ACCELERATION OF A PROCEDURE TO GENERATE FRACTAL CURVES OF A GIVEN DIMENSION THROUGH THE PROBABILISTIC CHARACTERIZATION OF EXECUTION TIME

MANUEL CEBRIAN
Manuel.Cebrian@ii.uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid

ALFONSO ORTEGA
Alfonso.Ortega@ii.uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid

MANUEL ALFONSECA
Manuel.Alfonseca@ii.uam.es
Escuela Politécnica Superior
Universidad Autónoma de Madrid

ABSTRACT
In a previous work [Ortega *et al.* 03], the authors have described the use of grammatical evolution to automatically generate L Systems (LS) representing fractal curves with a pre-determined fractal dimension. The experiments presented in this paper prove that the efficiency of this procedure is variable, with very different execution times for different executions with the same fractal dimension target. The paper shows that the probabilistic distribution of execution times belongs to a well-known family of random variables: heavy-tailed distributions. This analysis explains the erratic performance of the algorithm and suggests the use of a technique that corrects this variability and improves the efficiency about one order of magnitude.
Acknowledgements: *This paper has been sponsored by the Spanish Ministry of Science and Technology, project numbers TIC2001-0685-C02-01 and TIC 2002-01948.*

INTRODUCTION

Our procedure to generate fractal curves with a required dimension consists of three parts: a) representation of fractals by means of L Systems (LS); b) computation of the fractal dimension from the grammar; c) application of a grammar-evolution based genetic algorithm to get a grammar representing a fractal with the required dimension.

LS provide a powerful tool to represent fractals in the class of recursive transformations. The iterator may be represented by means of production rules, while the initiator corresponds to the axiom. The fractal curve is generated by the sequence of words derived from the axiom, by means of a *representation scheme: vector graphics or turtle graphics*. In a previous work [Alfonseca and Ortega 01] we have described an algorithm that estimates the fractal dimension of a non-trivial set of these fractals from their equivalent LS by means of symbolic manipulation, without the need of graphical procedures.

In [Ortega *et al.* 03] we applied Grammar Evolution (GE) [O'Neill and Ryan 2001] to obtain the LS equivalent to a fractal with the required dimension. The proposed procedure has a clear biological inspiration acting on three different levels: a genotype (a vector of integers) an intermediate level, equivalent to proteins (LS) and a phenotype (the fractal curve). See Figure 1.

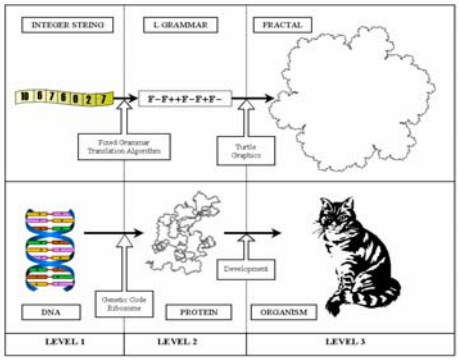

Figure 1. Parallels between our GE approach and biological evolution.

Finding by hand one fractal of a given dimension is easy. Our algorithm, however, generates an arbitrary number of different fractals with a required dimension, though with a widely varying efficiency [Ortega *et al.* 03]. Sometimes the objective is reached in the first generation. For less standard dimensions, the number of generations required is usually large, sometimes extremely large. Moreover, the standard deviation of the number of generations required (in separate execution runs, starting with different random seeds) is very large (see Table 1). This means that the performance of our implementation is very variable and may be quite low.

Table 1. Number of generations to reach the target in a set of GE tests.

Dimension	Angle	# tests	# generations to reach target
1.1	60	4	119 to 72122
1.2	45	8	188 to 11173
1.3	45	9	50 to 18627
1.25	60	15	1 to 2422
1.2618595...	60	4	1 to 2
1.4	60	10	33 to 1912
1.5	45	11	52 to 11138
1.6	45	5	275 to 3944
1.7	60	8	18 to 1221
1.8	60	13	69 to 3659
2	45	5	1

In the current work, we consider the random variable G_{dim} (the number of generations needed to generate a fractal with dimension *dim*). We show that G_{dim} belongs to a well-known class of moment-less distributions (they have infinite average and variance): the class of heavy-tailed (HT) distributions. This is the reason of the great deviation in execution times and of the loss in performance. First we present a few experimental evidences that gave rise to this study. Then we apply existing techniques, described in the statistical literature, to test our hypothesis that G_{dim} is HT. Finally, we will modify our algorithm by

introducing *re-starts* in the executions, thus reducing significantly the variance and increasing the performance of the algorithm about one order of magnitude.

HT DISTRIBUTIONS

HT distribution have asymptotic tails of the Pareto-Lévy form:

$$P\{X > x\} \sim C.x^{-\alpha},\ x > 0 \tag{1}$$

where α is a constant (For HT distributions, $0 < \alpha < 2$). In our case, X is $G_{\dim}$ and x is a positive numeric value (a number of generations). In this way, $P\{G_{\dim} > x\}$ is interpreted as the probability that the algorithm requires more than x generations to get its goal. The symbol C in the expression is a normalizing constant to make $P\{0 < G_{\dim} \leq \infty\} = 1$.

We can compare this *polynomic* decay of probability with that of a standard normal curve, which is *exponential*:

$$P\{X > x\} \sim \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \tag{2}$$

Constant α is called *stability index*, and determines the existence of the distribution moments, since it is possible to prove that:

$$\alpha = \sup\left\{b > 0 : E\left|X^b\right| < \infty\right\} \tag{3}$$

This means that moments with exponent less than α are finite, while those greater or equal to α are infinite. The average is the moment with exponent one, thus if $\alpha \leq 1$ the random variable X has no average (it is infinite) or higher order moments (such as variance). The variance is a function of the moment with exponent two, thus if $1 < \alpha < 2$ random variable X does have an average, but not a variance. For a detailed treatment of the previous considerations, see [Samorodnitsky *et al.* 94, Zolotarev 86].

EXPERIMENTAL EVIDENCE

In this section we show experimental evidence that distribution $G_{\dim}$ is HT. This implies that the probability of extremely large execution times is very large in comparison with what would happen with a standard distribution.

First we observe the behavior of a few statistics for several values of *dim*. Then we show a Table that proves that the standard deviations of $G_{\dim}$ are greater than their averages (HT distributions have no standard deviation (nor average, in some cases), so when we mention these statistic measurements, they are assumed to refer to the sample).

Table 2. Summary of averages and standard deviations for several target dimensions, computed with a sample of 1000 experiments.

	1.3	**1.5**	**1.8**
Mean	1919	1560	2815
standard deviation	1898	1660	1866

This algorithmic phenomenon is atypical, and evidences that execution time depends strongly on the random seed. This behavior is typical of HT distributions. We have mentioned that HT distributions usually have no variance, (it is infinite). A way to test this is by checking its convergence speed.

Figure 2 shows the convergence of the variance of the sample for several values of *dim*. We observe that the sample variance has a strong fluctuation in all dimensions and does not converge when the size of the sample is increased, or at least it converges much more slowly than a normal distribution. This increases our suspicion that our data have an infinite variance.

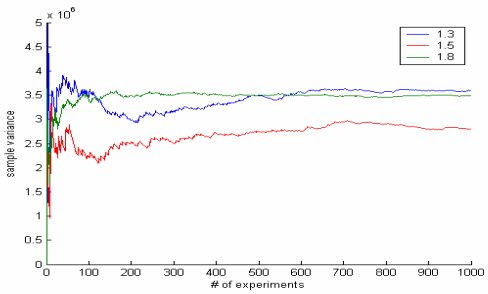

Figure 2. Variance for dim, showing its slow speed of convergence.

To complete our visual evidence, Figure 3 shows the cumulative probabilities for several values of *dim*. Looking also at Table 2, we observe that the probability that an execution takes longer than five times its standard deviation is over 10%, which implies that executions much longer than the average have a probability far higher than what would be expected for standard distributions.

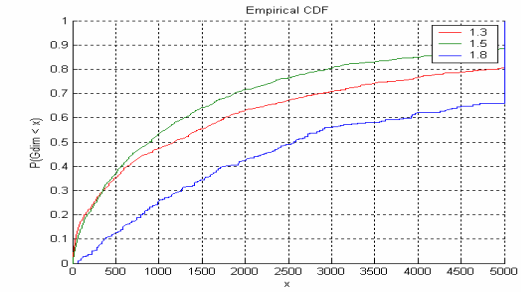

Figure 3. Empirical cumulative density function for several dimensions on samples with size 1000. After 5000 generations the execution was stopped.

ESTIMATION OF THE INDEX OF STABILITY

Once convinced experimentally that G_{dim} is a HT distribution, we estimate its *stability index* α using Hall's method [Hall 1982], a maximum likelihood estimator. Let $X_{n,1} \le X_{n,2} \le \ldots \le X_{n,n}$ be the ordered values (numbers of generations) in n experiments. Let $r < n$ be a truncation value used to take into account only the extreme observations. We get the following estimator:

Hall also determined the optimal value of r, but it is a function of unknown parameters of the distribution, therefore in practice values in the range $[n/4,\ n/3]$ are used. Table 3 shows our estimations of the stability indexes for the extreme values of the recommended interval and helps to give strength to our hypothesis: even testing the extreme values in the range, we still obtain values of α <2 in every case, which is consistent with the definition. Almost all these values are less than 1, which means that G_{dim} neither has a finite average.

Table 3. Estimation using Hall's method of the stability index of the distribution. The sample size used was 1000.

	***n*/3**	***n*/4**
1.3	0.626	0.511
1.5	0.917	0.800
1.8	1.078	0.854

EXPLOITING HT BEHAVIOUR

In HT distributions, events very far from the average have non-negligible probability and therefore should be taken into account. Since G_{dim} is HT, extremely long executions could happen. Even more important: the fact that execution has taken up to now a high number of generations (with respect to the average) gives no assurance that the end is near, however much we wait.

Looking at Figure 3, we can also reach the conclusion that the probability of executions below the average is great (over 50%). From both these two ideas we can deduce that we shouldn't be too patient with long executions, as there is a high probability of finding a short execution if we try again. This is done by choosing a threshold U in the number of generations; once it has been crossed, the algorithm is *re-started*. The new execution is different from the previous one, as it will use a different sequence of random numbers.

The optimal value for U can be determined if the distribution is known in analytical form, but in our case we only have experimental data and we have determined it empirically. To do this, we have repeated the experiments with different threshold values.

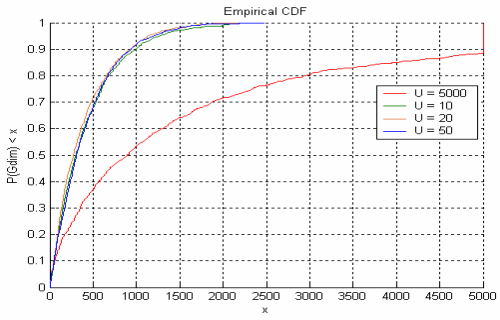

Figure 4. Cumulated density functions for several values of the threshold, with 1.5 fractal dimension goal. The HT nature disappears for low values of U.

In Figure 4 one can see that, for $U = 5000$, over 10% executions have not finished after 5000 generations. For $U = 10$, however, 90% executions are ended in less than 1000 generations. If we compute the *stability index* α for $U = 10$, we find it in the interval [2.01, 2.35], always greater than two, which proves empirically that the random variable $G_{\dim}$ is no longer HT.

Analyzing the 1.3 fractal dimension experiment, we find that the average number of generations for $U = 5$ is 228, while the average without re-starts (i.e. with a re-start threshold of 5000) is 1919, which means an improvement of about one order of magnitude in the execution time. We also saw that all the executions end before 1200 generations, which entails a big variance improvement (213 versus 1898 when measuring standard deviations).

CONCLUSIONS AND FUTURE RESEARCH

This paper presents a technique based on a probabilistic analysis to increase the performance and reduce the variance of an algorithm based on GE. The technique described transforms a HT random variable into another with standard distribution, by introducing filters that eliminate the heavy events (executions) and only allow the lighter ones. This technique may be used in different fields, whenever a similar probabilistic behavior of the execution time is detected.

At this point we are interested in finding the formal base that explains the HT properties of the random variable representing the execution time. We want to find if this behavior is something typical of the combinatorial problems explored by GE. [Mandelbrot 60] suggests that a HT behavior indicates that the search space is *self-similar*, independently of the search algorithm.

We intend to test whether the fact that our variable is HT depends on the topology of the search space for the problem under analysis, or is a consequence of its having been explored with GE. More work is needed for this: we must find problems solved by means of grammatical evolution that may be solved with other techniques, and study their execution time. In this way, we would add some formalism to this new branch in the family of genetic algorithms.

REFERENCES

M. Alfonseca and A. Ortega, “Determination of Fractal Dimensions from Equivalent L Systems”, *IBM J. Res. & Dev.* 45, No. 6, 797– 805 (2001).

Hall, P.: On some simple estimates of an exponent of regular variation, *J. Roy. Statist. Soc.* 44 (1982), 37–42.

Mandelbrot, Benoit, B. (1960). The Pareto-Lévy law and the distribution of income. *International Economic Review* 1, 79–106.

M. O’Neill and C. Ryan, “Grammatical Evolution”, *IEEE Trans. Evolutionary Computation* 5, No. 4, 349 –358 (2001).

A.Ortega, A.Abu Dalhoum, M.Alfonseca: Grammatical evolution to design fractal curves with a given dimension, *IBM Journal of Research and Development*, Vol. 47:4, p. 483-493, Jul. 2003.

M. O'Neill and C. Ryan, “Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language”**,** Kluwer Academic Publishers, 2003, ISBN 1-4020-7444-1.

Samorodnitsky, Gennady and Taqqu, Murad S. (1994). Stable Non-Gaussian Random Processes: Stochastic Models with In-finite Variance, *Chapman and Hall*, New York.

Zolotarev, V.: One-Dimensional Stable Distributions, Transl. Math. Monographs 65, *Amer. Math. Soc.*, 1986. Translation from the original 1983 Russian ed.