ON THE GAP STRUCTURE OF SEQUENCES OF POINTS ON A CIRCLE

by

Lyle Ramshaw

STAN-CS-77-637
NOVEMBER 1977

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

On the Gap Structure of Sequences of Points on a Circle

Lyle Ramshaw

Computer Science Department
Stanford University
Stanford, California 94305

Abstract.

Considerable mathematical effort has gone into studying sequences of points in the interval $[0,1)$ which are evenly distributed, in the sense that certain intervals contain roughly the correct percentages of the first n points. This paper explores the related notion in which a sequence is evenly distributed if its first n points split a given circle into intervals which are roughly equal in length, regardless of their relative positions. The sequence $x_k = (\log_2(2k-1) \bmod 1)$ was introduced in this context by DeBruijn and Erdös. We will see that the gap structure of this sequence is uniquely optimal in a certain sense, and optimal under a wide class of measures.

This research was supported in part by National Science Foundation grant MCS 72-03752 A03, by the Office of Naval Research contract N00014-76-C-0330, by IBM Corporation, and a National Science Foundation Graduate Fellowship,

Consider sequences of points on the circumference of a circle of radius $1/2\pi$, or equivalently in the unit interval $[0,1)$. Such a sequence is called uniformly distributed if the percentage of the first n points which lie in any fixed interval approaches the length of that interval as n tends to infinity; this concept has been studied extensively [4]. We can arrive at a different notion of even distribution by considering instead the lengths of the gaps between elements of the sequence. For each n, the first n points of any sequence divide the circle into n intervals, and we shall study those sequences which make these intervals roughly equal in length, regardless of the order in which they occur around the circle. Putting this another way, we will study strategies for successively breaking a unit stick into smaller and smaller fragments, while attempting to arrange that the n stick fragments present at time n are as nearly equal in length as possible, for all n.

More formally, let us define an <u>n-state</u> to be a multiset containing n nonnegative real numbers which sum to one; the elements of the n-state specify the lengths of the sticks present at time n. An n-state S is a <u>legal predecessor</u> of an $(n+1)$-state T if there exists a number x in S such that $S-\{x\} \subseteq T$. It follows that the multiset $T-(S-\{x\})$ must consist of exactly two numbers whose sum is x; that is, T arises from S by breaking a stick of length x into two nonnegative fragments.

A <u>stickbreaking strategy</u> is then an infinite sequence of states $\langle S_n\rangle_{n\geq 1}$, where S_n is an n-state and a legal predecessor of S_{n+1} for each n. Every sequence of points on the circle defines a unique stickbreaking strategy, and every strategy can be generated by at least one sequence.

We now turn to the study of stickbreaking strategies, in an attempt to find those strategies $\langle S_n \rangle$ in which the elements of S_n are nearly equal for each n. There are many different precise notions lurking behind this fuzzy concept; for example, we might try to

$$\text{minimize} \quad \limsup_n \{n \cdot \max(S_n)\} \text{ , or}$$

$$\text{maximize} \quad \liminf_n \{n . \min(S_n)\} \text{ , or}$$

$$\text{minimize} \quad \limsup_n \left\{ \frac{\max(S_n)}{\min(S_n)} \right\} \quad .$$

DeBruijn and Erdős considered these three measures in [1], and proved that the best possible values for any stickbreaking strategy were $1/\ln 2$, $1/\ln 4$, and 2 respectively, where " $\ln$ " denotes " $\log_e$ ". They also discovered a particular strategy which simultaneously achieves the optimum in all three measures. This strategy is the one defined by the sequence $\langle x_k \rangle_{k \geq 1}$ with $x_k = (\lg(2k-1) \bmod 1)$, where " $\lg$ " denotes " $\log_2$ " and mod 1 denotes the fractional part; we will call this the <u>log stickbreaking strategy</u>. The n-states of the log strategy have the form

$$\left\{ \lg\left(\frac{n+1}{n}\right), \lg\left(\frac{n+2}{n+1}\right), \ldots, \lg\left(\frac{2n}{2n-1}\right) \right\}$$

for each n ; the strategy works, in some sense, because

$$\lg\left(\frac{n+1}{n}\right) = \lg\left(\frac{2n+2}{2n}\right) = \lg\left(\frac{2n+1}{2n}\right) + \lg\left(\frac{2n+2}{2n+1}\right) \quad .$$

Note, by the way, that the sequence $\langle x_k \rangle_{k>1}$ which defines the log strategy is not uniformly distributed, since for example the ratio

$$\frac{(\text{number of } k\text{'s such that } 1 \leq k \leq n \text{ and } 0 \leq x_k < 1/2)}{n}$$

does not approach a limit as $n \to \infty$. Thus, the sequences which are excellently distributed in our new stickbreaking sense need not be evenly distributed at all in the classical sense of uniform distribution.

The graph in Figure 1 depicts the log stickbreaking strategy in action. A vertical line has been drawn from the top of the figure down to the point $\langle x_k, k \rangle$ for $1 \leq k \leq 64$. A horizontal cut through the resulting picture at height n reflects the state of the log -strategy at time n . Knuth [3] has used this type of graph to display the intriguing distribution structure of the sequence $\langle y_k \rangle_{k>0}$ where $y_k = \left(\frac{k(1+\sqrt{5})}{2} \bmod 1\right)$.

We now want to build a more general framework in which to explore the optimality of log stickbreaking. Our first task is to find a partial order on n-states which captures the notion of a state's elements being "more nearly equal". Suppose that S and T are n-states containing s_1, s_2 and t_1, t_2 respectively, and suppose that $S - \{s_1, s_2\} = T - \{t_1, t_2\}$. It must then be the case that $s_1 + s_2 = t_1 + t_2$. If, in addition, we have $s_1 \geq t_1 \geq s_2$, it follows that either $s_1 \geq t_1 \geq t_2 \geq s_2$ or $s_1 \geq t_2 \geq t_1 \geq s_2$. In either situation, we would intuitively say that the elements of T are more nearly equal than those of S . In particular, we can go from S to T by robbing $(s_1 - t_1)$ units from the rich s_1 and giving them to the poor s_2. We will then say that T results from S by a <u>Robin Hood act.</u> More generally, an n-state S will be said to <u>majorize</u> an n-state T whenever T can be reached from S by a finite sequence of Robin Hood acts; thus, if S majorizes T , the elements of S are at least as unequal as the elements of T .

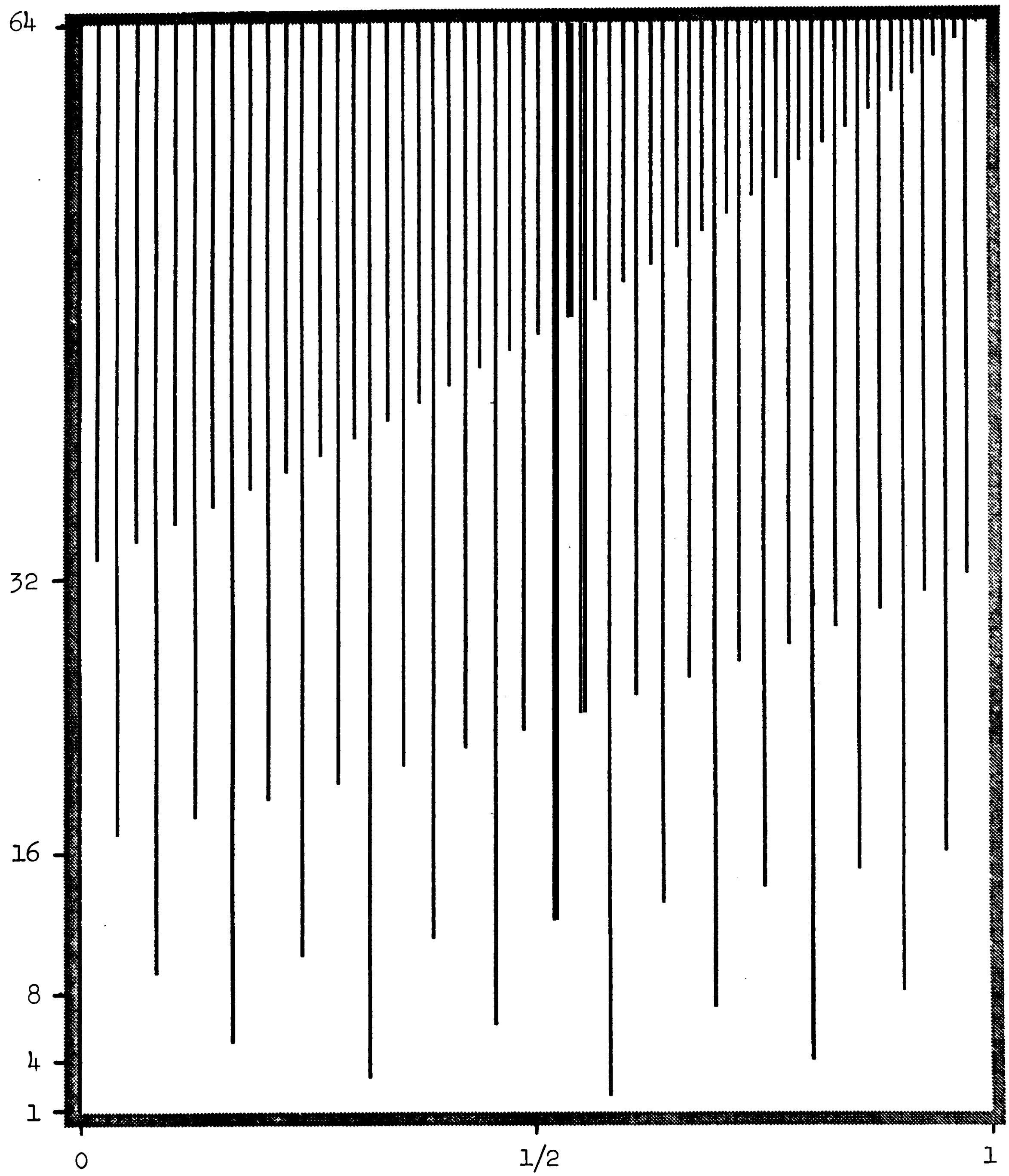

Figure 1

Majorization is a partial order on n-states; interestingly, we can get the same partial order in a different way. Let n-vectors be points in $\mathbb{R}^n$ whose components are all nonnegative and sum to one; an n-vector is an ordered version of an n-state. If $\sigma = \langle s_1, s_2, \ldots, s_n\rangle$ and $\tau = \langle t_1, t_2, \ldots, t_n\rangle$ are n-vectors, we will say that $\sigma \gtrsim \tau$ if, for all k in the range $0 \le k \le n$, we have $s_1 + s_2 + \cdots + s_k \ge t_1 + t_2 + \cdots + t_k$; in other words, $\sigma \gtrsim \tau$ when the partial sums of σ uniformly exceed those of τ. Now, with each n-state S, we can associate an n-vector $\sigma = \bar{S}$ whose components are the elements of S in nonincreasing order. It turns out that S majorizes T if and only if $\bar{S} \gtrsim \bar{T}$; a proof and still another characterization of this same partial order can be found in [2], sections 2.18 to 2.20.

Our first lemma shows that the relation $\sigma \gtrsim \tau$ holds more often than one might expect. One can view this result as a variant of Spitzer's Lemma [8].

Lemma 1. Let $\sigma = \langle s_1, s_2, \ldots, s_n\rangle$ and $\tau = \langle t_1, t_2, \ldots, t_n\rangle$ be n-vectors; for each k between 0 and n-1, let $\sigma^{(k)} = \langle s_{k+1}, s_{k+2}, \ldots, s_n, s_1, s_2, \ldots, s_k\rangle$ denote the sequence σ circularly shifted k places, and define $\tau^{(k)}$ analogously. Then, for some k in the range $0 < k < n$, we have $\sigma^{(k)} \gtrsim \tau^{(k)}$.

Proof. We want to shift those positions where τ is larger towards the right end. In fact, it is enough to choose k to maximize the quantity

$$\sum_{1 \le i \le k} (t_i - s_i) \ . \quad \square$$

We will use Lemma 1 in studying what can happen in a stickbreaking strategy between time n and time 2n. Define an n-slice to be a finite sequence of m-states $\langle S_m\rangle$ for $n \le m \le 2n$, where S_m is a legal predecessor of S_{m+1} for $n \leq m < 2n$. The behavior of any stickbreaking strategy over the interval $[n,2n]$ constitutes an n-slice, and any n-slice can be extended in many ways to a full stickbreaking strategy.

We can draw an n-slice as an oriented forest containing n trees and a total of $3n$ nodes. Each tree will depict the history over the slice of one of the n sticks which existed at time n , and each node will represent a stick. The nodes will be labelled $[\ell,m]$, where ℓ gives the stick's length, and m denotes the last time at which it remains unbroken. For each stick that-is still unbroken at time 2n we will write $m = *$, and the node will have no offspring. If $m \neq *$, then $n \leq m < 2n$ and the node has exactly two offspring representing its fragments when broken. For example, each n-slice of the log strategy defines the forest in Figure 2.

If an n-slice contains states with several sticks of the same size, that is, with elements of multiplicity greater than one, it may be possible to draw several different forests which represent that same n-slice. A simple example is the 2-slice $\{\{2/3,1/3\},\{1/3,1/3,1/3\},\{1/3,1/3,1/6,1/6\}\}$. Each portrayal of an n-slice as a forest will be called an <u>interpretation</u>. Of course, every legal n-slice must have at least one interpretation.

Note that each of the trees in the above unique interpretation of a slice of the log strategy contains exactly three nodes. A tree with only a single node represents a stick which survives unbroken from time n to time 2n ; call such sticks <u>atoms</u>. Call an n-slice <u>atomless</u> if it has at least one atomless interpretation. The following lemma shows that all the best slices are atomless.

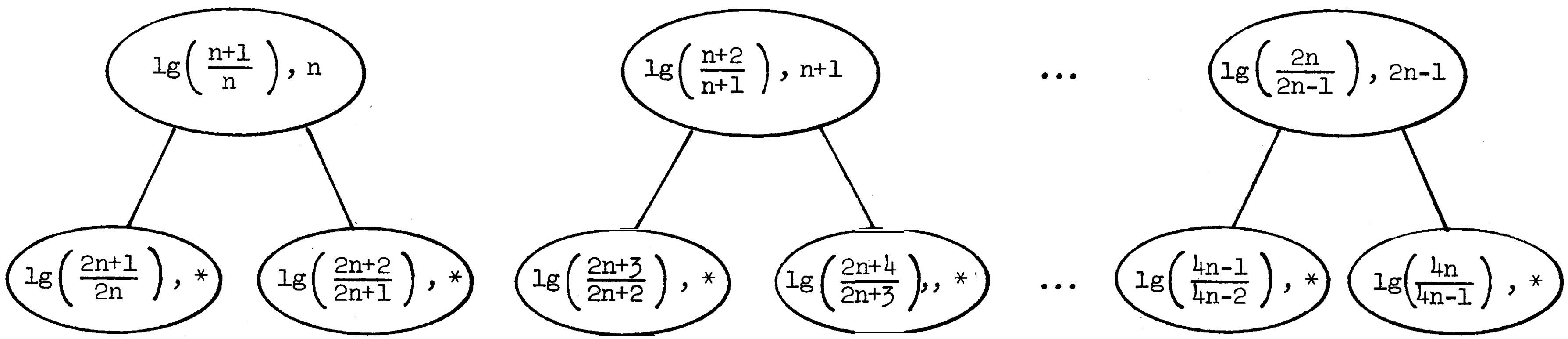
$\lg\left(\frac{n+1}{n}\right)$, n
$\lg\left(\frac{n+2}{n+1}\right)$, n+1
...
$\lg\left(\frac{2n}{2n-1}\right)$, 2n-1
$\lg\left(\frac{2n+1}{2n}\right)$, *
$\lg\left(\frac{2n+2}{2n+1}\right)$, *
$\lg\left(\frac{2n+3}{2n+2}\right)$, *
$\lg\left(\frac{2n+4}{2n+3}\right)$,, *
...
$\lg\left(\frac{4n-1}{4n-2}\right)$, *
$\lg\left(\frac{4n}{4n-1}\right)$, *

Figure2

Lemma 2. If $\langle S_m \rangle_{n \le m \le 2n}$ is any n-slice, there exists an atomless n-slice $\langle T_m \rangle_{n \le m \le 2n}$ such that S_m majorizes T_m for all m. That is, any n-slice can be uniformly improved upon by an n-slice with an atomless interpretation.

Proof. Let $\langle S_m \rangle$ be an n-slice, and fix a particular interpretation of $\langle S_m \rangle$ which has at least one atom. By induction, it suffices to show that there exists an n-slice (T_m) which uniformly improves upon $\langle S_m \rangle$, and an interpretation of $\langle T_m \rangle$ with one less atom.

Choose any atom of $\langle S_m \rangle$, and let its length be a. Since the atom is represented by a tree with a single node, there must be some other tree in the interpretation of $\langle S_m \rangle$ with at least five nodes. That tree must include a leaf node at level p where $p > 2$. Thus, $\langle S_m \rangle$ must have the form shown in Figure 3, where the triangles indicate arbitrary trees whose roots have the lengths shown. Note that, if $p = 2$, the nodes labelled ℓ_1 and ℓ_{p-1} are actually identical.

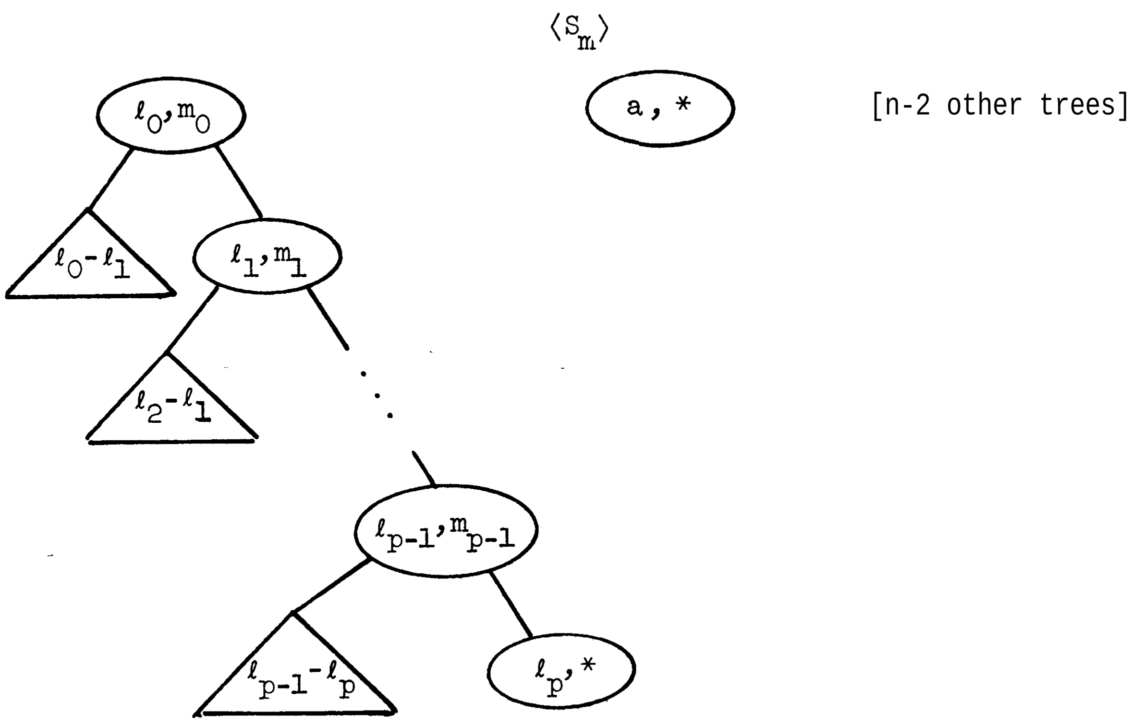

Figure3

The construction of the desired n-slice $\langle T_m \rangle$ divides into two cases, depending upon the size of a . Since $\ell_0 \geq \ell_1 > \ldots \geq \ell_{p-1} \geq \ell_p$, at least one of the inequalities $a \geq \ell_{p-1}$ and $a \leq \ell_1$ must hold. Suppose first that $a > \ell_{p-1}$. In this case, we can improve upon the slice $\langle S_m \rangle$ by breaking a and leaving ℓ_{p-1} alone, as shown in Figure 4. Let $\langle T_m \rangle$ be the n-slice defined by this interpretation; clearly $\langle T_m \rangle$'s interpretation has one less atom than the given interpretation of $\langle S_m \rangle$. Now, for $m \leq m_{p-1}$ the state T_m is identical to S_m. For $m > m_{p-1}$, we can go from S_m to T_m , by replacing the pair $\{a, \ell_p\}$ with $\{\ell_{p-1}, a+\ell_p-\ell_{p-1}\}$. Since $a \geq \ell_{p-1} \geq \ell_p$, this replacement constitutes-a Robin Hood act. Hence, S_m majorizes T_m for all m , and the first case is complete.

$\langle T_m \rangle$ if $a \geq \ell_{p-1}$

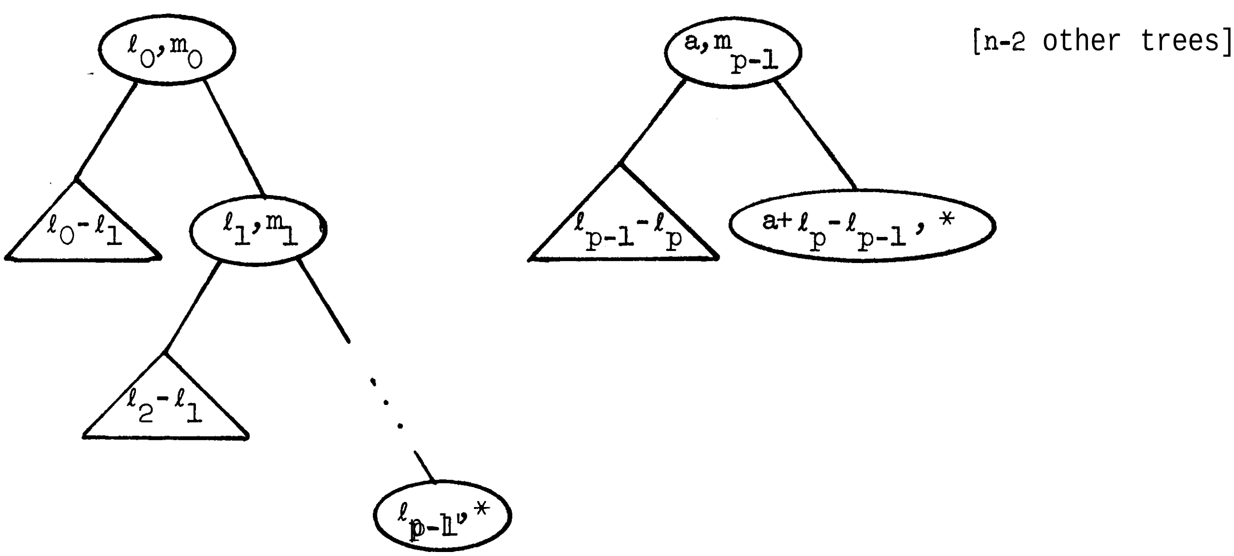

Figure4

On the other hand, suppose that $a \le \ell_1$. In this case, we can improve upon $\langle S_m \rangle$ by adjusting what happens early in the slice, instead of late. In particular, we can change the lengths of the initial intervals and get a as the result of a break, as shown in Figure 5. Once again, let $\langle T_m \rangle$ be defined by this interpretation, and note that we have reduced the number of atoms by one. Now, for $m > m_0$, we have $T_m = S_m$. For $m \le m_0$, we can get from S_m to T_m by replacing the pair $\{a, \ell_0\}$ with $\{\ell_1, a+\ell_0-\ell_1\}$; since $\ell_0 \ge \ell_1 \ge a$, this is again a Robin Hood act, and the proof is complete. $\square$

$\langle T_m \rangle$ if $a \le \ell_1$

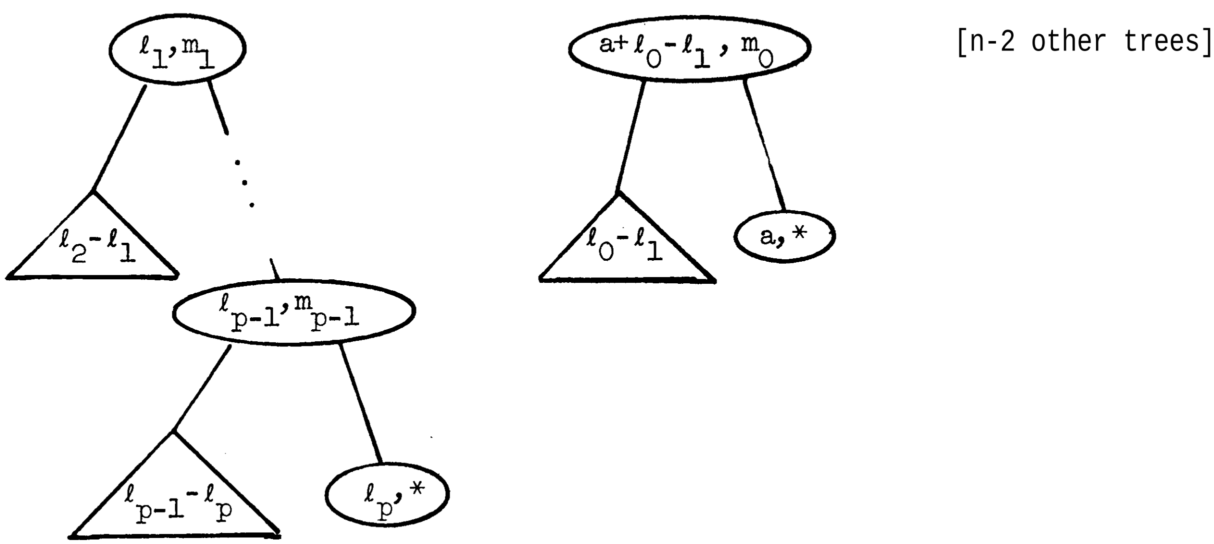

Figure5

We can now show that a rather wide class of stickbreaking slices has a weak form of optimality. In particular, we will define an n-slice $\langle S_m \rangle$ to be perfect if each break in the slice breaks the currently largest stick exactly in half, and if the slice is atomless. Let $\langle S_m \rangle_{n \le m \le 2n}$ be a perfect n-slice, and let $\overline{S_n} = \langle s_1, s_2, \ldots, s_n \rangle$ be the n-vector whose components are the sizes of its sticks at time n in nonincreasing order. Note that, since $\langle S_m \rangle$ is atomless, we must have $s_n \ge s_1/2$. Hence, the (n+k) -vectors $\overline{S_{n+k}}$ for $0 \le k \le n$ must be given by

$$\overline{S_{n+k}} = \left\langle s_{k+1}, s_{k+2}, \ldots, s_n, \frac{s_1}{2}, \frac{s_1}{2}, \frac{s_2}{2}, \frac{s_2}{2}, \ldots, \frac{s_k}{2}, \frac{s_k}{2} \right\rangle$$

Conversely, if $s_1 \ge s_2 \ge \ldots \ge s_n$ are any nonnegative numbers whose sum is one, and if $s_n \ge s_1/2$, there is a unique associated perfect n-slice whose states are specified as above. Our next theorem shows that all these n-slices have a certain optimality.

Theorem 1. Let $\langle S_m \rangle_{n \le m \le 2n}$ be a perfect n-slice, and let $\langle T_m \rangle_{n \le m \le 2n}$ be an arbitrary n-slice. Then, for some k in the range $0 < k < n$, the state T_{n+k} will majorize S_{n+k} ; that is, at some time the slice $\langle T_m \rangle$ must do at least as poorly as $\langle S_m \rangle$.

Proof. First, if-every interpretation of $\langle T_m \rangle$ contains atoms, we can use Lemma 2 to construct a uniformly superior atomless slice. Hence, we may assume without loss of generality that $\langle T_m \rangle$ has an atomless interpretation.

Under this interpretation, every stick represented by an element of T_n is broken exactly once during the course of the slice $\langle T_m \rangle$.

Number the elements of T_n in the order in which they are broken, $T_n = \{t_1, t_2, \ldots, t_n\}$, and let t_i' and t_i'' be the lengths of the fragments of t_i, for $1 \le i \le n$. Furthermore, choose the names to make $t_i' \ge t_i''$. Then, consider the vectors

$$\begin{aligned}
\tau_n &= \langle t_1, t_2, \ldots, t_n \rangle \\
&\vdots \\
\tau_{n+k} &= \langle t_{k+1}, \ldots, t_n, t_1', t_1'', t_2', t_2'', \ldots, t_k', t_k'' \rangle \\
&\vdots \\
\tau_{2n} &= \langle t_1', t_1'', t_2', t_2'', \ldots, t_n', t_n'' \rangle \quad .
\end{aligned}$$

Note that the components of τ_{n+k} are exactly the elements of T_{n+k}, but not necessarily in sorted order.

Now, recall that the perfect slice $\langle S_m \rangle$ takes the form

$$\begin{aligned}
\overline{S_n} &= \langle s_1, s_2, \ldots, s_n \rangle \\
&\vdots \\
\overline{S_{n+k}} &= \left\langle s_{k+1}, \ldots, s_n, \frac{s_1}{2}, \frac{s_1}{2}, \frac{s_2}{2}, \frac{s_2}{2}, \ldots, \frac{s_k}{2}, \frac{s_k}{2} \right\rangle \\
&\vdots \\
\overline{S_{2n}} &= \left\langle \frac{s_1}{2}, \frac{s_1}{2}, \frac{s_2}{2}, \frac{s_2}{2}, \ldots, \frac{s_n}{2}, \frac{s_n}{2} \right\rangle \quad .
\end{aligned}$$

By applying Lemma 1 to the n-vectors τ_n and $\overline{S_n}$, we deduce that there must exist some k in the range $0 \le k < n$, such that $\tau_n^{(k)} \gtrsim \overline{S_n}^{(k)}$; that is, such that

$$\langle t_{k+1}, \ldots, t_n, t_1, \ldots, t_k \rangle \gtrsim \langle s_{k+1}, \ldots, s_n, s_1, \ldots, s_k \rangle \quad .$$

This is almost enough information to conclude that, in fact, $\tau_{n+k} \gtrsim \overline{S_{n+k}}$; that is, that

$$\langle t_{k+1}, \ldots, t_n, t_1', t_1'', \ldots, t_k', t_k'' \rangle \gtrsim \langle s_{k+1}, \ldots, s_n, \frac{s_1}{2}, \frac{s_1}{2}, \ldots, \frac{s_k}{2}, \frac{s_k}{2} \rangle$$

The only partial sums that haven't been handled are those which include a t_i' but not the corresponding t_i''. Note, however, that we do know that

$$t_{k+1} + \cdots + t_n + t_1 + \cdots + t_{i-1} \geq s_{k+1} + \cdots + s_n + s_1 + \cdots + s_{i-1} \quad \text{and}$$

$$t_{k+1} + \cdots + t_n + t_1 + \cdots + t_{i-1} + t_i \geq s_{k+1} + \cdots + s_n + s_1 + \cdots + s_{i-1} + s_i \ .$$

We can deal with the remaining partial sums by averaging these two inequalities, and then using the additional-fact that $t_i' \geq t_i''$ implies $t_i' \geq t_i/2$. Thus, $\tau_{n+k} \gtrsim \overline{S_{n+k}}$.

Finally, note that τ_{n+k} is simply a rearrangement of $\overline{T_{n+k}}$ into a possibly non-sorted order. Thus, we must also have $\overline{T_{n+k}} \gtrsim \overline{S_{n+k}}$, since the sum of the largest j components of any vector is certainly at least as large as the sum of the leftmost j components. It follows that T_{n+k} majorizes S_{n+k}. □

In light of Theorem 1, it might seem to be rather hopeless to find a sense in which any particular stickbreaking strategy is uniquely optimal. In fact, Theorem 1 shows that stickbreaking is a rather zero-sum proposition; a strategy does well at some times by doing correspondingly poorly at other times. And different strategies do well at different times. To progress further in our study of stickbreaking, we must be willing to compare m-states and n-states where $m \neq n$. That is, we must extend the majorization partial order to deal with multisets of different sizes.

One possibility is to generalize majorization by using Lorenz curves. These curves are used in economics for studying inequity in distributions of income or wealth [7]. In our context, we will define the Lorenz curve of an n-state S_n to be the function $\hat{S}_n : [0,1] \to [0,1]$ with $\hat{S}_n(r)$ given by the sum of the rn largest elements of S_n, If rn is not an integer, we will define the value of $\hat{S}_n(r)$ by interpolating linearly between the nearest two values of r which make rn integral. In particular, if $\overline{S_n} = \langle s_1, s_2, \ldots, s_n \rangle$, then

$$S_n(k/n) = \sum_{1 \le i \le k} s_i \quad \text{for } 0 < k < n ,$$

and $\hat{S}_n(r)$ for other r is found by piecewise linear interpolation. (Warning: these Lorenz curves are "upside down" in comparison to the Lorenz curves of economics.)

The Lorenz curve of a state is a piecewise linear and concave function, which assumes the values 0 and 1 at 0 and 1 respectively. Furthermore, the discontinuities in the derivative of the function occur only at rational points. Conversely, any function with these properties is the Lorenz cuve of an infinite family of states. For example, the identity function is the Lorenz curve of the n-state $\{1/n, 1/n, \ldots, 1/n\}$ for each n .

Suppose that S_n and T_n are two n-states. Recall that S_n majorizes T_n if and only if $\overline{S_n} \gtrsim \overline{T_n}$, In terms of their Lorenz curves, the latter condition states that $\hat{S}_n(r) \ge \hat{T}_n(r)$ for r in $\{0, 1/n, 2/n, \ldots, 1\}$. But since Lorenz curves are linear in each region $[k/n, (k+1)/n]$, we can conclude that S_n majorizes T_n if and only if $\hat{S}_n(r) \ge \hat{T}_n(r)$ for all r in $[0,1]$. This latter condition is a natural partial order on Lorenz curves; we will say that $\hat{S}_n \geq \hat{T}_n$ when $\hat{S}_n(r) \geq \hat{T}_n(r)$ for all r. We can now extend majorization to relate states of different sizes by defining an m-state S_m to <u>majorize</u> an n-state T_n exactly when $\hat{S}_m \geq \hat{T}_n$. Note that this more general majorization is not quite a partial order on the set of all states, since two distinct states with the same Lorenz curve would each majorize the other.

We could arrive at the same generalization without using Lorenz curves. In order to compare an m-state S_m and n-state T_n, we could divide each element of S_m into n equal pieces, and each element of T_n into m equal pieces. This would generate two (mn)-states, which we could compare by the old methods. Since this refining process does not change the associated Lorenz curves, this idea leads to the same generalization of majorization that we found above.

The Lorenz curves of the log strategy have a particularly simple form. In fact, let L_n denote the state of the log strategy at time n, and define the <u>envelope</u> to be the graph of the function $\lg(1+r)$ on the unit interval, Then, $\hat{L}_n$ is exactly the function which piecewise linearly interpolates the envelope at the points $\{0, 1/n, 2/n, \ldots, 1\}$. This gives a good graphical intuition for the behavior of the $\hat{L}_n$; for example, we can now see that L_{kn} majorizes L_n for every k and n.

According to our definitions, no slice of the log strategy is perfect. But we can construct for each n a unique perfect n-slice which begins with the state L_n; it is only necessary to note that the biggest element of L_n is less than twice as large as the smallest. Let the perfect n-slice so defined be written $\underline{\langle P_{n,m}\rangle_{n<m\leq 2n}}$, where $P_{n,m}$ is the state at time m. Note

that $\hat{P}_{n,m}$ also has a simple structure; in particular, $\hat{P}_{n,m}$ interpolates the envelope over the n intervals defined by the $(n+1)$ points

$$\left\{0\ ,\ \frac{1}{m}\ ,\ \frac{2}{m}\ ,\ \ldots\ ,\ \frac{2n-m-1}{m}\ ,\ \frac{2n-m}{m}\ ,\ \frac{2n-m+2}{m}\ ,\ \frac{2n-m+4}{m}\ ,\ \ldots\ ,\ \frac{m-2}{m}\ ,\ 1\right\}\ .$$

We finally have enough information to characterize the log stick-breaking strategy in a non-trivial way.

Theorem 2. Log stickbreaking is the unique strategy with the property that none of its Lorenz curves anywhere exceed the envelope. In more detail, if an arbitrary strategy $\langle S_m \rangle_{m \geq 1}$ remains on or below the envelope everywhere before time $2n$, it must actually equal the log strategy until time n.

Proof. Suppose that $\langle S_m \rangle_{m \geq 1}$ does lie on or under the envelope before time $2n$; that is, $\hat{S}_m(r) \leq \lg(1+r)$ for $0 \leq r \leq 1$ and $1 \leq m < 2n$. Equivalently, we could assume that $\hat{L}_m \geq \hat{S}_m$ for $1 \leq m < 2n$. Apply Theorem 1 to the perfect slice $\langle P_{n,m} \rangle_{n \leq m \leq 2n}$ and the n-slice $\langle S_m \rangle_{n \leq m \leq 2n}$. The theorem allows us to conclude that there exists a k in the range $0 \leq k < n$ such that S_{n+k} majorizes $P_{n,n+k}$. Hence, we have

$$\hat{L}_{n+k} \geq \hat{S}_{n+k} \geq \hat{P}_{n,n+k}\ .$$

The graph in Figure 6 illustrates the situation for $k = 1$ and $n = 2$.

Now, consider what S_{n+k-1} could be like; it must arise from combining two elements of S_{n+k}. But from the above relation, we know that the smallest two elements of S_{n+k} must sum to precisely $\lg\left(\frac{n+k}{n+k-1}\right)$. Furthermore, since $\hat{S}_{n+k-1}$ must fit on or under the envelope, the state

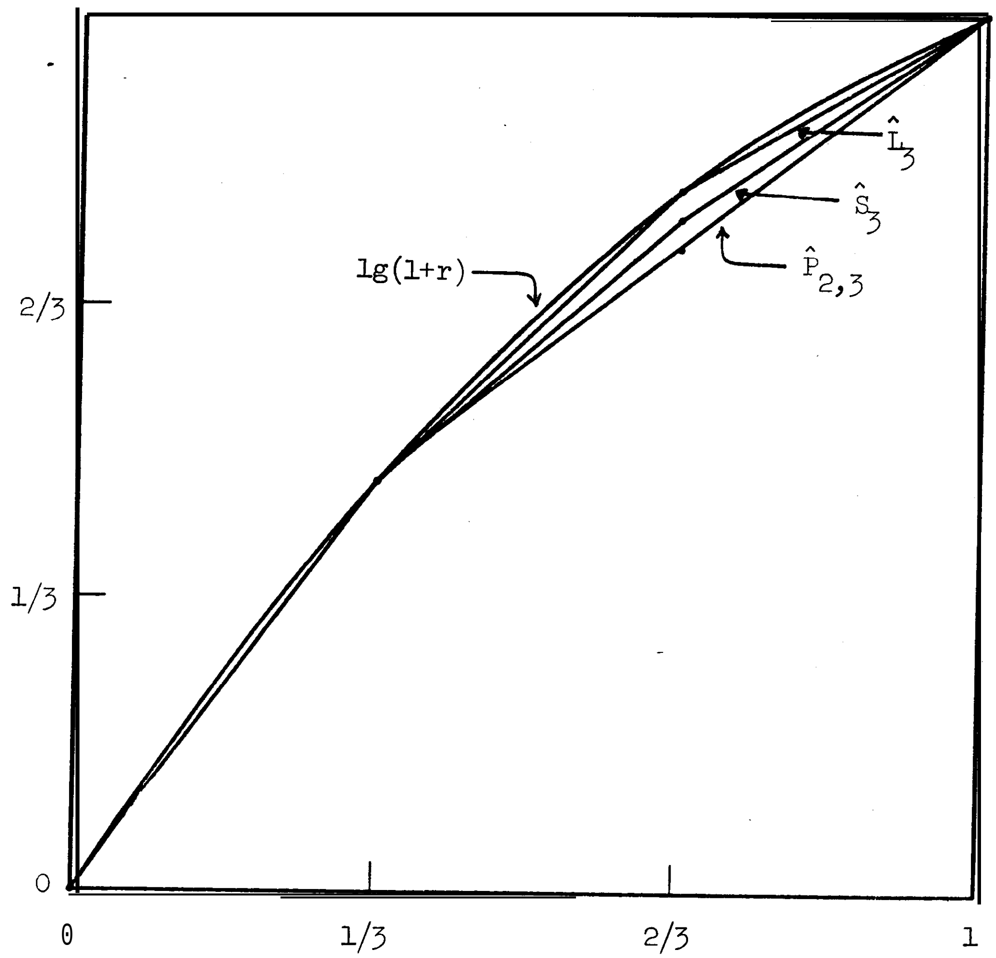

Figure 6

S_{n+k-1} cannot afford any element larger than $\lg\left(\frac{n+k}{n+k-1}\right)$; the only choice Is to combine the smallest two elements of S_{n+k} . Hence, we have $\hat{L}_{n+k-1} \geq \hat{S}_{n+k-1} \geq \hat{P}_{n,n+k-1}$.

Continuing inductively, we eventually conclude that $\hat{L}_n \geq \hat{S}_n \geq \hat{P}_{n,n} = \hat{L}_n$ and thus $S_n = L_n$. Pushing the same argument even further, we find that the history continues to be forced, and that $S_m = L_m$ for $1 \leq m \leq n$. □

Next, we want to use this characterization to show that log stick-breaking is actually uniquely optimal in some sense. Define an m-state S_m

to be <u>decent</u> if, for every stickbreaking strategy $\langle T_k \rangle_{k>1}$, there exists an infinite number of indices k such that T_k majorizes S_m. Intuitively, a decent state is not too bad, since every strategy must do at least as poorly infinitely often. The next theorem shows that the envelope marks the dividing line between decent and indecent states.

<u>Theorem 3</u>. Let $\hat{S}_m$ be the Lorenz curve of an m-state S_m. If $\hat{S}_m(r) < \lg(1+r)$ for all r in the open interval $(0,1)$, then the state S_m is decent. If there exists an r in $(0,1)$ where $\hat{S}_m(r) > \lg(1+r)$, then S_m is not decent.

<u>Proof</u>. The second implication is easier. If $\hat{S}_m$ actually exceeds the envelope at some point, then no state of the log strategy can possibly majorize S_m. Hence, S_m cannot be decent.

For the first implication, let S_m be an m-state whose Lorenz curve liesstrictly under the envelope except at 0 and 1. Our first goal is to prove the existence of perfect slices all of whose states majorize S_m. Consider the states $P_{n,k}$ for large n and $n \le k \le 2n$. Each curve $\hat{P}_{n,k}(r)$ is a piecewise linear interpolate of $\lg(1+r)$. Furthermore, as n tends to infinity, the lengths of the chords involved tend to zero, uniformly in k. Hence, the $\hat{P}_{n,k}(r)$ converge to the envelope $\lg(1+r)$ uniformly in r and k. Finally, since all Lorenz curves are concave, we can check that any k-state T_k majorizes S_m merely by checking that $\hat{T}_k(r) \ge \hat{S}_m(r)$ for r in the finite set $\{0, 1/m, 2/m, \ldots, 1\}$. Therefore, by choosing n sufficiently large, we can guarantee that the states $P_{n,k}$ majorize S_m for all k in the range $n \le k \le 2n$.

-Fix an n which is sufficiently large by this criterion, and let $\langle T_k \rangle_{k \geq 1}$ be any strategy which challenges the decency of S_m. By applying Theorem 1 to the perfect n-slice $\langle P_{n,k} \rangle_{n < k \leq 2n}$ and the n-slice $\langle T_k \rangle_{n \leq k \leq 2n}$, we deduce that there exists some k in the range $n < k < 2n$ such that T_k majorizes $P_{n,k}$. Since majorization is transitive, T_k will also majorize S_m. Finally, since the above works for all sufficiently large n, we find that the strategy $\langle T_k \rangle_{k \geq 1}$ majorizes S_m infinitely often; hence S_m is decent. □

Unfortunately, the above theorem does-not settle the really interesting cases! In particular, we would like to know whether or not the states L_n of the log strategy are decent. The author rather suspects that they are, but that question seems difficult to resolve. Instead, let us resort to the following definition, Call an n-state S_n <u>nearly decent</u> if $\overline{S_n}$ as a vector in $\mathbb{R}^n$ is an accumulation point of the set of $\overline{T_n}$ for decent T_n. That is, a state $S_n = \{s_1, s_2, \ldots, s_n\}$ is nearly decent when arbitrarily small perturbations of the s_i exist which make the state decent. The usefulness of this definition lies in the following theorem.

<u>Theorem 4</u>. An n-state S_n is nearly decent if and only if its Lorenz curve $\hat{S}_n$ never exceeds the envelope.

<u>Proof</u>. Once again it is convenient to do the easy half first. Suppose that $\hat{S}_n$ actually exceeds the envelope at some point. Then, it must in fact exceed the envelope at some point of the form k/n for $0 < k < n$; that is, we have $\hat{S}(k/n) = \lg(1 + k/n) + \varepsilon$ for some $\varepsilon > 0$.

Now, $\hat{S}_n(k/n)$ equals the sum of the k largest elements of S_n . A sufficiently small neighborhood of the n-vector $\overline{S_n}$ in $\mathbb{R}^n$ will therefore contain only n-vectors whose k largest components also sum to something strictly greater than $\lg(1 + k/n)$, Applying Theorem 3, we conclude that no state in this neighborhood can be decent, thus S_n is not even nearly decent.

Conversely, suppose that $\hat{S}_n$ lies everywhere on or under the envelope. Note that it can actually touch the envelope only at a finite number of points of the form k/n . Let $V = \{k \mid 0 < k < n \text{ and } \hat{S}_n(k/n) = \lg(1 + k/n)\}$, To prove that S_n is nearly decent, we want to find a family of decent n-states whose n-vectors converge to $\overline{S_n}$ in $\mathbb{R}^n$. We will construct these n-states by constructing their Lorenz curves; and we will do the latter by distorting $\hat{S}_n$ a little in the neighborhood of the points k/n for k in V. But what is "a little"?

First, note that for each k in V we must have

$$\hat{S}_n\left(\frac{k+1}{n}\right) - \hat{S}_n\left(\frac{k}{n}\right) < \hat{S}_n\left(\frac{k}{n}\right) - \hat{S}_n\left(\frac{k-1}{n}\right) ;$$

that is, the stick corresponding to the interval $[k/n, (k+1)/n]$ must be strictly smaller than the one corresponding to $[(k-1)/n, k/n]$. This follows since $\hat{S}_n$ actually touches the curving and concave envelope at k/n . Let the slack in this inequality be denoted Δ_k , and let

$$\rho = \frac{1}{2} \min_{k \in V} \Delta_k .$$

For ε in the range $0 < \varepsilon \leq \rho$, define the function $\hat{T}_{n,\varepsilon}$ at the points k/n by

$$\hat{T}_{n,\varepsilon}(k/n) = \begin{cases} \hat{S}_n(k/n) - \varepsilon & \text{if } k \in V \\ \hat{S}_n(k/n) & \text{if } 0 \le k \le n \text{ and } k \notin V , \end{cases}$$

and extend $\hat{T}_{n,\varepsilon}$ to the unit interval by linear interpolation. The tricky point now is to show that $\hat{T}_{n,\varepsilon}$ is concave. It suffices to check that the slope does not increase at each corner between linear segments. Consider the corner k/n ; if k is not in V , the change from $\hat{S}_n$ to $\hat{T}_{n,\varepsilon}$ only makes things better. If k is in V , the change to $\hat{T}_{n,\varepsilon}$ can at most affect the difference between the lengths of the sticks corresponding to $[k/n, (k+1)/n]$ and $[(k-1)/n, k/n]$ by 2ε . Since $2\varepsilon \le \Delta_k$, the change from $\hat{S}_n$ to $\hat{T}_{n,\varepsilon}$ does not destroy concavity.

Thus, for $0 < \varepsilon \le \rho$, the function $\hat{T}_{n,\varepsilon}$ is a valid Lorenz curve for an associated n-state $T_{n,\varepsilon}$. Note that the stick lengths of $T_{n,\varepsilon}$ each differ by at most ε from the corresponding stick lengths of S_n . Hence, as ε goes to zero, $\overline{T_{n,\varepsilon}}$ converges to $\overline{S_n}$ in $\mathbb{R}^n$. Since each $T_{n,\varepsilon}$ lies strictly below the envelope on $(0,1)$, we deduce from Theorem 3 that each $T_{n,\varepsilon}$ is decent; therefore, S_n is nearly decent, □

<u>Corollary</u>. The log stickbreaking strategy is the unique strategy all of whose states are nearly decent.

<u>Proof'</u>. This follows immediately from Theorems 2 and 4. □

This Corollary is the promised demonstration that log stickbreaking is uniquely optimal in some sense. To wrap things up, we will use this general optimality to show that log stickbreaking is also optimal in a fairly wide class of real-valued measures; in particular, this class will include the three measures studied by Erdős and DeBruijn.

A real-valued functional ν on the set of all states will be called a <u>monotone measure</u> if it has the following two properties:

(i) If an m-state S_m majorizes an n-state T_n, then $\nu(S_m) > \nu(T_n)$,

(ii) For each fixed n, $\nu(S_n) = \nu(\{s_1, s_2, \ldots, s_n\})$ is jointly continuous in the s_i.

Our earlier discussion of majorization shows that property (i) is equivalent to the following pair of conditions together:

(i') Performing a Robin Hood act never increases the value of ν.

(i'') Two states with the same Lorenz curve must have the same value of ν.

This latter pair of conditions is often easier to verify,

If the author's suspicions are correct and the states L_n of the log strategy are actually decent as well as nearly decent, then the continuity requirement, property (ii), could be dropped.

Many intuitively reasonable yardsticks of stickbreaking performance can be phrased as monotone measures. Here is a list of examples which begins with the three covered by DeBruijn and Erdős; let S_n be an n-state with $\overline{S_n} = \langle s_1, s_2, \ldots, s_n \rangle$.

(1) $$\nu(S_n) = n \cdot \min(S_n) = \hat{S}_n'(0)$$

(2) $$\nu(S_n) = -n \cdot \min(S_n) = -\hat{S}_n'(1).$$

(3) $$\nu(S_n) = \frac{\max(S_n)}{\min(S_n)} = \frac{\hat{S}_n'(0)}{\hat{S}_n'(1)}.$$

(4) $\nu(S_n) = (n^{p-1}) \sum_{1 \le i \le n} s_i^p = \int_0^1 (\hat{S}_n'(r))^p \, dr$

for fixed $p > 1$, especially $p = 2$.

(5) $\nu(S_n) = \hat{S}_n(r)$ for fixed r in $(0,1)$.

(6) $\nu(S_n) = \int_0^1 \hat{S}_n(r)dr = \frac{1}{2n} \cdot \sum_{1 \le i \le n} (2n-2i+1)s_i$.

(7) Generalizing 5 and 6, we can have

$$\nu(S_n) = \int_0^1 \hat{S}_n(r)dF(r)$$

for any nondecreasing $F: [0,1] \to \mathbb{R}$.

Given any particular monotone measure, we can rate the performance of a stickbreaking strategy $\langle S_n \rangle_{n \ge 1}$ by $\limsup_n \nu(S_n)$, where small values of this lim sup are desirable. Our final result is that log stickbreaking has the optimal lim sup in any monotone measure.

Theorem 5. If ν is any monotone measure and $\langle S_n \rangle_{n \ge 1}$ is any stickbreaking strategy, then

$$\limsup_n \nu(S_n) > \limsup_k \nu(L_k) = \sup_k \nu(L_k)$$

Proof. Fix an arbitrary $k \ge 1$; we want to show that

$$\limsup_n \nu(S_n) \ge \nu(L_k)$$

Since L_k is nearly decent, there exists a sequence of decent k-states

$\langle T_{k,p} \rangle_{p \geq 1}$ such that $\overline{T_{k,p}}$ converges to $\overline{L_k}$ in $\mathbb{R}^k$. By property (ii), the real numbers $\nu(T_{k,p})$ must converge to $\nu(L_k)$.

Now, each k-state $T_{k,p}$ is decent; hence there exists an infinite number of indices n such that S_n majorizes $T_{k,p}$. Therefore,

$$\limsup_n \nu(S_n) \geq \nu(T_{k,p})$$

for every p . Letting p go to infinity, we deduce

$$\limsup_n \nu(S_n) \geq \nu(L_k) \qquad \text{for each } k \text{ , and}$$

thus

$$\limsup_n \nu(S_n) > \sup_k \nu(L_k) .$$

Finally, the above argument with $S_n = L_n$ shows that

$$\limsup_n \nu(L_n) \geq \sup_k \nu(L_k) ,$$

hence these two quantities must in fact be equal. Alternatively, we could have deduced their equality at once by recalling that L_{kn} majorizes L_n for every k and n . □

Acknowledgments.

The author would like to thank Tom Cover, who suggested the problem and christened it "stickbreaking"; Frances Yao and Andrew Yao, who rediscovered some of the results of [1]; Don Knuth, who simplified the proof of Lemma 1; and Leo Guibas, who volunteered his helpful insights throughout.

References

[1] DeBruijn, N. G. and Erdös, P.: "Sequences of points on a circle," <u>Indagationes Mathematicae</u> 11 (1949); 14-17.

[2] Hardy, G. H., Littlewood, J. E., and Pólya, G.: <u>Inequalities</u>, Cambridge University Press, 1952.

[3] Knuth, D. E.: <u>The Art of Computer Programming</u>, Volume 3, Sorting Searching, Addison-Wesley, 1973; pages xiv, 510 and 511.

[4] Kuipers, L. and Niederreiter, H.: <u>Uniform Distribution of Sequences</u>, John Wiley & Sons, 1974.

[5] Ostrowski, A.: "Eine Verschärfung des Schubfächerprinzips in einem linearen Intervall," <u>Archiv der Mathematik</u> 8 (1957); 1-10, 330,

[6] Ostrowski, A.: "Zum Schubfächerprinzip in einem linearen Intervall," <u>Jahresbericht der Deutschen Mathematiker Vereinigung</u> 60 (1957); 33-39.

[7] Samuelson, P. A.: <u>Economics</u> (Tenth edition), McGraw-Hill, 1976;

[8] Spitzer, F.: 'A Combinatorial Lemma and its Applications to Probability Theory," <u>American Mathematical Society Transactions</u> 82 (1956); 323-339.