Phase changes in random recursive structures and algorithms*

Hsien-Kuei Hwang
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan
hkhwang@sinica.edu.tw

June 2, 2003

Abstract

A brief survey, based mainly on my recent work with coauthors, is given of the different types of phase changes (or transitions) appearing in random discrete structures and in analysis of algorithms with a recursive character.

Phase-transitions are important tools
because they make it easy to see
two things in one way or
one thing in two ways.
(quoted from the page A book called "n").

One of the most widely known phenomena of phase changes (of matters) is that water may change its state (to ice or to steam) when the underlying temperature varies. For mathematical functions (or structures or objects), we refer to "phase change" when there is a change of properties under varying parameters. When the phase change phenomenon is observed or discovered, the main problems are usually:

– Where does the phase change?

– How to describe the change or transition of phase?

– Why does the change occur? Is there any intuitive interpretation?

– Are there further phase changes in the transition range? and why? how?

The simplest example of a phase change[1] is the classical central limit theorem where the standard normal distribution is used to bridge the two extremes "event unlikely to happen" and "event happens almost always." This viewpoint offers several advantages. First, it makes the usual statement of central limit theorems more concrete and physical; second, its quantitative refinement from the 0-1 law or the law of large numbers becomes transparent; third, it makes the notion of "scaling window" clearer since intuitively the higher the resolution of a telescope, the tinier image or object one can

*Most materials of this paper appeared in my Chinese survey paper [24].

[1]We use mostly the term "phase change" instead of the more common "phase transition" in this paper since there is an obvious notion of discreteness in our problems.

perceive; finally, further refinement of the convergence in distribution may lead to further phase changes.

We survey in this paper, based on our study in the last few years, two different classes of phase changes:

1. phase changes related to the Poisson law; and
2. phase changes related to quicksort.

We will describe the phase changes, the tools used to derive the results, and intuitive interpretations if possible. Some open problems will also be indicated.

1 Phase changes related to the Poisson distribution

Among discontinuous distributions, the Poisson series is of first importance.
— Sir Ronald Aylmer Fisher (1890–1962)

The Poisson distribution usually appears in the form of law of rare events or small numbers (both being described as misnomers, however, by Feller [13]). It is one of the simplest discrete distributions used in modelling real-life problems; typical examples include the number of shark attacks in each summer, the number of students in a class with the same birthday, the number of times of winning the jackpot for the lottery, the number of typos per page made by a secretary, the number of phone calls received by a telephone operator, the number of flaws in a bolt of fabric; see [1, 13, 17, 30] for more information. To probablists, the so-called "misfortunes never come single" may also have a natural connection to Poisson law. The wide-spread use of the Poisson distribution lies partly in its simple definition:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad (k = 0, 1, \dots),$$

where $\lambda > 0$.

Among the many interesting properties of the Poisson distribution, we list the following ones that are closely related to our discussions; see [19]. All asymptotics refer to $\lambda \to \infty$.

1. For $m \geq 0$ and $\lambda - m \gg \sqrt{\lambda}$,
$$P(X \leq m) \sim e^{-\lambda}\frac{\lambda^m}{m!} \cdot \frac{1}{1 - m/\lambda}; \tag{1}$$
2. For $m = \lfloor \lambda + x\sqrt{\lambda} \rfloor$, where $x = o(\lambda^{1/6})$,
$$P(X \leq m) \sim \Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t;$$
3. For $m - \lambda \gg \sqrt{\lambda}$,
$$P(X \leq m) \sim 1.$$

These simple approximations reflect the trichotomous limiting behaviors of the Poisson distribution function, which is also inherent in many structures.

We briefly interpret these results. When m is small (first case), we can rewrite (1) as

$$P(X \leq m) \sim \frac{P(X = m)}{1 - m/\lambda},$$

meaning that the largest term $P(X = m)$ has a significant contribution to the distribution function (or $P(X \le m)$ behaves essentially like $P(X = m)$ for small m); when m lies around the mean value λ, the Poisson distribution is well approximated by a normal distribution; when m goes further to the right, the Poisson distribution approaches 1 in the limit. This reconfirms the phase change interpretation of the central limit theorems given above. In particular, the phase change occurs at $m \sim \lambda$ and the standard normal distribution is used to describe the phase transition. It also introduces another important notion: *the discovery (or observation) of new phenomena relies heavily on the efficiency of the tools used* since proving central limit theorems is usually more sophisticated than, say the zero-one law. Such a notion will appear repeatedly later in this paper. Also if one is interested in more refined approximations, then more tools and efforts are needed.

1.1 Maxima in $[0,1]^d$: from Poisson to constant

Multidimensional data have no total ordering. A natural partial order is the following "dominance" relation: given two points $A = (a_1, \dots, a_d)$ and $B = (b_1, \dots, b_d)$, where $d \ge 1$, we say that A *dominates* B if $a_i > b_i$ for all $i = 1, \dots, d$. The *maxima* or *maximal points* of a sample of points are those points not dominated by any other point. This simple partial ordering is widely used in diverse fields such as engineering, economics, operational research, sociology, etc.; see [3, 4] and the references therein. For example, if student A outperforms another student B in all subjects, then A is usually considered to be better than B. In such a case, there is no special advantage in using dominance. However, if one student performs best in one subject and worst in all others, then how should this student be classified? good or bad? From the dominance relation, this student is one of the maxima, and thus should not be ranked as very poor or bad. This viewpoint offers a more positive perspective for such students.

Assume now we take n iid points in $[0,1]^d$. How many maxima are there? Let $M_{n,d}$ denote the expected number of maxima. Then one easily sees that

$$\begin{aligned} M_{n,d} &= n \int_{[0,1]^d} (1 - x_1 \cdots x_d)^{n-1} \, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\ &= \sum_{1 \le k \le n} \binom{n}{k} (-1)^{k+1} k^{1-d} \qquad (n, d \ge 1). \end{aligned} \tag{2}$$

Question: When d varies with n, how does the asymptotic behaviors of $M_{n,d}$ change?

Before addressing this question, we naturally ask "why study this problem?" One concrete reason is that for practical considerations, n and d are always finite, and the order of d as a function of n is not uniquely determined. For example, if $n = 10^4$ and $d = 10$. Is $d = n^{1/4}$ or $\lceil \log^2 n \rceil$ or simply $O(1)$? This is a common situation in "uniform asymptotics," where a second parameter is varying with the major asymptotic parameter and one is interested in finding approximations that are uniform with respect to the second parameter (at least in some range).

Note that $M_{n,d}$ can be computed recursively by

$$M_{n,d} = M_{n-1,d} + \frac{M_{n,d-1}}{n}.$$

If we look closely at (2), we see that $(-1)^k$ plays a special role in cancelling the contribution of terms. For example, we have, by (2), $M_{n,1} = 1$ and

$$M_{n,2} = H_n := \sum_{1 \le j \le n} j^{-1} \sim \log n.$$

Thus, although individual terms can grow as large as $\binom{n}{\lfloor n/2 \rfloor} \asymp 2^n n^{-1/2}$ (exponential), the resulting sum is merely logarithmic. This also means that the practical usefulness of (2) is limited due to the *exponential cancellation.*

A more useful expression is the following integral expression (see [2, 14])

$$M_{n,d} = \frac{1}{2\pi i} \oint_{|z|=r<1} z^{-d} \prod_{1 \le j \le n} \frac{1}{1 - z/j} \, \mathrm{d}z \qquad (n, d \ge 1).$$

Observe that the product in the integrand can be decomposed as

$$\prod_{1 \le j \le n} \frac{1}{1 - z/j} = \frac{1}{1-z} \prod_{2 \le j \le n} \frac{1}{1 - z/j} = \frac{e^{(H_n - 1)z}}{1-z} g_n(z),$$

where $g_n(z)$ is analytic for $|z| < 2$. Thus the integrand has a simple pole at $z = 1$ and a saddlepoint at $(d-1)/(H_n - 1)$, and the saddlepoint approaches the simple pole when d is around $\log n$. From this integral representation, we can deduce, by complex-analytic tools, that (see [21, 20] for the tools needed)

$$M_{n,d} \sim \Gamma(2 - \rho) P(X_n < d), \tag{3}$$

uniformly for all possible forms of variations of d, where $\rho = \min\{1, (d-1)/\log n\}$, Γ is the Gamma function, and X_n follows a Poisson($\log n$) distribution. This expression can be re-written, by the asymptotic approximations (1)–(3) of the Poisson distribution, as follows.

$$\frac{M_{n,d}}{n} \sim \begin{cases} \dfrac{(\log n)^{d-1}}{n(d-1)!} \Gamma\left(1 - \dfrac{d}{\log n}\right), & \text{if } \log n - d \gg \sqrt{\log n}; \\ \Phi\left(\dfrac{d - \log n}{\sqrt{\log n}}\right), & \text{if } |d - \log n| = o((\log n)^{2/3}); \\ 1, & \text{if } d - \log n \gg \sqrt{\log n}. \end{cases}$$

These more transparent approximations are also intuitively clear: when d is very large, almost all points are maximal (when the number of subjects is increasing, it is becoming less likely for one student to dominate another student). But the fact that the "phase change" occurs at $d \approx \log n$ is not easy to guess intuitively. Also a natural question is: "why $M_{n,d}/n$ is so close to a Poisson distribution?" Is there a more intuitive interpretation? Finally, is there a more probabilistic (instead of complex-analytic) proof for (3)? See [2, 3, 4] for more results and references on probabilistic properties of maxima.

1.2 Irreducibles in polynomials: from Poisson to negative binomial

Given a finite field F_q, where q is a prime power. Assume that all q^n monic polynomials of degree n are equally likely. Let Y_n denote the number of irreducible factors (counted with multiplicity) in the prime factorization of a random polynomial. Defining $P_n(y) = q^n E(y^{Y_n})$, then (see [16])

$$\sum_{n \ge 0} P_n(y) z^n = \prod_{j \ge 1} \left(1 - y z^j\right)^{-I_j}, \tag{4}$$

where

$$\sum_{n \ge 1} I_n z^n = \sum_{j \ge 1} \frac{\mu(j)}{j} \log \frac{1}{1 - q z^j},$$

$\mu(n)$ being the Möbius function: $\mu(n) = 0$ if n is not square-free; $\mu(n) = (-1)^k$ if $n = p_1 \cdots p_k$ with distinct prime numbers $p_1, \dots, p_k$. From this generating function, we have the recurrence $P_0(y) = 1$ and

$$P_n(y) = n^{-1} \sum_{1 \le k \le n} P_{n-k}(y) \sum_{j|k} j I_j y^{k/j} \qquad (n \ge 1).$$

By a detailed analysis of the generating function (4) in the z- and y-plane, we deduce that (see [22])

$$P(Y_n = m) \sim \begin{cases} h\left(\dfrac{m-1}{\log n}\right) \dfrac{n(\log n)^{m-1}}{(m-1)!}, & \text{if } m \ge 1,\ q\log n - m \gg \sqrt{q \log n}; \\ c_1(q) \dfrac{(\log n)^{m-1}}{n(m-1)!} m^{q/2} e^{x^2/4} D_{-q}(-x), & \text{if } x := \dfrac{m - q\log n}{\sqrt{q \log n}} = o((\log n)^{1/6}); \\ c_2(q) \dfrac{(m - q\log n)^{m-1}}{(q-1)!} (n-m)^{q-1} q^{-m}, & \text{if } n - m \to \infty,\ m - q\log n \gg \sqrt{\log n}, \end{cases}$$

where $h(z)$ is analytic in $|z| < q$, $c_1(q), c_2(q)$ are two positive constants independent of n and m, and $D_{-\nu}(x)$ denotes the parabolic cylinder functions

$$D_{-\nu}(x) = \frac{e^{-x^2/4}}{\Gamma(\nu)} \int_0^\infty t^{\nu-1} e^{-xt - t^2/2} \, \mathrm{d}t \qquad (\nu > 0).$$

[In particular, $D_0(x) = e^{-x^2/4}$ and $D_{-1}(x) = \sqrt{2\pi}\, e^{x^2/4} \Phi(-x)$.]

In words, the phase change occurs for m near $q \log n$: when m is small, the probability that the number of polynomials with m total irreducible factors behaves asymptotically like a Poisson($\log n$) distribution; when m is large the probability is roughly like a negative binomial distribution, the transition being well described by the parabolic cylinder function. See [22] for more precise and uniform estimates, as well as a uniform approximation in terms of the convolution law of a Poisson and a negative binomial distributions.

The analytic context encountered here is more complicated than the previous one (for $M_{n,d}$) and consists of a saddlepoint and a pole of order q in the integrand. Thus the appearance of $D_{-q}(x)$ is quite expected; see [5, 31].

An intuitive interpretation of the result is that when Y_n is large, most of the irreducible factors are of degree 1. Thus we can write $Y_n \stackrel{d}{=} Y_n' + Z_n$, where Y_n' and Z_n count the number of irreducible factors of degree ≥ 2 and $= 1$, respectively, and prove that the Poisson behavior comes from Y_n' and that of negative binomial from Z_n; see [22] for more details.

See also [20, 21] for problems in combinatorial structures and in number theory with similar asymptotic behaviors (from Poisson to geometric).

1.3 Consecutive records in iid sequences: from Poisson to non-Poisson

The *records* (or record-breakings) of a given sequence are the elements whose values are larger than all previous ones.

Question: Given iid continuous random variables $X_1, \dots, X_n$, let $Y_{n,r}$ denote the number of times r consecutive records occur, where $r \ge 1$. What is the asymptotic distribution of $Y_{n,r}$?

The problem for $r = 1$ is an old one and much has been known since Rényi (see the survey paper [29]); in particular,

$$E(y^{Y_{n,1}}) = \prod_{1 \le j \le n} \left(1 + \frac{y-1}{j}\right) \qquad (n \ge 1).$$

From this one can show that (see [23])

$$Y_{n,1} \sim \text{Poisson}(\log n) \sim N(\log n, \log n),$$

where $N(a,b)$ denotes a normal variate with mean a and variance b.

When $r \geq 2$, the probability generating functions $F_{n,r}(y)$ of $Y_{n,r}$ satisfy the recurrence (see [10]): $F_{n,r}(y) = 1$ for $n < r$, and

$$F_{n,r}(y) = \frac{n+y-1}{n} F_{n-1,r}(y) + (1-y) \sum_{2 \leq j \leq r} \frac{(n-j)F_{n-j,r}(y)}{n(n-1)\cdots(n-j+1)},$$

for $n \geq r$.

From this recurrence, we obtain the following differential equation for the bivariate generating function $f(z,y) := \sum_{n \geq 0} E(y^{Y_{n,r}}) z^n$:

$$(1-z)f^{(r)} = (r-(1-y)(1-z))f^{(r-1)} + (1-y) \sum_{1 \leq j \leq r} (z+j) f^{(j)},$$

where $f^{(j)} = (\partial z^j / \partial^j) f(z,y)$. Then a detailed study of this differential equation leads to (see [10])

$$E(y^{Y_{n,r}}) = \phi_r(y) \left(1 + \frac{1-y}{r-1} n^{1-r} + O(n^{-r})\right) \qquad (r \geq 2),$$

where $\phi_r(y)$ is an entire probability generating function. In particular, $\phi_2(y) = e^{y-1}$ and $\phi_r(y) \neq e^{y-1}$ for $r \geq 3$. It follows that $Y_{n,2} \sim \text{Poisson}(1)$ and $Y_{n,r} \sim Y_r$, where Y_r is not Poisson for $r \geq 3$.

But what is Y_r? Is there a more explicit characterization? Only for $r = 3$ do we know that $E(y^{Y_3})$ is expressible in terms of the confluent hypergeometric functions; see [10]. More transparent representations for the probability generating function $E(y^{Y_r})$ for higher values of r remain open.

Of course, no matter how we characterize Y_r, the probability $P(Y_r = 0)$ tends to 1 very fast as r grows, meaning that it is harder to find an r-consecutive record for higher values of r. For example, $P(Y_{10} = 0) = 0.99999\,97213\ldots$ and $P(Y_{15} = 0) = 0.99999\,99999\,88\ldots$ Although this example is somewhat factitious when compared with the usual phase change phenomena, the problems and challenges it offers are typical and representative.

2 Phase changes related to Quicksort

It [Quicksort] alone was enough to make a man famous.
— Edsger W. Dijkstra (1930–2002)

Quicksort was invented by Hoare some 40 years ago (see [18]). It has been one of the most widely used sorting algorithms and was selected to be among the top ten algorithms in the 20th century having "the greatest influence on the development and practice of science and engineering"; see [12]. The popularity of the quicksort is mainly due to its *simplicity* and *efficiency*; see [9] for more references.

The simplest quicksort works as follows. To sort n elements (in increasing order), first take a random element as the *pivot* (referred to as "bound" in [18]) and then partition the $n-1$ remaining elements into two groups with the values of the elements smaller and larger, respectively, than the value of the pivot. Then the same procedure is applied recursively to both groups until the subfiles are sorted (namely, with one or less element).

To understand the stochastic behavior of this algorithm, the simplest model is to assume that the file to be sorted is a sequence of iid random variables with the same continuous distribution, and then to compute the number of comparisons (or element exchanges, partitioning stages, etc.) used.

But the reader may wonder if the model used is too ivory-towered? This is a general issue and may be viewed from different angles. First, the iid model, although possibly too idealized, gives often results that also hold in more general, complex models, which are usually less tractable mathematically. Second, this model is independent of machines, languages, programs, etc. Third, it is simple enough yet one can usually derive meaningful mathematical results of wide generality.

If we accept such an iid model, then the cost, denoted by Y_n, of quicksort satisfies in general the recurrence (assuming that randomness of each subproblem is also preserved): $Y_0 := 0$, and

$$Y_n \stackrel{d}{=} Y_{I_n} + Y^*_{n-1-I_n} + T_n \qquad (n \geq 1),$$

where $Y_n \stackrel{d}{=} Y_n^*$, $I_n = \text{Uniform}[0, 1, \ldots, n-1]$ in the simplest case and T_n denotes the cost used to split the original problem into two smaller subproblems. Here (Y_n), (Y_n^*), and (Y_n, I_n) are independent. We studied two types of phase changes (see [9, 26]):

1. Type I: the limit law of Y_n changes from normal to non-existence if we fix T_n (say $T_n \equiv 1$) and vary I_n;

2. Type II: the limit law of Y_n changes from normal to non-normal if we fix I_n (say $I_n = \text{Uniform}[0, 1, \ldots, n-1]$) and vary T_n.

These two types of phase changes are similar to those in statistical physics where type I phase transition is discrete in nature and type II continuous; see also [7] for another type of phase transition for quicksort.

2.1 Type I phase change: from normal to non-existence

A simple instance of Type I phase change is the following quicksort using median-of-$(2t+1)$: instead of choosing the pivot uniformly at random, choose the pivot as the median of a sample of $2t+1$ elements, where $t \geq 1$, and then partition the remaining elements as quicksort proper described above. The same median-of-$(2t+1)$ procedure is applied to subproblems of sizes $\geq 2t+1$.

Let Y_n denote the total number of steps the median-of-$(2t+1)$ partitioning procedure is used. Then Y_n satisfies $Y_n = 0$ for $0 \leq n \leq 2t$ and

$$Y_n \stackrel{d}{=} Y_{I_n} + Y^*_{n-1-I_n} + 1 \qquad (n \geq 2t+1), \tag{5}$$

where $Y_n \stackrel{d}{=} Y_n^*$ (with (Y_n), (Y_n^*), and (Y_n, I_n) independent) and

$$P(I_n = k) = \frac{\binom{k}{t}\binom{n-1-k}{t}}{\binom{n}{2t+1}} \qquad (k = 0, \ldots, n-1).$$

The interesting result for Y_n is (see [9]): *for* $1 \leq t \leq 58$

$$\frac{Y_n - E(Y_n)}{\sqrt{\text{Var}(Y_n)}} \xrightarrow{d} N(0,1);$$

and for $t > 58$, *the limiting distribution of* $(Y_n - E(Y_n))/\sqrt{\text{Var}(Y_n)}$ *does not exist.*

One naturally wonders why the limit law changes its nature? and why 58? Analytically, 58 is closely related to the magnitude of the second largest zeros (in real part) of the indicial equation

$$(z+t)\cdots(z+2t) = \frac{(2t+2)!}{(t+1)!}. \tag{6}$$

Denote the real part of the second largest zeros by α ($z = 2$ being the dominant zero). If $\alpha < 1.5$ or equivalently $1 \le t \le 58$, then

$$\begin{cases} E(Y_n) = \mu n + O(n^{1/2-\varepsilon}); \\ \mathrm{Var}(Y_n) \sim \sigma^2 n, \end{cases}$$

and the asymptotic normality of Y_n is intuitively well expected; when $\alpha > 1.5$ (or $t > 58$),

$$\begin{cases} E(Y_n) \sim \mu n + P_1(\log n) n^{\alpha-1}; \\ \mathrm{Var}(Y_n) \sim P_2(\log n) n^{2\alpha-2}, \end{cases}$$

where $P_1(u)$ and $P_2(u)$ are bounded periodic functions. Note that $2\alpha - 2 > 1$ so that the variance is larger than the mean. Also for higher central moments, we have

$$E\left(\frac{Y_n - E(Y_n)}{\sqrt{\mathrm{Var}(Y_n)}}\right)^k \sim P_k(\log n) \qquad (k \ge 3),$$

where the P_k's are bounded periodic functions and do not lead to simple forms (they are very messy indeed). However, by proper applications of the Frechet-Shohat moment convergence theorem, we can show that the limit law of $(Y_n - E(Y_n))/\sqrt{\mathrm{Var}(Y_n)}$ does not exist; see [8].

Is there a more intuitive, apart from the preceding analytic, interpretation? Of course, for this problem 58 is the product of refined analysis, so it is unlikely that a simple intuitive argument is sufficient to properly describe the change of the limit laws before and after 58. However, we can give a rough description of the underlying process at play and see why the limit laws undergo a phase change.

By the recursive definition (5) of Y_n, the calculation of the distribution of Y_n is reduced to those of smaller values, which in turn are reduced to those of the degenerate random variables Y_j, $1 \le j \le 2t$. Thus Y_n can be written as the linear combination of 1 and the sum of many degenerate random variables; and thus when t grows, the variance is increasing and then the limit law changes its nature from a certain value of t on.

This rough sketch can be made slightly more precise. By definition,

$$Y_n = \begin{cases} 0, & \text{if } 1 \le n \le 2t; \\ 1, & \text{if } 2t+1 \le n \le 3t+1; \\ \{1,2\}, & \text{if } 3t+2 \le n \le 4t+2; \\ \{2,3\}, & \text{if } 4t+3 \le n \le 5t+3; \\ \cdots. \end{cases}$$

The regularity is extended each time in a window of length $t+1$. Such a block-wise recursiveness is the source of the periodicity in the moments of Y_n (centered or not), and the periodicity in turn is the wellhead for the phase change. We can show that if the periodicity in the second order term of the mean can somehow be removed or smoothed out, then the limit law exists and is non-normal; see [8].

The phase change of Y_n at $t = 58$ is just the tip of an iceberg. We can systematically produce phase changes at other values; see [8, 9]. For example, consider the random variables Z_n defined by $Z_0 = 0$, $Z_n = 1$ for $1 \le n \le m-1$, where $m \ge 3$, and

$$Z_n \stackrel{d}{=} Z^{[1]}_{I_n(1)} + \cdots + Z^{[m]}_{I_n(m)} + 1 \qquad (n \ge m),$$

where the $Z_n^{[i]}$'s are independent, identical copies of Z_n, and

$$P(I_n(1) = i_1, \ldots, I_n(m) = i_m) = \binom{n}{m-1}^{-1},$$

for all nonnegative tuples $(i_1, \ldots, i_m)$ satisfying $i_1 + \cdots + i_m = n - m + 1$. Such Z_n's represent the storage requirement of random m-ary search trees, and it is proved in [8] that Z_n *is asymptotically normally distributed for* $3 \leq m \leq 26$*, and that the limit distribution of* $(Z_n - E(Z_n))/\sqrt{\mathrm{Var}(Z_n)}$ *does not exist for* $m > 26$.

For this class of problems, the main approach we use is the method of moments, together with the development of the so-called "asymptotic transfers" (linking the asymptotics of the non-homogeneous part of a recurrence to that of the recurrence). Traditionally, the method of moments is a primitive approach to proving a limit law; its use has been limited due to its "brute-force" nature. But for recursively defined random variables, the application of the method of moments proved fruitful. The main features are: (*i*) all moments (centered or not) satisfy the same type of recurrence, so that all asymptotic information needed is reduced to the derivation of the "asymptotic transfers," making possible the systematic use of the method; (*ii*) further refinement of the method leads not only to the optimal Berry-Esseen bound (or Kolmogorov distance) but also to local limit theorems, which turn out to exhibit further phase changes. For example, for Y_n defined in (5), we can prove that (see [25]) for $1 \leq t \leq 43$

$$\sup_x \left| P\left(\frac{Y_n - E(Y_n)}{\sqrt{\mathrm{Var}(Y_n)}} < x \right) - \Phi(x) \right| = O(n^{-1/2});$$

and for $44 \leq t \leq 58$

$$\sup_x \left| P\left(\frac{Y_n - E(Y_n)}{\sqrt{\mathrm{Var}(Y_n)}} < x \right) - \Phi(x) \right| = O(n^{-3(\alpha-3/2)}),$$

where α is, as described before, the real part of the second largest zero(s) of the indicial equation (6). The rates are, up to implied constants in the O-symbols, optimal in each case. See Table 1 for approximate values of α and $3(\alpha - 3/2)$ when t varies from 44 to 58.

t	α	$3(3/2-\alpha)$	t	α	$3(3/2-\alpha)$
44	1.33764	0.48705	52	1.43798	0.18603
45	1.35210	0.44368	53	1.44843	0.15469
46	1.36594	0.40217	54	1.45850	0.12449
47	1.37920	0.36238	55	1.46820	0.09537
48	1.39192	0.32422	56	1.47757	0.06728
49	1.40413	0.28760	57	1.48661	0.04015
50	1.41586	0.25241	58	1.49534	0.01395
51	1.42713	0.21858	59	1.50378	< 0

Table 1: *Numeric values of* α *and* $3(3/2-\alpha)$ *for* t *from 44 to 58.*

The proof is much more involved and relies on the uniform bounds

$$|E(Y_n - N(E(Y_n), \mathrm{Var}(Y_n)))^k| \leq k!A^k \times \begin{cases} n^{k/3}, & \text{if } 1 \leq t \leq 43; \\ n^{k(\alpha-1)}, & \text{if } 44 \leq t \leq 58, \end{cases}$$

for all $n, k \geq 1$, where $A > 0$ is a sufficiently large constant. This result reflects again that the discovery (or observation) of new phenomena relies heavily on the efficiency of the tools used.

See also [6, 27, 28] for other approaches to the limit laws of Z_n.

2.2 Type II phase change: normal to non-normal

Consider the random variables Y_n defined by $Y_0 = 0$ and

$$Y_n \stackrel{d}{=} Y_{I_n} + Y^*_{n-1-I_n} + T_n \qquad (n \geq 1),$$

where $I_n = \text{Uniform}[0, 1, \ldots, n-1]$, T_n is given and the Y_n^*'s are identical copies of Y_n with (Y_n), (Y_n^*), and (Y_n, I_n) independent.

Question: How does the limit law change under varying T_n?

Intuitively, if each T_n is not large, then Y_n is roughly the sum of many small independent random variables, which, according to the classical law of errors, is expected to be asymptotically normally distributed for large n. On the other hand, if T_n is large, then Y_n is dictated by some large T_n's, and one expects a non-normal limit law if it exists. This intuition can be rephrased in more vivid terms: when T_n is small, one can think of a democratic system where each vote or individual has more or less the same contribution to the whole system, and the system can be maintained in a "normal" way; on the other hand, if some or few individuals have excessive influence to the system (like dictatorship or totalitarian), then the system has larger variance and is likely to become "abnormal".

More precisely, we can show that when $E(T_n) = O(n^{1/2}L(n))$, where $L(n)$ is slowly varying at infinity, then, under some regularity conditions, Y_n is asymptotically normally distributed. On the other hand, when $E(T_n) \gg n^{1/2}$, then the limit law of Y_n, under suitable assumptions, exists and is non-normal. We see that $n^{1/2}$ is the dividing line separating normal and non-normal limit laws.

Also we can further apply the refined method of moments (see [25]) to show that $n^{1/3}$ is the threshold separating good and bad convergence rates.

Such a framework applies to a large number of concrete problems in data structures and algorithms; see [26].

See Devroye [11] for a different approach using Stein's method.

3 Conclusions

Phase changes are ubiquitous. Their real meaning and description rely heavily on observer's tools for handling different scales and uniformities. The questions we highlighted in the Introduction fall into different levels, some easy and some hard. Researchers usually have to try several different viewpoints, approaches to think of deeper structural characteristics, to connect the similarities of different objects, and to unveil possibly the universality of phenomena. Phase changes are highly interesting not only because of their physical concreteness, but also due to these diverse perspectives, which are usually fascinating and challenging.

Acknowledgement

I thank Ralph Neininger for useful comments.

References

[1] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, New York, 1989

[2] Z.-D. Bai, C.-C. Chao, H.-K. Hwang and W.-Q. Liang, On the variance of the number of maxima in random vectors and its applications, *Annals of Applied Probability*, **8** (1998), 886–895.

[3] Z.-D. Bai, H.-K. Hwang, W.-Q. Liang and T.-H. Tsai, Limit theorems for the number of maxima of random samples from planar regions, *Electronic Journal of Probability*, **6** (2001), Paper 3, 41 pages.

[4] Z.-D. Bai, H.-K. Hwang and T.-H.Tsai, Berry-Esseen bounds for the number of maxima in planar regions, *Electronic Journal of Probability*, accepted for publication; available at algo.stat.sinica.edu.tw.

[5] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Second Edition, Dover Publications Inc., New York (1986).

[6] B. Chauvin and N. Pouyanne, m-ary search trees when $m \geq 27$: a strong asymptotics for the space requirement, preprint (2002); available at fermat.math.uvsq.fr/$\sim$chauvin/mary.ps.

[7] H.-H. Chern and H.-K. Hwang, Transitional behaviors of the average cost of quicksort with median-of-$(2t+1)$, *Algorithmica*, **29** (2001), 44–69.

[8] H.-H. Chern and H.-K. Hwang, Phase changes in random m-ary search trees and generalized quicksort, *Random Structures and Algorithms*, 19 (2001), 316–358.

[9] H.-H. Chern, H.-K. Hwang and T.-H. Tsai, An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms, *Journal of Algorithms*, **44** (2002), 177-225.

[10] H.-H. Chern, H.-K. Hwang and Y.-N. Yeh, Distribution of the number of consecutive records, *Random Structures and Algorithms*, **17** (2000), 169–196.

[11] L. Devroye, Limit laws for sums of functions of subtrees of random binary search trees, *SIAM Journal on Computing*, **32** (2002), 152–171.

[12] J. Dongarra and F. Sullivan, Guest editors' introduction: the top 10 algorithms, *Computing in Science and Engineering*, **2**:1 (2000), 22-23.

[13] W. Feller, *An Introduction to Probability Theory and its Applications*, Volume I, Wiley, 1968.

[14] P. Flajolet, G. Labelle, L. Laforest and B. Salvy, Hypergeometrics and the cost structure of quadtrees, *Random Structures and Algorithms*, **7** (1995), 117-144.

[15] P. Flajolet and A. Odlyzko, Singularity analysis of generating functions, *SIAM Journal on Discrete Mathematics*, **3** (1990), 216–240.

[16] P. Flajolet, M. Soria, Gaussian limiting distributions for the number of components in combinatorial structures, *Journal of Combinatorial Theory, Series A*, **53** (1990), 165-182.

[17] F. A. Haight, *Handbook of the Poisson Distribution*, John Wiley & Sons, New York, 1967.

[18] C. A. R. Hoare, Quicksort, *Computer Journal*, **5** (1962), 10–15.

[19] H.-K. Hwang, Asymptotic estimates of elementary probability distributions, *Studies in Applied Mathematics*, **99** (1997), 393-417.

[20] H.-K. Hwang, A Poisson $*$ geometric convolution law for the number of components in unlabelled combinatorial structures, *Combinatorics, Probability and Computing*, **7** (1998), 89–110.

[21] H.-K. Hwang, Sur la répartition des valeurs des fonctions arithmétiques: le nombre de facteurs premiers d'un entier, *Journal of Number Theory*, **69** (1998), 135–152.

[22] H.-K. Hwang, A Poisson $*$ negative binomial convolution law for random polynomials over finite fields, *Random Structures and Algorithms*, **13** (1998), 17–47.

[23] H.-K. Hwang, Asymptotics of Poisson approximation to random discrete distributions: an analytic approach, *Advances in Applied Probability*, **31** (1999), 448–491.

[24] H.-K. Hwang, Phase changes in random recursive structures and algorithms, *NSC Natural Science Newsletter*, **14 (3)** (2002), 74–80; available at www.nsc.gov.tw/nat/info/J_v14n3.files/78-84.pdf (in Chinese).

[25] H.-K. Hwang, Second phase changes in random m-ary search trees and generalized quicksort: convergence rates, *Annals of Probability*, **31** (2003), 609–629.; available at algo.stat.sinica.edu.tw.

[26] H.-K. Hwang and R. Neininger, Phase change of limit laws in the quicksort recurrences under varying toll functions, *SIAM Journal on Computing*, **31** (2002), 1687-1722.

[27] H. M. Mahmoud and B. Pittel, Analysis of the space of search trees under the random insertion algorithm, *Journal of Algorithms*, **10** (1989), 52-75.

[28] R. Neininger and L. Rüschendorf, A general contraction theorem and asymptotic normality in combinatorial structures, submitted for publication (2001); available at neyman.mathematik.uni-freiburg.de/homepages/neininger/asynorm.pdf.

[29] V. B. Nevzorov, Records, *Theory of Probability and Its Applications*, **32** (1987), 201–228.

[30] B. Schmuland, Shark attacks and the Poisson approximation; available at www.stat.ualberta.ca/people/schmu/preprints/poisson.pdf.

[31] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, Boston, MA (1989).