Journal of Statistical Mechanics:
Theory and Experiment

SPECIAL ISSUE IN MEMORY OF VLADIMIR RITTENBERG

Sequential and continuous time stick-breaking

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER:

Sequential and continuous time stick-breaking*

Peter F Arndt

Department for Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 63/73, 14195 Berlin, Germany
E-mail: arndt@molgen.mpg.de

Received 14 January 2019
Accepted for publication 22 April 2019
Published 18 June 2019

Online at stacks.iop.org/JSTAT/2019/064003
https://doi.org/10.1088/1742-5468/ab1dd8

Abstract. The repeated breaking of a linear object, for example a stick, is a fundamental process which underlies numerous natural phenomena. Here we compare two distinct ensembles of stick-breaking: (i) a stick is broken with a certain rate over time; and (ii) a stick is broken a finite number of times. Both ensembles are deduced from appropriate integral equations and related to each other. The analyses performed here and the comparison of the two ensembles enables us to better understand the stick-breaking process by itself.

Keywords: dynamical processes, evolutionary processes, fracture, bioinformatics

* This paper is dedicated to the memory of Vladimir Rittenberg.

1742-5468/19/064003+8$33.00

Contents

1. Introduction

The analysis and understanding of complex phenomena that are associated with mechanical failure and fragmentation of objects is of great importance in basic research and applied material science. In engineering for example, the geometry and material composition of macroscopic objects is significant for the functioning as well as the manufacturing process and accordingly well studied and optimized. Here, methods borrowed from a diverse range of fields from molecular dynamics [1] to finite element methods [2] are used.

In statistical physics however, the interest is more focused on quantities which do not refer to macroscopic or microscopic details of such a system. Models are hence more general and can subsequently be used to describe a multitude of other seemingly unrelated phenomena (see [3] and references therein). A prominent subject in models of fragmentation is the distribution of a conserved quantity—such as the mass, energy, or momentum—among the pieces of a disintegrated object. Such a model may be used to describe the mass distribution of a meteorite shower [4] but has also been used to describe the distribution of resources among competing species in an environmental niche [5].

The stick-breaking model is conceptually one of the simplest models of fragmentation and describes the breaking of a one dimensional object or interval at random positions. Furthermore, since the stick-breaking process can be used to construct a Dirichlet or beta measure, it attracted a lot of attention in the mathematical community and is used to construct priors in Bayesian analysis [6–8].

Often the breaking of an object is described as a process that is continuous in space and time [3, 9, 10]. In this framework, breaks occur with a certain rate over time as specified by the model. Due to the probabilistic nature of this process, the total number of breaks after a certain time is not fixed but follows a specific distribution, a Poisson

J. Stat. Mech. (2019) 064003

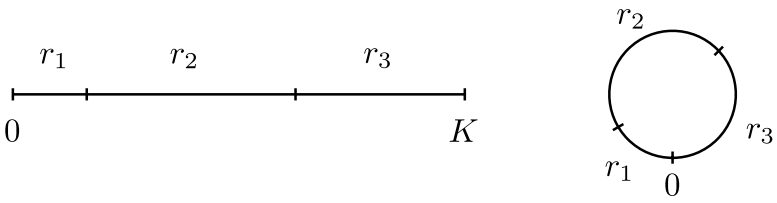

Figure 1. A linear interval of length K with $b = 2$ breaks (left panel). Its circular representation after identifying its beginning and end and introducing an auxiliary break at 0 (right panel).

distribution in the simplest case of uniform breakage. However, when observing broken objects in nature, the rates and times of the fragmentation process are often not accessible and only the total number of broken pieces can be observed as final products of the process. For the theoretical analysis we therefore need to consider at least two different ensembles of broken sticks, one continuous time ensemble in which the rate and time of breaking is fixed, and one sequential stick-breaking ensemble, in which the total number of breaks is fixed.

In this article we aim to explore these two alternative ensembles and point out differences between them. First we will shortly introduce both ensembles, i.e. the one in continuous time and the one in the number of breaks. Next, we will deduce the length distribution of the resulting smaller sticks, which follows an exponential function in the continuous time framework, but is polynomial in the framework of sequential breaks. After computing key quantities, as for instance the mean length of broken sticks, we further recover the length distribution of broken sticks in the continuous time ensemble by summing up such length distributions for appropriate numbers of breaks. We also show that considering a collection of sticks with a uniform distributed number of breaks gives rise to a scale-free distribution of broken stick lengths, which has been observed in natural phenomena [11, 12].

2. The stick-breaking process

Consider an interval or stick of length K that is randomly broken at positions, which are chosen uniformly along the stick (see figure 1, left panel). This process will generate a collection of smaller intervals whose lengths add up to K.

After introducing periodic boundary conditions and identifying the beginning and the end of the interval as well as introducing an auxiliary break at 0 (figure 1, right panel) the system is invariant under rotations and therefore the length distribution of small intervals will be the same for all pieces irrespective on where they are located and whether they include one of the two ends of the original stick or not.

2.1. The distribution of stick lengths after a finite time of random breaking

In a continuous time model of a breaking stick an interval of initial length K is assumed to break with a certain rate in time. Assuming that breaks occur homogeneously we can denote this rate per length and time interval by μ, i.e. the probability that a break will

J. Stat. Mech. (2019) 064003

occur in an infinitesimal small length interval $\mathrm{d}K$ in an infinitesimal small time interval $\mathrm{d}t$ is given by $\mu\,\mathrm{d}K\,\mathrm{d}t$. In this formulation breaks will occur after exponentially distributed waiting times with mean $\bar{\Delta}t = 1/(K\mu)$. Observing a large ensemble of breaking sticks one will therefore find sticks with different number of breaks.

If we denote the length distribution of small pieces after breaking for a time t by $m(r,t)$ then this quantity follows the differential equation

$$\frac{\partial m(r,t)}{\partial t} = -\mu r m(r,t) + 2\mu \int_r^K m(s,t)\,\mathrm{d}s, \tag{1}$$

where the first term describes the loss of sticks of length r due to breaks which will occur with rate μr in time. The second term encodes the gain of sticks of length r due to breaks of longer sticks of length s occurring at a distance r from one of its end. With initial length distribution $m(r,0) = \delta(r-K)$ and δ being the Kronecker delta function, this differential equation can be solved by

$$m(r,t) = \begin{cases} (2\mu t + \mu^2 t^2 (K-r))\exp(-\mu t r) & \text{for } r < K \\ \exp(-\mu t K) & \text{for } r = K \end{cases} \tag{2}$$

as deduced previously [10, 12]. As expected the length distribution of the resulting broken sticks after a certain time exhibits an exponential tail as well as an exponentially vanishing delta peak at $r = K$ representing the presence of unbroken sticks in this ensemble.

2.2. The distribution of stick lengths after a finite number of random breaks

Let us now consider a different ensemble of breaking sticks, one where the number of breaks b is fixed. We denote the length distribution of broken sticks after b breaks with initial length K by $m(r,b)$. This length distribution can be recursively computed using the following integral equation which involves the same length distribution for $b-1$ breaks only:

$$m(r,b) = m(r,b-1) - \frac{r}{K} m(r,b-1) + \frac{2}{K}\int_r^K m(s,b-1)\,\mathrm{d}s, \tag{3}$$

i.e. the difference $m(r,b) - m(r,b-1)$ is again given by two terms. The first describes the loss of a stick of length r if it is broken anywhere in between, which happens with probability r/K and the other term reflects the gain of a piece if a piece of length $s > r$ is broken at one out of two possible sites.

This length distribution for an unbroken stick, $b = 0$, of length K is clearly

$$m(r,0) = \delta(r-K). \tag{4}$$

With equation (3) we compute that the length distribution of sticks after one break is uniform

$$m(r,1) = \frac{2}{K}, \tag{5}$$

where the factor 2 reflects the fact that we now have two smaller sticks. In the general case for all $b > 1$, the above recursion is solved by

J. Stat. Mech. (2019) 064003

$$m(r,b) = \frac{b(b+1)}{K}\left(1-\frac{r}{K}\right)^{b-1} \tag{6}$$

a polynomial function in the length r. With this distribution in hand we can check that the total number of pieces is given by

$$\int_0^K m(r,b)\,\mathrm{d}r = b+1 \tag{7}$$

and that the total length of these pieces is

$$\int_0^K r\,m(r,b)\,\mathrm{d}r = K \tag{8}$$

as expected.

2.3. An alternative derivation of the length distribution after a finite number of breaks

It is instructive to also deduce the above length distribution $m(r,b)$ in equation (6) in a different way. Consider an interval of length K (see figure 1) and let us first focus on the first interval, which is flanked on its right side by the break with the smallest coordinate (denoted by r_1 in the figure). The cumulative probability, that the length of this first interval, r_1, is smaller than a given length r, $\mathrm{prob}(r_1 < r)$, is one minus the probability of all b breaks falling in the interval (r,K) [13]. Therefore

$$\mathrm{prob}(r_1 < r) = 1 - \left(\frac{K-r}{K}\right)^b. \tag{9}$$

The probability density function for the length of the first interval is the derivative of this function with respect to r and therefore:

$$m_1(r,b) = \frac{b}{K}\left(1-\frac{r}{K}\right)^{b-1}. \tag{10}$$

The other b intervals stemming from the b breaks are statistically equivalent to the first one as discussed above. Their length distribution is therefore finally the one given in equation (6). The mean length of a single interval is

$$\bar{r} = \int_0^K r\,m_1(r,b)\,\mathrm{d}r = \frac{K}{b+1} \tag{11}$$

clearly reflecting that all intervals are in fact equivalent. The variance of this distribution can be computed to be

$$\int_0^K r^2\,m_1(r,b)\,\mathrm{d}r - \bar{r}^2 = \frac{K^2}{b+1}\left(\frac{2}{b+2}-\frac{1}{b+1}\right). \tag{12}$$

2.4. Relationship between the two ensembles

The two discussed ensembles can be related to each other. If a stick of length K is dynamically broken by a random process with rate μ per length interval and time as

J. Stat. Mech. (2019) 064003

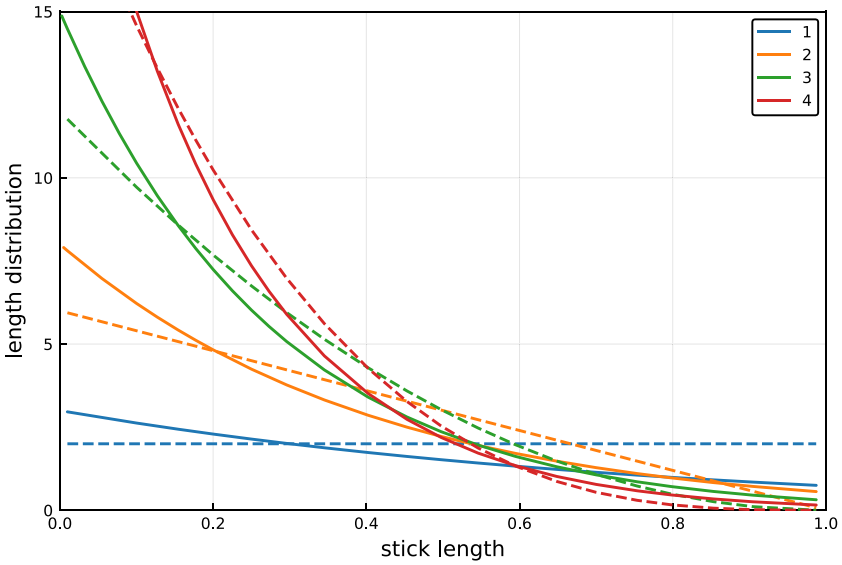

Figure 2. The distributions of stick length $m(r, b)$ for a defined number of breaks (dashed lines) and $m(r, t)$ for given times (continuous lines). The initial interval has length $K = 1$. The number of breaks b for $m(r, b)$ and scaled times $\bar{b} = K\mu t$ for $m(r, t)$ are color coded. Note that curves with the same color have on average equal numbers of breaks.

described above, then the number of breaks b after a given time t follows a Poisson distribution

$$\frac{(\mu K t)^b}{b!} \exp(-\mu K t), \tag{13}$$

where the mean number of breaks is $\bar{b} = \mu K t$ and increases linearly in time. Therefore the length distribution of sticks after time t, $m(r, t)$, and with b breaks, $m(r, b)$ are related by the equation

$$m(r, t) = \sum_{b=0}^{\infty} m(r, b) \frac{(\mu K t)^b}{b!} \exp(-\mu K t), \tag{14}$$

which holds true for the above distributions in equations (2) and (6). The two length distributions for various numbers of breaks or given times are compared in figure 2. The times are conveniently scaled, such that curves of the same color have on average equal number of breaks and therefore the mean length of a single break is equal. Interestingly the curves for the stick-breaking process in continuous time have more weight for smaller stick length r than its counterpart for a defined number of breaks. This is due to the presence of sticks with more than $\bar{b} = K\mu t$, see equation (13), in this ensemble. Similarly, the presence of sticks with less than the mean number of breaks leads to more weight in this distribution for large r once b is larger than one.

2.5. Collections of broken sticks

The random stick-breaking process was previously discussed [12] because it could explain the power-law distribution of exactly matching substrings in genomic sequences

J. Stat. Mech. (2019) 064003

as observed in [14]. A similar power-law distribution with exponent -3 can be found by collecting sticks with different number of breaks, i.e. one with 1 break, one with 2 breaks, one with 3 breaks, and so on. The resulting length distribution is

$$m(r) = \sum_{b=1}^{\infty} m(r,b) = \sum_{b=1}^{\infty} \frac{b(b+1)}{K} \left(1 - \frac{r}{K}\right)^{b-1} = \frac{2K^2}{r^3}. \tag{15}$$

Also higher order power-laws, corresponding to scenarios with different distributions of breaks as described in [15] can be derived

$$\sum_{b=1}^{\infty} b\, m(r,b) = \sum_{b=1}^{\infty} \frac{b^2(b+1)}{K} \left(1 - \frac{r}{K}\right)^{b-1} = \frac{2K^2(3K-2r)}{r^4} \tag{16}$$

or more general $\sum_{b=1}^{\infty} b^{\kappa}\, m(r,b) \sim r^{-(\kappa+3)}$ for large r.

3. Summary

In this article we considered two ensembles of stick-breaking. One, in which a stick is broken with a constant rate over time and one, in which the stick is broken a definite number of times. The first ensemble is often considered describing natural phenomena [3]. However, the second ensemble, where the number of breaks is fixed, is more appropriate in other situations, for instance when considering the differences of a random sample of size b when the samples are arranged in order of their magnitude as already discussed in [13, 16]. Here we relate these two ensembles with each other leading to a more comprehensive understanding of stick-breaking processes.

Acknowledgments

I graciously thank Vladimir Rittenberg for his guidance and his continuous inquiry *'Was ist neu bei dir?'* which still resonates with me and encourages me to explore new phenomena in natural sciences. I am also grateful to Vladimir for supporting a research fellowship at the Scuola Internazionale Superiore di Studi Avanzati in Trieste, Italy, which led me take Italian lessons. In addition, I acknowledge many fruitful discussions with Florian Massip and Misha Sheinman for this project. I would like to thank the Institut National de la Recherche Agronomique, Université Paris-Saclay, Jouy-en-Josas, France and the Broad Institute of MIT and Harvard, Cambridge MA, USA for their hospitality while completing this work.

References

[1] Markus B 2008 *Atomistic Modeling of Materials Failure* (Boston, MA: Springer)
[2] Shukla A 2006 *Dynamic Fracture Mechanics* (Singapore: World Scientific)
[3] Krapivsky P L, Redner S and Ben-Naim E 2010 *A Kinetic View of Statistical Physics* (Cambridge: Cambridge University Press) pp 172–98
[4] Oddershede L, Meibom A and Bohr J 1998 *Europhys. Lett.* **43** 598
[5] MacArthur R H 1957 *Proc. Natl Acad. Sci. USA* **43** 293

J. Stat. Mech. (2019) 064003

[6] Sethuraman J 1994 *Stat. Sin.* **4** 639
[7] Paisley J, Blei D and Jordan M I 2012 *Proc. of the 15th Int. Conf. on Artificial Intelligence, Statistics (La Palma)* ed N D Lawrence and M Girolami pp 850–8 (jmlr.org)
[8] Paisley J, Zaas A, Woods C W, Ginsburg G S and Carin L 2010 *ICML'10 Proc. of the 27th Int. Conf. on Machine Learning (Omnipress, Haifa) ed* J Fürnkranz and T Joachims pp 847–54
[9] Saito O 1958 *J. Phys. Soc. Japan* **13** 198
[10] Ziff R M and McGrady E D 1985 *J. Phys. A: Math. Gen.* **18** 3027
[11] Oddershede L, Dimon P and Bohr J 1993 *Phys. Rev. Lett.* **71** 3107
[12] Massip F and Arndt P F 2013 *Phys. Rev. Lett.* **110** 148101
[13] Pyke R 1965 *J. R. Stat. Soc.* B **27** 395
[14] Gao K and Miller J 2011 *PloS One* **6** e18464
[15] Massip F, Sheinman M, Schbath S and Arndt P F 2015 *Mol. Biol. Evol.* **32** 524
[16] Pearson K 1902 *Biometrika* **1** 390

J. Stat. Mech. (2019) 064003