Cambridge University Press
978-1-107-03157-9 - Analytic Combinatorics in Several Variables
Robin Pemantle and Mark C. Wilson
Excerpt
More information

PART I

Combinatorial Enumeration

1

Introduction

1.1 Arrays of Numbers

The main subject of this book is an array of numbers

$$\{a_{r_1,\ldots,r_d} : r_1, \ldots, r_d \in \mathbb{N}\} .$$

This is usually written as $\{a_{\boldsymbol{r}} : \boldsymbol{r} \in \mathbb{N}^d\}$, where as usual $\mathbb{N} = \{0, 1, 2, \ldots\}$. The numbers $a_{\boldsymbol{r}}$ may be integers, real numbers, or even complex numbers. We always use d to denote the dimension of the array. The variables r, s, and t are reserved as synonyms for r_1, r_2, and r_3, respectively, to avoid subscripts in examples of dimensions up to three.

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 $a_{\boldsymbol{r}} := \binom{|\boldsymbol{r}|}{r_1 \cdots r_d}$, with $|\boldsymbol{r}| := \sum_{j=1}^d r_j$, in which case $a_{\boldsymbol{r}}$ counts sequences with r_1 1's, r_2 2's, and so forth up to r_d occurrences of the symbol d. Another frequent source of these arrays is in probability theory. Here, 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 ends at the integer point s.

How might one understand an array of numbers? There might be a simple, explicit formula. The multinomial coefficients, for example, are given by ratios of factorials. As Stanley[1] (1997) points out in the introduction, a formula of this brevity seldom exists; when it does, we don't need fancy techniques to describe the array. Often, if a formula exists at all, it will not be in closed form but will have a summation in it. As Stanley says, "There are actually formulas in the literature (nameless here forevermore) for certain counting functions whose evaluation requires listing all of the objects being counted! Such a 'formula' is completely worthless." (Example 1.1.4 page 2) Less egregious are the formulae containing functions that are rare or complicated and whose properties are not

[1] Much of the presentation in this first section is heavily influenced by Stanley – see the notes to this chapter.

immediately familiar to us. It is not clear how much good it does to have this kind of formula.

Another way of describing arrays of numbers is via recursions. The simplest recursions are finite linear recursions, such as the recursion

$$a_{r,s} = a_{r-1,s} + a_{r,s-1}$$

for the binomial coefficients. A recursion for $a_{\boldsymbol{r}}$ in terms of values $\{a_{\boldsymbol{s}} : \boldsymbol{s} < \boldsymbol{r}\}$ whose indices precede $\boldsymbol{r}$ in the coordinate-wise partial order may be pretty unwieldy, perhaps requiring evaluation of a complicated function of all $a_{\boldsymbol{s}}$ with $\boldsymbol{s} < \boldsymbol{r}$. However, if the recursion is of bounded complexity, such as a linear recursion $a_{\boldsymbol{r}} = \sum_{\boldsymbol{j} \in F} c_{\boldsymbol{j}} a_{\boldsymbol{r}-\boldsymbol{j}}$ for some finite set $\{c_{\boldsymbol{j}} : \boldsymbol{j} \in F\}$ of constants, then the recursion gives a polynomial time algorithm for computing $a_{\boldsymbol{r}}$. Still, even in this case, the estimation of $a_{\boldsymbol{r}}$ is not at all straightforward. Thus, although we look for recursions to help us understand number arrays, recursions rarely provide definitive descriptions.

A third way of understanding an array of numbers is via an estimate. If one uses Stirling's formula

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

one obtains an estimate for binomial coefficients

$$a_{r,s} \sim \left(\frac{r+s}{r}\right)^r \left(\frac{r+s}{s}\right)^s \sqrt{\frac{r+s}{2\pi r s}} \tag{1.1.1}$$

and a similar estimate for multinomial coefficients. If number theoretic properties of $a_{\boldsymbol{r}}$ are required, then we are better off sticking with the formula $(r+s)!/(r!\,s!)$, but when the approximate size of $a_{\boldsymbol{r}}$ is paramount, then the estimate (1.1.1) is better.

A fourth way to understand an array of numbers is to give its generating function. The *generating function* for the array $\{a_{\boldsymbol{r}}\}$ is the series $F(\boldsymbol{z}) := \sum_{\boldsymbol{r}} a_{\boldsymbol{r}} \boldsymbol{z}^{\boldsymbol{r}}$. Here $\boldsymbol{z}$ is a d-dimensional vector of indeterminates $(z_1, \ldots, z_d)$, and $\boldsymbol{z}^{\boldsymbol{r}}$ denotes the monomial $z_1^{r_1} \cdots z_d^{r_d}$. In our running example of multinomial coefficients, the generating function

$$F(\boldsymbol{z}) = \sum_{\boldsymbol{r}} \binom{|\boldsymbol{r}|}{r_1 \ \cdots \ r_d} z_1^{r_1} \cdots z_d^{r_d}$$

is written more compactly as $1/(1 - r_1 - \cdots - r_d)$. Stanley calls the generating function "the most useful but the most difficult to understand" method for describing a sequence or array.

One reason a generating function is useful is that the algebraic form of the function is intimately related to recursions for $a_{\boldsymbol{r}}$ and combinatorial decompositions for the objects enumerated by $a_{\boldsymbol{r}}$. Another reason is that estimates (and exact formulae if they exist) may be extracted from a generating function. In other words, formulae, recursions, and estimates all ensue once a generating function is known.

1.2 Generating Functions and Asymptotics

We employ the usual asymptotic notation, as follows. If f, g are real valued functions, then the statement "$f = O(g)$" is shorthand for the statement "$\limsup_{x \to x_0} |f(x)|/|g(x)| < \infty$." It must be made clear at which value, x_0, the limit is taken; if f and g depend on parameters other than x, it must also be made clear which is the variable being taken to the limit. Most commonly, $x_0 = +\infty$; in the statement $a_n = O(g(n))$, the limit is always taken at infinity. The statement "$f = o(g)$" is shorthand for $f(x)/g(x) \to 0$, again with the limiting value of x specified. Lastly, the statement "$f \sim g$" means $f/g \to 1$ and is equivalent to "$f = (1 + o(1)) \cdot g$" or "$f - g = o(g)$"; again, the variable and its limiting value must be specified. Two more useful notations are $f = \Omega(g)$, which just means $g = O(f)$, and $f = \Theta(g)$, which means that both $f = O(g)$ and $g = O(f)$ are satisfied. An ***asymptotic expansion***

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

for a function f in terms of a sequence $\{g_j : j \in \mathbb{N}\}$ satisfying $g_{j+1} = o(g_j)$ is said to hold if for every $M \geq 1$, $f - \sum_{j=0}^{M-1} g_j = O(g_M)$. This is equivalent to $f - \sum_{j=0}^{M-1} g_j = o(g_{M-1})$. Often we slightly extend the notion of an asymptotic series by saying that $f \sim \sum_{j=0}^{\infty} a_n g_n$ even when some a_n vanish, as long as $f - \sum_{n=0}^{M-1} a_n g_n = O(g_M)$ and infinitely many of the $\{a_n\}$ do not vanish.

A function f is said to be ***rapidly decreasing*** if $f(n) = O(n^{-K})$ for every $K > 0$, ***exponentially decaying*** if $f(n) = O(e^{-\gamma n})$ for some $\gamma > 0$, and ***super-exponentially decaying*** if $f(n) = O(e^{-\gamma n})$ for every $\gamma > 0$.

Example 1.2.1 Let $f \in C^\infty(\mathbb{R})$ be a smooth real function defined on a neighborhood of zero. Thus it has a Taylor expansion whose n^{th} coefficient is $c_n := f^{(n)}(0)/n!$. If f is not analytic, then this expansion may not converge to f (e.g., if $f(x) = e^{-1/x^2}$ then $c_n \equiv 0$) and may even diverge for all nonzero x, but we always have Taylor's remainder theorem:

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

for some $\xi \in [0, M]$. This proves that

$$f \sim \sum_n c_n x^n$$

is always an asymptotic expansion for f near zero. ■

All these notations hold in the multivariate case as well, except that if the limit value of z is infinity, then a statement such as $f(z) = O(g(z))$ must also specify how z approaches the limit. Our chief concern is with the asymptotics of $a_{\mathbf{r}}$ as $\mathbf{r} \to \infty$ in a given direction. More specifically, by a ***direction***, we mean an element of $(d-1)$-dimensional projective space whose class contains a d-tuple of positive real numbers. Often we parametrize positive projective vectors by the corresponding unit vector $\hat{\mathbf{r}} := \mathbf{r}/|\mathbf{r}|$. It turns out that a typical asymptotic formula for $a_{\mathbf{r}}$ is $a_{\mathbf{r}} \sim C|\mathbf{r}|^{\alpha} \mathbf{z}^{-\mathbf{r}}$, where $|\mathbf{r}|$ is the sum of the coordinates of $\mathbf{r}$, and the d-tuple $\mathbf{z}$ and the multiplicative constant C depend on $\mathbf{r}$ only through $\hat{\mathbf{r}}$. In hindsight, formulae such as these make it natural to consider $\mathbf{r}$ projectively and take $\mathbf{r}$ to infinity in prescribed directions. In its original context, the above quote from Stanley referred chiefly to univariate arrays, i.e., the case $d = 1$. As is seen in Chapter 3, it is indeed true that 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 $n \to \infty$. [Another notational aside: we use $f(z)$ and a_n instead of $F(z)$ and a_r in one variable to coincide with notation in the univariate literature.]

To estimate a_n when f is known, begin with Cauchy's integral formula:

$$a_n = \frac{1}{2\pi i} \int z^{-n-1} f(z)\, dz\,. \tag{1.2.1}$$

The integral is a complex contour integral on a contour encircling the origin, and one may apply complex analytic methods to estimate the integral. The necessary knowledge of residues and contour shifting may be found in an introductory complex variables text such as Conway (1978) or Berenstein and Gay (1991), although one obtains a better idea of univariate saddle point integration from Henrici (1988) or Henrici (1991).

The situation for multivariate arrays is nothing like the situation for univariate arrays. In 1974, when Bender published his review article (Bender, 1974) on asymptotic enumeration, the asymptotics of multivariate generating functions was largely a gap in the literature. 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. (page 512)

In the 1980s and 1990s, a small body of results was developed by Bender, Richmond, Gao, and others, giving the first partial answers to questions of asymptotics of generating functions in the multivariate setting. The first article to concentrate on extracting asymptotics from multivariate generating functions was Bender (1973), already published at the time of Bender's survey, but the seminal work is Bender and Richmond (1983). The hypothesis is that F has a singularity of the form $A/(z_d - g(\mathbf{x}))^q$ on the graph of a smooth function g, for some real exponent q, where $\mathbf{x}$ denotes $(z_1, \dots, z_{d-1})$. 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 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

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

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

$$f_n(\boldsymbol{x}) \sim C_n g(\boldsymbol{x}) h(\boldsymbol{x})^n \tag{1.2.2}$$

and that sequences of generating functions obeying (1.2.2) satisfy a central limit theorem and a local central limit theorem.

These results always produce Gaussian (central limit) behavior. The applicability of the entire GF-sequence method is limited to the single, although important, case where the coefficients $a_{\boldsymbol{r}}$ are nonnegative and possess a Gaussian limit. The work of Bender and Richmond (1983) has been greatly expanded upon, but always in a similar framework. For example, it has been extended to matrix recursions (Bender, Richmond, and Williamson, 1983), and the applicability has been extended from algebraic to algebraico-logarithmic singularities of the form $F \sim (z_d - g(\boldsymbol{x}))^q \log^\alpha(1/(z_d - g(\boldsymbol{x})))$ (Gao and Richmond, 1992). The difficult step is always deducing asymptotics from the hypotheses $f_n \sim C_n\, g \cdot h^n$. Thus some publications in this stream refer to such an assumption in their titles (Bender and Richmond, 1999), and the term "quasi-power" has been coined for such a sequence $\{f_n\}$.

1.3 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 results would exist, well known and neatly packaged, that gave asymptotic estimates for the probabilities in question. At that time, the most recent and complete reference on asymptotic enumeration was Odlyzko's 1995 survey (Odlyzko, 1995). Only six of its more than 100 pages are devoted to multivariate asymptotics, mainly to the GF-sequence results of Bender et al. Odlyzko's section on multivariate methods closes with a call for further work in this area. Evidently, in the multivariate case, a general asymptotic formula or method was not known, even for the simplest imaginable class, namely rational functions. This stands in stark contrast to the univariate theory of rational functions, which is trivial (see Chapter 3). The relative difficulty of the problem in higher dimensions is perhaps unexpected. The connections to other areas of mathematics such as Morse theory are, however, quite intriguing, and these, more than anything else, have caused us to pursue this line of research long after the urgency of the original motivating problems had faded.

Odlyzko (1995) describes why he believes multivariate coefficient estimation to be difficult. First, the singularities are no longer isolated, but form $(d-1)$-dimensional hypersurfaces. Thus, he points out, "Even rational multivariate functions are not easy to deal with." Second, the multivariate analogue of the one-dimensional residue theorem is the considerably more difficult theory of Leray (1959). This theory was later fleshed out by Aĭzenberg and Yuzhakov (1983), who spent a few pages in their 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 (1993), although at the time of Odlyzko's survey, none of the works based on multivariate residues such as Lichtin (1991) and Bertozzi and McKenna (1993) had resulted in any kind of systematic application of these methods to enumeration.

The focus of this book is a recent vein of research, begun in Pemantle and Wilson (2002) and continued in Pemantle and Wilson (2004); Lladser (2003); Wilson (2005); Lladser (2006); Raichev and Wilson (2008); Pemantle and Wilson (2008); DeVries (2010); Pemantle and Wilson (2010), and Raichev and Wilson (2012b), as well as several others (Baryshnikov and Pemantle, 2011; DeVries, van der Hoeven, and Pemantle, 2012). This research extends ideas that are present to some degree in Lichtin (1991) and Bertozzi and McKenna (1993), using complex methods that are genuinely multivariate to evaluate coefficients via the multivariate Cauchy formula

$$a_{\mathbf{r}} = \left(\frac{1}{2\pi i}\right)^d \int_T \mathbf{z}^{-\mathbf{r}-\mathbf{1}} F(\mathbf{z})\, d\mathbf{z}\,. \tag{1.3.1}$$

By avoiding symmetry-breaking decompositions such as $F = \sum f_n(z_1, \ldots, z_{d-1}) z_d^n$, one hopes the methods will be more universally applicable and the formulae more canonical. In particular, the results of Bender et al. and the results of Bertozzi and McKenna (1993) are seen to be two instances of a more general result estimating the Cauchy integral via topological reductions of the cycle of integration. These topological reductions, although not fully automatic, are algorithmically decidable in large classes of cases. An ultimate goal, stated in Pemantle and Wilson (2002) and Pemantle and Wilson (2004), 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. 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 dominant singularity must be a pole. This translates to the requirement that the function be meromorphic in a neighborhood of a certain polydisk (see the remark following Pemantle and Wilson [2008, Theorem 3.16] for exact hypotheses), which means that it has a representation, at least locally, as a quotient of analytic functions. Nevertheless, as illustrated in Pemantle and Wilson (2008) and in the present book, meromorphic functions cover a good number of combinatorially interesting examples.

Throughout these notes, we reserve the variable names

$$F = \frac{G}{H} = \sum_{\boldsymbol{r}} a_{\boldsymbol{r}} \boldsymbol{z}^{\boldsymbol{r}}$$

for the meromorphic function F expressed (locally) as the quotient of analytic functions G and H. We assume this representation to be in lowest terms. What this means about the common zeros of G and H will be clearer once stratifications have been discussed. The variety $\{\boldsymbol{z} : H(\boldsymbol{z}) = 0\}$ at which the denominator H vanishes is called the ***singular variety*** and is denoted by $\mathcal{V}$. We now describe the method briefly (more details are provided in Chapter 8).

(i) Use the multidimensional Cauchy integral (1.3.1) to express $a_{\boldsymbol{r}}$ as an integral over a d-dimensional torus 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 holomorphy of the integrand.
(iii) Deform the cycle to lower the modulus of the integrand as much as possible; use Morse theoretic methods to characterize the minimax cycle in terms of *critical points*.
(iv) Use algebraic methods to find the critical points; these are points of $\mathcal{V}$ that depend on the direction $\hat{\boldsymbol{r}}$ of the asymptotics and are saddle points for the magnitude of the integrand.
(v) Use topological methods to locate one or more *contributing* critical points $\boldsymbol{z}_j$ and replace the integral over T by an integral over *quasi-local* cycles $C(\boldsymbol{z}_j)$ near each $\boldsymbol{z}_j$.
(vi) Evaluate the integral over each $C(\boldsymbol{z}_j)$ by a combination of residue and saddle point techniques.

When successful, this approach leads to an asymptotic representation of the coefficients $a_{\boldsymbol{r}}$ of the following sort. The set of directions $\boldsymbol{r}$ is partitioned into finitely many cones K. On the interior of each cone, there is a continuously varying set $\texttt{contrib}(\boldsymbol{r}) \subseteq \mathcal{V}$ that depends on $\boldsymbol{r}$ only through the projective vector $\hat{\boldsymbol{r}}$ and formulae $\{\Phi_{\boldsymbol{z}} : \boldsymbol{z} \in \texttt{contrib}\}$ that involve $\boldsymbol{r}$ and $\boldsymbol{z}(\hat{\boldsymbol{r}})$. Uniformly, as $\boldsymbol{r}$ varies over compact projective subsets of such a cone,

$$\begin{aligned} a_{\boldsymbol{r}} &\sim \frac{1}{(2\pi i)^d} \int_{[T]} \boldsymbol{z}^{\boldsymbol{r}-\boldsymbol{1}} F(\boldsymbol{z})\, d\boldsymbol{z} \\ &= \frac{1}{(2\pi i)^d} \sum_{\boldsymbol{z} \in \texttt{contrib}} \boldsymbol{z}^{-\boldsymbol{r}-\boldsymbol{1}} F(\boldsymbol{z})\, d\boldsymbol{z} \\ &\sim \sum_{\boldsymbol{z} \in \texttt{contrib}} \Phi_{\boldsymbol{z}}(\boldsymbol{r}) . \end{aligned} \tag{1.3.2}$$

The first line of this is steps (i) and (ii). In the second line, the set `contrib` is a subset of the set `critical` of critical points in step (iii). The set `critical` is easy to compute (see step [iv]), whereas determining membership in the subset `contrib` can be challenging (see step [v]). The explicit formulae $\Phi_z(\boldsymbol{r})$ in the last line are computed in step (vi), sometimes relatively easily (Chapter 9) and sometimes with more difficulty (Chapter 10 and especially Chapter 11).

1.4 Outline of the Remaining Chapters

The book is divided into three parts, the third of which is the heart of the subject: deriving asymptotics in the multivariate setting once a generating function is known. Nevertheless, some discussion is required of how generating functions are obtained, what meaning can be read into them, what are the chief motivating examples and applications, and what did we know how to do before the recent spate of research described in Part III. Another reason to include these topics is to make the book into a somewhat self-contained reference. A third is that in obtaining asymptotics, one must sometimes return to the derivation for a new form of the generating function, turning an intractable generating function into a tractable one by changing variables, re-indexing, aggregating, and so forth. Consequently, the first three chapters comprising Part I form a crash course in analytic combinatorics. Chapter 2 explains generating functions and their uses, introducing formal power series, their relation to combinatorial enumeration, and the combinatorial interpretation of rational, algebraic, and transcendental operations on power series. Chapter 3 is a review of univariate asymptotics. Much of this material serves as mathematical background for the multivariate case. Although some excellent sources are available in the univariate case, for example, Wilf (2006), van Lint and Wilson (2001), and Flajolet and Sedgewick (2009), none of these is concerned with providing the brief yet reasonably complete summary of analytic techniques that we provide here. It seems almost certain that someone trying to understand the main subject of these notes will profit from a review of the essentials of univariate asymptotics.

Carrying out the multivariate analyses described in Part III requires a fair amount of mathematical background. Most of this is at the level of graduate coursework, ideally already known by practicing mathematicians but in reality forgotten, never learned, or not learned in sufficient depth. The required background is composed of small to medium-sized chunks taken from many areas: undergraduate complex analysis, calculus on manifolds, saddle point integration (both univariate and multivariate), algebraic topology, computational algebra, and Morse theory. Many of these background topics would be a full semester's course to learn from scratch, which of course is too much material to include here, but we also want to avoid the scenario in which a reference library is required each time a reader picks up this book. Accordingly, we have included substantial background material.