Entropy and Divergence Statistics for Markovian Processes and Related Structural and Statistical Properties of Pseudorandom Number Generators

Habilitationsschrift
zur Erlangung der Venia Docendi für

Mathematik
unter besonderer Berücksichtigung der
Wahrscheinlichkeitstheorie und Statistik

an der Naturwissenschaftlichen Fakultät
der Paris–Lodron–Universität Salzburg

eingereicht von
Stefan Wegenkittl

Salzburg, im Februar 2003

Für Alex

> Liebe ist stets der Anfang des Wissens, so wie Feuer der Anfang des Lichts ist. (Thomas Carlyle)

Foreword

The presented cumulative habilitation thesis consists of 9 articles in the order given below. All articles have been published in international mathematical journals or in (strictly refereed) proceedings of important international conferences dealing with the topic of statistical tests for randomness. The numbering scheme is identical to that in my CV. The list is followed by a short introduction to the main topics and a description of the author's contributions with respect to the single articles.

List of Publications:

[3] S. Wegenkittl. A generalized ϕ-divergence for asymptotically multivariate normal models. *Journal of Multivariate Analysis*, **83**:288–302, 2002.

[4] S. Wegenkittl. Entropy estimators and serial tests for ergodic chains. *IEEE Transactions on Information Theory*, **47**(6):2480–2489, Sep 2001.

[5] S. Wegenkittl. Gambling tests for pseudorandom number generators. *Mathematics and Computers in Simulation*, **55**(1–3):281–288, 2001.

[7] P. L'Ecuyer, R. Simard, and S. Wegenkittl. Sparse Serial Tests of Uniformity for Random Number Generators. *SIAM Journal on Scientific Computing*, **24**(2):652–668, 2002.

[8] S. Wegenkittl and M. Matsumoto. Getting rid of correlations among pseudorandom numbers: Discarding versus tempering. *ACM Transactions on Modeling and Computer Simulation*, **9**(3):282–294, 1999.

[9] S. Wegenkittl. Are there hyperbola in the scatter plots of inversive congruential pseudorandom numbers? *J. Computational And Applied Mathematics*, **95**(1-2):117–125, 1998.

[11] H. Leeb and S. Wegenkittl. Inversive and linear congruential pseudorandom number generators in empirical tests. *ACM Transactions on Modeling and Computer Simulation*, **7**(2):272–286, 1997.

[13] S. Wegenkittl. Monkeys, gambling, and return times: Assessing pseudorandomness. In P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 625–631. IEEE Press, 1999.

[16] J. Eichenauer-Herrmann, E. Herrmann, and S. Wegenkittl. A survey of quadratic and inversive congruential pseudorandom numbers. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996*, number 127 in Lecture Notes in Statistics, pages 66–97. Springer, New York, 1997.

The main mathematical results covered in the first part of this habilitation thesis are grouped around the notion of entropy of discrete time and space Markov chains and around statistics for measuring entropy or related properties in such processes.

In [3], I have established a connection between two previously separated lines of research: quadratic forms in weak inverses of asymptotically multivariate normal distributed vectors are joined with the concept of ϕ-divergence statistics for multinomial distributed data. As a result, the *generalized ϕ-divergence* is introduced and the asymptotic distribution under multivariate normal models is studied. This new class is shown to comprise many well-known goodness-of-fit statistics for multinomial data, that is, for data arising from counting occurrences of events in i.i.d. processes. Similar analysis can now be applied to the frequency count of visits to states of a discrete ergodic Markov chain not belonging to the class of i.i.d. processes. As an example, the generalized I-divergence for a two-state Markov chain is discussed.

The process of overlapping d-tuples of successive states of an order κ Markov chain (if $\kappa \leq d$, such a process is an ordinary Markov chain by itself) deserves special attention. If the underlying chain is an independent process ($\kappa = 0$), two essentially different kind of asymptotics can be studied, namely the *dense* and the *sparse* case. The corresponding asymptotics of so-called *overlapping ϕ-divergences* and of several related statistics are considered in [7] which also contains an empirical study done by R. Simard and P. L'Ecuyer.

If the underlying process is an arbitrary ergodic chain with $\kappa \leq d$, we can use the same type of statistics in order to estimate the entropy of the chain. This entropy estimator is established in [4] and related to estimators based on return-time statistics. In addition, an open conjecture of U. Maurer on the convergence of the expected logarithm of the return-time is corrected and proved.

Article [5] is an extension of the work in [3] and [7]. It introduces the notion of a *Gambling Test* which can be viewed as a tricky implementation of the divergence statistic of a linear mapping of a (very) high-dimensional counter vector. The covariances needed for the parameterization of the test statistic are derived. As an example, the application to a sequence of random numbers is considered, the corresponding covariances and the values of the test statistics are calculated numerically by the aid of *Mathematica* and `C++` programs.

All these theoretical results, which belong to the field of mathematical statistics, have successfully been applied in the empirical assessment of pseudorandomness. Pseudorandom number generators (PRNGs) are deterministic algorithms that generate sequences of numbers which are used in the place of samples from various stochastic processes. Such generators play an important role in stochastic simulation, (Quasi) Monte Carlo methods, and randomized algorithms. These are situations where one wants to simulate a stochastic experiment in a fast, reliable, and repeatable way.

Assessment of the desired structural and statistical properties of PRNGs aims at revealing (unwanted) correlations which could bias the results in an application. For a good PRNG, it should practically be impossible to assess the deterministic origin of the numbers without exhausting large parts of the period (deterministic PRNGs are *finite* state machines and therefore yield semi-periodic output). The past three decades brought ever more demanding applications so that both, the PRNGs themselves and the structural and statistical assessment procedures, had to be refined. The following articles in the second part of this thesis all contain related contributions.

Articles [8] and [9] analyze structural properties of inversive and sparse-linear generators. Scatter plots of inversive pseudorandom numbers show hyperbola-like clusters of points. In [9] this most eye-catching structural element is analytically described. [8] addresses a problem of several modern fast huge-period linear generators: due to implementational considerations, only a few bits of the state space determine the next state. This introduces a correlation structure which may easily show up in rather simple simulations.

The defect and two solutions to the problem (discarding and tempering) are discussed.

In the invited overview article [13], the gap between theoretical results and practical application is addressed by a taxonomy of tests and generators. The article summarizes practical observations and experiences made during the assessment of several families of PRNGs.

More detailed and specific results are contained in the articles [7], [8], [11], and [16], which address the empirical assessment of PRNGs using divergence and return-time based statistics. The main work in these articles is the translation of theoretical results (such as the asymptotics of statistics and the structural properties of PRNGs) into practical test designs and into valid interpretation of the results. Article [11] is the first systematic study of differences of linear and inversive methods with respect to their ability to mimic random point processes in low-dimensional hypercubes. Similar methods have been applied to quadratic and some additional inversive PRNGs in [16]. This line of research is continued in [7], where a systematic relationship between the period length of the generator and the number of samples that can safely be used in stochastic simulation is established for several families of generators. In [8], empirical testing with the *Gambling Test* is used to reveal the impact of the theoretically foreseen defect of the PRNG in a stochastic simulation. Also, the improvement of the statistical quality by the tempering method is shown.

The numerical results have been computed with a `C++` software package which was co-developed by P. Hellekalek, H. Leeb, K. Entacher, O. Lendl, G. Wesp, and the author.

Acknowledgements:

The author's research was supported by the Austrian Science Fund (FWF), project no. S8303-MAT. The scientific environment at the University of Salzburg was the pLab research group directed by Peter Hellekalek and the scientific work of Ferdinand Österreicher, who introduced me to the notion of divergence statistics. Thank you, Peter, for making everything possible, and thank you, Ferdinand, for your support!