Bimodality and phase transitions in the profile variance of random binary search trees

MICHAEL DRMOTA
Institut für Diskrete Mathematik und Geometrie
Technische Universität Wien
Wiedner Hauptstrasse 8-10/118
1040 Wien
Austria

HSIEN-KUEI HWANG*
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

January 22, 2004

Abstract

We show that the variance of the profiles (number of nodes at each level) of random binary search trees exhibits asymptotically four phase transitions and a bimodal or "two-humped" behavior, in contrast to the unimodality of the mean value of the profiles. Precise asymptotic approximations are derived. The same types of phenomena also hold for the profiles of random recursive trees.

1 Introduction

Profiles (number of nodes having the same distance to the root) are informative shape characteristics of trees. They are directly related to the total path length (the sum of the distances of all nodes to the root) and depth (the distance of a random node to the root) on the one hand, and can be used to derive effective bounds for the height and width on the other hand. In terms of branching process language, profiles correspond to the number of descendants in each generation; they also have more concrete algorithmic interpretations such as breadth-first search and applications; see Devroye and Robson (1995), Louchard and Szpankowski (1995), Chern and Hwang (2001). We study in this paper the variance of the profiles in random binary search trees (abbreviated as BSTs). Part of our aims is to clarify Figure 1 by more precise mathematical terms.

Binary search trees. A BST $\mathcal{T}$ is a binary tree constructed from a given sequence of keys, say $\mathcal{A} := \{a_1, \ldots, a_n\}$ as follows. If $n = 0$, then $\mathcal{T}$ is empty and, for convenience, we regard $\mathcal{T}$ as consisting of only a node called *external node*. If $n \geq 1$, then the first key a_1 is placed at the root (called an *internal node*). The remaining keys are compared successively to the root key, and are directed to the left (or right) branch if they are smaller (or larger), and keys directed to the same branch are constructed recursively as a BST. By construction, a query operation like "$x \in \mathcal{T}$?" can be easily carried out in BSTs, thus the name.

*Partly supported by a Research Award of the Alexander von Humboldt Foundation.

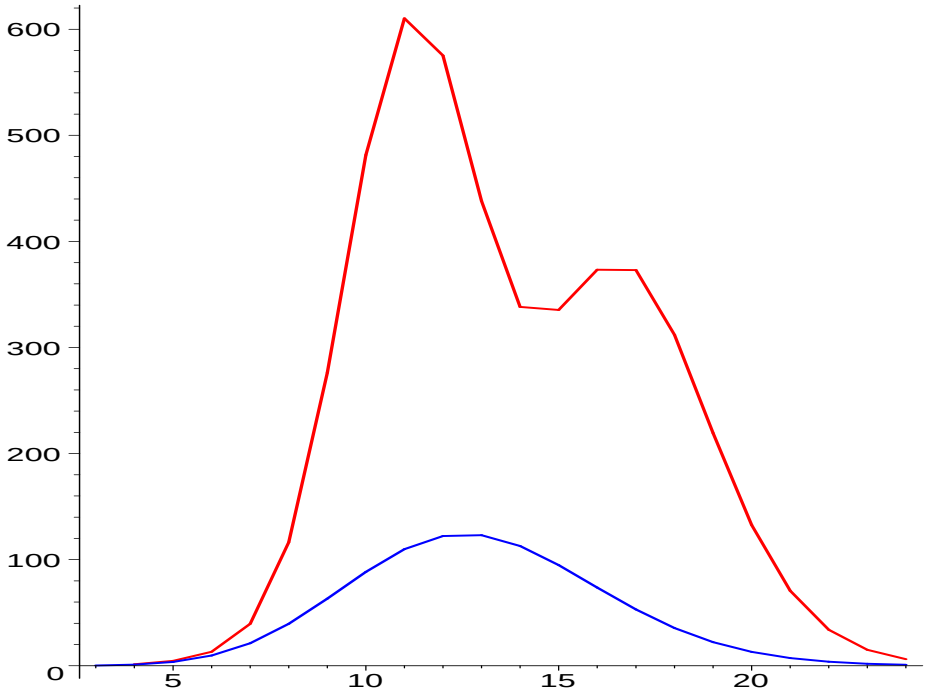

Figure 1: *Profiles of BSTs: exact mean and variance of the number* $X_{1000,k}$ *of external nodes at level* k *in random binary search trees of* 1000 *nodes.*

BSTs are one of the simplest and widely used data structures in Computer Algorithms. They also appeared, under different guises, in other contexts such as branching processes, population genetics, diffusion models and evolutionary trees; see Aldous and Shields (1988), Aldous (1996), Barlow et al. (1997), Majumdar and Krapivsky (2003). The large number of diverse extensions and variants add significantly to their importance in practice, in algorithm design, and in theory.

Random BSTs. Assume that the given input is a finite sequence of independent, and identically distributed random variables with a common continuous distribution. The BST constructed from this random sequence is called a *random BST*. Since only the rank and the order of the keys are relevant, an equivalent model is to assume that the input is a random permutation when all $n!$ permutations of n elements are equally likely.

Many properties of random BSTs have been studied in the literature; see Gonnet and Baeza-Yates (1990), Mahmoud (1992), Knuth (1998), Devroye (2003), Hwang and Neininger (2002) for more information.

Profiles of random BSTs. We are concerned with the random variables $X_{n,k}$, defined to be the number of external nodes at level k (the root being at level 0) in a random binary search tree of n nodes. It is known that

$$\mathbb{E}(X_{n,k}) = \frac{2^k}{n!}\, s(n,k) \qquad (0 \le k \le n), \tag{1}$$

where the $s(n,k)$'s denote the signless Stirling numbers of the first kind:

$$\sum_{0\le k\le n} s(n,k) w^k = w^{\overline{n}} \qquad (n \ge 0),$$

with $w^{\overline{n}}$ denoting the rising factorial $w^{\overline{n}} := \prod_{0\le j<n}(w+j)$; see Lynch (1965), Knuth (1998), Brown and Shubert (1984), Mahmoud and Pittel (1984), Pittel (1984), Louchard (1987), Devroye (1988). Thus the asymptotic behaviors of $\mathbb{E}(X_{n,k})$ can be derived from known results for Stirling numbers $s(n,k)$; see Hwang (1995), Temme (1993).

In particular, the asymptotic behaviors of $\mathbb{E}(X_{n,k})$ for varying k are well approximated by a normal distribution, with mode near $k \approx 2\log n$; see Chauvin et al. (2001, 2003) for more precise properties. Note that the sequence $\{\mathbb{E}(X_{n,k})\}_k$ for fixed n is unimodal, by the simple fact that the generating polynomial $\sum_k \mathbb{E}(X_{n,k})w^k$ has only real zeros; see Comtet (1974), Hammersley (1951).

Known results beyond mean. Almost sure convergence of $X_{n,k}/\mathbb{E}(X_{n,k})$ and other type of results are derived in Chauvin et al. (2001), Jabbour-Hattab (2001); see also the recent paper Chauvin et al. (2003). Pittel (1984) derived the expression

$$\mathbb{E}(X_{n,k}^2) = \frac{2^k}{n!}\sum_{1\le t\le n}\frac{1}{(2\pi i)^2}\oint\!\!\oint \frac{(\sqrt{8}x/y-1)^{\overline{t-1}}(x^2+t)^{\overline{n-t}}}{yx^{2k-1}\sqrt{1-y^2}}\,\mathrm{d}x\,\mathrm{d}y,$$

and then showed that

$$\mathbb{E}(X_{n,k}^2) = O((\log n)^{3/2} n^{2(\alpha-\alpha\log(\alpha/2)-1)}) \qquad (2-\sqrt{2}\le\alpha\le 2+\sqrt{2}),$$

for any $\varepsilon>0$, where, *here and throughout this paper*, $\alpha := k/\log n$.

Global description of the phase transitions. The aim of this paper is to derive more precise asymptotic approximations to the variance $\mathbb{V}(X_{n,k})$ for all ranges of interest. We show that the asymptotic behavior of $\mathbb{V}(X_{n,k})$ exhibits phase transitions at the four points $\alpha = 3\pm 2\sqrt{2}$ and $\alpha = 2\pm\sqrt{2}$ (not viewable from Figure 1 though). The rough picture of $\mathbb{V}(X_{n,k})$ is as follows; see Theorem 2.

– When α is small or large, more precisely, $0\le\alpha\le 3-2\sqrt{2}-\varepsilon$ or $\alpha\ge 3+2\sqrt{2}+\varepsilon$, then the variance is of the same order as the mean

$$\mathbb{V}(X_{n,k}) \sim \mathbb{E}(X_{n,k}^2) \asymp \mathbb{E}(X_{n,k});$$

– When α lies in the middle range, namely, $2-\sqrt{2}+\varepsilon\le\alpha\le 2+\sqrt{2}-\varepsilon$, then the variance is of the order of $(\mathbb{E}(X_{n,k}))^2$

$$\mathbb{V}(X_{n,k}) \sim \varphi(\alpha)(\mathbb{E}(X_{n,k}))^2, \tag{2}$$

where

$$\varphi(\alpha) := \frac{\Gamma(\alpha)^2\alpha^2(2\alpha-1)}{\Gamma(2\alpha)(4\alpha-\alpha^2-2)} - 1, \tag{3}$$

Γ being the Gamma function;

– When α lies in the two intermediate ranges $3-2\sqrt{2}+\varepsilon\le\alpha\le 2-\sqrt{2}-\varepsilon$ and $2+\sqrt{2}+\varepsilon\le\alpha\le 3+2\sqrt{2}-\varepsilon$, then the variance is larger in order than the mean and the mean square

$$\mathbb{E}(X_{n,k}), (\mathbb{E}(X_{n,k}))^2 \ll \mathbb{V}(X_{n,k}) \sim \mathbb{E}(X_{n,k}^2).$$

Note that $\mathbb{E}(X_{n,k}) = o(1)$ for $\alpha<\alpha_-$ and $\alpha>\alpha_+$, where $\alpha_- \approx 0.37336\ldots$ and $\alpha_+\approx 4.31107\ldots$ are the two zeros of the equation $e^{(z-1)/z} = z/2$ (sometimes called the binary search tree constants; see §5.13, Finch, 2003). Also $\mathbb{E}(X_{n,k}) \ll (\mathbb{E}(X_{n,k}))^2$ for $\alpha_-<\alpha<\alpha_+$.

To bridge the asymptotic estimates in neighboring ranges, we need more uniform estimates. We show that the transition is well dictated by a *parabolic cylinder function* when α crosses $3\pm 2\sqrt{2}$, and by a *normal distribution function* when α crosses the other two transitional points.

The valley. The approximation (2) in the middle range is insufficient for describing the behaviors of the variance when $\alpha \approx 2$ since $\varphi(2) = \varphi'(2) = 0$. More precise approximations are thus needed and we derive an asymptotic expansion for $\mathbb{V}(X_{n,k})$ in the middle range. In particular, the visible valley in Figure 1 is roughly due to the estimates

$$\mathbb{V}(X_{n,\lfloor 2\log n+O(1)\rfloor}) \asymp \frac{n^2}{(\log n)^3},$$
$$\mathbb{V}\left(X_{n,\left\lfloor 2\log n\pm\sqrt{2\log n}\right\rfloor}\right) \asymp \frac{n^2}{(\log n)^2}.$$

Indeed, we show that

$$\max_{k\geq 0}\mathbb{V}(X_{n,k}) \sim \frac{21-2\pi^2}{24\pi e}\cdot\frac{n^2}{(\log n)^2}.$$

See Section 5 for a more precise description of the valley, including an explanation of why the left "hump" is higher than the right one.

Numerically, the valley for $\mathbb{V}(X_{n,k})$ appears only when $n \geq 357$.

A "false valley". While the valley near $2\log n$ may be quite expected (see Chauvin et al., 2001, 2003), the function $\varphi(\alpha)$ also satisfies $\varphi(1) = \varphi'(1) = 0$, suggesting that there may be a second valley near $\alpha \sim 1$. We show that this is indeed a "false valley" since the decrease of the variance in the logarithmic term is well "smoothed out" by other larger factors; see Corollary 5.

Why the valley? Structurally, the valley for the variance near $k = 2\log n + O(\sqrt{\log n})$ indicates that there is a better concentration of external nodes near these levels, and indeed almost all external nodes lie at these levels, each level having about $n/\sqrt{\log n}$ nodes; see also Chauvin et al. (2001). Similarly, the "false valley" near $k = \log n + O(\sqrt{\log n})$ may be ascribable to the structural change of number of internal nodes near there.

Methodology. Our approach is mostly analytic and relies on integral representations for the second moments. The basic idea is to consider the bivariate generating function, say $F_2(z, w)$ of $E(X_{n,k}(X_{n,k}-1))$, which satisfies a differential equation of first order. Solving the differential equation yields an integral representation for F_2, from which we apply Cauchy's integral expression and complex-analytic tools, including singularity analysis, saddlepoint method, and some uniform asymptotic methods (for handling the coalescence of a saddlepoint and an algebraic singularity). The approach is of some generality and may be applied to other $\log$-class of trees (see Bergeron et al., 1992, Devroye, 1999).

Universality? The above interesting phenomena naturally suggest the question: are the phase transitions and bimodality unique for BSTs? or is there some sort of universality for such phenomena? We will briefly examine recursive trees in Section 7, and show that the profile variance also exhibits a bimodality near $\log n$ and two phase transitions. Similar behaviors are expected for other ($\log$-) class of trees like m-ary search trees, fringe-balanced BSTs (see Devroye, 1999), but the precise description and general prediction are expected to be more involved.

Limiting distribution? It is known that (see Chauvin et al. 2003)

$$\frac{X_{n,k}}{\mathbb{E}(X_{n,k})} \to X_{\alpha/2} \qquad (2-\sqrt{2} < \alpha < 2+\sqrt{2}),$$

in probability, where $X_z \stackrel{d}{=} zU^{2z-1}X_z + z(1-U)^{2z-1}X'_z$, U being uniformly distributed in the unit interval and $X'_z \stackrel{d}{=} X_z$; see also Jabbour-Hattab (2001). But $X_{1/2} = X_1 = 1$, so that the limit distribution for $(X_{n,k} - \mathbb{E}(X_{n,k}))/\sqrt{\mathbb{V}(X_{n,k})}$ in the two special cases $\alpha \sim 2$ and $\alpha \sim 1$ remains open.

Similar bimodal behaviors are observed (by Monte Carlo simulations) for higher absolute, central moments of $X_{n,k}$ for BSTs, but it does not seem easy to even conjecture the possible form of the limiting distribution of, say $X_{n,\lfloor 2\log n\rfloor}$ (suitably normalized); the valley near there and the periodicity of $\{2\log n\}$ seem to complicate the problem proper.

Profiles of another class of trees (which we may roughly term as "$\sqrt{n}$-class", in contrast to our "$\log n$-class" of trees) have received much recent interests and well clarified (see Aldous, 1993, Drmota and Gittenberger, 1997, Pitman 1999, Kersting, 1998), but many properties of the profiles for the $\log n$-class of trees (of which BST is a prototype) remain unknown and very challenging.

Outline of the paper. This paper is organized as follows. We first derive the basic recurrence for the profiles in the next section, and then the solution to the generating function of m-th moments. In particular, an exact solution for the second factorial moment is given. We then state our main results on phase transitions and bimodality in Section 3. Proofs are given in later sections, and recursive trees are briefly examined in Section 7.

2 Generating functions and integral representations

We give here a self-contained approach to computing the moments of $X_{n,k}$. Define the bivariate generating function

$$P_k(z,y) := \sum_{n\ge 0} \mathbb{E}(y^{X_{n,k}})z^n \qquad (k \ge 0).$$

Then, by the recursive construction,

$$X_{n,k} \stackrel{d}{=} X_{I_n,k-1} + X_{n-I_n-1,k-1},$$

where $P(I_n = j) = n^{-1}$ for $0 \le j \le n-1$. Thus P_k can be computed recursively by

$$\begin{cases} P_0(z,y) &= y + \dfrac{z}{1-z}, \\ P_{k+1}(z,y) &= 1 + \displaystyle\int_0^z P_k^2(t,y)\,\mathrm{d}t \qquad (k \ge 0). \end{cases} \tag{4}$$

Explicit solutions (beyond the iterative integral forms) for this system of equations for all k seem intractable; we consider instead the moments of $X_{n,k}$ by expanding P_k as follows.

$$P_k(z,y) := \sum_{m\ge 0} \frac{M_{m,k}(z)}{m!}\,(y-1)^m,$$

so that $M_{m,k}(z) = \sum_n \mathbb{E}(X_{n,k}(X_{n,k}-1)\cdots(X_{n,k}-m+1))z^n$ and they satisfy, by (4),

$$M'_{m,k+1}(z) = \frac{2}{1-z}\,M_{m,k}(z) + \sum_{1\le j<m} \binom{m}{j} M_{j,k}(z)M_{m-j,k}(z), \tag{5}$$

for $k \ge 0$ and $m \ge 1$, with $M_{0,k}(z) = 1/(1-z)$ and $M_{m,k}(0) = 0$ $(k \ge 1)$.

More explicit representations for the $M_{m,k}$'s can be derived by considering the generating function

$$F_m(z,w) := \sum_{k\geq 0} M_{m,k}(z) w^k,$$

which satisfies, by (5), $F_m(0,w) = 0$ and

$$\frac{\partial}{\partial z} F_m(z,w) = \frac{2w}{1-z} F_m(z,w) + \sum_{1\leq j<m} \binom{m}{j} \sum_{k\geq 0} M_{j,k}(z) M_{m-j,k}(z) w^{k+1}.$$

Solving this first-order differential equation yields

$$F_1(z,w) = (1-z)^{-2w}, \tag{6}$$

and for $m \geq 2$

$$F_m(z,w) = \sum_{1\leq j<m} \binom{m}{j} (1-z)^{-2w} \int_0^z (1-t)^{2w} \sum_{k\geq 0} M_{j,k}(t) M_{m-j,k}(t) w^{k+1}\, \mathrm{d}t. \tag{7}$$

From (6), it follows that

$$M_{1,k}(z) = \frac{2^k}{k!} \log^k \frac{1}{1-z} \qquad (k \geq 0),$$

which implies (1), and then, by (7),

$$F_2(z,w) = 2w(1-z)^{-2w} \int_0^z (1-t)^{2w} I_0\left(4\sqrt{w} \log \frac{1}{1-t}\right) \mathrm{d}t, \tag{8}$$

where

$$I_0(z) = \sum_{k\geq 0} \frac{z^{2k}}{k!k!4^k}$$

is the modified Bessel function of order zero (see §9.6, Abramowitz and Stegun,1965).

Before going further, we derive an explicit formula for $\mathbb{E}(X_{n,k}(X_{n,k}-1))$.

Lemma 1. *The second factorial moments of $X_{n,k}$ can be computed by*

$$\mathbb{E}(X_{n,k}(X_{n,k}-1)) = \frac{2^k}{n!} \sum_{0\leq j<k} \binom{2j}{j} 2^j \sum_{k+j-1\leq m<n} s(n-1,m) \binom{m-2j-1}{k-j-1}. \tag{9}$$

Proof. First observe that

$$\begin{aligned}
\int_0^1 (1-t)^{2w} I_0\left(4\sqrt{w} \log \frac{1}{1-t}\right) \mathrm{d}t &= \sum_{j\geq 0} \frac{4^j}{j!j!} w^j \int_0^1 y^{2w} \log^{2j}(1/y)\, \mathrm{d}y \\
&= \sum_{j\geq 0} \binom{2j}{j} \frac{(4w)^j}{(2w+1)^{2j+1}} \\
&= (4w^2 - 12w + 1)^{-1/2},
\end{aligned} \tag{10}$$

provided that $|\frac{16w}{(2w+1)^2}| < 1$. Assume for the moment that w lies in that region. Then, similarly as above,

$$\begin{aligned}
&(1-z)^{-2w}\int_z^1 (1-t)^{2w} I_0\left(4\sqrt{w}\log\frac{1}{1-t}\right)\,\mathrm{d}t\\
&\quad = (1-z)\sum_{k\geq 0}\frac{(2k)!}{k!k!}(4w)^k \sum_{0\leq j\leq 2k}\frac{(-\log(1-z))^j}{j!}\cdot\frac{1}{(2w+1)^{2k+1-j}}\\
&\quad = \sum_{k\geq 0}\binom{2k}{k}(4w)^k\frac{1}{2\pi i}\oint_{|t|=c<|2w+1|}\frac{t^{-2k-1}(1-z)^{1-t}}{2w+1-t}\,\mathrm{d}t.
\end{aligned}$$

But the residue of the integrand at $t = 2w+1$ equals $(2w+1)^{-2j-1}(1-z)^{-2w-1}$. It follows that

$$\begin{aligned}
F_2(z,w) &= \frac{2w}{2\pi i}\oint_{|t|=c}\frac{(1-z)^{1-t}}{(t-2w-1)\sqrt{t^2-16w}}\,\mathrm{d}t \qquad (c>|2w+1|)\\
&= \frac{2w}{2\pi i}\oint_{|y|=c}\frac{(1-z)^{1-1/y}}{(1-(2w+1)y)\sqrt{1-16wy^2}}\,\mathrm{d}y \qquad (c<\varepsilon),
\end{aligned}$$

for properly chosen integration contours. The restriction for w can now be dropped.

By Cauchy's integral representation

$$\mathbb{E}(X_{n,k}(X_{n,k}-1)) = \frac{2^k}{(2\pi i)^2}\oiint w^{-k}\frac{\binom{n-2+1/y}{n}}{(1-(w+1)y)\sqrt{1-8wy^2}}\,\mathrm{d}y\,\mathrm{d}w.$$

Thus we have

$$\mathbb{E}(X_{n,k}(X_{n,k}-1)) = 2^k\sum_{0\leq\ell<k}\binom{2\ell}{\ell}\frac{2^\ell}{2\pi i}\oint_{|z|=c>1}\frac{\binom{n+z-2}{n}}{z^{2\ell+1}(z-1)^{k-\ell}}\,\mathrm{d}z, \tag{11}$$

from which (9) follows. ∎

3 Phase transitions and bimodality

Notation. For convenience, we use the symbol $[\![a,b]\!]$ to denote the interval $[a+K/\sqrt{\log n}, b-K/\sqrt{\log n}]$ for a sufficiently large K; The one-sided conventions $[a,b]\!]$ and $[\![a,b]$ stand for $[a, b-K/\sqrt{\log n}]$ and $[a+K/\sqrt{\log n}, b]$, respectively. The generic symbols K and ε always represent large and small, respectively, constants whose values may vary from one occurrence to another. Throughout this paper, $\alpha = \alpha_{n,k} = k/\log n$.

3.1 Asymptotics of $\mathbb{E}(X_{n,k})$

For completeness, we first state two known expansions for $\mathbb{E}(X_{n,k})$ that will be needed.

Theorem 1. *Uniformly for* $1 \leq k \leq K\log n$,

$$\mathbb{E}(X_{n,k}) = \frac{(2\log n)^k}{nk!\Gamma(\alpha)}\left(1+O\left((\log n)^{-1}\right)\right); \tag{12}$$

and uniformly for $k \to \infty$, $k \le K \log n$,

$$\mathbb{E}(X_{n,k}) \sim \frac{n^{\alpha-\alpha\log(\alpha/2)-1}}{\sqrt{2\pi k}\,\Gamma(\alpha)} \sum_{j\ge 0} c_j k^{-j}, \tag{13}$$

for some coefficients c_j.

Proof. (Sketch) The proof of both approximations starts from (1) and then uses the uniform approximation

$$\sum_k \mathbb{E}(X_{n,k}) w^k = 2^k \binom{n+w-1}{n} = 2^k \frac{n^{w-1}}{\Gamma(w)} \left(1 + O(n^{-1})\right),$$

uniformly for $|w| \le K$ (by the singularity analysis of Flajolet and Odlyzko, 1990). Then

$$\mathbb{E}(X_{n,k}) = \frac{2^k}{2\pi i} \oint_{|w|=\alpha} w^{-k-1} \frac{n^{w-1}}{\Gamma(w)} \left(1 + O(n^{-1})\right)\, \mathrm{d}w,$$

and (12) follows by expanding $1/\Gamma(w)$ at $w = \alpha = k/\log n$, and by estimating the error terms properly; see Hwang (1995) for details. The proof for (13) uses the usual saddlepoint method and is similar. ∎

From (12), we see that the asymptotics of $\mathbb{E}(X_{n,k})/n$ is roughly dictated by a Poisson distribution with mean $2\log n$. In particular, it is unimodal (at least for $0 \le k \le K \log n$), and there is no change of asymptotic behavior in the main range of interests ($k \le K \log n$).

3.2 Asymptotics of $\mathbb{E}(X_{n,k}^2)$

For the second moment and the variance, the situation becomes completely different. We give our first approximations to $\mathbb{E}(X_{n,k}^2)$ by splitting the range $[0, K]$ into five non-overlapping intervals.

Global silhouette. For simplicity of presentation, we drop the error terms in the following estimates, and we define two constants

$$C_\pm := \frac{\sqrt{2} \pm 1}{2\sqrt{\pi\sqrt{2}}\,\Gamma(3 \pm 2\sqrt{2})}.$$

Theorem 2. (***I***) *If* $\alpha \in [0, 3 - 2\sqrt{2}]\!]$, *then*

$$\mathbb{E}(X_{n,k}^2) \sim \mathbb{E}(X_{n,k}) \left(1 + \frac{\alpha}{\sqrt{\alpha^2 - 6\alpha + 1}}\right); \tag{14}$$

(***II***) *if* $\alpha \in [\![3 - 2\sqrt{2}, 2 - \sqrt{2}]\!]$, *then*

$$\mathbb{E}(X_{n,k}^2) \sim C_- \frac{2^k n^{2-2\sqrt{2}} \left(3 - 2\sqrt{2}\right)^{-k}}{\sqrt{k - (3 - 2\sqrt{2})\log n}}; \tag{15}$$

(***III***) *if* $\alpha \in [\![2 - \sqrt{2}, 2 + \sqrt{2}]\!]$, *then*

$$\mathbb{E}(X_{n,k}^2) \sim (\mathbb{E}X_{n,k})^2 \frac{\Gamma(\alpha)^2 \alpha^2 (2\alpha - 1)}{\Gamma(2\alpha)(4\alpha - \alpha^2 - 2)}; \tag{16}$$

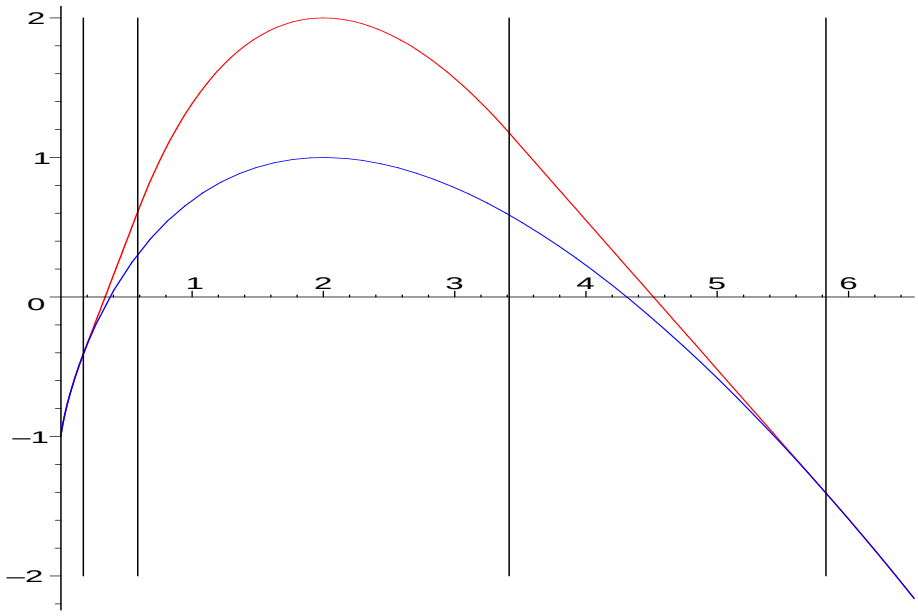

Figure 2: *A plot of the limiting curve for* $\log \mathbb{E}(X_{n,k}^2)/\log n$ *(upper curve) and for* $\log \mathbb{E}(X_{n,k})/\log n$ *(lower curve) for* α *in each interval (horizontal coordinate). The intervals are also explicitly depicted by vertical lines.*

(***IV***) *if* $\alpha \in [\![2+\sqrt{2}, 3+2\sqrt{2}]\!]$*, then*

$$\mathbb{E}(X_{n,k}^2) \sim C_+ \frac{2^k n^{2+2\sqrt{2}} \left(3+2\sqrt{2}\right)^{-k}}{\sqrt{(3+2\sqrt{2})\log n - k}}; \tag{17}$$

(***V***) *finally, if* $\alpha \in [\![3+2\sqrt{2}, K]$*, then*

$$\mathbb{E}(X_{n,k}^2) \sim \mathbb{E}(X_{n,k}) \left(1 + \frac{\alpha}{\sqrt{\alpha^2 - 6\alpha + 1}}\right). \tag{18}$$

A more transparent approximation is as follows; see Figure 2 for a plot.

Corollary 1 (Phase transitions). *The growth order of* $\mathbb{E}(X_{n,k}^2)$ *satisfies*

$$\frac{\log \mathbb{E}(X_{n,k}^2)}{\log n} \to \begin{cases} \alpha - \alpha\log(\alpha/2) - 1, & \text{if } \alpha \in [0, 3-2\sqrt{2}] \\ 2 - 2\sqrt{2} - 2\alpha\log(1-2^{-1/2}), & \text{if } \alpha \in [3-2\sqrt{2}, 2-\sqrt{2}] \\ 2(\alpha - \alpha\log(\alpha/2) - 1), & \text{if } \alpha \in [2-\sqrt{2}, 2+\sqrt{2}] \\ 2 + 2\sqrt{2} - 2\alpha\log(1+2^{-1/2}), & \text{if } \alpha \in [2+\sqrt{2}, 3+2\sqrt{2}] \\ \alpha - \alpha\log(\alpha/2) - 1, & \text{if } \alpha \in [3+2\sqrt{2}, K]. \end{cases}$$

By continuity, the (almost) open boundaries $[\![$ and $]\!]$ in all cases are replaced by the closed ones $[$ and $]$, respectively, as will become clear.

Transitional behaviors. These quick (and rough) estimates leave open the asymptotics of the second moment in the transitional ranges $k = (3\pm 2\sqrt{2})\log n + O(\sqrt{\log n})$ and $k = (2\pm\sqrt{2})\log n + O(\sqrt{\log n})$, which will be handled by more uniform asymptotic tools.

Let $D_{-\nu}(x)$ denote the parabolic cylinder function (see Ch. 19, Abramowitz and Stegun, 1965), which can be defined by

$$D_{-\nu}(x) = \frac{e^{-x^2/4}}{\Gamma(\nu)} \int_0^\infty u^{\nu-1} e^{-xu-u^2/2}\, \mathrm{d}u \qquad (\nu > 0), \tag{19}$$

and let $\Phi(x)$ denote the standard normal distribution function. Note that $\Phi(x)$ is itself a special case of the parabolic cylinder functions

$$D_{-1}(x) = \sqrt{2\pi}\, e^{x^2/4}\Phi(-x).$$

Theorem 3. *All asymptotic estimates below hold uniformly for $t = o((\log n)^{1/6})$. (i) If $\alpha = 3 - 2\sqrt{2} + (\sqrt{2}-1)t/\sqrt{\log n}$, then*

$$\mathbb{E}(X_{n,k}^2) \sim 2^{-1/2} C_- e^{t^2/4} D_{-1/2}(-t) k^{-1/4}\, n^{2-2\sqrt{2}-2\alpha\log(1-1/\sqrt{2})}; \tag{20}$$

(ii) if $\alpha = 2 - \sqrt{2} + \sqrt{1-2^{-1/2}}t/\sqrt{\log n}$, then

$$\mathbb{E}(X_{n,k}^2) \sim 2^{1/4} C_- \Phi(-t) k^{-1/2} n^{2-2\sqrt{2}-2\alpha\log(1-1/\sqrt{2})}; \tag{21}$$

(iii) if $\alpha = 2 + \sqrt{2} + \sqrt{1+2^{-1/2}}t/\sqrt{\log n}$, then

$$\mathbb{E}(X_{n,k}^2) \sim 2^{1/4} C_+ \Phi(t) k^{-1/2} n^{2+2\sqrt{2}-2\alpha\log(1+1/\sqrt{2})};$$

(iv) finally, if $\alpha = 3 + 2\sqrt{2} + (\sqrt{2}+1)t/\sqrt{\log n}$, then

$$E(X_{n,k}^2) \sim 2^{-1/2} C_+ e^{t^2/4} D_{-1/2}(t) k^{-1/4}\, n^{2+2\sqrt{2}-2\alpha\log(1+1/\sqrt{2})}.$$

In all cases, the dropped error terms are of the form

$$1 + O\left(\frac{1+|t|^3}{\sqrt{\log n}}\right).$$

These estimates complete the gap left open in Theorem 2; furthermore, one can easily check that in the overlapping ranges ($K \le |t| = o((\log n)^{1/6})$) the approximations in both Theorems coincide by the following asymptotic estimates (see §19.7, Abramowitz and Stegun, 1965)

$$\begin{cases} D_{-\nu}(x) \sim x^{-\nu} e^{-x^2/4} & (x \to \infty), \\ D_{-\nu}(-x) \sim \dfrac{\sqrt{2\pi}}{\Gamma(\nu)} x^{\nu-1} e^{x^2/4} & (x \to \infty). \end{cases} \tag{22}$$

Bimodality. Everything up to now is only unimodal. Bimodality of the variance appears in the middle range $\alpha \in [\![2-\sqrt{2}, 2+\sqrt{2}]\!]$.

First, from Theorem 2, we readily obtain the following estimate.

Corollary 2. *The variance of $X_{n,k}$ satisfies*

$$\mathbb{V}(X_{n,k}) \sim \varphi(\alpha)(\mathbb{E}X_{n,k})^2,$$

for $\alpha \in [\![2-\sqrt{2}, 2+\sqrt{2}]\!]$, where φ is defined in (3), and $\mathbb{V}(X_{n,k}) \sim \mathbb{E}(X_{n,k}^2)$ for all other ranges.

Observe that

$$\varphi(1) = \varphi(2) = \varphi'(1) = \varphi'(2) = 0; \tag{23}$$

thus the estimate (3) is insufficient for an asymptotic equivalent for the variance in the central range $k = (2 + o(1)) \log n$ and in the somewhat unexpected range $k = (1 + o(1)) \log n$. We need stronger approximations.

Theorem 4. *If $\alpha \in [\![2 - \sqrt{2}, 2 + \sqrt{2}]\!]$, then*

$$\mathbb{V}(X_{n,k}) \sim n^{2(\alpha - \alpha \log(\alpha/2) - 1)} \sum_{j \geq 1} \frac{\upsilon_j(\alpha)}{(\log n)^j}, \tag{24}$$

for some coefficients $\upsilon_j(\alpha)$; see (38) and (40) below.

In particular, $\upsilon_1(\alpha) = \varphi(\alpha)/(2\pi\alpha\Gamma(\alpha)^2)$ also satisfies property (23), and $\upsilon_2(\alpha)$ satisfies $\upsilon_2(1) = \upsilon_2(2) = 0$.

Corollary 3. *If $\alpha = 2 + t/\log n$, where $t = o(\log n)$, then*

$$\mathbb{V}(X_{n,k}) = \frac{p_1(t)}{720\pi} \cdot \frac{n^{2(\alpha - \alpha \log(\alpha/2) - 1)}}{(\log n)^3} \left(1 + O\left(\frac{1 + |t|}{\log n}\right)\right),$$

uniformly in t, where $p_1(t)$ is a quadratic polynomial defined by

$$\begin{aligned} p_1(t) := {} & 15(21 - 2\pi^2)t^2 - 30\left(4\pi^2(1 - \gamma) + 24\zeta(3) + 42\gamma - 69\right) t \\ & - 2\pi^4 - 30\left(4\gamma^2 - 8\gamma + 11\right)\pi^2 + 180\left(7\gamma^2 - 23\gamma + 29\right) - 1440\zeta(3)(1 - \gamma), \end{aligned} \tag{25}$$

where γ denotes Euler's constant.

The reason of writing the corollary in its form is that the variation of the order of $\mathbb{V}(X_{n,k})$ when $k = (2 + o(1)) \log n$ becomes more transparent. Thus, if $\alpha = 2 + t/\log n$, where $t = o((\log n)^{3/2})$, then

$$\mathbb{V}(X_{n,k}) \sim \frac{p_1(t)}{720\pi} \cdot \frac{n^2}{(\log n)^3} \exp\left(-\frac{t^2}{2 \log n}\right),$$

uniformly in t. From this we can derive approximations to the scale of the two "humps" and the valley seen in Figure 1.

Corollary 4. *The largest value of $\mathbb{V}(X_{n,k})$ is asymptotically achieved at $k = \lfloor 2 \log n \pm \sqrt{2 \log n} \rfloor$, and*

$$\max_{k \geq 0} \mathbb{V}(X_{n,k}) \sim \frac{21 - 2\pi^2}{48\pi e} \cdot \frac{n^2}{(\log n)^2};$$

on the other hand,

$$\min_{|k - 2 \log n| = O(\sqrt{\log n})} \mathbb{V}(X_{n,k}) \geq (C + o(1)) \frac{n^2}{(\log n)^3}, \tag{26}$$

where

$$C = \frac{4\pi^6 + 378\pi^4 - 9090\,\pi^2 - 38205 - 8640\zeta(3)^2 + 19440\zeta(3) - 38205}{720\pi(21 - 2\pi^2)}.$$

The smallest value of $\mathbb{V}(X_{n,k})$, for $k = 2 \log n + O(\sqrt{\log n})$, is asymptotically achieved only for the subsequence of n for which $\{2 \log n\} \to 1 - t_0$, where

$$t_0 = -2(1 - \gamma) + \frac{3(8\zeta(3) - 9)}{21 - 2\pi^2} \approx 0.62126\ldots$$

satisfies $p_1'(t_0) = 0$.

Thus *the variance can vary from* $n^2/(\log n)^2$ *to* $n^2/(\log n)^3$ *when* $k = 2\log n + O(\sqrt{\log n})$*, and these are precisely the orders of the peak and the valley, respectively, as shown in Figure 1.*

Our analysis here says that the two peaks are asymptotically of the same order, although Figure 1 may lead one to guess that the left peak is higher. We will see that this is indeed true by further examining the sign of the next order term; see Section 5 for more details.

A "false valley".

Corollary 5. *If* $\alpha = 1 + t/\log n$*, where* $t = o((\log n)^{2/3})$*, then*

$$\mathbb{V}(X_{n,k}) \sim \frac{4^t \varpi(t)}{720\pi} \cdot \frac{n^{2\log 2}}{(\log n)^3}\, e^{-t^2/\log n},$$

uniformly in t*, where* $\varpi(t)$ *is defined by*

$$\begin{aligned}\varpi(t) := {} & 60(12-\pi^2)t^2 + 120\left(\pi^2\gamma - 12\gamma - 6\zeta(3) + 12\right)t \\ & - \pi^4 - 60\left(\gamma^2+2\right)\pi^2 + 720\left(\gamma^2 - 2\gamma + \zeta(3)\gamma + 3\right).\end{aligned}$$

One sees that although the order of the variance can reach that of $\mathbb{E}(X_{n,k}^2)/(\log n)^2$ (when $k = \log n + O(1)$) as in the case $k = 2\log n + O(1)$, there is no new "valley" generated when $k = \log n + O(\sqrt{\log n})$ since the logarithmically smaller terms are "smoothed out" by an exponentially large factor 4^t.

4 Phase transitions: Proof of Theorem 2

For more methodological interest and for shedding more light on how the different ranges arise, we give in this section two proofs of Theorem 2. The first relies essentially on the exact expression (9), which has some elementary flavor, although the main estimate needed relies on saddlepoint method. The second uses (8) and is analytic in nature; it can be easily extended to derive asymptotic expansions.

4.1 A direct approach

We give in this section the sketch of an approach to proving Theorem 2 using (9). The basic idea is first to find a good uniform estimate for the sum

$$S_{n,k,j} := \sum_{k+j-1\le m<n} \frac{s(n-1,m)}{n!}\binom{m-2j-1}{k-j-2} \qquad (0\le j<k);$$

then we evaluate the sum

$$\mathbb{E}(X_{n,k}(X_{n,k}-1)) = 2^k \sum_{0\le j<k}\binom{2j}{j}2^j S_{n,k,j}, \tag{27}$$

by different means according to the range of α.

In this subsection, we always write $\alpha = k/\log n$ and $\lambda = j/\log n$.

Lemma 2. *Define*

$$f(z) = f(\alpha,\lambda;z) := z - 2\lambda\log z - (\alpha-\lambda)\log(z-1),$$

and

$$z_0 = z_0(\alpha, \lambda) := \frac{\alpha+\lambda+1}{2} + \sqrt{\left(\frac{\alpha+\lambda+1}{2}\right)^2 - 2\lambda}.$$

If $1+\varepsilon \le z_0 \le K$, *then*

$$S_{n,k,j} \sim \frac{n^{f(z_0)-2}}{z_0\Gamma(z_0-1)\sqrt{2\pi f''(z_0)\log n}},$$

uniformly in k *and* j.

Proof. We start from the integral representation (see (11))

$$\begin{aligned} S_{n,k,j} &= \frac{1}{2\pi i}\oint_{|z|=z_0} \frac{\binom{n+z-2}{n}}{z^{2j+1}(z-1)^{k-j}}\,\mathrm{d}z \\ &= \frac{1}{2\pi i}\oint_{|z|=z_0} \frac{n^{z-2}}{\Gamma(z-1)z^{2j+1}(z-1)^{k-j}}\left(1+O(n^{-1})\right)\,\mathrm{d}z, \end{aligned}$$

by singularity analysis. Observe that z_0 is the saddlepoint at which $f'(z_0) = 0$, and that the second derivative of f

$$f''(z) = \frac{2\lambda}{z^2} + \frac{\alpha-\lambda}{(z-1)^2}$$

remains strictly positive in the range of interest. The required result follows from applying the saddlepoint method to the integral

$$\frac{1}{2\pi i}\oint_{|z|=z_0} \frac{e^{\log n f(z)}}{z\Gamma(z-1)}\,\mathrm{d}z. \quad \blacksquare$$

Middle range. Consider first **Case** (**III**): $\alpha \in [\![2-\sqrt{2}, 2+\sqrt{2}]\!]$. In this case terms with large j's are dominant. Thus, we set $r := k-j \ge 1$. By applying Lemma 2 with $\lambda = \alpha - r/\log n$,

$$z_0 = 2\alpha - \frac{2r(\alpha-1)}{(2\alpha-1)\log n} + O\left((\log n)^{-2}\right),$$

and

$$f(\alpha,\lambda;z_0) = 2\alpha + r\frac{2\log(2\alpha)-\log(2\alpha-1)}{\log n} + O\left((\log n)^{-2}\right),$$

we get

$$S_{n,k,j} \sim \left(\frac{4\alpha^2}{2\alpha-1}\right)^r \frac{n^{f(\alpha,\alpha;z_0)-2}}{z_0\Gamma(z_0-1)\sqrt{2\pi\log n/(2\alpha)}};$$

also

$$2^j\binom{2j}{j} \sim \frac{8^{k-r}}{\sqrt{k\pi}}.$$

These estimates lead to

$$\begin{aligned} \mathbb{E}(X_{n,k}(X_{n,k}-1)) &\sim \frac{16^k}{\sqrt{k\pi}}\sum_{r\ge1}\left(\frac{4\alpha^2}{8(2\alpha-1)}\right)^r \frac{n^{f(\alpha,\alpha;2\alpha)-2}}{z_0\Gamma(z_0-1)\sqrt{2\pi\log n/(2\alpha)}} \\ &= \frac{\alpha^2}{\Gamma(2\alpha-1)(4\alpha-\alpha^2-2)}\cdot\frac{4^k e^{2k}(\log n)^{2k}}{2\pi k n^2 k^{2k}} \\ &\sim \frac{\alpha^2(2\alpha-1)}{\Gamma(2\alpha)(4\alpha-\alpha^2-2)}\cdot\left(\frac{(2\log n)^k}{n\,k!}\right)^2. \end{aligned}$$

Intermediate ranges. For **Case** (**IV**): $\alpha \in [\![2+\sqrt{2}, 3+2\sqrt{2}]\!]$, no terms are asymptotically negligible; we thus sum all terms up and obtain

$$\mathbb{E}(X_{n,k}(X_{n,k}-1)) \sim \frac{2^k}{\sqrt{2}\pi n^2 \log n} \sum_{1\le j<k} \frac{n^{F(\lambda)}}{\sqrt{\lambda} z_0 \Gamma(z_0-1)\sqrt{f''(z_0)}},$$

where $F(\lambda) := \lambda \log 8 + f(\alpha, \lambda; z_0(\alpha,\lambda))$. Since $f'(\alpha,\lambda,z_0(\alpha,\lambda)) = 0$, we get $F'(\lambda) = \log 8 - 2\log z + \log(z-1)$; and, consequently, $F'(\lambda_0) = 0$ for $z_0(\alpha,\lambda_0) = 2(2+\sqrt{2})$, which implies that $\lambda_0 = \sqrt{2}(3+2\sqrt{2}-\alpha)$. It follows that $F(\lambda_0) = 4+2\sqrt{2}-\alpha\log(3+2\sqrt{2})$, $F''(\lambda_0) = -\sqrt{2}/(5+4\sqrt{2}+\alpha)$, and

$$f''(\alpha,\lambda_0;2(2+\sqrt{2})) = \frac{1}{4}(17\sqrt{2}-24)(5+4\sqrt{2}+\alpha);$$

we obtain, by standard application of the saddlepoint method,

$$\begin{aligned}\mathbb{E}(X_{n,k}(X_{n,k}-1)) &\sim \frac{2^k}{\sqrt{2}z_0\Gamma(z_0-1)\pi n^2\log n}\sqrt{\frac{2\pi\log n}{-\lambda_0 F''(\lambda_0) f''(z_0)}}\, n^{F(\lambda_0)} \\ &= \frac{2^k n^{2+2\sqrt{2}}(3+2\sqrt{2})^{-k}}{\sqrt{2\pi\log n}(2-\sqrt{2})\Gamma(3+2\sqrt{2})\sqrt{\sqrt{2}(3+2\sqrt{2}-\alpha)}}.\end{aligned}$$

This proves (17). The proof for **Case** (**II**) is similar.

Extremal ranges. **Case** (**V**): $\alpha \in [\![3+2\sqrt{2}, K]\!]$. In this case, the terms with small j are dominant. For every finite $j \ge 0$, we have ($z_0 = \alpha+1$)

$$\begin{aligned}S_{n,k,j} &\sim \frac{1}{2\pi i}\oint_{|z|=z_0} \frac{n^{z-2}}{\Gamma(z-1)z^{2k+1}}\left(\frac{z-1}{z^2}\right)^j \mathrm{d}z \\ &\sim \left(\frac{\alpha}{(\alpha+1)^2}\right)^j \frac{n^{\alpha-1-\alpha\log\alpha}}{(\alpha+1)\Gamma(\alpha)\sqrt{2\pi\log n/\alpha}} \\ &\sim \left(\frac{\alpha}{(\alpha+1)^2}\right)^j \frac{\alpha(\log n)^k}{(\alpha+1)\Gamma(\alpha)nk!}.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}(X_{n,k}(X_{n,k}-1)) &\sim \frac{(2\log n)^k\alpha}{n\,k!\,(\alpha+1)\Gamma(\alpha)}\sum_{j\ge0}\binom{2j}{j}\left(\frac{2\alpha}{(\alpha+1)^2}\right)^j \\ &= \frac{\alpha(2\log n)^k}{(\alpha+1)\Gamma(\alpha)nk!}\left(1-\frac{8\alpha}{(\alpha+1)^2}\right)^{-1/2} \\ &= \frac{\alpha(2\log n)^k}{\Gamma(\alpha)\sqrt{\alpha^2-6\alpha+1}nk!}.\end{aligned}$$

Case (**I**) is similar.

4.2 An analytic approach

This approach relies on (8) and the convergence or divergence of the integral

$$\int_0^1 (1-t)^{2w} I_0\left(4\sqrt{w}\log\frac{1}{1-t}\right)\,\mathrm{d}t, \tag{28}$$

plays a crucial rôle in determining the different ranges.

We first give the main idea of this approach using mostly heuristic reasoning; the technical justification and detailed estimates of the error terms will be provided later.

A sketch of proof. We need the asymptotics of the modified Bessel function (see §9.6, Abramowitz and Stegun, 1965)

$$I_0(z) = \frac{e^z}{\sqrt{2\pi z}}\left(1+O(|z|^{-1})\right), \tag{29}$$

the O-term being uniform for $|z|\to\infty$ in the region $-\pi/2 < \arg(z) \le \pi/2$.

Small or large α. First if the integral (28) is convergent, then (see (10))

$$\begin{aligned} F_2(z,w) &\sim 2w(1-z)^{-2w}\int_0^1 (1-t)^{2w} I_0\left(4\sqrt{w}\log\frac{1}{1-t}\right)\,\mathrm{d}t \\ &= \frac{2w}{\sqrt{4w^2-12w+1}}(1-z)^{-2w}; \end{aligned} \tag{30}$$

so we expect that (by singularity analysis and then by saddlepoint method)

$$\begin{aligned} \mathbb{E}(X_{n,k}(X_{n,k}-1)) &= [w^k z^n]F_2(z,w) && (31)\\ &\sim [w^k]\frac{2wn^{2w-1}}{\sqrt{4w^2-12w+1}\,\Gamma(2w)} \\ &\sim \frac{\alpha}{\sqrt{\alpha^2-6\alpha+1}\,\Gamma(\alpha)}\cdot\frac{(2\log n)^k}{nk!}, && (32) \end{aligned}$$

where $\alpha>0$ has to satisfy $\alpha^2-6\alpha+1>0$. This gives rise to the first two ranges $\alpha\in[0,3-2\sqrt{2})$ and $\alpha\in(3+2\sqrt{2},K]$, and the estimates (14) and (18).

Middle range. On the other hand, if the integral (28) diverges, then by (29)

$$\begin{aligned} F_2(z,w) &\sim 2w(1-z)^{-2w}\int_0^z (1-t)^{2w} I_0\left(4\sqrt{w}\log\frac{1}{1-t}\right)\,\mathrm{d}t \\ &\sim \frac{2w(1-z)^{-2w}}{\sqrt{8\pi\sqrt{w}}}\int_0^z \left(\log\frac{1}{1-t}\right)^{-1/2}(1-t)^{2w-4\sqrt{w}}\,\mathrm{d}t \\ &\sim \frac{w}{\sqrt{2\pi\sqrt{w}}(4\sqrt{w}-2w-1)}\left(\log\frac{1}{1-z}\right)^{-1/2}(1-z)^{-4\sqrt{w}+1}. \end{aligned} \tag{33}$$

Thus we expect that (again by singularity analysis and then by saddlepoint method)

$$\begin{aligned}\mathbb{E}(X_{n,k}(X_{n,k}-1)) &\sim [w^k]\frac{wn^{4\sqrt{w}-2}(\log n)^{-1/2}}{\sqrt{2\pi\sqrt{w}}(4\sqrt{w}-2w-1)\Gamma(4\sqrt{w}-1)}\\ &\sim \frac{\alpha^2}{(4\alpha-\alpha^2-2)\Gamma(2\alpha-1)}\cdot\frac{(2\log n)^{2k}}{n^2(k!)^2},\end{aligned}$$

which yields the second pairs of transitional points since

$$4\alpha-\alpha^2-2>0 \text{ iff } \alpha\in(2-\sqrt{2},2+\sqrt{2}).$$

Intermediate ranges. Observe that the error term in (30) is of the form (by (29))

$$\begin{aligned}&(1-z)^{-2\Re(w)}\int_z^1(1-t)^{2w}I_0\left(4\sqrt{w}\log\frac{1}{1-t}\right)\,\mathrm{d}t\\ &\quad=O\left(\frac{(1-z)^{-4\sqrt{w}+1}}{|4\sqrt{w}-2w-1|}\left(\log\frac{1}{1-z}\right)^{-1/2}\right),\end{aligned}$$

(see also (33)) whose contribution to $\mathbb{E}(X_{n,k}(X_{n,k}-1))$ is roughly of the order

$$[w^k]\frac{n^{4\sqrt{w}-2}}{4\sqrt{w}-2w-1}(\log n)^{-1/2}=O\left(\frac{(2\log n)^{2k}}{n^2k!^2}\right),$$

essentially the same order as $(\mathbb{E}(X_{n,k}))^2$.

Thus we can use the estimate (30) when k lies in the intervals of **Cases** (**II**) and (**IV**); but instead of applying the saddlepoint method as in **Cases** (**I**) and (**V**), we use again singularity analysis since the singularities at $w=\frac{3}{2}\pm\sqrt{2}$ in (30) is dominating.

Consider **Case** (**II**). Let $\beta:=3/2-\sqrt{2}$. We have, by (30),

$$\begin{aligned}\mathbb{E}(X_{n,k}(X_{n,k}-1)) &\sim [w^k]\frac{2w}{\sqrt{4w^2-12w+1}}\cdot\frac{n^{2w-1}}{\Gamma(2w)}\\ &\sim \frac{2\beta n^{2\beta-1}}{\sqrt{8\sqrt{2}\beta}\,\Gamma(2\beta)}\,[w^k]\frac{n^{2(w-\beta)}}{\sqrt{1-w/\beta}}\\ &\sim \frac{n^{2-2\sqrt{2}}}{\sqrt{2\pi\sqrt{2}}\,\Gamma(3-2\sqrt{2})}\left(\frac{3}{2}-\sqrt{2}\right)^{-k+1/2}(k-2\beta\log n)^{-1/2},\end{aligned}$$

which, in view of (12), implies (15).

Case (**IV**) is similar.

4.3 Technical justification and error estimates

We start from deriving a different integral representation for F_2 suitable for all ranges.

Lemma 3.

$$F_2(z,w)=\frac{2w}{\pi}\int_{-1}^1\frac{(1-z)^{-2w}-(1-z)^{-4\sqrt{w}v+1}}{\sqrt{1-v^2}(2w+1-4\sqrt{w}v)}\,\mathrm{d}v. \tag{34}$$

Note that this representation is well-defined for all w (including at the zeros of the factors in the denominator).

Proof. By the integral representation for $I_0(z)$ (see p. 376 of Abramowitz and Stegun, 1964)

$$I_0(z) = \frac{1}{\pi}\int_0^\pi e^{z\cos t}\,\mathrm{d}t,$$

and by (8), we have

$$\begin{aligned} F_2(z,w) &= \frac{2w}{\pi}(1-z)^{-2w}\int_0^z (1-t)^{2w}\int_0^\pi (1-t)^{-4\sqrt{w}\cos y}\,\mathrm{d}y\,\mathrm{d}t \\ &= \frac{2w}{\pi}(1-z)^{-2w}\int_{-1}^1 \frac{1}{\sqrt{1-v^2}}\int_0^z (1-t)^{2w-4\sqrt{w}v}\,\mathrm{d}t\,\mathrm{d}v, \end{aligned}$$

which yields (34). ∎

Note that when $w \notin [3/2-\sqrt{2}, 3/2+\sqrt{2}]$ we can split the integral (34) and obtain

$$\begin{aligned} F_2(z,w) &= \frac{2w}{\pi}(1-z)^{-2w}\int_{-1}^1 \frac{\mathrm{d}v}{\sqrt{1-v^2}(2w+1-4\sqrt{w}v)} \\ &\quad + \frac{2w}{\pi}\int_{-1}^1 \frac{(1-z)^{-4\sqrt{w}v+1}}{\sqrt{1-v^2}(4\sqrt{w}v-2w-1)}\,\mathrm{d}v \\ &= \frac{2w(1-z)^{-2w}}{\sqrt{4w^2-12w+1}} + \frac{2w}{\pi}\int_{-1}^1 \frac{(1-z)^{-4\sqrt{w}v+1}}{\sqrt{1-v^2}(4\sqrt{w}v-2w-1)}\,\mathrm{d}v. \end{aligned}$$

Roughly, when k lies in the middle range, the main contribution comes from the second integral, which becomes asymptotically negligible for k outside that range.

Proposition 1. *Uniformly for $\alpha \le K$,*

$$[w^k z^n]F_2(z,w) = [w^k]\frac{2w}{\pi}\int_{-1}^1 \frac{1}{\sqrt{1-v^2}(2w+1-4\sqrt{w}v)}\left(\frac{n^{2w-1}}{\Gamma(2w)} - \frac{n^{4\sqrt{w}v-2}}{\Gamma(4\sqrt{w}v-1)}\right)\mathrm{d}v + T_1, \quad (35)$$

where

$$T_1 = O\left(\frac{(2\log n)^k}{n^2 k!}\sqrt{k}\log n + \frac{(2\log n)^{2k}}{n^3 k!^2}\,k\log n\right).$$

Proof. By singularity analysis (see Flajolet and Odlyzko, 1990), we have

$$[z^n](1-z)^{-\omega} = \frac{n^{\omega-1}}{\Gamma(\omega)}\left(1 + \frac{\omega(\omega-1)}{2n} + O\left(n^{-2}\right)\right),$$

uniformly for $|\omega| \le K$. Note that if $4\sqrt{w}v \sim 2w+1$, then

$$[z^n]\frac{(1-z)^{-2w} - (1-z)^{-4\sqrt{w}v+1}}{2w+1-4\sqrt{w}v} = O\left(n^{2\Re(w)-1}\log n\right).$$

Thus

$$[z^n]F_2(z,w) = \frac{2w}{\pi}\int_{-1}^1 \frac{1}{\sqrt{1-v^2}(2w+1-4\sqrt{w}v)}\left(\frac{n^{2w-1}}{\Gamma(2w)} - \frac{n^{4\sqrt{w}v-2}}{\Gamma(4\sqrt{w}v-1)}\right)\mathrm{d}v + T_2,$$

where

$$T_2 = O\left(n^{2\Re(w)-2}\log n + n^{4\Re(\sqrt{w})-3}\log n\right).$$

Now by Cauchy's integral formula

$$\begin{aligned}[w^k]T_2 &= O\left(r_1^{-k}n^{2r_1-2}\log n + r_2^{-k}n^{4\sqrt{r}-3}\log n\right)\\ &= O\left(n^{\alpha-\alpha\log(\alpha/2)-2}\log n + n^{2(\alpha-\alpha\log(\alpha/2)-1)-2}\log n\right),\end{aligned}$$

by taking $r_1 = \alpha/2$ and $r_2 = (\alpha/2)^2$. Thus (35) follows. ∎

Cases (I) and (V). Consider first **Case (I)**. With the uniform estimate (35) at hand, we obtain the leading term in (14) by expanding the factor

$$H(w) := \frac{2w}{\sqrt{4w^2-12w+1}\,\Gamma(2w)},$$

at $w = \alpha/2$ and then use saddlepoint method; see Hwang (1995) for similar details. It remains to show, again by (35), that the integral

$$T_3 := \frac{2}{\pi}\cdot\frac{1}{2\pi i}\oint_{|w|=r} w^{-k}\int_{-1}^{1}\frac{n^{4\sqrt{w}v-2}}{\sqrt{1-v^2}(2w+1-4\sqrt{w}v)\Gamma(4\sqrt{w}v-1)}\,\mathrm{d}v\,\mathrm{d}w,$$

where $r := (\alpha/2)^2$, satisfies

$$T_3 = O\left(\frac{(2\log n)^{2k}}{n^2 k!^2}\right), \tag{36}$$

uniformly for $\alpha \in [0, 3-2\sqrt{2}]$. Indeed, we prove that this estimate holds uniformly for $|\alpha-(2\pm\sqrt{2})| \geq K\sqrt{\log n}$.

By the elementary inequality $1-\cos t \geq 2t^2/\pi^2$ for $|t| \leq \pi$, we have

$$n^{4\Re(\sqrt{w})v} = n^{4\sqrt{r}v\cos(t/2)} \leq n^{2\alpha v-\alpha v t^2/\pi^2} = e^{2kv-kvt^2/\pi^2} \qquad (|t|\leq\pi),$$

so that the major contribution to T_3 comes from the ranges

$$1-\varepsilon \leq v \leq 1 \quad \text{and} \quad \{w = re^{it} \,:\, |t|\leq\varepsilon\},$$

the integrals over the remaining ranges being bounded above by

$$O\left(n^{2(\alpha-\alpha\log(\alpha/2)-1)-\varepsilon}\right).$$

Thus when $|2r+1-4\sqrt{r}| = |\alpha^2-4\alpha+2|/2 \geq \varepsilon$

$$\begin{aligned}T_3 &= O\left(\frac{(\alpha/2)^{-2k}n^{-2}}{|\alpha^2-4\alpha+2|}\int_{|t|\leq\varepsilon} e^{2k-kt^2/\pi^2}\int_0^{(\log n)^{-3/5}} u^{-1/2}e^{-2ku}\,\mathrm{d}u\,\mathrm{d}t\right)\\ &= O\left(\frac{(\alpha/2)^{-2k}e^{2k}n^{-2}}{|\alpha^2-4\alpha+2|k}\right),\end{aligned}$$

from which we obtain (36). By examining further the second order terms (see (39) below), we can take $\varepsilon = K/\sqrt{\log n}$. This proves (14).

The estimate (18) is similar.

Cases (II) and (IV). Consider first **Case (II)**. Since there is a singularity at $w = \beta := 3/2 - \sqrt{2}$, we apply again singularity analysis to the integral

$$T_4 := \frac{1}{2\pi i} \oint_{|w|=r} H(w) w^{-k-1} n^{2w-1} \, \mathrm{d}w$$
$$= \frac{1}{2\pi i} \oint_{|w|=r} \frac{h(w)}{\sqrt{\beta - w}} \, w^{-k} n^{2w-1} \, \mathrm{d}w, \tag{37}$$

where $0 < r < \beta$ and

$$h(w) := \frac{2}{\Gamma(2w)} \sqrt{\frac{\beta - w}{4w^2 - 12w + 1}},$$

the principal branch being taken so that $h(w) > 0$ for $0 < w < \beta$. The integration circle is then deformed into the one shown in Figure 3, where the smaller circle (left) is described by $|w - \beta| = 1/k$.

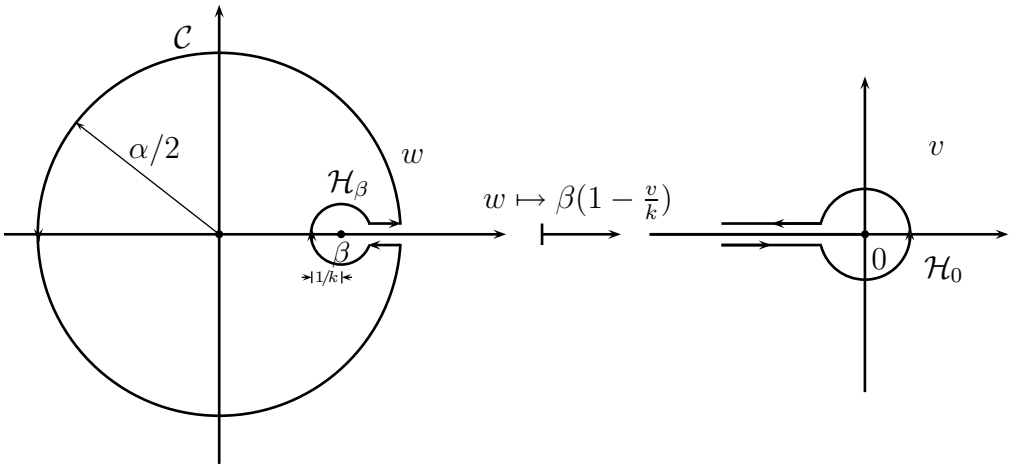

Figure 3: *The Hankel type contours used for proving the estimate in* **Case** (***II***).

The contribution to T_4 from the outer circle $\mathcal{C}$ is easily seen to be of order

$$O\left(\frac{n^{\alpha - \alpha \log(\alpha/1) - 1}}{\sqrt{(2\beta - \alpha) \log n}} \right).$$

For the integral along the contour $\mathcal{H}_\beta$, we make the change of variables $w \mapsto \beta(1 - v/k)$, so that $\mathcal{H}_\beta$ is transformed into $\mathcal{H}_0$ (also shown in Figure 3). Then

$$T_4 = k^{-1/2} \beta^{-\beta+1/2} n^{2\beta-1} \cdot \frac{1}{2\pi i} \int_{\mathcal{H}_0} h\left(\beta\left(1 - v/k\right)\right) v^{-1/2} e^{v(1-2\beta/\alpha)} \left(1 + O(|v|^2 k^{-1})\right) \, \mathrm{d}v$$
$$+ O\left(\frac{n^{\alpha - \alpha \log(\alpha/1) - 1}}{\sqrt{(2\beta - \alpha) \log n}} \right)$$
$$= \frac{h(\beta)}{\sqrt{\pi(1 - 2\beta/\alpha)}} k^{-1/2} \beta^{-\beta+1/2} n^{2\beta-1} \left(1 + O\left(\frac{1}{(\alpha - 2\beta)^2 k} \right) \right),$$

from which (15) follows since $\beta^{1/2} = 1 - 2^{-1/2}$ and $h(\beta) = 2^{-3/4}/\Gamma(2\beta)$; see Flajolet and Odlyzko (1990) for similar details. The error term yields exactly the left boundary $\alpha \geq (3 - 2\sqrt{2}) \log n + K\sqrt{\log n}$; the right boundary $(2 - \sqrt{2}) \log n - K\sqrt{\log n}$ comes from (35).

For the estimate (17), the proof is similar. Note that since $H(w)$ has a singularity at $w = \beta$, we have to start from (35) and then proceed similarly.

Middle range. We use again (35). The same observation that the major contribution comes from $v \sim 1$ and w near the positive real line is still needed since there may be removable singularity for some v. The integrals are estimated similarly as above, and we need only a more precise approximation to T_3. Since an asymptotic expansion for T_3 is derived in the next section, we drop the details for deriving (16) here to avoid repetition.

5 An asymptotic expansion for $\mathbb{V}(X_{n,k})$ in the middle range

We first prove in this section the following expansion for $\mathbb{E}(X_{n,k}^2)$.

Lemma 4. *If $\alpha \in [\![2-\sqrt{2}, 2+\sqrt{2}]\!]$, then*

$$\mathbb{E}(X_{n,k}^2) \sim n^{2(\alpha-\alpha\log(\alpha/2)-1)} \sum_{j\ge 1} \frac{\eta_j(\alpha)}{(\log n)^j}, \tag{38}$$

for some coefficients $\eta_j(\alpha)$.

Proof. Since

$$\mathbb{E}(X_{n,k}^2) = \mathbb{E}(X_{n,k}(X_{n,k}-1)) + O(\mathbb{E}(X_{n,k})),$$

and by the estimate (12) and the analysis in the last section, we need to evaluate the integral

$$T_5 := \frac{1}{2\pi i} \int_{\substack{|w|=(\alpha/2)^2 \\ |\arg(w)|\le\varepsilon}} w^{-k-1} n^{4\sqrt{w}-2} \int_0^{\varepsilon} G(w,u) u^{-1/2} n^{-4\sqrt{w}u} \,\mathrm{d}u \,\mathrm{d}w,$$

where

$$G(w,u) := \frac{2w}{\pi\sqrt{2-u}(4\sqrt{w}(1-u)-2w-1)\Gamma(4\sqrt{w}(1-u)-1)}.$$

By applying Laplace's method (or Watson's lemma; see Wong, 1989) for the inner integral, we obtain

$$T_5 \sim \sum_{j\ge 0} \frac{\Gamma(j+1/2)}{(4\log n)^{j+1/2}} \cdot \frac{1}{2\pi i} \int_{\substack{|w|=(\alpha/2)^2 \\ |\arg(w)|\le\varepsilon}} g_j(w) w^{-k-j/2-5/4} n^{4\sqrt{w}-2} \,\mathrm{d}w,$$

where ($\kappa(w) := 4\sqrt{w}-2w-1$)

$$\begin{aligned} g_j(w) &:= [u^j]G(w,u) \\ &= \frac{\sqrt{2}\,w}{\pi\Gamma(4\sqrt{w}-1)} \sum_{0\le m\le j} \frac{(4\sqrt{w})^{j-m}}{\kappa(w)^{j-m+1}} \sum_{0\le \ell\le m} \binom{2\ell}{\ell} 8^{-\ell} [u^{m-\ell}] \frac{\Gamma(4\sqrt{w}-1)}{\Gamma(4\sqrt{w}-1-4\sqrt{w}u)}. \end{aligned}$$

Then a straightforward application of saddlepoint method leads to (38). Note that

$$\eta_j(\alpha) = O\left(|4\alpha-\alpha^2-2|^{-2j-1}\right), \tag{39}$$

when $\alpha \to 2\pm\sqrt{2}$ (from inside the interval $(2-\sqrt{2}, 2+\sqrt{2})$), implying that the asymptotic expansion (38) is meaningful in the region $[\![2-\sqrt{2}, 2+\sqrt{2}]\!]$. ∎

Note that the asymptotic expansion (38) can also be derived in a more straightforward way by starting from (8) and applying the expansion for the modified Bessel function (see §9.7, Abramowitz and Stegun, 1965).

Proof of Theorem 4. From the asymptotic expansion (13), we obtain

$$(\mathbb{E}(X_{n,k}))^2 \sim n^{2(\alpha-\alpha\log(\alpha/2)-1)} \sum_{j\ge 1} \frac{\xi_j(\alpha)}{(\log n)^j}, \tag{40}$$

for some coefficients $\xi_j(\alpha)$. Then combining (40) and (38) leads to (24) with $\upsilon_j(\alpha) = \eta_j(\alpha) - \xi_j(\alpha)$. ∎

Calculations of the coefficients. The coefficients in the expansions (38) and (40) can be easily computed with the assistance of any symbolic softwares, but are very challenging by hand. For example, when $\alpha = 2 + t/\log n$, we can rewrite (24) as

$$\mathbb{V}(X_{n,k}) \sim \sum_{j\ge 1} \frac{p_j(t)}{(\log n)^{j+2}}, \tag{41}$$

where $p_j(t)$ is a polynomial of degree $j+1$ given by $p_j(t) := \sum_{0\le \ell\le j+1} \upsilon_{j+2-\ell}^{(\ell)}(2) t^\ell/\ell!$. Since the coefficients of $(\log n)^{-1}$ and $(\log n)^{-2}$ are both zero in the expansion, we need explicit coefficients of $\upsilon_j(\alpha)$ for $j = 1, 2, 3$ in order to get the form for $p_1(t)$.

In particular, writing $q(x) := -x^2 + 4x - 2$ and $\bar\alpha := 2\alpha - 1$, we have

$$\begin{aligned}
\eta_1(\alpha) &= \frac{\alpha}{2\pi q(\alpha)\Gamma(\bar\alpha)},\\
\eta_2(\alpha) &= \frac{-1}{12\pi q(\alpha)^3\Gamma(\bar\alpha)} \big(-6\alpha^2 q(\alpha)^2\psi'(\bar\alpha) + 6\alpha^2 q(\alpha)^2\psi(\bar\alpha)^2\\
&\qquad -24\alpha(\alpha-1)q(\alpha)\psi(\bar\alpha) + \alpha^4 + 16\alpha^3 - 52\alpha^2 + 32\alpha + 4\big),\\
\eta_3(\alpha) &= \frac{\left(\begin{array}{l}
36\alpha^4 q(\alpha)^4\psi(\bar\alpha)^4 + 48\alpha^3 q(\alpha)^3 q_1(\alpha)\psi(\bar\alpha)^3\\
-12\alpha^2 q(\alpha)^2\left[18\alpha^2 q(\alpha)^2\psi'(\bar\alpha) - q_2(\alpha)\right]\psi(\bar\alpha)^2\\
+48\alpha q(\alpha)\left[3\alpha^3 q(\alpha)^3\psi''(\bar\alpha) - 3\alpha^2 q_1(\alpha)q(\alpha)^2\psi'(\bar\alpha) - q_3(\alpha)\right]\psi(\bar\alpha)\\
-36\alpha^4 q(\alpha)^4\psi'''(\bar\alpha) + 48\alpha^3 q(\alpha)^3 q_1(\alpha)\psi''(\bar\alpha)\\
-12\alpha^2 q(\alpha)^2 q_2(\alpha)\psi'(\bar\alpha) + 108\alpha^4 q(\alpha)^4\psi'(\bar\alpha)^2 + q_4(\alpha)
\end{array}\right)}{144\pi\alpha q(\alpha)^5\Gamma(\bar\alpha)}
\end{aligned}$$

where ψ is the logarithmic derivative of the Gamma function and

$$\begin{aligned}
q_1(\alpha) &= \alpha^2 - 10\alpha + 8\\
q_2(\alpha) &= \alpha^4 + 16\alpha^3 + 32\alpha^2 - 112\alpha + 76\\
q_3(\alpha) &= 7\alpha^5 + 3\alpha^4 - 68\alpha^3 + 108\alpha^2 - 52\alpha - 4\\
q_4(\alpha) &= \alpha^8 + 320\alpha^7 - 856\alpha^6 - 1600\alpha^5 + 8920\alpha^4 + 11264\alpha^3 + 4640\alpha^2 + 256\alpha + 16.
\end{aligned}$$

And the first three $\xi_j(\alpha)$'s are given by (see (13))

$$\begin{aligned}
\xi_1(\alpha) &= \frac{1}{2\pi\alpha\Gamma(\alpha)^2}, \quad \xi_2(\alpha) = -\frac{6\alpha^2\psi'(\alpha) - 6\alpha^2\psi(\alpha)^2 - 1}{12\pi\alpha^2\Gamma(\alpha)^2},\\
\xi_3(\alpha) &= \frac{\left(\begin{array}{l}
-18\alpha^4\psi'''(\alpha) + 24\alpha^3(3\alpha\psi(\alpha) - 2)\psi''(\alpha)\\
+72\alpha^4\psi'(\alpha)^2 - 12\alpha(12\alpha^2\psi(\alpha)^2 - 12\alpha\psi(\alpha) + 1)\psi'(\alpha)\\
+36\alpha^4\psi(\alpha)^4 - 48\alpha^3\psi(\alpha)^3 + 12\alpha^2\psi(\alpha)^2 + 1
\end{array}\right)}{144\pi\alpha^3\Gamma(\alpha)^2}.
\end{aligned}$$

The exact forms of ξ_j and η_j are less important; the special property we need is that (see Figure 4)

$$\upsilon_1(i) = \upsilon_1'(i) = \upsilon_2(i) = 0 \qquad (i = 1, 2).$$

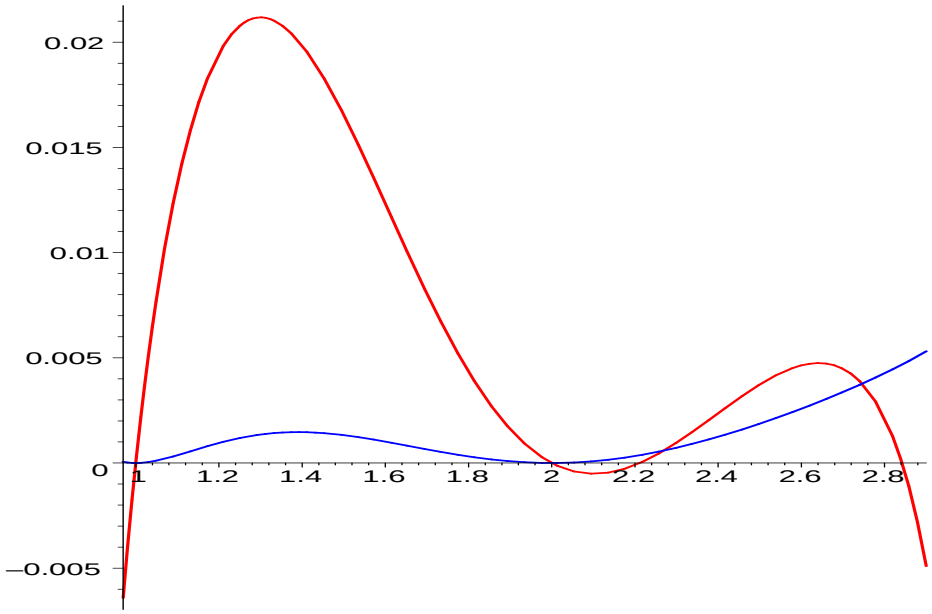

Figure 4: *A plot of the two functions $v_1(\alpha)$ (smaller amplitude) and $v_2(\alpha)$, where the horizontal line is the value of α. There are additional zeros for the latter besides 1 and 2, but they are minor.*

Order of the two "humps". By the expansions (41) and

$$2(\alpha - \alpha\log(\alpha/2) - 1)\log n = \sum_{j\ge 2} \frac{(-1)^{j-1}t^j}{j(j-1)2^{j-2}(\log n)^{j-1}},$$

when $\alpha = 1 + t/\log n$, we have, when $t = o((\log n)^{2/3})$,

$$\mathbb{V}(X_{n,k}) \sim \frac{n^2}{720\pi(\log n)^3} e^{-t^2/(2\log n)} \left(p_1(t) + \frac{p_2(t)}{\log n} + \cdots\right).$$

Since $p_1(t)$ is a quadratic polynomial, the asymptotic maximum of the right-hand side is easily seen to be reached at $t = \pm\sqrt{2\log n} + O(1)$, and

$$\mathbb{V}(X_{n,k}) = e^{-1}\frac{n^2}{(\log n)^3}\left(\frac{21-2\pi^2}{24\pi}\log n \mp \frac{\sqrt{2}(21-2\pi^2)}{72\pi}\sqrt{\log n} + O(1)\right),$$

for $t = \pm\sqrt{2\log n} + O(1)$. This roughly explains why the left "hump" is higher than the right "hump".

Expansions for $\alpha = 1 + o(1)$ are similar.

The valley. When $k = \lfloor 2\log n \rfloor$, we have

$$\mathbb{V}(X_{n,k}) \sim \frac{p_1(\{2\log n\})}{720\pi}\cdot\frac{n^2}{(\log n)^3}.$$

Since $p_1(t)$ is concave upward, the minimum of $p_1(\{2\log n\})$ is asymptotically achieved at the subsequence of n for which $\{2\log n\} \to 1 - t_0$.

Note that the range in (26) where $\mathbb{V}(X_{n,k}) \ge (C+o(1))n^2/(\log n)^3$ can be extended from $O(\sqrt{\log n})$ to t_n, where $t_n \to \infty$ is given by $t_n^2 e^{-t_n^2/(2\log n)} = C$, which (expressible in terms of Lambert's W-function) satisfies asymptotically,

$$t_n = \sqrt{2\log n \log\log n}\left(1 + \frac{\log\log\log n + O(1)}{\log\log n}\right).$$

6 Transitional behaviors

We prove Theorem 3 in this section. By analogy, we prove only the first two estimates (20) and (21).

The first phase transition at $3-2\sqrt{2}$. Recall that $D_\nu(z)$ denotes the parabolic cylinder functions (see (19) and Ch. 19, Abramowitz and Stegun, 1965). Define $\beta = 3/2 - \sqrt{2}$. To describe the transitional behavior (20) of $\mathbb{E}(X_{n,k}^2)$ near the point $\alpha = 3-2\sqrt{2}$, it suffices to evaluate the integral T_4 defined in (37) and prove the following estimate.

Lemma 5. *If $\alpha = 2\beta + \sqrt{2\beta}t/\sqrt{\log n}$, then*

$$T_4 = \frac{h(\beta)\sqrt{\beta}}{\sqrt{2\pi}}\, e^{t^2/4} D_{-1/2}(-t) k^{-1/4} (\alpha/2)^{-k} n^{\alpha-1} \left(1 + O\left(\frac{|t|+|t|^3}{\sqrt{k}}\right)\right), \tag{42}$$

uniformly for $t = o((\log n)^{1/6}$.

Estimates uniformly valid in a wider interval of α can be derived by standard tools for handling coalescence of algebraic singularities and saddlepoints; see Bleistein and Handelsman (1975). We content here with the above estimates using the following simpler method of proof.

Proof. Assume first that $\alpha < 2\beta$. By the change of variables $w \mapsto \alpha(1+iv/\sqrt{k})/2$, we deduce that

$$\begin{aligned} T_4 = h(\alpha/2)(\alpha/2)^{-k+1/2} n^{\alpha-1} k^{-1/4} \cdot \frac{1}{2\pi} \fint_{-\varepsilon\sqrt{k}}^{\varepsilon\sqrt{k}} \frac{e^{-v^2/2}}{\sqrt{\Delta\sqrt{k} - iv}} \left(1 + O\left(\frac{|v|+|v|^3}{\sqrt{k}}\right)\right) \mathrm{d}v \\ + O\left((\alpha/2)^{-k} n^{\alpha-1-\varepsilon}\right), \end{aligned}$$

where $\Delta := 2\beta/\alpha - 1$ and $\fint$ means that an indentation (upward) of the integration path is needed if $\Delta = 0$. For the integral on the right-hand side, we use the integral representation (see p. 688, Abramowitz and Stegun, 1964)

$$\frac{1}{2\pi} \fint_{-\infty}^{\infty} \frac{e^{-v^2/2}}{\sqrt{x - iv}}\, \mathrm{d}v = \frac{e^{x^2/4}}{\sqrt{2\pi}}\, D_{-1/2}(x) \qquad (x \in \mathbb{R}),$$

and the estimates (22). The estimate (42) then follows by the expansion

$$\Delta\sqrt{k} = -t + O\left(t^2 (\log n)^{-1/2}\right). \quad \blacksquare$$

The second phase transition at $\alpha = 2-\sqrt{2}$. From the proof of (38), we have

$$\begin{aligned} T_5 &= \frac{1}{2\pi i} \int_{\substack{|w|=(\alpha/2)^2 \\ |\arg(w)|\le\varepsilon}} w^{-k-1} n^{4\sqrt{w}-2} \frac{g_0(w)\sqrt{\pi}}{\sqrt{4\sqrt{w}\log n}} \left(1 + O\left(\frac{1}{|\kappa(w)|\log n}\right)\right) \mathrm{d}w \\ &= \frac{n^{-2}}{\sqrt{\log n}} \cdot \frac{1}{2\pi i} \int_{\substack{|u|=\alpha/2 \\ |\arg(u)|\le\varepsilon}} \frac{g(u)}{u - (1-2^{-1/2})}\, u^{-2k} n^{4u} \left(1 + O\left(\frac{1}{|4u-2u^2-1|\log n}\right)\right) \mathrm{d}u, \end{aligned}$$

where

$$g(u) := \frac{\sqrt{2u}(u - (1-2^{-1/2}))}{\sqrt{\pi}(4u - 2u^2 - 1)\Gamma(4u-1)}.$$

We need to prove the following estimate, which implies (21).

Lemma 6. *If* $\alpha = 2 - \sqrt{2} + \sqrt{1 - 2^{-1/2}}t/\sqrt{\log n}$, *then*

$$T_5 = g(\alpha/2)e^{t^2/2}\Phi(-t)(\log n)^{-1/2}n^{2-\alpha-2\alpha\log(\alpha/2)}\left(1 + O\left(\frac{1+|t|^3}{\sqrt{\log n}}\right)\right), \tag{43}$$

uniformly for $t = o((\log n)^{1/6}$

Proof. The proof follows, *mutatis mutandis*, the same pattern as for (42), starting with the change of variables $v = \alpha(1 + iv/\sqrt{2k})/2$. The main difference is that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-v^2/2}}{iv - x}\,\mathrm{d}v = e^{x^2/2}\Phi(-x) \qquad (x \in \mathbb{R}),$$

where the integration path has to be indented suitably downward when $x = 0$. Note that

$$g(1 - 2^{-1/2}) = \frac{\sqrt{2 - \sqrt{2}}}{2\sqrt{2\pi}\Gamma(3 - 2\sqrt{2})}. \quad \blacksquare$$

7 Profiles of recursive trees

We briefly discuss the profiles of random recursive trees in this section.

One way of constructing a random recursive tree of n nodes is as follows. One starts from a root node holding the key 1; at stage i ($i = 2, \ldots, n$) a new node holding i is attached uniformly at random to one of the previous nodes. The process stops after node n is inserted. By construction, the values of the nodes along any path from the root to a node forms an increasing sequence. For a survey of probabilistic properties of recursive trees, see Smythe and Mahmoud (1995).

Let $X_{n,k}$ denote the number of *internal nodes* at level k in a random recursive tree of n nodes. Then (see van der Hofstad et al., 2002)

$$X_{n,k} \stackrel{d}{=} X_{\mathsf{unif}[1,n-1],k-1} + X^*_{n-\mathsf{unif}[1,n-1],k},$$

where $X^*_{n,k}$ is an independent copy of $X_{n,k}$ and unif$[1, n-1]$ takes any of the values in $\{1, \ldots, n-1\}$ with equal probability $1/(n-1)$.

From this recursive decomposition, we deduce that

$$\begin{cases} P_0(z,y) & = 1 + \dfrac{yz}{1-z}, \\ P_{k+1}(z,y) & = 1 + z\exp\left(\displaystyle\int_0^z \frac{P_k(t,y)-1}{t}\,\mathrm{d}t\right) \quad (k \geq 0), \end{cases}$$

where $P_k(z,y) := \sum_n \mathbb{E}(y^{X_{n,k}})z^n$. Adopting the same set of symbols used for BSTs, we obtain

$$F_1(z,w) = z(1-z)^{-1-w},$$

so that

$$\mathbb{E}(X_{n,k}) = \frac{s(n,k+1)}{(n-1)!}.$$

Similarly, for the second factorial moment,

$$F_2(z,w) = 2\sqrt{w}z(1-z)^{-1-w}\int_0^z (1-t)^{w-1}I_1\left(2\sqrt{w}\,\log\frac{1}{1-t}\right)\,\mathrm{d}t,$$

where

$$I_1(z) := \frac{1}{2}\sum_{m\geq 0}\frac{z^{2m+1}}{m!(m+1)!4^m}$$

denotes the modified Bessel function of first order. Note that for $|w| > 4$

$$2\sqrt{w}\int_0^1 (1-t)^{w-1} I_1\left(2\sqrt{w}\,\log\frac{1}{1-t}\right)\,\mathrm{d}t = \frac{4}{w\sqrt{1-4/w}(1+\sqrt{1-4/w})}.$$

The same set of tools used for BSTs also applies here; the analytic context is indeed much simpler since it is known that (see van der Hofstad et al., 2002)

$$\mathbb{E}(X_{n,k}^2) = \sum_{0\leq j\leq k}\binom{2j}{j}\frac{s(n,k+j+1)}{(n-1)!};$$

compare (9).

The asymptotic behaviors of $\mathbb{E}(X_{n,k}^2)$ can be summarized as follows. Let $\alpha := k/\log n$.

– If $\alpha \in [0,2]$, then

$$\mathbb{E}(X_{n,k}^2) \sim \frac{(\log n)^{2k}}{(1-\alpha/2)\Gamma(2\alpha+1)k!^2}; \tag{44}$$

– if $\alpha = 2 + t/\sqrt{\log n}$, then

$$\mathbb{E}(X_{n,k})^2 \sim \frac{1}{24\sqrt{\pi}}\,\Phi(t)k^{-1/2}4^{-k}n^4,$$

uniformly for $t = o((\log n)^{1/6})$;

– if $\alpha \in [\![2,4]\!]$, then

$$\mathbb{E}(X_{n,k}^2) \sim \frac{1}{24\sqrt{\pi(4\log n - k)}}\,4^{-k}n^4;$$

– if $\alpha = 4 + 2t/\sqrt{\log n}$, then

$$\mathbb{E}(X_{n,k}^2) \sim \frac{1}{24\sqrt{2\pi}}\,e^{t^2/2}D_{-1/2}(t)k^{-1/4}4^{-k}n^4,$$

uniformly for $t = o((\log n)^{1/6})$;

– if $\alpha \in [\![4,K]$, then

$$\mathbb{E}(X_{n,k}^2) \sim \left(1 + \frac{4}{\alpha\sqrt{1-4/\alpha}(1+\sqrt{1-4/\alpha})}\right)\frac{(\log n)^k}{\Gamma(\alpha+1)k!}.$$

From (44) and the following estimate for the mean

$$\mathbb{E}(X_{n,k}) \sim \frac{(\log n)^k}{\Gamma(\alpha+1)k!} \qquad (\alpha \in [0,K]),$$

we obtain, for $\alpha \in [0,2]$,

$$\mathbb{V}(X_{n,k}) \sim \varphi(\alpha)\frac{(\log n)^{2k}}{k!k!},$$

where

$$\varphi(\alpha) = \frac{1}{(1-\alpha/2)\Gamma(2\alpha+1)} - \frac{1}{\Gamma(\alpha+1)^2}.$$

The function $\varphi(\alpha)$ satisfies $\varphi(1) = \varphi'(1) = 0$, and the same type of bimodal behavior occurs when $\alpha = 1 + O(1/\sqrt{\log n})$, with the variance varying from $n^2/(\log n)^3$ to $n^2/(\log n)^2$ there. Finer results as those for BSTs can be derived; we omit all details here. Interestingly, $\mathbb{V}(X_{n,k})$ starts to exhibit the bimodal behavior for $n = 33$, much smaller than that for BSTs.

8 Conclusions

In this paper, we added several new aspects to the usual description of the profiles of BSTs as some fig-like shape $\diamondsuit$. In addition to the new phenomena of phase transitions and bimodality exhibited by the variance of the profiles, several parameters on trees have close connections to profiles, especially the mean values. For example, one of the most studied parameters is the hight H_n, defined to be the length of the longest path from the root. The second moments of $X_{n,k}$ for BSTs were originally studied to derive better bounds for some estimates required for the hight; see Pittel (1984). Indeed, already the mean of $X_{n,k}$ can be used to derive useful estimate on the average height as follows; cf. Devroye (1987). By the inequality

$$\mathbb{P}(H_n \geq k) \leq \sum_{j\geq k} \mathbb{E}(X_{n,j}),$$

and (1), we obtain

$$\begin{aligned}\sum_{j\geq k} \mathbb{P}(H_n \geq j) &\leq \sum_{j\geq k} \frac{2^j}{n!}(j-k+1)s(n,j)\\ &= \frac{2^k}{2\pi i}\oint_{|w|=\alpha} w^{-k}\frac{(w+1)\cdots(w+n-1)}{n!(1-2/w)^2}\,\mathrm{d}w\\ &= O\left(k^{-1/2}n^{\alpha-\alpha\log(\alpha/2)-1}\right).\end{aligned}$$

Choose $k_0 = \alpha_+ \log n - \alpha' \log\log n + O(1)$, where $\alpha_+ > 1$ solves the equation $e^{(z-1)/z} = z/2$ and $\alpha' = \alpha_+/(2\alpha_+ - 2)$, so that $\sum_{j\geq k_0} P(H_n \geq j) = O(1)$. And it follows that

$$\begin{aligned}\mathbb{E}(H_n) &\leq \sum_{j\leq k_0} \mathbb{P}(H_n \geq j) + \sum_{j\geq k_0} \mathbb{P}(H_n \geq j)\\ &\leq k_0 + O(1).\end{aligned}$$

This upper bound is, up to the constant of the second-order term, the right order; see Drmota (2003) and the references there for further information. Unfortunately, the second moment of $X_{n,k}$ is not sufficient to prove tight lower bound for $\mathbb{E}(H_n)$; see Pittel (1984). See also Chern and Hwang (2001) for other applications of the mean profile.

Acknowledgement

Part of the work of the second author was finished while he was visiting Institut für Stochastik und Mathematische Informatik, J. W. Goethe-Universität (Frankfurt); he thanks the Institute for hospitality and support.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions (with Formulas, Graphs and Mathematical Tables)*, Dover, New York, 1965.

[2] D. Aldous, The continuum random tree. II. An overview, in *Stochastic analysis (Durham, 1990)*, 23–70, LMS Lecture Note Series, 167, Cambridge University Press, Cambridge, 1991.

[3] D. Aldous, Probability distributions on cladograms, in *Random Discrete Structures*, Edited by D. Aldous and R. Pemantle, pp. 1–18, Springer, 1996.

[4] D. Aldous and P. Shields, A diffusion limit for a class of randomly-growing binary trees, *Probability Theory and Related Fields*, **79** (1988), 509–542.

[5] M. T. Barlow, R. Pemantle and E. A. Perkins, Diffusion-limited aggregation on a tree, *Probability Theory and Related Fields*, **107** (1997) 1–60.

[6] F. Bergeron, P. Flajolet, and B. Salvy, Varieties of increasing trees, in *CAAP '92* (Rennes, 1992), pp. 24–48, *Lecture Notes in Computer Science*, 581, Springer, Berlin, 1992.

[7] N. Bleistein and R. A. Handelsman, *Asymptotic Expansions of Integrals*, Holt, Rinehart and Winston, 1975.

[8] G. G. Brown and B. O. Shubert, On random binary trees, *Mathematics of Operations Research*, **9** (1984), 43–65.

[9] B. Chauvin, M. Drmota and J. Jabbour-Hattab, The profile of binary search trees, *Annals of Applied Probability*, **11** (2001), 1042–1062.

[10] B. Chauvin, T. Klein, J.-F. Marckert, and A. Rouault, Martingales, embedding and tilting of binary trees, preprint, 2003.

[11] 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.

[12] L. Devroye, Branching processes in the analysis of the heights of trees, *Acta Informatica*, **24** (1987), 277–298.

[13] L. Devroye, Applications of the theory of records in the study of random trees, *Acta Informatica*, **26** (1988), 123–130.

[14] L. Devroye, Universal limit laws for depths in random trees, *SIAM Journal on Computing*, *28* (1999), 409–432.

[15] L. Devroye, Limit laws for sums of functions of subtrees of random binary search trees, *SIAM Journal on Computing*, *32* (2003), 152–171.

[16] L. Devroye and J. M. Robson, On the generation of random binary search trees, *SIAM Journal on Computing*, **24** (1995), 1141–1156.

[17] M. Drmota and B. Gittenberger, On the profile of random trees, *Random Structures and Algorithms*, **10** (1997), 421–451.

[18] S. R. Finch, *Mathematical Constants*, Cambridge University Press, 2003.

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

[20] G. H. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, 2nd edition, Addison-Wesley, Wokingham, UK, 1991.

[21] J. M. Hammersley, The sum of products of the natural numbers, *Proceedings of the London Mathematical Society*, **1** (1951), 435–452.

[22] H.-K. Hwang, Asymptotic expansions for the Stirling numbers of the first kind, *Journal of Combinatorial Theory, Series A*, **71** (1995), 343–351.

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

[24] J. Jabbour-Hattab, Martingales and large deviations for the binary search trees, *Random Structures and Algorithms*, **19** (2001), 112–127.

[25] G. Kersting, On the height profile of a conditioned Galton-Watson tree, preprint (1998).

[26] D. E. Knuth, *The Art of Computer Programming, Volume 3, Sorting and Searching*, Second Edition, Addison-Wesley, Reading, MA, 1998.

[27] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *RAIRO Informatique Théorique et Applications*, **21** (1987), 479–495.

[28] G. Louchard and W. Szpankowski, Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm, *IEEE Transactions on Information Theory*, **41** (1995), 478–488.

[29] W. C. Lynch, More combinatorial properties of certain trees, *Computer Journal*, **7** (1965), 299–302.

[30] H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[31] H. Mahmoud and B. Pittel, On the most probable shape of a binary search tree grown from a random permutation, *SIAM Journal on Algebraic and Discrete Methods*, **5** (1984), 69–81.

[32] S. N. Majumdar and P. L. Krapivsky, Extreme value statistics and traveling fronts: various applications, *Physica A*, **318** (2003) 161–170.

[33] J. Pitman, The SDE solved by local times of a Brownian excursion or bridge derived from the height profile of a random tree or forest, *Annals of Probability*, **27** (1999), 261–283.

[34] B. G. Pittel, On growing random binary trees, *Journal of Mathematical Analysis and Applications*, **103** (1984), 461–480.

[35] R. T. Smythe and H. M. Mahmoud, A survey of recursive trees, *Theory of Probabability and Mathematical Statistics*, **51** (1995), 1–27.

[36] N. M. Temme, Asymptotic estimates of Stirling numbers, *Studies in Applied Mathematics*, **89** (1993), 233–243.

[37] R. van der Hofstad, G. Hooghiemstra and P. van Mieghem, On the covariance of the level sizes in random recursive trees, *Random Structures and Algorithms*, **20** (2002), 519–539.

[38] R. Wong, *Asymptotic Approximations of Integrals*, SIAM, Philadelphia, 2001.