniform asymptotic expansion*** on D if the implied constants can be chosen independently of the sequence $\boldsymbol{r}$ as long as $\hat{\boldsymbol{r}} \in D$.

Example 1.7. In Chapter 9 we shall derive the result

$$\binom{r+s}{s} \sim \frac{(r+s)^{(r+s)}}{r^r s^s} \sqrt{\frac{r+s}{2\pi rs}}$$

for all $r, s > 0$ as $(r, s) \to \infty$ with $r/(r+s)$ and $s/(r+s)$ remaining bounded and away from 0. This gives the first term of an asymptotic series which is uniform provided r/s and s/r are bounded away from 0, with all terms in the series varying smoothly with direction. Because of our restrictions on r/s, this asymptotic series can be expressed in terms of the asymptotic scale

$$g_j(r, s) = \frac{(r+s)^{(r+s)}}{r^r s^s} \sqrt{\frac{r+s}{rs}} \, (r+s)^{-j},$$

an asymptotic scale involving decreasing powers s^{-j} of s, or an asymptotic scale involving decreasing powers r^{-j} of r. Note that this multivariate asymptotic approximation is not uniform for all real directions: for instance, if $r = 0$ then $\binom{r+s}{s} = 1$ for all s. ◁

Remark. Throughout this book, we typically use $f(z)$ and a_n instead of $F(z)$ and $a_{\boldsymbol{r}}$ when dealing with the univariate case.

As we will see in Chapter 3, the generating function $f(z)$ for a univariate sequence $\{a_n : n \in \mathbb{N}\}$ leads, almost automatically, to asymptotic estimates for

a_n as $n \to \infty$. To estimate a_n when its generating function f is known, we begin with Cauchy's integral formula

$$a_n = \frac{1}{2\pi i} \int_C z^{-n-1} f(z)\, dz\,. \tag{1.3}$$

Equation (1.3) represents a_n by a complex contour integral on a sufficiently small circle C around the origin, and one may apply complex analytic methods to obtain an asymptotic estimate. The necessary knowledge of residues and contour shifting may be found in an introductory complex variables text such as [Con78b; BG91], with a particularly nice treatment of univariate saddle point integration found in [Hen88; Hen91]. In particular, the singularities of $f(z)$ play a large role in characterizing asymptotic behavior.

The situation for multivariate arrays is nothing like the situation for univariate arrays. In 1974, when Bender published his review article [Ben74] on asymptotic enumeration, the literature on asymptotics of multivariate generating functions was in its infancy. Bender's concluding section urges research in this area:

> Practically nothing is known about asymptotics for recursions in two variables even when a generating function is available. Techniques for obtaining asymptotics from bivariate generating functions would be quite useful.

In the 1980s and 1990s, a small body of results was developed by Bender, Richmond, Gao, and others, giving the first partial answers to asymptotic questions for multivariate generating functions. The first paper to concentrate on extracting asymptotics from multivariate generating functions was [Ben73], already published at the time of Bender's survey, but the seminal paper is [BR83]. The authors work under the hypothesis that F has a singularity of the form $A/(z_d - g(\boldsymbol{x}))^q$ on the graph of a smooth function g, for some real exponent q, where $\boldsymbol{x}$ denotes $(z_1, \ldots, z_{d-1})$. They show, under appropriate further hypotheses on F, that the probability measure μ_n one obtains by renormalizing $\{a_{\boldsymbol{r}} : r_d = n\}$ to sum to 1 converges to a multivariate normal distribution when appropriately rescaled. Their method, which we call the ***GF-sequence method***, is to break the d-dimensional array $\{a_{\boldsymbol{r}}\}$ into a sequence of $(d-1)$-dimensional slices and consider the sequence of $(d-1)$-variate generating functions

$$f_n(\boldsymbol{x}) = \sum_{\boldsymbol{r}: r_d = n} a_{\boldsymbol{r}} \boldsymbol{x}^{\boldsymbol{r}}\,.$$

They show that, asymptotically as $n \to \infty$,

$$f_n(\boldsymbol{x}) \sim C_n g(\boldsymbol{x}) h(\boldsymbol{x})^n \tag{1.4}$$

and that sequences of generating functions obeying (1.4) satisfy a central limit theorem and a local central limit theorem.

The GF-sequence method is limited to the single, though important, case where the coefficients $a_{\boldsymbol{r}}$ are nonnegative and possess Gaussian (central limit) behavior. The work of [BR83] has been greatly expanded upon, but always in a similar framework. For example, it has been extended to matrix recursions [BRW83] and, in [GR92; BR99], from algebraic to algebraico-logarithmic singularities of the form $F \sim (z_d - g(\boldsymbol{x}))^q \log^\alpha(1/(z_d - g(\boldsymbol{x})))$. The difficult step under these hypotheses is deducing asymptotics from the *quasi-power* hypothesis (1.4).

1.2 New multivariate methods

The research presented in this book grew out of several problems encountered by the first author, concerning bivariate and trivariate arrays of probabilities. One might have thought, based on the situation for univariate generating functions, that there would be well-known, neatly packaged results yielding asymptotic estimates for the probabilities in question. At that time, the most recent and complete reference on asymptotic enumeration was a 1995 survey of Odlyzko [Odl95]. As mentioned in the preface, only six of the over 100 pages of the survey are devoted to multivariate asymptotics, mainly to the GF-sequence results of Bender et al., and its section on multivariate methods closes with a call for further work in this area. Evidently, a general asymptotic method was not known in the multivariate case, even for the simplest non-trivial class of rational functions.

This stands in stark contrast to the univariate theory of rational functions, which is trivial in combinatorial applications (see Chapter 3). The relative difficulty of the problem in higher dimensions is perhaps unexpected, but connections to other areas of mathematics such as Morse theory are quite intriguing. These connections, as much as anything else, have caused us to pursue this line of research long after the urgency of the original motivating problems had faded.

Odlyzko [Odl95] describes why he believes multivariate coefficient estimation to be difficult. First, generating function singularities are no longer isolated, but generally form $(d-1)$-dimensional hypersurfaces, so even multivariate rational functions have an infinite set of singularities. Second, the multivariate analogue of the one-dimensional residue theorem is the considerably more difficult theory of Leray residues [Ler59]. This theory is fleshed out in the text of Aizenberg and Yuzhakov [AY83], who also spend a few pages [AY83, Section 23] on generating functions and combinatorial sums. Further progress in using multivariate residues to evaluate coefficients of generating functions was made by Bertozzi and McKenna [BM93], though at the time of Odlyzko's survey none of the papers based on multivariate residues such as [Lic91; BM93] had resulted in any kind of systematic application of these methods to enumeration. It is interesting to note that several of these early works, such as [BM93; KY96], are centered on queueing theory applications.

The focus of this book is a more recent vein of research, begun in [PW02], continued in its infancy in [PW04; Lla03; Wil05; Lla06; RW08; RW11; PW08; DeV10; PW10], and now comprising a stable and ever-growing component of enumerative combinatorics. This research extends ideas that are present to some degree in [Lic91; BM93; KY96], using complex methods that are genuinely multivariate to evaluate coefficients via the multivariate Cauchy formula

$$a_{\boldsymbol{r}} = \left(\frac{1}{2\pi i}\right)^d \int_T \boldsymbol{z}^{-\boldsymbol{r}-\boldsymbol{1}} F(\boldsymbol{z})\, d\boldsymbol{z}\,, \tag{1.5}$$

where T is a suitable product of circles in each variable. We hope that by avoiding the symmetry-breaking decompositions of the GF-sequence method we will obtain methods that are more universally applicable. In particular, much of this past work can be viewed as instances of a more general result estimating the Cauchy integral via topological reductions of the cycle T of integration. These topological reductions, while not fully automatic, are algorithmically decidable in many cases. The ultimate goal, now well on its way to fruition [Mel21, Chapter 7], is to develop software to automate all of the computation.

We can by no means say that the majority of multivariate generating functions fall prey to these new techniques. Nevertheless, as illustrated in this text and a steadily increasing number of papers, we can treat a large number of combinatorially interesting examples. The class of functions to which the methods described in this book may be applied is larger than the class of rational functions, but similar in spirit: the function must have singularities, and the singularities dictating asymptotics must be poles. This translates to the requirement that the function be meromorphic in a neighborhood of a certain polydisk, which means that it has a representation, at least locally, as a quotient of analytic functions.

Throughout this book, we reserve the symbols F, P, and Q for a meromorphic function F expressed as the quotient P/Q of analytic functions with a convergent series expansion

$$F(\boldsymbol{z}) = \frac{P(\boldsymbol{z})}{Q(\boldsymbol{z})} = \sum_{\boldsymbol{r}} a_{\boldsymbol{r}} \boldsymbol{z}^{\boldsymbol{r}} .$$

Although this introduction has focused on power series expansions, we will develop the theory for convergent Laurent expansions, allowing the index $\boldsymbol{r}$ to range over $\mathbb{Z}^d$. The set $\mathcal{V}$ of singularities of F, which is crucial to the asymptotic analysis, is known as its *singular variety*. For instance, if P and Q are coprime polynomials then the singular variety is the algebraic set $\mathcal{V} = \{\boldsymbol{z} \in \mathbb{C}^d : Q(\boldsymbol{z}) = 0\}$.

We now briefly describe the ACSV approach to computing multivariate asymptotics. A more detailed overview is provided in Chapter 7.

(i) Use the multidimensional Cauchy integral (1.5) to express $a_{\boldsymbol{r}}$ as an integral over a d-dimensional torus (product of circles) T in $\mathbb{C}^d$.

(ii) Observe that T may be replaced by any cycle homologous to $[T]$ in $H_d(\mathcal{M})$, where $\mathcal{M}$ is the domain of analyticity of the integrand.

(iii) Deform the cycle T to lower the modulus of the integrand as much as possible. Morse-theoretic arguments imply that local maxima are characterized by the set $\texttt{critical}(\boldsymbol{r})$ of *critical points* of $\mathcal{V}$, which depend only on the direction $\hat{\boldsymbol{r}}$ of $\boldsymbol{r}$ as $\boldsymbol{r} \to \infty$ and are saddle points for the magnitude of the integrand.

(iv) Use algebraic methods to encode the elements of $\texttt{critical}(\boldsymbol{r})$ by a finite collection of equalities and inequalities (defined by polynomials when F is rational).

(v) Use topological methods to find certain minimax cycles $C(\boldsymbol{w})$ near each critical point $\boldsymbol{w}$, termed *quasi-local cycles*, such that the homology class $[T]$ can be represented by a sum $\sum_{\boldsymbol{w}} n_{\boldsymbol{w}} C(\boldsymbol{w})$ with each $n_{\boldsymbol{w}} \in \mathbb{Z}$.

(vi) Refine the set of critical points to the set $\texttt{contrib}(\boldsymbol{r})$ of *contributing points* that maximize the modulus of the Cauchy integrand among the critical points $\boldsymbol{w}$ with $n_{\boldsymbol{w}} \neq 0$. In the vast majority of cases for which we have explicit asymptotic results, it is the case that $n_{\boldsymbol{w}} \in \{0, \pm 1\}$.

(vii) Asymptotically approximate integrals over the $C(\boldsymbol{w})$ as $\boldsymbol{w}$ ranges over the set of contributing points, using a combination of residue and saddle point techniques.

When successful, this approach leads to an asymptotic representation of the coefficients $a_{\boldsymbol{r}}$ that is uniform as $\boldsymbol{r}$ varies on the interior of finitely many cones that partition $\mathbb{R}^d$. As $\hat{\boldsymbol{r}}$ varies over compact subsets in the interior of such cones,