Extreme value statistics and traveling fronts: Application to computer science

Satya N. Majumdar[1] and P. L. Krapivsky[2]
[1]*Laboratoire de Physique Quantique (UMR C5626 du CNRS), Université Paul Sabatier, 31062 Toulouse Cedex, France*
[2]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
(Received 17 September 2001; published 27 February 2002)

We study the statistics of height and balanced height in the binary search tree problem in computer science. The search tree problem is first mapped to a fragmentation problem that is then further mapped to a modified directed polymer problem on a Cayley tree. We employ the techniques of traveling fronts to solve the polymer problem and translate back to derive exact asymptotic properties in the original search tree problem. The second mapping allows us not only to rederive the already known results for random binary trees but to obtain exact results for search trees where the entries arrive according to an arbitrary distribution, not necessarily randomly. Besides it allows us to derive the asymptotic shape of the full probability distribution of height and not just its moments. Our results are then generalized to *m*-ary search trees with arbitrary distribution.

DOI: 10.1103/PhysRevE.65.036127 PACS number(s): 02.50.−r, 89.20.Ff, 89.75.Hc

I. INTRODUCTION

The techniques developed in statistical physics, particularly in the theory of spin glasses, have been recently applied to a variety of problems in theoretical computer science [1]. These include various optimization problems such as the traveling salesman problem [2], graph partitioning [3], satisfiability problems [4], the knapsack problem [5], the vertex covering problem [6], error correcting codes [7], number partitioning problems [8], matching problems, [9] and many others [10]. The purpose of this paper is to study analytically certain problems in a different area of theoretical computer science known as sorting and searching [11]. The standard techniques of spin glass theory are not directly suitable for these problems. Instead, we employ the techniques developed to study the propagation of traveling fronts in various nonlinear systems [12–18].

The ''sorting and searching'' is an important area of computer science that deals with the following basic question: How to organize or sort out the incoming data so that the computer takes the minimum time to search for a given data if required later? Amongst various search algorithms, the binary search turns out to be one of the most efficient [11]. To understand this algorithm, let us start with a simple example. Suppose the incoming data string consists of the twelve months of the year appearing in the following order: July, September, December, May, April, February, January, October, November, March, June, and August. Suppose later we need to look for the month of August in this data string. Consider first the sequential search where the computer starts from the first element (July), checks if it is the right month and if not, moves to the next element of the string (September), checks the element there and continues in this fashion until it finds the right month. In the example above, to find the month ''August,'' the computer has to make 12 comparisons. Thus, the sequential search algorithm is rather inefficient as it typically takes a search time of order N, where N is the number of entries in the data string.

In a binary search, on the other hand, the typical search time scales as $\ln N$ [11]. The binary search is implemented by organizing the data string on a tree according to the following algorithm. An order is first chosen for the incoming data, e.g., it can be alphabetical or chronological (January, February, March, etc.). Let us choose the chronological order. Now the first element of the input string (July) is put at the root of a tree (see Fig. 1). The next element of the string is September. One compares with the root element (July) and sees that September is bigger than July (in chronological order). So one assigns September to a daughter node of the root in the right branch. On the other hand, if the new element were less than the root, it would have gone to the daughter node of the left branch. Then the next element is December. We compare at the root (July) and decide that it has to go to the right, then we compare with the existing right daughter node (September) and decide that December has to go to the node that is the right daughter of September. The process continues till all the elements are assigned their nodes on the tree. For the particular data string in the above example, we finally get the unique tree as shown in Fig. 1. Such a tree is called a binary search tree (BST).

Once this tree is constructed, the subsequent search, say for the month of August, takes much less number of comparisons. We start with the root (July). Since the sought after element August is bigger than July, we know that it must be on the right branch of the two daughter subtrees. This already eliminates searching roughly half the elements which

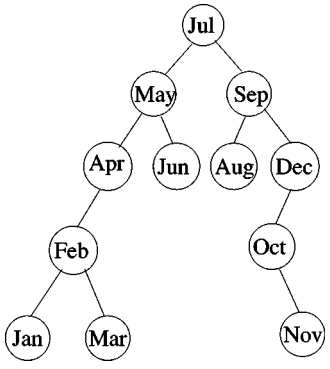

FIG. 1. The binary search tree corresponding to the data string in the order: July, September, December, May, April, February, January, October, November, March, June, and August.

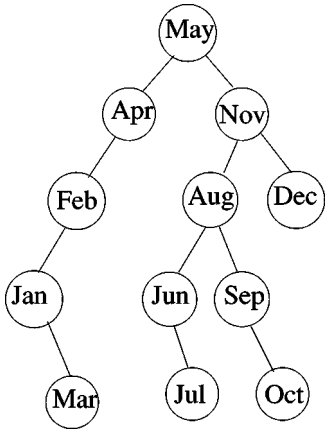

FIG. 2. The binary search tree corresponding to the data string in the order: May, November, August, April, December, February, June, September, July, January, October, and March.

are on the left subtree. We next encounter the key September. Since August is less than September, we go to the left and thus we do not need to search anymore the right branch of the subtree rooted at September. Once we go to the left, we find the required key August.

Thus, the BST algorithm requires only three comparisons as opposed to 12 operations in the sequential search. Since the typical search time is proportional to the depth of an element in the tree and since the typical depth D is related to the total size N via $2^D \approx N$, the search time scales as $\ln N$, making the BST algorithm one of the most efficient search algorithms.

If the incoming data string had a different order of appearance, one would have obtained a different BST. For example, suppose the months appear in a different order as: May, November, August, April, December, February, June, September, July, January, October, and March. For this data string, the same algorithm of constructing a binary tree as before gives a tree of different shape (Fig. 2). Each permutation of the incoming data string leads to a different binary tree and there are $N!$ possible binary trees for any incoming data string with N entries. Usually the incoming data string appears in a random fashion. This would indicate that each of the $N!$ possible binary trees occurs with equal probability. Such trees, each generated with equal probability, are called ''random binary search trees'' (RBST). Of course, if the incoming data is not completely random, the probability measure over the space of trees will not be uniform. The results derived in this paper will be applicable not only to RBST but also to more general BST's with arbitrary measure.

Each BST has several observables (such as the depth or the height of a tree) associated with it that quantify the efficiency of the underlying search algorithm. Hence the knowledge of the statistics of such observables are of central importance. Here are a few observables:

D_N is the depth (distance from the root) of the last inserted element in a given BST of size N. For example, $D_N = 3$ for the tree in Fig. 1 (counting the depth of the root element as 1). Each BST has a different D_N, so D_N is a random variable. The average depth $\langle D_N \rangle$ (averaged over the probability measure of the trees) gives a measure of the average time required to insert a new element in a tree [11].

H_N is the height of a given tree, defined as the depth of the farthest node from the root. For example, $H_N = 5$ for the BST in Fig. 1. Clearly H_N is also a random variable and $\langle H_N \rangle$ gives a measure of the *maximum* possible time that could be required to insert an element, i.e., a measure of the *worst* case scenario [19–22].

h_N is the balanced height defined as the *maximum* depth from the root up to which the tree is *fully* balanced, i.e., all the nodes up to this depth are fully occupied [11]. In the tree shown in Fig. 1, $h_N = 3$, whereas $h_N = 2$ in the tree in Fig. 2. Hence h_N is also a random variable whose statistics is important.

Some of the observables mentioned above such as the height H_N and the balanced height h_N are of extremal nature, i.e., they are the maximum or the minimum of a set of *correlated* random variables. In this paper, we limit ourselves only to such extreme observables of the binary tree. While the statistics of the extremum of a set of *uncorrelated* random variables is well understood [23–25], little is known about the same for correlated variables [26]. However, in the present problem the random variables are correlated in a special hierarchical way that facilitates analysis. We will see that the extreme variables in the BST problem satisfy nonlinear recursion relations that admit traveling front solutions in some suitable variables. A lot is known about the speed and the shape of such fronts appearing in various nonlinear systems [12–18]. Below, we will use these techniques to study the statistics of extreme variables in the binary tree problem.

Some of our results for the RBST were already known that we will mention as we go along. However, the approach used here is quite different from those used by the computer scientists. Computer scientists tend to establish upper and lower bounds to the quantity of interest (typically the average value or the variance of the observable) and then tighten the bounds [21]. If the bounds coincide, one obtains an exact result [21]. Our approach, on the other hand, is a typical physicist's approach. The methods we use may not always be rigorous in the strict mathematical sense, but they lead to exact asymptotic results in a physically transparent way. Moreover, our approach allows us not only to reproduce already known asymptotics for the average height and the average balanced height of the RBST, but also to obtain information about the variance and even the asymptotic shapes of the full probability distributions. Besides, our method goes beyond the RBST and yields exact results for trees generated with arbitrary distributions.

Our approach utilizes two exact mappings that can be summarized as follows. Following Devroye [27], we first map the RBST problem to a random fragmentation problem where an object of initial length N breaks randomly into two fragments, each of which further breaks randomly into two parts, and so on. The fragmentation problem is interesting on its own right as it appears in the context of various physical problems such as the energy cascades in turbulence [28], rapture processes in earthquakes [29], financial crashes in stock markets [30], and the stress propagation in granular medium [31]. Some of the extremal problems in the random

fragmentation problem were studied recently by Hattori and Ochiai [32] and by us [33]. The method used in our previous paper [33] allowed us to obtain exact asymptotic results for the average of the maximal piece of the 2^n fragments after n iterations. The statistics of this maximal piece is closely related to the statistics of the height in the RBST. However, this method was not easy to extend to the cases beyond the random fragmentation, i.e., when the break point is chosen from an arbitrary distribution, not necessarily uniform. We will see later that a fragmentation problem with a given break point distribution corresponds to a BST problem where the incoming entries to the tree appear according to a specific distribution and not just randomly.

In this paper, we show that the fragmentation problem, with arbitrary break point distribution, can further be mapped onto a modified directed polymer (MDP) problem on a Cayley tree. The MDP problem differs from the conventional directed polymer (DP) problem on a Cayley tree studied by Derrida and Spohn [34] due to the presence of a special constraint in the MDP. Derrida and Spohn were mostly interested in the finite temperature spin glass transition in the DP problem. Our problem reduces to a zero temperature problem, albeit with a special constraint. We then solve this MDP problem using traveling front techniques and translate back to derive exact asymptotic results for the original BST problem. We will see that the statistics of the height H_N of the BST problem is related (via the two successive mappings) to the statistics of the minimum or the ground state energy of the MDP problem. On the other hand, the statistics of the balanced height h_N will be related to that of the maximum energy of the directed polymer (a quantity of little interest in statistical physics framework). This second mapping also allows us to obtain exact results for nonrandom BST problem.

The paper is organized as follows. In Sec. I, we set up notation, review known results for the RBST problem, and summarize the results obtained in this paper. Section II contains the exact mapping of the BST problem to the fragmentation problem. In Sec. III, we map the fragmentation problem to the MDP problem. In Sec. IV, we derive the exact nonlinear recursion relations in the MDP problem and analyze them using the traveling front techniques. The main results for the RBST problem are also derived in this section. In Sec. V, we go beyond the random trees and derive exact results for the fragmentation problem with arbitrary break point distribution. Section VI contains the generalization to the case of m-ary trees with arbitrary distributions. We finally conclude in Sec. VII with a summary and outlook.

II. BINARY SEARCH TREES: OLD AND NEW RESULTS

Let us label the incoming data string of N elements by integers $1, \ldots, N$. For example, if the data string consists of the 12 months of the year, we can label, say the month of January by 1, the month of February by 2, and so on. In that example, $N=12$. A specific data string will then be isomorphic to a corresponding ordered sequence of these integers. For example, the particular sequence of months in Fig. 1 reduces to the ordered sequence (7,9,12,5,4,2,1,10,11,3,6,8). A different sequence of months will correspond to a different permutation of these integers. Each such sequence or permutation will then correspond to a separate BST, constructed by the algorithm explained in the Introduction. In RBST, all these $N!$ sequences (and their corresponding trees) occur with equal probability.

We will focus on the statistics of the extreme variables associated with these trees, in particular, the height H_N and the balanced height h_N of a BST as defined in the Introduction. Each BST has a unique value of H_N and h_N. Since the trees occur with a given probability distribution (which is uniform in case of RBST), both H_N and h_N are random variables. Of interest are the statistics of these variables such as the average, variance, or even the full probability distributions of H_N and h_N.

The RBST problem has been studied for a long time by computer scientists and we now mention a few known results. Devroye [21] proved that for large N, the average height of a RBST $\langle H_N \rangle \approx \alpha_0 \ln N$ where the constant $\alpha_0 = 4.311\,07\ldots$. Hattori and Ochiai conjectured that the true asymptotic behavior of $\langle H_N \rangle$ has an additional subleading double logarithmic correction,

$$\langle H_N \rangle \approx \alpha_0 \ln N + \alpha_1 \ln(\ln N), \tag{1}$$

and they determined the constant $\alpha_1 \approx -1.75$ numerically [32]. Using traveling front techniques we confirmed the above asymptotics and computed the correction term analytically, $\alpha_1 = -3\alpha_0/[2(\alpha_0-1)] = -1.953\,03\ldots$ [33]. The same result was simultaneously proved by Reed [35]. Based on numerical data, Robson conjectured [36] that the variance is bounded. Recently, Drmota [37] has proved that all moments $\langle (H_N - \langle H_N \rangle)^m \rangle$ are bounded.

For the balanced height h_N of RBST, Devroye showed that the leading asymptotic behavior of the average balanced height is given by $\langle h_N \rangle \approx \alpha_0' \ln N$ where $\alpha_0' = 0.3733\ldots$ [21,27]. Indeed, α_0 in $\langle H_N \rangle$ and α_0' in $\langle h_N \rangle$ turn out to be the two solutions of the same transcendental equation $(2e/\alpha)^\alpha = e$ [21,27]. This suggests some kind of duality between the height and the balanced height. We will show later that the correct asymptotic behavior of $\langle h_N \rangle$ is given by

$$\langle h_N \rangle \approx \alpha_0' \ln N + \alpha_1' \ln(\ln N), \tag{2}$$

where relation $\alpha_1' = -3\alpha_0'/[2(\alpha_0'-1)]$ holds again. Drmota has recently proved that all the moments of h_N are also bounded [37] as in the case of H_N.

Note that all the results mentioned above are for RBST with fixed size N. Recently by using a rate equation approach we studied the statistics of height and balanced height for randomly growing binary trees where the average size of a tree grows with time linearly $\langle N(t) \rangle \sim t$ [38]. The expected height and balanced height for large random binary trees were found to have exactly the same asymptotic formulas (1) and (2), provided one replaces N by $\langle N(t) \rangle$ in these equations. This approach is thus reminiscent of the grand canonical approach in statistical mechanics with the time t playing the role of the chemical potential that can be chosen to fix the average size. In this paper, we focus only on the canoni-

cal approach, i.e., trees with fixed given size N, since this is more familiar in theoretical computer science.

We will exploit a two stage mapping ''BST problem $\rightarrow$ fragmentation problem $\rightarrow$ MDP problem'' and use the traveling front technique to analyze the MDP problem. This technique allows to rederive in a physically transparent way all results for the RBST mentioned above and provides a lot of important results. We will show that the constants α_0 and $\alpha_0{}'$ are simply related to the velocities of traveling fronts. The subleading correction terms can also be derived analytically. The traveling front approach also predicts ''concentration of measure'' of the variables H_N and h_N. This means that the asymptotic probability distributions of these variables are highly localized around their respective averages. As a result, a typical value of $H_N \sim \langle H_N \rangle$ and the spread in H_N is of order $O(1)$ in the large N limit. Naturally the variance and higher cumulants of both H_N and h_N are bounded. We also derive an asymptotically exact nonlinear integral equation for the full probability distributions of H_N and h_N. While we could not solve this nonlinear equation in closed form, we could derive the behaviors at the tails of these highly localized distribution functions. We will also see that within this approach the variables H_N and h_N map, respectively, onto the minimum and maximum energy of a directed polymer and hence the observed duality between them is rather natural.

The main advantage of the present approach is that it allows us to go beyond the random trees and obtain exact asymptotic results for the statistics of H_N and h_N for BST's with arbitrary distributions. This is the main result of the present paper. Besides, we also generalize basic results to m-ary search trees with arbitrary distributions.

III. MAPPING OF THE BST PROBLEM TO A FRAGMENTATION PROBLEM

In order to derive the asymptotics of the statistics of the height and the balanced height in the BST problem, it is convenient to first map this problem to a fragmentation problem following Devroye [21,27]. To illustrate how this mapping works, let us consider again the example in Fig. 1 where the months (or the corresponding integers from 1 to 12) appear in the particular sequence (7,9,12,5,4,2,1, 10,11,3,6,8). The first element (which in this example is 7) is chosen randomly from the available $N=12$ elements in the case of RBST. Once this element is chosen, the remaining elements will belong either to the interval $(1-6)$ or $(8-12)$, which are subsequently completely disconnected from each other. Thus choosing the first element is equivalent to breaking the original interval $(1-12)$ into two intervals, the left $(1-6)$ and the right $(8-12)$ at the break point 7 which is chosen randomly. Now consider the next element. It will either belong to the left or the right interval. In the particular example we are discussing, the next element 9 belongs to the right interval $(8-12)$. This new element then divides the right interval $(8-12)$ again into two parts: the left containing only the single element (8) and the right $(10-12)$. These two new intervals subsequently become completely independent of each other. The third element (12) breaks subsequently the interval $(10-12)$ into two parts: the left part $(10-11)$ and the right part which is empty. Similarly the fourth element (5) breaks the interval $(1-6)$ into two parts: the left part $(1-4)$ and the right part (6) and so on.

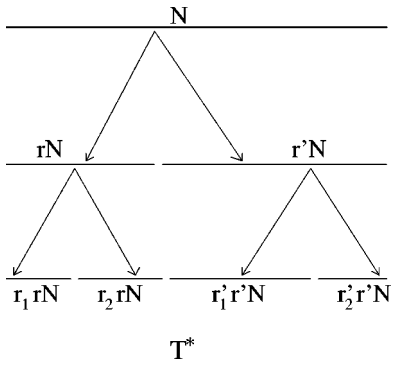

FIG. 3. The fragmentation process has itself a tree structure (denoted by T^*), shown here up to level 2. In the first step an interval of length N is broken into two pieces of lengths rN and $r'N$ such that $r+r'=1$. Each of those pieces is further broken into two halves satisfying the constraints $r_1+r_2=1$ and $r_1'+r_2'=1$. At level n, there will be 2^n pieces.

Thus, one can think of the construction of the RBST as a dynamical fragmentation process where one starts with a stick of initial length N and breaks it randomly into two parts: a left part of length rN and a right part of length $r'N$ with the constraint $r+r'=1$, where r is a random number distributed uniformly over the interval $[0-1]$. At the next step, one breaks each of these intervals again into two parts. At any stage of breaking, the random variable r characterizing the break point of an interval is chosen independently from interval to interval. They are also independent from stage to stage. After n steps of breaking, there are 2^n intervals. Note that this fragmentation process has itself a tree structure and can be represented by a branching process as depicted in Fig. 3.

A search tree of fixed size N is completed when in the corresponding fragmentation process, the lengths of all intervals are less than 1 because this means that all the elements of the incoming data string have already been incorporated onto the search tree. Although in the fragmentation problem we have continuous intervals whereas in the RBST the intervals consist of discrete integers, it does not really matter since one can associate the integer part of a break point to a particular integer element of the RBST. For example, if the first break point in the fragmentation problem is 7.3, this means that in the RBST problem, the first element (the root) is integer 7.

Let us first consider the height H_N of the RBST. By definition, H_N is the distance from the root (depth) of the farthest element in the RBST. The RBST stops growing beyond H_N as all the incoming N elements have been incorporated in the tree. Thus when the RBST attains the depth H_N, in the corresponding fragmentation problem, the length of every interval is less than 1. Denote by $l_1, \ldots, l_{2^n}$ the lengths of 2^n intervals after n steps of breaking. Clearly, the probability $\mathrm{Prob}[H_N < n]$ in the RBST problem is the same as the prob-

ability that all 2^n intervals in the fragmentation problem have lengths less than 1,

$$\text{Prob}[H_N < n] = \text{Prob}[l_1 < 1, \ldots, l_{2^n} < 1]. \tag{3}$$

The right-hand side of Eq. (3) is also the probability that the maximum of the lengths of the 2^n pieces is less than 1 in the fragmentation problem.

We next consider the balanced height h_N of the RBST. By definition, h_N is the depth up to which the RBST is fully saturated and balanced. Beyond this depth, some parts of the RBST stop growing (see Fig. 1 where $h_N=3$). This means that in the corresponding random fragmentation process, as long as the step number of breaking is less than h_N, the lengths of all the intervals must still be bigger than 1, so that each such interval can incorporate a new element. Thus the probability $\text{Prob}[h_N > n]$ in the RBST is the same as the probability that all 2^n intervals in the fragmentation problem have lengths bigger than 1,

$$\text{Prob}[h_N > n] = \text{Prob}[l_1 > 1, \ldots, l_{2^n} > 1]. \tag{4}$$

The right-hand side of Eq. (4) is also the probability that the minimum of the lengths of the 2^n pieces is bigger than 1 in the fragmentation problem.

In the RBST, the new elements in the tree arrive randomly. The corresponding fragmentation problem is also random in the sense that at each stage an interval l is broken into two parts of lengths rl and $r'l$ with $r+r'=1$ where the random variable r is chosen each time independently and is distributed uniformly over $[0-1]$. One can, of course, generalize this random fragmentation problem where the variable r is chosen independently each time but with an arbitrary distribution over $[0-1]$, not necessarily uniform. This would correspond to a BST problem where the new elements arrive with a specified distribution. In general, at any stage of breaking, the joint probability distribution of r and r' can be written as

$$\text{Prob}[r,r'] = \phi(r)\,\phi(r')\,\delta(r+r'-1). \tag{5}$$

The delta function ensures that the total length is conserved at every stage of breaking. The joint distribution is written in a symmetric way to ensure that both r and r' have the same effective distribution that is given by $\eta(r) = \text{Prob}(r) = \int_0^1 \text{Prob}[r,r']dr' = \phi(r)\,\phi(1-r)$. The function $\phi(r)$ must be chosen such that the induced distribution $\eta(r)$ satisfies the conditions, $\int_0^1 \eta(r)dr = 1$ and $\int_0^1 r\,\eta(r)dr = 1/2$. The first condition ensures normalizability of the single point distribution $\eta(r)$ and the second condition comes from the strict constraint $r+r'=1$ that indicates $\langle r\rangle = \langle r'\rangle = 1/2$. In the case of random breaking, the function $\phi(r)=1$ and consequently the induced distribution $\eta(r)=1$ for $0 \leqslant r \leqslant 1$. A simple example of a nonrandom break point distribution is given by, $\phi(r)=\sqrt{6}r$ with the induced distribution $\eta(r) = 6r(1-r)$ that satisfies the two constraints [33].

Apart from connection to the BST problem, the random fragmentation problem is interesting on its own rights as it arises in various contexts such as the energy cascades in turbulence [28], rapture processes in earthquakes [29], financial crashes in stock markets [30], and in the stress propagation in granular medium [31]. In our previous paper [33], we had studied the asymptotic laws governing the probability distribution of the maximal lengths of the intervals after n steps of breaking in the random fragmentation problem using traveling front techniques. The same differential equation that describes the Laplace transform of this distribution was also studied independently by Drmota via a different method [37]. Both these methods work well for the random problem [where $\eta(r)=1$] but seem difficult to extend to the general case when the break point in the fragmentation process is chosen from an arbitrary induced distribution $\eta(r)$ [33]. It turns out, however, that the fragmentation problem with general $\eta(r)$ can further be mapped to a MDP problem as presented in the next section. This further mapping followed by the traveling front analysis then allows us to obtain exact asymptotic results for the general case with arbitrary $\eta(r)$.

IV. MAPPING OF THE FRAGMENTATION PROBLEM TO A MODIFIED DIRECTED POLYMER PROBLEM

In this section, we further map the fragmentation problem onto a MDP problem on a Cayley tree. This MDP problem turns out to be slightly different from the conventional DP problem studied in statistical mechanics due to the presence of a special constraint. Nevertheless, asymptotic properties in the MDP problem can be derived analytically using the traveling front techniques.

To understand this mapping, consider the set of 2^n intervals in the fragmentation problem after n steps of breaking, starting from the initial length N. Let l_k denote the length of the kth interval where $k=1, \ldots, 2^n$. From Fig. 3, it is clear that the length of any typical piece l_k can be expressed as the product

$$l_k = N \prod_{i=1}^{n} r_i, \tag{6}$$

where r_i's are the set of independent random variables encountered in getting the final piece of length l_k after n steps of breaking the original interval of length N. Note that in the tree T^* in Fig. 3, there is a unique path connecting the original interval (the root element of T^*) to the kth interval at stage n and the set of random variables r_i's encountered in going from the root of T^* to the kth piece at stage n defines this unique path. Alternately, we can associate an energy variable $\epsilon_i = -\ln r_i \geqslant 0$ to the bonds connecting this path and the set of energies ϵ_i's also uniquely characterize the path (see Fig. 4). Taking logarithm in Eq. (6), we see that the total energy E_k of a path (starting at the root and ending at the k interval at the stage n) becomes

$$E_k = \ln\left(\frac{N}{l_k}\right) = -\sum_{i=1}^{n} \ln r_i = \sum_{i=1}^{n} \epsilon_i. \tag{7}$$

This path then represents a typical configuration of a directed polymer (directed in the downward direction) with energy given by Eq. (7) where ϵ_i's are random bond energies. Note

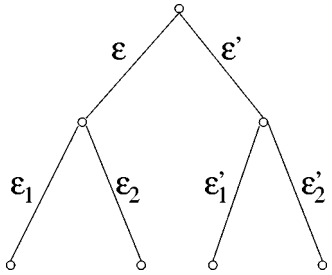

FIG. 4. The MDP on a Cayley tree. This tree is isomorphic to the tree of the fragmentation process shown in Fig. 3. Each bond energy ϵ is related to the corresponding fraction r via $\epsilon=-\ln r$. The bond energies are correlated due to the constraints $e^{-\epsilon}+e^{-\epsilon'}=1$, $e^{-\epsilon_1}+e^{-\epsilon_2}=1$, $e^{-\epsilon_1'}+e^{-\epsilon_2'}=1$, etc.

that up to level n, there are a total number of 2^n different paths each having different total energies $E_1, \ldots, E_{2^n}$.

In the conventional DP problem, the bond energies ϵ_i's are completely uncorrelated. To understand why they are correlated in the present problem, recall that when an interval is broken into two parts the random variables r and r' characterizing the lengths of the two daughter intervals satisfy the length conservation constraint, $r+r'=1$. Translated into the DP problem, the corresponding bond energies $\epsilon=-\ln r$ and $\epsilon'=-\ln r'$ associated with the two bonds emanating downwards from a given node must satisfy the constraint

$$e^{-\epsilon}+e^{-\epsilon'}=1. \tag{8}$$

This constraint holds at every branching point of the tree (see Fig. 4). This correlation makes the MDP problem slightly different from the conventional DP problem.

The joint distribution $p(\epsilon,\epsilon')$ of the energies of the two bonds emanating from the common node and the induced effective single bond distribution $\rho(\epsilon)$ are obtained from Eq. (5) to give

$$p(\epsilon,\epsilon')=\phi(e^{-\epsilon})\phi(e^{-\epsilon'})e^{-\epsilon-\epsilon'}\delta(e^{-\epsilon}+e^{-\epsilon'}-1),$$

$$\rho(\epsilon)\equiv\int_0^\infty p(\epsilon,\epsilon')d\epsilon'=\phi(e^{-\epsilon})\phi(1-e^{-\epsilon})e^{-\epsilon}. \tag{9}$$

For example, for the RBST we have $\phi(r)=1$, and therefore $\rho(\epsilon)=e^{-\epsilon}$. Note that in the conventional DP problem, the joint distribution $p(\epsilon,\epsilon')$ would simply be the product of the single point distributions, $p(\epsilon,\epsilon')=\rho(\epsilon)\rho(\epsilon')$ since they are independent. The MDP problem, however, lacks this factorization property.

Having set up the notation we turn to the variables H_N and h_N in the original BST problem. What do the distributions of H_N and h_N correspond to in the MDP problem? First, consider the height distribution $\mathrm{Prob}[H_N<n]$. From Eqs. (3) and (7), one finds

$$\mathrm{Prob}[H_N<n]=\mathrm{Prob}[l_1<1,\ldots,l_{2^n}<1]$$
$$=\mathrm{Prob}[E_1>\ln N,\ldots,E_{2^n}>\ln N], \tag{10}$$

where E_k's ($k=1,\ \ldots,\ 2^n$) are respectively the total energies of the all possible 2^n paths going from the root to the leaves at the nth level in the DP problem. The probability in the last line in Eq. (10) is also the same as the probability $\mathrm{Prob}[\min\{E_1,\ldots,E_{2^n}\}>\ln N]$. Thus the height distribution $\mathrm{Prob}[H_N<n]$ in the BST problem is precisely related to the distribution of the minimum (ground state) energy of the MDP problem, a quantity of considerable interest in statistical physics.

Let us next consider the balanced height h_N. Using Eqs. (4) and (7), it follows similarly that

$$\mathrm{Prob}[h_N>n]=\mathrm{Prob}[l_1>1,\ldots,l_{2^n}>1]$$
$$=\mathrm{Prob}[E_1<\ln N,\ldots,E_{2^n}<\ln N], \tag{11}$$

which is also the probability that the maximum energy $\max\{E_1,\ldots,E_{2^n}\}$ is less than $\ln N$. Thus the balanced height distribution $\mathrm{Prob}[h_N>n]$ in the BST problem is related to the distribution of the maximum energy in the MDP problem, a quantity that is usually not of much interest in statistical mechanics.

A. Statistics of the height or the minimum energy

In this subsection we analyze the asymptotic statistics of the height H_N in the BST problem or equivalently that of the minimum energy in the MDP problem. Let $P_n(x)=\mathrm{Prob}[\min\{E_1,\ldots,E_{2^n}\}>x]$, where E_k's with $k=1,\ldots,2^n$ are the energies of the 2^n polymer paths from the root to the nth level. It is then easy to write a recursion relation for $P_n(x)$,

$$P_{n+1}(x)=\int_0^\infty\int_0^\infty P_n(x-\epsilon)P_n(x-\epsilon')p(\epsilon,\epsilon')d\epsilon\,d\epsilon', \tag{12}$$

where $p(\epsilon,\epsilon')$ is the joint distribution of the two bond energies as given by Eq. (9). Equation (12) has been derived by analyzing different possibilities for the energies of the bonds emanating from the root and using the fact that the two subsequent daughter trees are statistically independent. Note that in the conventional DP problem, the corresponding recursion relation would be simplified using the factorization property of the joint distribution $p(\epsilon,\epsilon')$ and one would get [40,41]

$$P_{n+1}(x)=\left[\int_0^\infty P_n(x-\epsilon)\rho(\epsilon)d\epsilon\right]^2. \tag{13}$$

We have to solve the recursion relation (12) subject to the initial condition

$$P_0(x)=\begin{cases}1, & x\leq 0,\\ 0, & x>0,\end{cases} \tag{14}$$

and the boundary conditions

$$P_n(x)\to\begin{cases}1, & x\to-\infty,\\ 0, & x\to\infty.\end{cases} \tag{15}$$

The recursion relation (12) is nonlinear and in general difficult to solve exactly. However, its asymptotic properties can be derived analytically. As n increases, the solution $P_n(x)$ in Eq. (12) looks like a $(1-0)$ front (i.e., $P_n(x)\sim 1$ for small x but falls off rapidly to 0 for large x) advancing in the positive direction. This suggests that for large n, Eq. (12) admits a traveling front solution, $P_n(x)=F(x-x_n)$ where x_n denotes the location of the front and the shape of the front is described by the fixed point scaling function F that becomes independent of n. This implies that the width of the front is of order $O(1)$, i.e., it saturates in the large n limit. The traveling front ansatz also indicates that the front advances with a uniform velocity, i.e., $x_n \approx v n$, to leading order for large n where the velocity v is yet to be determined. Substituting this traveling front ansatz, $P_n(x)=F(x-vn)$ for large n in Eq. (12), we find that the fixed point function $F(y)$ satisfies the nonlinear integral equation,

$$F(y-v)=\int_0^\infty \int_0^\infty F(y-\epsilon)F(y-\epsilon')p(\epsilon,\epsilon')d\epsilon\, d\epsilon', \tag{16}$$

where the velocity v is still undetermined and $F(y)$ satisfies the boundary conditions

$$F(y)\to\begin{cases}1 & \text{as } y\to-\infty,\\ 0 & \text{as } y\to\infty.\end{cases} \tag{17}$$

Let us first analyze Eq. (16) in the tail region $y\to-\infty$. Plugging $F(y)=1-f(y)$ in Eq. (16) and neglecting terms of order $O(f^2)$ we find that $f(y)$ satisfies

$$f(y-v)=2\int_0^\infty f(y-\epsilon)\rho(\epsilon)d\epsilon, \tag{18}$$

where we have used the relation $\rho(\epsilon)=\int_0^\infty p(\epsilon,\epsilon')d\epsilon'$. This linear equation (18) clearly admits an exponential solution $f(y)=\exp(\lambda y)$ provided the inverse decay rate λ is related to the velocity v via the dispersion relation

$$v(\lambda)=-\frac{1}{\lambda}\ln\left[2\int_0^\infty e^{-\lambda\epsilon}\rho(\epsilon)d\epsilon\right]. \tag{19}$$

For a given induced distribution $\rho(\epsilon)$, the function $v(\lambda)\to -\ln(2)/\lambda$ as $\lambda\to 0$ and $v(\lambda)\to 0$ as $\lambda\to\infty$ with a single maximum at a finite λ^* determined via

$$\left.\frac{dv}{d\lambda}\right|_{\lambda^*}=0. \tag{20}$$

Thus for all λ such that $\int_0^\infty e^{-\lambda\epsilon}\rho(\epsilon)d\epsilon<1/2$, the corresponding velocity $v(\lambda)>0$. While any such λ is in principle allowed, a particular velocity is actually asymptotically selected by the front. This velocity selection mechanism has been observed in a large class of nonlinear problems with a traveling front solution [12–18,33,39–41]. It is known that as long as the initial condition is sharp [as in the present case in Eq. (14)], the extreme value is chosen. From this general front selection principle, we infer that in our present problem, the front finally selects the velocity $v(\lambda^*)$ where λ^* is given by the solution of Eq. (20). Thus the asymptotic front position, to leading order for large n, is given by

$$x_n\approx v(\lambda^*)n. \tag{21}$$

While the leading behavior of the front position x_n is given exactly by Eq. (21), it turns out that it has an associated slow logarithmic correction. This logarithmic correction to the front velocity was first derived by Bramson in the context of a reaction-diffusion equation [14], and was subsequently found in many other systems with a traveling front [17,18,33,39,40]. In Appendix A, we present a detailed derivation of this correction term following the approach of Brunet and Derrida [39]. The main result of this exercise is that the asymptotic front position for large n is given by

$$x_n\approx v(\lambda^*)n+\frac{3}{2\lambda^*}\ln n. \tag{22}$$

One can even calculate the next correction term by employing a more sophisticated approach [18] but we omit these results here. One important point to note is that while the velocity $v(\lambda^*)$ and λ^* are nonuniversal as they depend explicitly on the distribution $\rho(\epsilon)$, the prefactor 3/2 of the logarithmic correction in Eq. (22) is actually universal and is precisely the first excited state energy of a quantum harmonic oscillator (see Appendix A).

Let us now translate back these results to see what they mean for the height distribution in the original BST problem. From Eq. (10), it is clear that the cumulative height distribution for large n is given by

$$\text{Prob}[H_N<n]=P_n(\ln N)\approx F(\ln N-x_n), \tag{23}$$

where the front position x_n is given by Eq. (22) and the function $F(y)$ is given by the solution of Eq. (16). Since the function $F(y)$ has the shape of a front with center at $y=0$ and width of order $O(1)$, its derivative $F'(y)$ is a localized function around $y=0$ with width of order $O(1)$. From Eq. (23) it then follows that the height distribution $\text{Prob}[H_N=n]$ is also localized around its average value $\langle H_N\rangle$ with a variance $V(H_N)\sim O(1)$. Thus H_N has a concentration of measure around its average value $\langle H_N\rangle$ that is given by the value of n that corresponds to the zero of the argument of the function $F(y)$, i.e., when $x_n=\ln N$. Using $x_n=\ln N$ in Eq. (22) and solving for the required value of n for large N, we obtain one of our main results

$$\langle H_N\rangle=\frac{1}{v(\lambda^*)}\ln N-\frac{3}{2\lambda^* v(\lambda^*)}\ln(\ln N), \tag{24}$$

where $v(\lambda)$ and λ^* are given respectively by Eqs. (19) and (20). This is the first result for the fragmentation problem with arbitrary break-point distribution going beyond the uniform case or equivalently for the BST problem where the elements in the tree arrive with an arbitrary distribution and not just randomly.

It is useful to exemplify the above general results. For the original RBST problem, $\phi(r)=1$ or $\rho(\epsilon)=e^{-\epsilon}$ [see Eq. (9)]. Substituting $\rho(\epsilon)=e^{-\epsilon}$ into Eq. (19) we get

$$v(\lambda)=-\frac{1}{\lambda}\ln\left[\frac{2}{\lambda+1}\right], \tag{25}$$

which has a single maximum at $\lambda^*=3.311\,07\ldots$ with $v(\lambda^*)=0.231\,96\ldots$. Substituting λ^* and $v(\lambda^*)$ into Eq. (24) we arrive at Eq. (1) with $\alpha_0=4.311\,07\ldots$ and $\alpha_1=-1.953\,02\ldots$, in agreement with Refs. [33,35].

Consider another example, $\phi(r)=\sqrt{6}r$, a problem that couldn't be solved by the techniques used in our previous short paper [33]. This corresponds to the fragmentation problem where the induced distribution of the break point is $\eta(r)=6r(1-r)$. In the MDP problem, it corresponds to the induced bond energy distribution

$$\rho(\epsilon)=6e^{-2\epsilon}(1-e^{-\epsilon}). \tag{26}$$

Substituting this form in Eq. (19), we get

$$v(\lambda)=-\frac{1}{\lambda}\ln\left[\frac{12}{(\lambda+2)(\lambda+3)}\right], \tag{27}$$

which has a unique maximum at $\lambda^*=3.924\,08\ldots$ and this maximum velocity is given by $v(\lambda^*)=0.313\,22\ldots$. Substituting these results into the general formula (24) we recover Eq. (1) with $\alpha_0=3.192\,58\ldots$ and $\alpha_1=-1.220\,38\ldots$.

The traveling front approach also gives the full probability distribution of the height variable in the BST and not just its exact average value as in Eq. (24). Indeed, we have seen that cumulative height distribution is given by Eq. (23) where the function $F(y)$ is the solution of the boundary value problem (16),(17). While we have not been able to solve the nonlinear integral equation (16) exactly, one can easily deduce the extreme behavior of $F(y)$. We have already seen that in the tail region $y\to-\infty$, the function $F(y)$ saturates to 1 exponentially fast, $1-F(y)\sim\exp[\lambda^* y]$, where λ^* is the solution of Eq. (20). One can also deduce the asymptotic behavior of $F(y)$ when $y\to\infty$ (Appendix B) for arbitrary distribution $\rho(\epsilon)$. Thus the asymptotic behaviors of the function $F(y)$ read

$$F(y)\approx\begin{cases}1-Ae^{\lambda^* y}, & y\to-\infty,\\ 2\displaystyle\int_y^\infty \rho(y'+v)\,dy', & y\to\infty,\end{cases} \tag{28}$$

where A is a constant, λ^* is found from Eq. (20), and $v=v(\lambda^*)$. In particular, for the RBST where $\rho(\epsilon)=e^{-\epsilon}$, $\lambda^*=3.311\,07$, and $v(\lambda^*)=0.231\,96$, one has

$$F(y)\approx\begin{cases}1-Ae^{3.311\,07y}, & y\to-\infty,\\ 1.585\,96e^{-y}, & y\to\infty.\end{cases} \tag{29}$$

In conclusion, the height distribution is a localized function around its average value $\langle H_N\rangle$ given by Eq. (24). For any unbounded distribution $\rho(\epsilon)$, the height distribution decays at in the tail regions according to Eq. (28). For bounded distributions, however, $F(y)$ vanishes for sufficiently large y. Recall that the distribution of the minimum of a set of uncorrelated random variables is known to have a universal superexponential decay for large value [23]. However, it was shown in Ref. [41] that in the conventional DP problem the distribution of the minimum energy of a polymer violates this Gumbel law due to hierarchical correlations between the energies of different paths. From Eq. (28) it is clear that in the MDP problem the forward tail is nonuniversal since it depends explicitly on the distribution $\rho(\epsilon)$. Generally, the forward tail is not superexponential thus clearly violating the Gumbel statistics.

B. Statistics of the balanced height or the maximum energy

The analysis of the statistics of balanced height $\langle h_N\rangle$ follows more or less the same approach as in the case of height variable, except that one is now concerned with the distribution of maximum energy in the MDP problem. Let $R_n(x)=\mathrm{Prob}[\max\{E_1,E_2,\ldots,E_{2^n}\}<x]$ where E_k's with $k=1,2,\ldots,2^n$ are the energies of the 2^n polymer paths from the root to the nth level. Then $R_n(x)$ satisfies the same recursion relation as the $P_n(x)$ in Eq. (12),

$$R_{n+1}(x)=\int_0^\infty\int_0^\infty R_n(x-\epsilon)R_n(x-\epsilon')p(\epsilon,\epsilon')\,d\epsilon\,d\epsilon'. \tag{30}$$

The only difference is in the initial condition,

$$R_0(x)=\begin{cases}0, & x\leqslant 0,\\ 1, & x>0,\end{cases} \tag{31}$$

and in the boundary conditions,

$$R_n(x)\to\begin{cases}0, & x\to-\infty,\\ 1, & x\to\infty.\end{cases} \tag{32}$$

As in the case of Eq. (12), the recursion relation (30) admits a traveling front solution for large n, $R_n(x)=G(x-x_n^*)$ where x_n^* is the front position and the fixed point scaling function $G(x)$ describes the shape of the front. Unlike the $[1-0]$ front in the previous subsection, the front for $R_n(x)$ has a $[0-1]$ form advancing in the positive direction. The front again advances with asymptotically constant velocity v_1, i.e., the position of the front is $x_n^*\approx v_1 n$. Substituting $R_n(x)=G(x-v_1 n)$ in Eq. (30), we find that $G(y)$ satisfies the nonlinear integral equation

$$G(y-v_1)=\int_0^\infty\int_0^\infty G(y-\epsilon)G(y-\epsilon')p(\epsilon,\epsilon')\,d\epsilon\,d\epsilon'. \tag{33}$$

The velocity v_1 is still undetermined and the front shape $G(y)$ satisfies the boundary conditions, $G(y)\to 0$ as $y\to-\infty$ and $G(y)\to 1$ for $y\to\infty$. As in the previous subsection, we will analyze the Eq. (33) in the tail where $G(y)\to 1$; in

the present case, this means $y \to \infty$. Substituting $G(y)=1-g(y)$ in Eq. (33) and neglecting terms of order $O(g^2)$ we get the linear equation

$$g(y-v_1)=2\int_0^\infty g(y-\epsilon)\rho(\epsilon)d\epsilon. \tag{34}$$

Equation (34) admits asymptotically exponential solution, $g(y)=\exp(-\mu y)$ as $y \to \infty$, with

$$v_1(\mu)=\frac{1}{\mu}\ln\left[2\int_0^\infty e^{\mu\epsilon}\rho(\epsilon)d\epsilon\right]. \tag{35}$$

The dispersion relation in Eq. (35) has a single minimum at $\mu=\mu^*$ determined from relation

$$\left.\frac{dv_1}{d\mu}\right|_{\mu^*}=0. \tag{36}$$

By the general front selection mechanism, we infer that this minimum velocity will be selected by the front

$$x_n^* \approx v_1(\mu^*)n. \tag{37}$$

The associated slow logarithmic correction can also be worked out following the same calculation as in Appendix A and we finally get

$$x_n^* \approx v_1(\mu^*)n-\frac{3}{2\mu^*}\ln n. \tag{38}$$

Note that the correction term in Eq. (38) has a negative sign compared to the positive sign in Eq. (22).

In terms of the BST problem, it is clear from Eq. (11) that the cumulative balanced height distribution for large n is given by

$$\text{Prob}[h_N>n]=R_n(\ln N)\approx G(\ln N-x_n^*), \tag{39}$$

where the front position x_n^* is given by Eq. (38) and the function $G(y)$ is the solution of Eq. (33). As argued in the previous subsection, the derivative $G'(y)$ is a localized function around $y=0$ with width of order $O(1)$. Thus the balanced height distribution $\text{Prob}[h_N=n]$ is also localized around its average value $\langle h_N\rangle$ with a variance $V(h_N) \sim O(1)$. The average value reads

$$\langle h_N\rangle=\frac{1}{v_1(\mu^*)}\ln N+\frac{3}{2\mu^* v_1(\mu^*)}\ln(\ln N). \tag{40}$$

Consider again the same examples that were studied for the height variable in the previous subsection. For the RBST problem where $\phi(r)=1$ or equivalently $\rho(\epsilon)=e^{-\epsilon}$, Eq. (35) becomes

$$v_1(\mu)=\frac{1}{\mu}\ln\left[\frac{2}{1-\mu}\right], \tag{41}$$

which has a single minimum at $\mu^*=0.626\,63\ldots$ where $v_1(\mu^*)=2.678\,34\ldots$. Thus, Eq. (40) reduces to Eq. (2) with $\alpha_0'=0.373\,36\ldots$ and $\alpha_1'=0.893\,74\ldots$.

For the second example, $\phi(r)=\sqrt{6}r$ or equivalently $\rho(\epsilon)$ given by Eq. (26), Eq. (35) becomes

$$v_1(\mu)=\frac{1}{\mu}\ln\left[\frac{12}{(2-\mu)(3-\mu)}\right]. \tag{42}$$

The function $v_1(\mu)$ in Eq. (42) has a single minimum at $\mu^*=1.178\,64\ldots$ where $v_1(\mu^*)=1.766\,53\ldots$. Equation (40) again reduces to Eq. (2) with $\alpha_0'=0.566\,07\ldots$ and $\alpha_1'=0.720\,41\ldots$.

Finally we explain the duality between H_N and h_N in the BST problem. In the language of the MDP problem, these variables correspond to the minimum and maximum energy of a directed polymer in a random medium where the bond energies ϵ_i's have nonzero support only for $\epsilon_i \geqslant 0$. Changing the sign of the bond energies maps the minimum energy in the negative support problem into the negative of the maximum energy in the positive support problem. This fact is reflected in the relation between the two dispersion relations in Eqs. (19) and (35), $v(-\lambda)=v_1(\lambda)$. Thus λ^* and $-\mu^*$ are actually the two different roots of the same transcendental equation (20). Consequently, the constants α_0 and α_0' in Eqs. (1),(2) are merely two different roots of the same transcendental equation.

V. GENERALIZATION TO *m*-ARY SEARCH TREES WITH ARBITRARY DISTRIBUTIONS

The results obtained in the previous sections for the statistics of H_N and h_N of the BST's with arbitrary distributions can be generalized in a straightforward manner to the m-ary search trees. An m-ary search tree is constructed in the following way. One first collects the first $(m-1)$ elements of the incoming data string and arranges them together in the root of the tree in an ordered sequence $x_1<\ldots<x_{m-1}$. Next when the mth element x_m comes, one compares first with x_1. If $x_m<x_1$, the mth element is assigned to the root of the leftmost daughter tree. If $x_1<x_m<x_2$, then x_m goes to form the root of the second branch and so on. Each subsequent incoming element is assigned to either of the m branches according to the above rule. Note that the level of the tree will increase beyond a given node only when the node gets filled beyond its capacity of $(m-1)$ elements. Thus in the m-ary search tree, each node will contain at the most $(m-1)$ elements.

The mapping to the fragmentation problem goes through following the same line of arguments used for the binary tree in Sec. III. In this case, one starts with an interval of size N and breaks it into m pieces. Subsequently each piece is further broken into m pieces and so on. When an interval is broken into m pieces, each of the new pieces is a fraction of the original piece. The lengths of these m new pieces are characterized by a set of m random numbers $\{r_1,\ldots,r_m\}$ such that $\Sigma_{i=1}^m r_i=1$ thus enforcing the length conservation. For each interval a new set of r_i's are chosen from the same joint probability distribution

$$\text{Prob}[r_1, \ldots, r_m] = \delta\left(\sum_{i=1}^{m} r_i - 1 \right) \prod_{j=1}^{m} \phi(r_j). \quad (43)$$

As in the binary case, the distribution (43) is written in a symmetric form. Note that each new piece has the same effective induced distribution $\eta(r)$ given by the integral $\int_0^1 dr_2 \ldots \int_0^1 dr_m \, \text{Prob}[r, r_2, \ldots, r_m]$, or

$$\eta(r) = \phi(r) \int_0^1 \ldots \int_0^1 \delta\left(\sum_{i=2}^{m} r_i + r - 1 \right) \prod_{i=2}^{m} \phi(r_i) dr_i .$$

The function $\phi(r)$ must be chosen such that $\eta(r)$ satisfies the conditions, $\int_0^1 \eta(r) dr = 1$ and $\int_0^1 r \eta(r) dr = 1/m$.

The random m-ary search tree corresponds to a random fragmentation problem where each of the fractions $r_1, \ldots, r_{m-1}$ is chosen from a uniform distribution between 0 and 1, setting $r_m = 1 - \Sigma_{i=1}^{m-1} r_i$, and then keeping only those sets where $r_m \geqslant 0$. This is precisely the so-called ''uniform'' distribution used by Coppersmith *et al.* in the context of the q-model of force fluctuations in granular media [31]. In this case, $\phi(r)$ is a constant chosen in such a way that the joint distribution (43) is normalized. One finds [31]

$$\text{Prob}[r_1, \ldots, r_m] = (m-1)! \, \delta\left(\sum_{i=1}^{m} r_i - 1 \right). \quad (44)$$

The corresponding effective single point distribution $\eta(r)$ reads [31]

$$\eta(r) = (m-1)(1-r)^{m-2}. \quad (45)$$

Another interesting distribution is $\phi(r) \propto r$. In this case, the normalized joint distribution is given by (see Appendix C)

$$\text{Prob}[r_1, \ldots, r_m] = \Gamma(2m) \, \delta\left(\sum_{i=1}^{m} r_i - 1 \right) \prod_{i=1}^{m} r_i . \quad (46)$$

The corresponding effective distribution $\eta(r)$ can be deduced by recursive method (as shown in Appendix C) and we get

$$\eta(r) = (2m-1)(2m-2) r (1-r)^{2m-3}. \quad (47)$$

Note that for $m = 2$, it reduces to $\eta(r) = 6r(1-r)$ which was studied in detail for the binary case in Sec. III.

The m-piece fragmentation problem for the special case of uniform distribution (44) was studied in Ref. [33]. However, as in the binary case, this method is not easy to extend to handle the general distribution $\eta(r)$ including, for example, the distribution (47). To go beyond the uniform case, we first map the fragmentation problem into the MDP problem as in the binary case. One proceeds exactly as in the binary case by associating an energy $\epsilon_i = -\ln r_i$ to each bond of a directed polymer going from the root to the leaves of a Cayley tree, but now with m daughters emerging from each node. The energies of the m bonds emanating downwards from any given node are correlated due to the relation $\Sigma_{i=1}^{m} r_i = 1$ which translates into the constraint

$$\sum_{i=1}^{n} e^{-\epsilon_i} = 1. \quad (48)$$

As in the binary case, this constraint holds at every branching point of the tree. The joint distribution $p(\epsilon_1, \ldots, \epsilon_m)$ is found from Eq. (43) to give

$$p(\epsilon_1, \ldots, \epsilon_m) = \delta\left(\sum_{i=1}^{m} e^{-\epsilon_i} - 1 \right) \prod_{i=1}^{m} e^{-\epsilon_i} \phi_i(e^{-\epsilon_i}). \quad (49)$$

Also the induced bond energy distribution $\rho(\epsilon)$ is related to the induced fraction distribution $\eta(r)$ via

$$\rho(\epsilon) = \int_0^\infty \ldots \int_0^\infty p(\epsilon, \epsilon_2, \ldots, \epsilon_m) d\epsilon_2 \ldots d\epsilon_m = \eta(e^{-\epsilon}) e^{-\epsilon}. \quad (50)$$

On this m-branch Cayley tree, there are a total of m^n possible paths of the directed polymer going from the root to the leaves at the nth level. Following arguments similar to the binary case, the cumulative height distribution in the m-ary search tree is related exactly to the distribution of the minimum energy of the m^n polymer paths in the MDP problem via

$$\text{Prob}[H_N < n] = \text{Prob}[l_1 < 1, \ldots, l_{m^n} < 1] = \text{Prob}[E_1 > \ln N, \ldots, E_{m^n} > \ln N], \quad (51)$$

where E_k's ($k = 1, 2, \ldots, m^n$) are respectively the total energies of the all possible m^n paths. Similarly the cumulative distribution of the balanced height is related to the distribution of the maximum energy of the polymer paths via

$$\text{Prob}[h_N > n] = \text{Prob}[l_1 > 1, \ldots, l_{m^n} > 1] = \text{Prob}[E_1 < \ln N, \ldots, E_{m^n} < \ln N]. \quad (52)$$

A. Statistics of the height

Let $P_n(x) = \text{Prob}[\min\{E_1, \ldots, E_{m^n}\} > x]$. This distribution satisfies the recursion relation

$$P_n(x) = \int_0^\infty \ldots \int_0^\infty p(\epsilon_1, \ldots, \epsilon_m) \prod_{i=1}^{m} P_{n-1}(x - \epsilon_i) d\epsilon_i , \quad (53)$$

where the joint distribution $p(\epsilon_1, \ldots, \epsilon_m)$ is given by Eq. (49). The recursion starts with the same initial condition as in Eq. (14). The rest of the analysis is exactly the same as in the binary case. Substituting a traveling front solution, $P_n(x) = F(x - vn)$ in Eq. (53) and then linearizing near the tail $y \to -\infty$, we find as in the binary case, $F(y) \sim 1 - e^{\lambda y}$ where the velocity v of the front is related to λ via the dispersion relation

$$v(\lambda) = -\frac{1}{\lambda} \ln\left[m \int_0^\infty e^{-\lambda \epsilon} \rho(\epsilon) d\epsilon \right], \quad (54)$$

where the induced distribution $\rho(\epsilon)$ is given by Eq. (50). The front velocity is then given by the maximum $v(\lambda^*)$ of the dispersion curve in Eq. (54) and is obtained by solving Eqs. (20) and (54). Similarly one can also work out the logarithmic correction to the front velocity and the asymptotic front position is given by the same formula in Eq. (22), only λ^* and $v(\lambda^*)$ are different from the binary case. Similarly the average height $\langle H_N \rangle$ for the m-ary search tree is also given by the same formula as in Eq. (24), only change is in the dispersion curve $v(\lambda)$.

Let us now present some specific results. For the uniform distribution, $\rho(\epsilon)=(m-1)[1-e^{-\epsilon}]^{m-2}e^{-\epsilon}$ as follows from Eqs. (45) and (50). Substituting this into the dispersion relation (54) yields

$$v(\lambda)=-\frac{1}{\lambda}\ln[m(m-1)B(\lambda+1,m-1)], \tag{55}$$

where $B(m,n)$ is the Beta function. For instance, for $m=3$ the velocity $v(\lambda)$ has a single maximum at $\lambda^*=3.489\,85\ldots$ with $v(\lambda^*)=0.404\,87\ldots$. Plugging these in the general formula (24) we again arrive at Eq. (1) with $\alpha_0=2.4698\ldots$ and $\alpha_1=-1.0616\ldots$.

Consider now the large m limit. Using asymptotic properties of the Beta function, one gets

$$\lambda^*\approx\ln m, \quad v(\lambda^*)\approx\ln(m/\lambda^*). \tag{56}$$

Therefore, when $m\to\infty$, the average height is given by Eq. (1) with

$$\alpha_0=\frac{1}{\ln(m/\lambda^*)}, \quad \alpha_1=-\frac{3}{2\lambda^*\ln(m/\lambda^*)}. \tag{57}$$

Similarly for the distribution (47), Eqs. (50) and (54) lead to the following dispersion relation:

$$v(\lambda)=-\frac{1}{\lambda}\ln[m(2m-1)(2m-2)B(\lambda+2,2m-2)].$$

For $m=3$, we get the maximum at $\lambda^*=4.178\,86\ldots$ with $v(\lambda^*)=0.532\,35\ldots$. The average height is given by Eq. (1) with $\alpha_0=1.878\,45\ldots$ and $\alpha_1=-0.674\,27\ldots$. The large m behavior turns out to be exactly the same as in the case of uniform distribution. One can work out the large m asymptotics for arbitrary distribution $\eta(r)$ (see Appendix D) and one gets the same asymptotics (56) as in the above examples. Therefore, the asymptotic behavior of $\langle H_N \rangle$ is universal (independent of the details of the distribution) in the large m limit.

B. Statistics of the balanced height

As in the binary case, we again utilize the distribution $R_n(x)=\mathrm{Prob}[\max\{E_1,E_2,\ldots,E_{m^n}\}<x]$. This distribution satisfies the recursion relation

$$R_n(x)=\int_0^\infty\ldots\int_0^\infty p(\epsilon_1,\ldots,\epsilon_m)\prod_{i=1}^{m}R_{n-1}(x-\epsilon_i)d\epsilon_i, \tag{58}$$

and the same initial and boundary conditions (31),(32) as in the binary case. Plugging a traveling front solution $R_n(x)=G(x-v_1 n)$ into Eq. (58) and linearizing in the tail region $y\to\infty$ according to $G(y)\approx 1-e^{-\mu y}$, we arrive at the dispersion relation

$$v_1(\mu)=\frac{1}{\mu}\ln\left[m\int_0^\infty e^{\mu\epsilon}\rho(\epsilon)d\epsilon\right], \tag{59}$$

where the induced distribution is given by Eq. (50). The front velocity is then selected by the minimum $v_1(\mu^*)$ of this dispersion relation. Proceeding as in the binary case, the asymptotic front position is given by the same general formula in Eq. (38), the only difference is that μ^* and $v_1(\mu^*)$ are different from the binary case. Finally the average balanced height $\langle h_N \rangle$ for the m-ary search trees is also given by the same general formula in Eq. (40), the only difference being the dispersion relation $v_1(\mu)$.

For the uniform distribution, Eq. (45), we reduce Eq. (59) to

$$v_1(\mu)=\frac{1}{\mu}\ln[m(m-1)B(1-\mu,m-1)]. \tag{60}$$

Equation (60) can also be obtained from Eq. (55) by changing the sign of $\lambda=-\mu$ as expected. For example, for $m=3$, the dispersion relation (60) has a unique minimum at $\mu^*=0.681\,89\ldots$ where $v_1(\mu^*)=3.902\,27\ldots$. Then the general formula (40) reduces to Eq. (2) with $\alpha_0'=0.256\,26\ldots$ and $\alpha_1'=0.563\,71\ldots$.

For the distribution (47), the dispersion relation reads

$$v_1(\mu)=\frac{1}{\mu}\ln[m(2m-1)(2m-2)B(2-\mu,2m-2)].$$

One has $\mu^*=1.286\,65\ldots$ and $v_1(\mu^*)=2.623\,34\ldots$ indicating in the particular case of $m=3$, so in this situation the averaged balanced height is given by Eq. (2) with $\alpha_0'=0.381\,19\ldots$ and $\alpha_1'=0.444\,40\ldots$.

One can also work out the large m behavior for arbitrary distribution $\eta(r)$ (Appendix D). Unlike the case of the height variable, the large m behavior in the case of balanced height is nonuniversal and depends explicitly on the small r behavior of the distribution $\eta(r)$. If $\eta(r)\sim r^a$ as $r\to 0$, then (see Appendix D) $\mu^*\approx a+1$ and $v_1(\mu^*)\approx(a+2)/(a+1)\ln m$. Both these quantities, and hence the average balanced height, depend on the parameter a. Therefore, the balanced height remains nonuniversal in the large m limit.

VI. CONCLUSIONS

In this paper we studied the statistics of height and balanced height in the BST problem by exploiting a two stage mapping "the BST problem → fragmentation problem → the MDP problem" and then using the traveling front techniques to solve the MDP problem. While the first mapping has been used previously to obtain exact asymptotic results for RBST problem, the second mapping allowed us to go beyond random trees and obtain exact asymptotic results for

BST's where the new entries arrive in the tree according to any arbitrary distribution, not necessarily randomly.

An interesting extension of the present work would be to see if the traveling front techniques can be applied to obtain the asymptotic statistics of observables that are not necessarily extreme. For example, in the context of the fragmentation problem, it would be interesting to compute the probability $P_n(k,N)$ that after n levels of fragmentation there will be k pieces (out of the total 2^n pieces) with lengths less than 1, given that the initial length is N. This probability interpolates between the two extreme limits $k=0$ and $k=2^n$. For $k=0$, this is the probability that all pieces have lengths bigger than 1 and hence is just the probability that the balanced height h_N is bigger than n [see Eq. (4)], as studied in this paper. On the other hand, for $k=2^n$, this is the probability that all pieces have lengths less than 1 that is precisely the cumulative distribution of the height variable H_N as in Eq. (3). It would be interesting to see if for any intermediate k ($0<k<2^n$) the probability $P_n(k,N)$ has a traveling wave structure as in the case of $k=0$ and $k=2^n$ and if so, how does the velocity v_k depend on k?

The fact that the traveling wave techniques, used previously in nonlinear physics, can be used successfully in computer science problems is not just interesting but it allows us to obtain the shape of the full distribution of height and not just its moments. It would be interesting to apply these techniques to more sophisticated search algorithms in computer science.

ACKNOWLEDGMENTS

We thank D. S. Dean for useful discussions. P.L.K. was partially supported by NSF(DMR9978902).

APPENDIX A: DERIVATION OF THE LOGARITHMIC CORRECTION TO THE FRONT POSITION

In this appendix, we present a detailed derivation of the logarithmic correction to the asymptotic front position. We employ the approach of Ref. [39] where such a correction was computed for a reaction diffusion equation. In the present context, our starting point is the recursion relation in Eq. (53) for the m-ary search trees. We first substitute $P_n(x)=1-f_n(x)$ in Eq. (53) and then neglect terms of order $O(f_n^2)$ in the regime $x\to-\infty$ to get a linear equation

$$f_{n+1}(x)=m\int_0^\infty f_n(x-\epsilon)\rho(\epsilon)d\epsilon, \tag{A1}$$

where $\rho(\epsilon)$ is the effective induced distribution $\rho(\epsilon)$ given by Eq. (50). Next we assume that for large n the front position is given by $x_n=vn+c(n)$, where both the velocity v and the functional form of the correction term $c(n)$ are yet to be determined. Following Ref. [39], we then assume that for large n the solution $f_n(x)$ of Eq. (A1) is given by the scaling form

$$f_n(x)=n^\gamma H\left(\frac{x-x_n}{n^\gamma}\right)e^{\lambda(x-x_n)}, \tag{A2}$$

where the exponent γ and the scaling function $H(y)$ are not yet known. We only know that $H(y)\to 0$ as $y\to\pm\infty$ [since $0\leqslant f_n(x)\leqslant 1$ for all x]. Also, since for large n, the prefactor n^γ in Eq. (A2) must go away, indicating that $H(y)\sim y$ as $y\to 0$.

Let us define $z_n=(x-x_n)/n^\gamma$. Then to leading order for large n, one has $z_{n+1}\approx z_n-(\gamma/n)z_n-vn^{-\gamma}$. Substituting z_{n+1} in Eq. (A2) and keeping only leading order terms we get for the left-hand side of Eq. (A1),

$$f_{n+1}(x)\approx n^\gamma e^{\lambda(x-x_n-v)}\left[H(z)-\frac{vH'(z)}{n^\gamma}-\frac{\gamma}{n}zH'(z)+\left(\frac{\gamma}{n}-\lambda\frac{dc}{dn}(n)\right)H(z)+\frac{v^2}{2n^{2\gamma}}H''(z)\right]. \tag{A3}$$

In the above equation, we used the shorthand notations $z_n=z$, $H'(z)=dH/dz$, and $H''(z)=d^2H/dz^2$.

Similarly, inserting Eq. (A2) into the right-hand side of Eq. (A1), expanding $H[(x-x_n-\epsilon)n^{-\gamma}]$ in Taylor series in $\epsilon e^{-\gamma}$, and keeping only leading order terms, we find the right-hand side of Eq. (A1)

$$f_{n+1}(x)\approx mn^\gamma e^{\lambda(x-x_n)}\left[\mu_0 H(z)-\frac{\mu_1}{n^\gamma}H'(z)+\frac{\mu_2}{2n^{2\gamma}}H''(z)\right], \tag{A4}$$

where $\mu_k=\int_0^\infty \epsilon^k e^{-\lambda\epsilon}\rho(\epsilon)d\epsilon$. Comparing the left-hand side given by Eq. (A3) and the right-hand side given by Eq. (A4), we recover, to leading order for large n, the dispersion relation

$$e^{-\lambda v}=m\int_0^\infty e^{-\lambda\epsilon}\rho(\epsilon)d\epsilon. \tag{A5}$$

As argued before, the front will choose the maximum velocity $v(\lambda^*)$ of the dispersion relation (A5). At $\lambda=\lambda^*$, $v'(\lambda^*)=0$. Differentiating Eq. (A5) with respect to λ we obtain $v(\lambda^*)\exp[-\lambda^* v(\lambda^*)]=m\mu_1$. Using this in Eq. (A3) shows that the term of order $n^{-\gamma}$ in Eq. (A3) cancels the corresponding term on the right-hand side in Eq. (A4). To ensure that remaining terms are of the same order, we must have $\gamma=1/2$ and $dc/dn=b/n$. The latter equation gives $c(n)=b\ln n$, where b is still undetermined. Employing these choices for γ and $c(n)$ and equating Eqs. (A3) and (A4), we obtain

$$(v^2-me^{\lambda^* v}\mu_2)H''(z)-zH'(z)+(1-2b\lambda^*)H(z)=0,$$

where $v=v(\lambda^*)$. This equation can be further simplified as follows. Differentiating Eq. (A5) twice with respect to λ and using $v'(\lambda^*)=0$ we get an additional relation, $v^2(\lambda^*)-m\mu_2\exp[\lambda^* v(\lambda^*)]=\lambda^* v''(\lambda^*)$. By inserting this into the above equation we finally arrive at the eigenvalue equation

$$-\lambda^* v''(\lambda^*)H''(z)+zH'(z)+(2b\lambda^*-1)H(z)=0. \tag{A6}$$

Note that $v(\lambda)$ has a maximum at $\lambda=\lambda^*$ indicating $v''(\lambda^*)<0$. Rescaling $z=\sqrt{-\lambda^* v''(\lambda^*)}\zeta$, we find that the

solution of Eq. (A6) that vanishes at $\zeta \to \infty$ is given by $H(\zeta) = Be^{-\zeta^2/4} D_{2b\lambda^*-2}(\zeta)$, where B is a constant and $D_p(\zeta)$ is the parabolic cylinder function of index p. The condition that $H(\zeta) \sim \zeta$ as $\zeta \to 0$ enforces the choice of the index $p = 2b\lambda^* - 2 = 1$ indicating $b = 3/2\lambda^*$. Note that the above solution describes precisely the wave function of the first excited state of a quantum harmonic oscillator and the factor 3/2 is the corresponding energy eigenvalue. Finally, the leading asymptotic behavior of the front position is given by

$$x_n = v(\lambda^*)n + \frac{3}{2\lambda^*} \ln n. \tag{A7}$$

A similar calculation can be carried out for the balanced height where one finds a dispersion relation $v_1(\mu)$ as given by Eq. (59) and front position is given by

$$x_n^* = v_1(\mu^*) - \frac{3}{2\mu^*} \ln n, \tag{A8}$$

where μ^* denotes the point where $v_1(\mu)$ has its unique minimum.

APPENDIX B: ASYMPTOTIC BEHAVIOR OF THE CUMULATIVE HEIGHT DISTRIBUTION

In this appendix, we derive the large y behavior of the cumulative height distribution $F(y)$. The function $F(y)$ is the solution of the boundary value problem (16) and (17). We already know that $1 - F(y) \sim e^{\lambda^* y}$ as $y \to -\infty$, where λ^* denotes the value of λ where the dispersion curve $v(\lambda)$ in Eq. (19) has its maximum. In order to derive the asymptotic behavior of $F(y)$ in the other limit $y \to \infty$, we first recast the integral equation (16) in a slightly different form. Let us first define the cumulative distribution function

$$Y(\epsilon, \epsilon') = \int_0^{\epsilon} \int_0^{\epsilon'} p(x_1, x_2)\, dx_1\, dx_2, \tag{B1}$$

where the joint distribution $p(x_1, x_2)$ is given by Eq. (9). Writing $p(\epsilon, \epsilon') = \partial^2 Y / \partial\epsilon\, \partial\epsilon'$ on the right-hand side of Eq. (16) and performing the integrations by part (first over ϵ and then over ϵ'), we finally arrive at the following equation:

$$F(y-v) = \int_0^{\infty} \int_0^{\infty} F'(y-\epsilon)\, F'(y-\epsilon') Y(\epsilon, \epsilon')\, d\epsilon\, d\epsilon' + 2\int_0^{\infty} F(y-\epsilon)\rho(\epsilon)\, d\epsilon - 1, \tag{B2}$$

where $F'(y) = dF/dy$ and we have used the boundary conditions of $F(y)$. Note that due to the concentration of measure, $F(y)$ has roughly the shape of the step function, $F(y) \approx \theta(-y)$ with the front located at $y = 0$. Thus the derivative roughly behaves as a negative delta function, $F'(y) \approx -\delta(y)$. First reconsider the limit $y \to -\infty$. In this limit, the arguments of the functions $F'(y)$ inside the integrands in the first term on the right-hand side in Eq. (B2) are always very large and negative, indicating that the contribution from this term is negligible as $y \to -\infty$. Neglecting the first term, one finds that the resulting linear equation admits the exponential solution $1 - F(y) \sim e^{\lambda y}$ where v depends on λ through the dispersion relation in Eq. (19). Thus one recovers the correct result in the $y \to -\infty$ limit.

Turn now to the complementary limit $y \to \infty$. Then the arguments of $F'(y)$ inside the integrands of the first term on the right-hand side of Eq. (B2) can be close to zero to pick up a substantial contribution. For large y, one can approximate $F'(y) \approx -\delta(y)$ inside the integrands on the right-hand side of Eq. (B2) and one then gets

$$F(y-v) \approx -1 + Y(y,y) + 2\int_0^{\infty} F(y-\epsilon)\rho(\epsilon)\, d\epsilon. \tag{B3}$$

$Y(y,y) \to 1$ and $F(y) \to 0$ as $y \to \infty$. To find the asymptotics of $F(y)$ we differentiate Eq. (B3) with respect to y and use $F'(y) \approx -\delta(y)$ in the second term. This gives

$$F'(y-v) \approx -2\rho(y) + 2\left.\frac{\partial Y(y, y_2)}{\partial y_2}\right|_{y_2 = y}. \tag{B4}$$

Using the definitions in Eqs. (B1) and (9) we find

$$\frac{\partial Y}{\partial y_2} = -e^{-y_2} \phi(e^{-y_2})\, \phi(1 - e^{-y_2})\, \theta(e^{-y_1} + e^{-y_2} - 1).$$

When $y_1 = y_2$ is large, the argument of the step function in the above equation is always negative, indicating that one can neglect the second term on the right-hand side of Eq. (B4). This gives $F'(y) \approx -2\rho(y+v)$. Hence the desired large y behavior of $F(y)$ is given by

$$F(y) \approx 2\int_y^{\infty} \rho(y' + v)\, dy', \tag{B5}$$

where $v = v(\lambda^*)$ is the maximum velocity associated with the dispersion relation (19).

Note that the constraint $e^{-\epsilon} + e^{-\epsilon'} = 1$ does not modify the form of the dispersion curve when compared to the unconstrained conventional DP problem [the only difference is that one has to first find the effective single point energy distribution $\rho(\epsilon)$ in the constrained case from Eq. (9)]. However, the above constraint does modify the large y behavior of the cumulative distribution $F(y)$. For example, Eq. (B4) is valid for the unconstrained problem as well. However, in the unconstrained case, $Y(y,y) = [\int_0^y \rho(\epsilon)\, d\epsilon]^2$. In that case one finds after taking the derivative, $F'(y) \approx -2\rho(y+v) \int_y^{\infty} \rho(\epsilon)\, d\epsilon$ indicating that for large y

$$F(y)|_{\text{unconstrained}} \approx 2\int_y^{\infty} dy' \rho(y' + v) \int_{y'}^{\infty} \rho(\epsilon)\, d\epsilon. \tag{B6}$$

For example, for the RBST where $\rho(\epsilon) = e^{-\epsilon}$, the large y asymptotics are $F(y) \sim e^{-y}$ (constrained case) and $F(y) \sim e^{-2y}$ (unconstrained case).

APPENDIX C: DERIVATION OF THE INDUCED DISTRIBUTION

In this Appendix we derive the induced distribution $\eta(r)$ [see Eq. (47)] starting from the joint distribution

$$\text{Prob}[r_1,\ldots,r_m]=A_m\delta\left(\sum_{i=1}^{m} r_i-1\right)\prod_{i=1}^{m} r_i. \qquad \text{(C1)}$$

The constant A_m in the above equation has to be chosen such that the joint distribution is normalized. The induced distribution $\eta(r)$ is obtained by fixing one of the fractions, say the first one, to the value r and then integrating over all other fractions. Thus by definition

$$\eta(r)=A_m r\int_0^1\ldots\int_0^1\delta\left(\sum_{i=2}^{m} r_i+r-1\right)\prod_{i=2}^{m} r_i\,dr_i. \qquad \text{(C2)}$$

Note that r_i's denote the lengths of m intervals with the total length equal to unity. Let us define a set of new variables, $x_2=r+r_2$, $x_3=x_2+r_3$, $\ldots$, $x_{m-1}=x_{m-2}+r_{m-1}$. Here x_i's denote the points separating adjacent intervals. Clearly then $x_{m-1}=1-r_m$ since the total length is unity. With these change of variables the integral in Eq. (C2) becomes

$$\eta(r)=A_m r\zeta_m(r), \qquad \text{(C3)}$$

where $\zeta_m(r)$ is given by

$$\zeta_m(r)=\int_r^1(x_2-r)dx_2\int_{x_2}^1(x_3-x_2)dx_3\ldots\ldots\int_{x_{m-2}}^1(x_{m-1}-x_{m-2})(1-x_{m-1})dx_{m-1}. \qquad \text{(C4)}$$

Thus $\zeta_m(r)$ satisfies the recursion relation

$$\zeta_m(r)=\int_r^1(x_2-r)\zeta_{m-1}(x_2)dx_2. \qquad \text{(C5)}$$

One directly computes $\zeta_2(r)=1-r$ and $\zeta_3(r)=(1-r)^3/6$ which suggests to seek a solution in the form $\zeta_m(r)=B_m(1-r)^{2m-3}$. Plugging the above expression in recursion (C5) yields

$$B_m=\frac{B_{m-1}}{(2m-3)(2m-4)}, \qquad \text{(C6)}$$

which is iterated to give $B_m=1/(2m-3)!$. Thus we obtain $\eta(r)=A_m r(1-r)^{2m-3}/(2m-3)!$. The normalization condition $\int_0^1\eta(r)dr=1$ then gives $A_m=\Gamma(2m)$ where $\Gamma(x)$ is the gamma function. Therefore

$$\eta(r)=(2m-1)(2m-2)r(1-r)^{2m-3}, \qquad \text{(C7)}$$

which is valid for all $m\geqslant 2$.

APPENDIX D: LARGE m RESULTS FOR ARBITRARY DISTRIBUTION

In this appendix we derive the large m behavior of $\langle H_N\rangle$ and $\langle h_N\rangle$ for m-ary search trees with arbitrary distribution $\eta(r)$. We start with the height variable and write the dispersion relation

$$e^{-\lambda v}=m\int_0^\infty e^{-\lambda\epsilon}\rho(\epsilon)d\epsilon=m\int_0^1 r^\lambda\eta(r)dr. \qquad \text{(D1)}$$

The constraint $\Sigma r_i=1$ leads to $\int_0^1 r\,\eta(r)dr=1/m$. Thus for large m, a generic distribution $\eta(r)$ will be concentrated near $r=0$. Consider a class of distributions that behave as $\eta(r)\approx C_m r^a e^{-b_m r}$ near the origin. For example, $C_m=m-1$, $a=0$, and $b_m=m-2$ for the uniform distribution (45). Similarly, $C_m=(2m-1)(2m-2)$, $a=1$ and $b_m=2m-3$ for the distribution (47). These two examples suggest that $C_m\sim m^{a+1}$ and $b_m\sim m$. Making use of the constraints $\int_0^1\eta(r)dr=1$ and $\int_0^1 r\,\eta(r)dr=1/m$ one indeed confirms the above asymptotics: $b_m\approx(a+1)m$ and $C_m\approx b_m^{a+1}/\Gamma(a+1)$.

We now consider the integral in Eq. (D1). Substituting the small r behavior of $\eta(r)$, performing the integral, and using the Stirling formula one gets

$$(b_m e^{-v})^\lambda\approx\frac{m}{\Gamma(a+1)}\sqrt{2\pi(\lambda+a)}\left(\frac{\lambda+a}{e}\right)^{\lambda+a}. \qquad \text{(D2)}$$

Taking the logarithm, differentiating with respect to λ, and setting $v'(\lambda^*)=0$ we determine λ^* and $v(\lambda^*)$. The leading contributions are given by Eq. (56). Therefore, the large m behavior of $\langle H_n\rangle$ is indeed universal.

We now turn to the large m behavior of the average balanced height $\langle h_N\rangle$. In this case, the appropriate dispersion relation is given by Eq. (59),

$$e^{\mu v_1}=m\int_0^\infty e^{\mu\epsilon}\rho(\epsilon)d\epsilon=m\int_0^1 r^{-\mu}\eta(r)dr. \qquad \text{(D3)}$$

Substituting the small r behavior, $\eta(r)\approx C_m r^a e^{-b_m r}$, and performing the integral we obtain

$$(b_m^{-1}e^{v_1})^\mu\approx\frac{m}{\Gamma(a+1)}\Gamma(a+1-\mu). \qquad \text{(D4)}$$

We will see that in the large m limit, $\mu^*\to a+1$. Hence we write $\mu\to a+1-\delta$, assume that $\delta\ll 1$, plug these in Eq. (D4) and take the logarithm to obtain

$$(a+1-\delta)(v_1-\ln b_m)\approx\ln\frac{m}{\Gamma(a+1)}-\ln\delta. \qquad \text{(D5)}$$

Differentiating Eq. (D5) with respect to δ and setting $v'(\delta^*)=0$ yields

$$\mu^*=a+1-\frac{a+1}{\ln m}+\cdots,$$

$$v(\lambda^*)=\frac{a+2}{a+1}\ln m+\cdots.$$

The parameter a appears in the leading order even in the large m limit. Consequently, $\langle h_N\rangle$ also depends on a and thence the balanced height is not universal in the large m limit.

[1] For a recent review, see O. Martin, R. Monasson, and R. Zecchina, Theor. Comput. Sci. **265**, 3 (2001).
[2] M. Mézard and G. Parisi, J. Phys. (Paris) **47**, 1285 (1986).
[3] Y. Fu and P.W. Anderson, J. Phys. A **19**, 1605 (1986).
[4] R. Monasson and R. Zecchina, Phys. Rev. E **56**, 1357 (1997); S. Cocco and R. Monasson, Phys. Rev. Lett. **86**, 1654 (2001).
[5] J. Inoue, J. Phys. A **30**, 1407 (1997).
[6] M. Weigt and A.K. Hartmann, Phys. Rev. Lett. **86**, 1658 (2001).
[7] N. Sourlas, Nature (London) **339**, 693 (1989); A. Montanari and N. Sourlas, Eur. Phys. J. B **18**, 107 (2000).
[8] Y. Fu, in *Lectures in the Science of Complexity*, edited by D.L. Stein (Addison-Wesley, Reading, MA, 1989); F.F. Ferreira and J.F. Fontanari, J. Phys. A **31**, 3417 (1998); S. Martens, Phys. Rev. Lett. **81**, 4281 (1998); T. Sasamoto, T. Toyoizumi, and H. Nishimori, e-print cond-mat/0106125.
[9] M. Mézard and G. Parisi, J. Phys. (Paris) **48**, 1451 (1987); D.J. Aldous, Prob. Theor. Rel. Fields, **93**, 507 (1992); D.J. Aldous, Random Struct. Algorithms **18**, 381 (2001); V.S. Dotsenko, J. Phys. A **33**, 2015 (2000); G. Parisi and M. Ratieville, Eur. Phys. J. B **22**, 229 (2001).
[10] M. Mézard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[11] D.E. Knuth, *The Art of Computer Programming, Sorting and Searching*, 2nd ed. (Addison-Wesley, Reading, MA, 1998), Vol. 3.
[12] R.A. Fisher, Ann. of Eugenics **7**, 355 (1937).
[13] A. Kolmogorov, I. Petrovsky, and N. Piscounov, Moscow Univ. Math. Bull. (Engl. Transl.) **1**, 1 (1937); translated and reprinted in P. Pelce, *Dynamics of Curved Fronts* (Academic Press, San Diego, 1988).
[14] M. Bramson, Commun. Pure Appl. Math. **21**, 531 (1978).
[15] Ya.B. Zeldovich, G.I. Barenblatt, V.B. Librovich, and G.M. Makhviladze, *The Mathematical Theory of Combustion and Explosions* (Consultants Bureau, New York, 1985).
[16] J.D. Murray, *Mathematical Biology* (Springer-Verlag, New York, 1989).
[17] W. van Saarloos, Phys. Rev. A **39**, 6367 (1989).
[18] U. Ebert and W. van Saarloos, Phys. Rev. Lett. **80**, 1650 (1998); Physica D **146**, 1 (2000).
[19] J.M. Robson, Austr. Comput. J. **11**, 151 (1979).
[20] P. Flajolet and A. Odlyzko, J. Comput. Syst. Sci. **25**, 171 (1982).
[21] L. Devroye, J. Assoc. Comput. Mach. **33**, 489 (1986).
[22] H.M. Mahmoud, *Evolution of Random Search Trees* (Wiley, New York, 1992).
[23] E.J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958).
[24] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed. (R.E. Krieger Publishing Co., Malabar, 1987).
[25] S.M. Berman, *Sojourns and Extremes of Stochastic Processes* (Wadsworth and Brooks/Cole, Stamford, CT, 1992).
[26] J.-P. Bouchaud and M. Mezard, J. Phys. A **30**, 7997 (1997).
[27] L. Devroye, Acta Inform. **24**, 277 (1987).
[28] M. Greiner, H.C. Eggers, and P. Lipa, Phys. Rev. Lett. **80**, 5333 (1998).
[29] W.I. Newman and A.M. Gabrielov, Int. J. Fract. **50**, 1 (1991).
[30] D. Sornette and A. Johansen, Physica A **261**, 581 (1998).
[31] S.N. Coppersmith, C.-h. Liu, S.N. Majumdar, O. Narayan, and T. Witten, Phys. Rev. E **53**, 4673 (1996).
[32] T. Hattori and H. Ochiai (unpublished).
[33] P.L. Krapivsky and S.N. Majumdar, Phys. Rev. Lett. **85**, 5492 (2000).
[34] B. Derrida and H. Spohn, J. Stat. Phys. **51**, 817 (1988).
[35] B. Reed, J. Assoc. Comput. Mach. (to be published).
[36] J. M. Robson, Theor. Comput. Sci. (to be published).
[37] M. Drmota, J. Assoc. Comput. Mach. (to be published).
[38] E. Ben-Naim, P.L. Krapivsky, and S.N. Majumdar, Phys. Rev. E **64**, 035101 (2001).
[39] E. Brunet and B. Derrida, Phys. Rev. E **56**, 2597 (1997).
[40] S.N. Majumdar and P.L. Krapivsky, Phys. Rev. E **62**, 7735 (2000).
[41] D.S. Dean and S.N. Majumdar, Phys. Rev. E **64**, 046126 (2001).