

Recurrence and Waiting Times in Stationary Processes, and Their Applications in Data Compression

Author: **Ioannis Kontoyiannis**

Technical Report number (Dept. of Statistics, Stanford University): **1998-7**

Date: **May 1998**

Abstract:

Over the past 25 years, the practical requirement for efficient data compression algorithms has generated a large volume of research covering the whole spectrum from practically implementable algorithms to deep theoretical results. One prominent example is the Lempel-Ziv algorithm for lossless data compression: Not only is it implemented on most computers used today, but also, attempts to analyze its performance have provided new problems in probability, information theory and ergodic theory, whose solutions reveal a series of interesting results about the entropy and the recurrence structure of stationary processes.

The main problems considered in this thesis are those of determining the asymptotic behavior of waiting times and recurrence times in stationary processes. These questions are motivated primarily by their important applications in data compression and the analysis of string matching algorithms in DNA sequence analysis. In particular, solving the waiting times problem also allowed us to solve a long-standing open problem in data compression: That of finding a practical extension of the Lempel-Ziv coding algorithm for lossy compression.

This thesis is divided into three parts. In the first part we generalize one of the central theoretical results in source coding theory: We prove a natural generalization of the celebrated Shannon-McMillan-Breiman theorem (as well as its subsequent refinements by Ibragimov and by Philipp and Stout) for real-valued processes and for the case when distortion is allowed. These results are inspired by, and provide the key technical ingredient in, our asymptotic analysis of recurrence and waiting times, in the second part. The main probabilistic tools used in establishing them are uniform almost-sure approximation, powerful techniques from large deviations, and classical second-moment blocking arguments.

In the second part we consider the problem of waiting times between stationary processes. We show that waiting times grow exponentially with probability one and, that their rate is given by the solution to an explicit variational problem in terms of the entropies of the underlying processes. Moreover, we show that, properly scaled, the deviations of the waiting times from their limiting exponent are asymptotically Gaussian (with a limiting variance explicitly identified), and we prove finer theorems (e.g., a law of the iterated logarithm and an almost sure invariance principle) that provide the exact rate of convergence in the above limit theorems. Corresponding results are proved for recurrence times, and dual results are stated and proved for certain longest-match lengths

between stationary processes.

Finally, in the third part, we use the insight gained by the waiting times results to find a practical extension of the Lempel-Ziv scheme for the case of lossy data compression. We propose a new lossy version of the so-called Fixed-Database Lempel-Ziv coding algorithm, which is of complexity "comparable" to that of the corresponding lossless scheme, and we prove that its compression performance is (asymptotically) optimal.