Acknowledgments

First and foremost, I would like to acknowledge my debt to my advisor, Professor Tom Cover. Being part of his research group during my five years at Stanford has influenced me deeply. I am grateful to him for creating an environment of unusual intellectual breadth, constant curiosity and challenge.

During the past two years I was very fortunate also to work closely with Professor Amir Dembo. Our frequent interactions and our later collaboration were very valuable learning experiences, and I am grateful for them. I would also like to thank Professors Mike Harrison and George Papanicolaou for being on my Orals committee and for all the help they offered me along the way.

My first classes in probability at Stanford were taught by Professor Tze Lai. His enthusiasm for the subject and his later interest (and encouragement) for my work are greatly appreciated.

Even when it did not have a specific purpose, informal interaction with several members of the faculty was a major source of learning and new ideas. In particular, Professors George Papanicolaou and Sam Karlin were always very generous with their time and knowledge.

Very special thanks to Professor Yurii Suhov of Cambridge University. During my Master's year he introduced me to the *real* world of mathematics, and, since then, he has remained an important force of motivation and a deep source of mathematical insight. Thanks also to Stephen Souras for turning my career around by introducing me to Shannon's work; more importantly, thanks for being the great friend he is.

My office-mates and my fellow ISL graduate students played an essential role in my education. I learnt a great deal from Paul Algoet, Vittorio Castelli, Tom Chung, Elza Erkip, Paul Fahn, Garud Iyengar, Don Kimber, Amos Lapidoth, Costis Maglaras, Erik Ordentlich, Joshua Singer, Rick Wesel and Assaf Zeevi. Especially I

want to thank Jim Hwang for helping render my first few months at Stanford tolerable. And many thanks to Denise Cuevas, for single-handedly managing to keep the whole process running smoothly.

I was very lucky to make very good friends while at Stanford. None of this would have been worth it without them. Hilary Teplitz, Christos Tryfonas and Vassilis Vassalos have won their place in these acknowledgments by exhibiting (well, occasionally, at least!) super-human amounts of patience and tolerance for my moods.

I owe very much to my family and my parents. Their love, support and encouragement have been unconditional, constant and boundless. I cannot imagine a better gift than what they have given me.

Financial support for the work reported in this thesis was provided, in part, by the following grants: NSF #NCR-9628193, JSEP #DAAH04-94-G-0058 and ARPA #J-FBI-94-218-2.

This thesis is dedicated to the precious memory of my mother.

Λάμπει μέσα της κείνο που αγνοεί. Μα ωστόσο λάμπει.

Contents

Abstract **v**

Acknowledgments **vii**

1 Introduction **1**

1.1 The Question of Recurrence 2

1.1.1 Recurrence and Entropy 3

1.1.2 Second-Order Results 5

1.1.3 Recurrence and Data Compression 6

1.2 Three More Questions 7

1.2.1 Waiting Times 7

1.2.2 Lossy Data Compression 9

1.2.3 Match Lengths and DNA Template Matching 11

1.3 History 12

1.4 About This Thesis 14

1.4.1 Theory and Applications 14

1.4.2 Organization 14

1.5 Notation 15

2 The Shannon-McMillan-Breiman Theorem, Generalizations, and Refinements **19**

2.1 Known Results 19

2.2 Allowing Distortion 24

2.3 Large Deviations 29

2.4 Uniform Approximation 33

3 Recurrence in Stationary Processes **35**
3.1 Introduction and Main Results 35
3.1.1 Match Lengths 39
3.1.2 Earlier Work 39
3.2 Strong Approximation 40

4 Waiting Times Between Stationary Processes **43**
4.1 Motivation 43
4.1.1 Earlier Work 44
4.1.2 The Strong Approximation Framework 45
4.2 Waiting Times Results 46
4.2.1 Waiting Times With No Distortion 46
4.2.2 Waiting Times Between Different Processes 48
4.2.3 Waiting Times Allowing Distortion 50
4.3 Match Lengths Results 52
4.4 Strong Approximation 56
4.5 Duality: Match Lengths 57

5 Efficient, Universal, Lossy Data Compression **61**
5.1 Introduction: Data Compression 62
5.2 Lempel-Ziv Coding and Waiting Times 65
5.2.1 The Idealized Coding Scenario 65
5.2.2 Waiting Times with Multiple Databases 67
5.3 Description of the Algorithm 69
5.3.1 The Algorithm 70
5.4 Algorithm Optimality 74
5.5 Redundancy, Complexity, Implementation 76
5.5.1 The Complexity-Redundancy Trade-off 76
5.5.2 Implementation and Simulation Results 77
5.6 Extensions 78
5.7 Proofs 82
5.7.1 Proof of Lemma 5.1 82
5.7.2 Proof of Lemma 5.2 83
5.7.3 Proof of Corollary 5.3 85

5.7.4 Proof of Theorem 5.2 86

6 Concluding Remarks **89**

6.1 Summary of Contributions 89

6.2 Extensions and Future Directions 91

A Some Technical Points **93**

A.1 Proof of Theorem 2.3 93

A.2 Proof of Lemma 2.1 95

A.3 Proof of Proposition 2.2 97

A.4 Choice of $s(n)$-types 98

Bibliography **99**