

36-754, Advanced Probability II
or
Almost None of the Theory of Stochastic
Processes

Cosma Shalizi

Spring 2007

Contents

Table of Contents	i
Comprehensive List of Definitions, Lemmas, Propositions, Theorems, Corollaries, Examples and Exercises	xxiii
Preface	1
I Stochastic Processes in General	2
1 Basics	3
1.1 So, What Is a Stochastic Process?	3
1.2 Random Functions	5
1.3 Exercises	7
2 Building Processes	15
2.1 Finite-Dimensional Distributions	15
2.2 Consistency and Extension	16
3 Building Processes by Conditioning	21
3.1 Probability Kernels	21
3.2 Extension via Recursive Conditioning	22
3.3 Exercises	25
II One-Parameter Processes in General	26
4 One-Parameter Processes	27
4.1 One-Parameter Processes	27
4.2 Operator Representations of One-Parameter Processes	31
4.3 Exercises	31
5 Stationary Processes	34
5.1 Kinds of Stationarity	34

5.2	Strictly Stationary Processes and Measure-Preserving Transformations	35
5.3	Exercises	37
6	Random Times	38
6.1	Reminders about Filtrations and Stopping Times	38
6.2	Waiting Times	39
6.3	Kac's Recurrence Theorem	41
6.4	Exercises	45
7	Continuity	46
7.1	Kinds of Continuity for Processes	46
7.2	Why Continuity Is an Issue	48
7.3	Separable Random Functions	50
7.4	Exercises	51
8	More on Continuity	52
8.1	Separable Versions	52
8.2	Measurable Versions	57
8.3	Cadlag Versions	58
8.4	Continuous Modifications	59
III	Markov Processes	60
9	Markov Processes	61
9.1	The Correct Line on the Markov Property	61
9.2	Transition Probability Kernels	62
9.3	The Markov Property Under Multiple Filtrations	65
9.4	Exercises	67
10	Markov Characterizations	70
10.1	Markov Sequences as Transduced Noise	70
10.2	Time-Evolution (Markov) Operators	72
10.3	Exercises	76
11	Markov Examples	78
11.1	Probability Densities in the Logistic Map	78
11.2	Transition Kernels and Evolution Operators for the Wiener Process	80
11.3	Lévy Processes and Limit Laws	82
11.4	Exercises	86
12	Generators	88
12.1	Exercises	93

13 Strong Markov, Martingales	95
13.1 The Strong Markov Property	95
13.2 Martingale Problems	96
13.3 Exercises	98
14 Feller Processes	99
14.1 Markov Families	99
14.2 Feller Processes	100
14.3 Exercises	106
15 Convergence of Feller Processes	107
15.1 Weak Convergence of Processes with Cadlag Paths (The Skorohod Topology)	107
15.2 Convergence of Feller Processes	109
15.3 Approximation of Ordinary Differential Equations by Markov Processes	111
15.4 Exercises	113
16 Convergence of Random Walks	114
16.1 The Wiener Process is Feller	114
16.2 Convergence of Random Walks	116
16.2.1 Approach Through Feller Processes	117
16.2.2 Direct Approach	119
16.2.3 Consequences of the Functional Central Limit Theorem	120
16.3 Exercises	121
IV Diffusions and Stochastic Calculus	123
17 Diffusions and the Wiener Process	124
17.1 Diffusions and Stochastic Calculus	124
17.2 Once More with the Wiener Process and Its Properties	126
17.3 Wiener Measure; Most Continuous Curves Are Not Differentiable	127
18 Stochastic Integrals Preview	130
18.1 Martingale Characterization of the Wiener Process	130
18.2 A Heuristic Introduction to Stochastic Integrals	131
19 Stochastic Integrals and SDEs	133
19.1 Integrals with Respect to the Wiener Process	133
19.2 Some Easy Stochastic Integrals, with a Moral	138
19.2.1 $\int dW$	138
19.2.2 $\int W dW$	139
19.3 Itô's Formula	141
19.3.1 Stratonovich Integrals	146
19.3.2 Martingale Representation	146
19.4 Stochastic Differential Equations	147

19.5	Brownian Motion, the Langevin Equation, and Ornstein-Uhlenbeck Processes	153
19.6	Exercises	155
20	More on SDEs	158
20.1	Solutions of SDEs are Diffusions	158
20.2	Forward and Backward Equations	160
20.3	Exercises	163
21	Small-Noise SDEs	164
21.1	Convergence in Probability of SDEs to ODEs	165
21.2	Rate of Convergence; Probability of Large Deviations	166
22	Spectral Analysis and L_2 Ergodicity	170
22.1	White Noise	171
22.2	Spectral Representation of Weakly Stationary Processes	174
22.2.1	How the White Noise Lost Its Color	180
22.3	The Mean-Square Ergodic Theorem	181
22.3.1	Mean-Square Ergodicity Based on the Autocovariance	181
22.3.2	Mean-Square Ergodicity Based on the Spectrum	183
22.4	Exercises	184
V	Ergodic Theory	186
23	Ergodic Properties and Ergodic Limits	187
23.1	General Remarks	187
23.2	Dynamical Systems and Their Invariants	188
23.3	Time Averages and Ergodic Properties	191
23.4	Asymptotic Mean Stationarity	194
23.5	Exercises	197
24	The Almost-Sure Ergodic Theorem	198
25	Ergodicity and Metric Transitivity	204
25.1	Metric Transitivity	204
25.2	Examples of Ergodicity	206
25.3	Consequences of Ergodicity	207
25.3.1	Deterministic Limits for Time Averages	208
25.3.2	Ergodicity and the approach to independence	208
25.4	Exercises	209
26	Ergodic Decomposition	210
26.1	Preliminaries to Ergodic Decompositions	210
26.2	Construction of the Ergodic Decomposition	212
26.3	Statistical Aspects	216
26.3.1	Ergodic Components as Minimal Sufficient Statistics	216

26.3.2	Testing Ergodic Hypotheses	218
26.4	Exercises	219
27	Mixing	220
27.1	Definition and Measurement of Mixing	221
27.2	Examples of Mixing Processes	223
27.3	Convergence of Distributions Under Mixing	223
27.4	A Central Limit Theorem for Mixing Sequences	225
VI	Information Theory	227
28	Entropy and Divergence	228
28.1	Shannon Entropy	229
28.2	Relative Entropy or Kullback-Leibler Divergence	231
28.2.1	Statistical Aspects of Relative Entropy	232
28.3	Mutual Information	234
28.3.1	Mutual Information Function	234
29	Rates and Equipartition	236
29.1	Information-Theoretic Rates	236
29.2	Asymptotic Equipartition	239
29.2.1	Typical Sequences	242
29.3	Asymptotic Likelihood	243
29.3.1	Asymptotic Equipartition for Divergence	243
29.3.2	Likelihood Results	243
29.4	Exercises	243
VII	Large Deviations	245
30	Large Deviations: Basics	246
30.1	Large Deviation Principles: Main Definitions and Generalities . .	246
30.2	Breeding Large Deviations	250
31	IID Large Deviations	256
31.1	Cumulant Generating Functions and Relative Entropy	257
31.2	Large Deviations of the Empirical Mean in \mathbb{R}^d	260
31.3	Large Deviations of the Empirical Measure in Polish Spaces . . .	263
31.4	Large Deviations of the Empirical Process in Polish Spaces . . .	264
32	Large Deviations for Markov Sequences	266
32.1	Large Deviations for Pair Measure of Markov Sequences	266
32.2	Higher LDPs for Markov Sequences	270

33 The Gärtner-Ellis Theorem	271
33.1 The Gärtner-Ellis Theorem	271
33.2 Exercises	275
34 Freidlin-Wentzell Theory	276
34.1 Large Deviations of the Wiener Process	277
34.2 Large Deviations for SDEs with State-Independent Noise	282
34.3 Large Deviations for State-Dependent Noise	283
34.4 Exercises	284
Bibliography	284

Definitions, Lemmas, Propositions, Theorems, Corollaries, Examples and Exercises

Chapter 1 Basic Definitions: Indexed Collections and Random Functions	3
Definition 1 A Stochastic Process Is a Collection of Random Variables	3
Example 2 Random variables	4
Example 3 Random vector	4
Example 4 One-sided random sequences	4
Example 5 Two-sided random sequences	4
Example 6 Spatially-discrete random fields	4
Example 7 Continuous-time random processes	4
Example 8 Random set functions	4
Example 9 One-sided random sequences of set functions	4
Example 10 Empirical distributions	5
Definition 11 Cylinder Set	5
Definition 12 Product σ -field	6
Definition 13 Random Function; Sample Path	6
Definition 14 Functional of the Sample Path	6
Definition 15 Projection Operator, Coordinate Map	6
Theorem 16 Product σ -field-measurability is equivalent to measurability of all coordinates	6
Definition 17 A Stochastic Process Is a Random Function	7
Corollary 18 Measurability of constrained sample paths	7
Example 19 Random Measures	7
Example 20 Point Processes	7
Example 21 Continuous random processes	7
Exercise 1.1 The product σ -field answers countable questions	7

Exercise 1.2 The product σ -field constrained to a given set of paths	7
--	---

Chapter 2 Building Infinite Processes from Finite-Dimensional Distributions

Definition 22 Finite-dimensional distributions	15
Theorem 23 Finite-dimensional distributions determine process distributions	16
Definition 24 Projective Family of Distributions	17
Lemma 25 FDDs Form Projective Families	17
Proposition 26 Randomization, transfer	18
Theorem 27 Daniell Extension Theorem	18
Proposition 28 Carathéodory Extension Theorem	19
Theorem 29 Kolmogorov Extension Theorem	19

Chapter 3 Building Infinite Processes from Regular Conditional Probability Distributions

Definition 30 Probability Kernel	21
Definition 31 Composition of probability kernels	22
Proposition 32 Set functions continuous at \emptyset	23
Theorem 33 Ionescu Tulcea Extension Theorem	23
Exercise 3.1 Lomnick-Ulam Theorem on infinite product measures	25
Exercise 3.2 Measures of cylinder sets	25

Chapter 4 One-Parameter Processes, Usually Functions of Time

Definition 34 One-Parameter Process	28
Example 35 Bernoulli process	28
Example 36 Markov models	28
Example 37 “White Noise” (Not Really)	28
Example 38 Wiener Process	29
Example 39 Logistic Map	29
Example 40 Symbolic Dynamics of the Logistic Map	29
Example 41 IID Samples	30
Example 42 Non-IID Samples	30
Example 43 Estimating Distributions	30
Example 44 Doob’s Martingale	30
Example 45 The One-Dimensional Ising Model	30
Example 46 Text	31
Example 47 Polymer Sequences	31
Definition 48 Shift Operators	31
Exercise 4.1 Existence of proto-Wiener processes	32
Exercise 4.2 Time-Evolution Semi-Group	32

Chapter 5 Stationary One-Parameter Processes

Definition 49 Strong Stationarity	34
Definition 50 Weak Stationarity	35
Definition 51 Conditional (Strong) Stationarity	35

Theorem 52 Stationarity is Shift-Invariance	35
Definition 53 Measure-Preserving Transformation	36
Corollary 54 Measure-preservation implies stationarity	36
Exercise 5.1 Functions of Stationary Processes	37
Exercise 5.2 Continuous Measure-Preserving Families of Trans- formations	37
Exercise 5.3 The Logistic Map as a Measure-Preserving Trans- formation	37
Chapter 6 Random Times and Their Properties	38
Definition 55 Filtration	38
Definition 56 Adapted Process	39
Definition 57 Stopping Time, Optional Time	39
Definition 58 \mathcal{F}_τ for a Stopping Time τ	39
Definition 59 Hitting Time	40
Example 60 Fixation through Genetic Drift	40
Example 61 Stock Options	40
Definition 62 First Passage Time	40
Definition 63 Return Time, Recurrence Time	41
Proposition 64 Some Sufficient Conditions for Waiting Times to be Weakly Optional	41
Lemma 65 Some Recurrence Relations for Kac's Theorem	42
Theorem 66 Recurrence in Stationary Processes	42
Corollary 67 Poincaré Recurrence Theorem	43
Corollary 68 "Nietzsche"	43
Theorem 69 Kac's Recurrence Theorem	44
Example 70 Counter-example for Kac's Recurrence Theorem	45
Exercise 6.1 Weakly Optional Times and Right-Continuous Fil- trations	45
Exercise 6.2 Discrete Stopping Times and Their σ -Algebras	45
Exercise 6.3 Kac's Theorem for the Logistic Map	45
Chapter 7 Continuity of Stochastic Processes	46
Definition 71 Continuity in Mean	46
Definition 72 Continuity in Probability, Stochastic Continuity	47
Definition 73 Continuous Sample Paths	47
Definition 74 Cadlag	47
Definition 75 Versions of a Stochastic Process, Stochastically Equiv- alent Processes	47
Lemma 76 Versions Have Identical FDDs	47
Definition 77 Indistinguishable Processes	48
Proposition 78 Continuous Sample Paths Have Measurable Extrema	48
Example 79 A Horrible Version of the proto-Wiener Process	49
Definition 80 Separable Function	50
Lemma 81 Some Sufficient Conditions for Separability	50
Definition 82 Separable Process	51

Exercise 7.1 Piecewise Linear Paths Not in the Product σ -Field . . .	51
Exercise 7.2 Indistinguishability of Right-Continuous Versions . . .	51
Chapter 8 More on Continuity	52
Definition 83 Compactness, Compactification	53
Proposition 84 Compact Spaces are Separable	53
Proposition 85 Compactification is Possible	53
Example 86 Compactifying the Reals	53
Lemma 87 Alternative Characterization of Separability	54
Lemma 88 Confining Bad Behavior to a Measure Zero Event . . .	54
Lemma 89 A Null Exceptional Set	55
Theorem 90 Separable Versions, Separable Modification	56
Corollary 91 Separable Modifications in Compactified Spaces . . .	57
Corollary 92 Separable Versions with Prescribed Distributions Exist	57
Definition 93 Measurable random function	58
Theorem 94 Exchanging Expectations and Time Integrals	58
Theorem 95 Measurable Separable Modifications	58
Theorem 96 Cadlag Versions	58
Theorem 97 Continuous Versions	59
Definition 98 Modulus of continuity	59
Lemma 99 Modulus of continuity and uniform continuity	59
Definition 100 Hölder continuity	59
Theorem 101 Hölder-continuous versions	59
Chapter 9 Markov Processes	61
Definition 102 Markov Property	61
Lemma 103 The Markov Property Extends to the Whole Future	61
Definition 104 Product of Probability Kernels	62
Definition 105 Transition Semi-Group	63
Theorem 106 Existence of Markov Process with Given Transition Kernels	63
Definition 107 Invariant Distribution	64
Theorem 108 Stationarity and Invariance for Homogeneous Markov Processes	65
Definition 109 Natural Filtration	65
Definition 110 Comparison of Filtrations	65
Lemma 111 The Natural Filtration Is the Coarsest One to Which a Process Is Adapted	65
Theorem 112 Markovianity Is Preserved Under Coarsening	66
Example 113 The Logistic Map Shows That Markovianity Is Not Preserved Under Refinement	66
Exercise 9.1 Extension of the Markov Property to the Whole Future	67
Exercise 9.2 Futures of Markov Processes Are One-Sided Markov Processes	67
Exercise 9.3 Discrete-Time Sampling of Continuous-Time Markov Processes	67

Exercise 9.4 Stationarity and Invariance for Homogeneous Markov Processes	67
Exercise 9.5 Rudiments of Likelihood-Based Inference for Markov Chains	68
Exercise 9.6 Implementing the MLE for a Simple Markov Chain	68
Exercise 9.7 The Markov Property and Conditional Independence from the Immediate Past	68
Exercise 9.8 Higher-Order Markov Processes	69
Exercise 9.9 AR(1) Models	69
Chapter 10 Alternative Characterizations of Markov Processes	70
Theorem 114 Markov Sequences as Transduced Noise	70
Definition 115 Transducer	71
Definition 116 Markov Operator on Measures	72
Definition 117 Markov Operator on Densities	72
Lemma 118 Markov Operators on Measures Induce Those on Densities	72
Definition 119 Transition Operators	73
Definition 120 L_∞ Transition Operators	73
Lemma 121 Kernels and Operators	74
Definition 122 Functional	74
Definition 123 Conjugate or Adjoint Space	74
Proposition 124 Conjugate Spaces are Vector Spaces	74
Proposition 125 Inner Product is Bilinear	74
Example 126 Vectors in \mathbb{R}^n	74
Example 127 Row and Column Vectors	74
Example 128 L_p spaces	75
Example 129 Measures and Functions	75
Definition 130 Adjoint Operator	75
Proposition 131 Adjoint of a Linear Operator	75
Lemma 132 Markov Operators on Densities and L_∞ Transition Operators	75
Theorem 133 Transition operator semi-groups and Markov processes	76
Lemma 134 Markov Operators are Contractions	76
Lemma 135 Markov Operators Bring Distributions Closer	76
Theorem 136 Invariant Measures Are Fixed Points	76
Exercise 10.1 Kernels and Operators	76
Exercise 10.2 L_1 and L_∞	77
Exercise 10.3 Operators and Expectations	77
Exercise 10.4 Bayesian Updating as a Markov Process	77
Exercise 10.5 More on Bayesian Updating	77

Chapter 11 Examples of Markov Processes	78
Definition 137 Processes with Stationary and Independent Increments	82
Definition 138 Lévy Processes	82
Example 139 Wiener Process is Lévy	82
Example 140 Poisson Counting Process	82
Theorem 141 Processes with Stationary Independent Increments are Markovian	83
Theorem 142 Time-Evolution Operators of Processes with Stationary, Independent Increments	83
Definition 143 Infinitely-Divisible Distributions and Random Variables	84
Proposition 144 Limiting Distributions Are Infinitely Divisible	84
Theorem 145 Infinitely Divisible Distributions and Stationary Independent Increments	84
Corollary 146 Infinitely Divisible Distributions and Lévy Processes	84
Definition 147 Self-similarity	85
Definition 148 Stable Distributions	86
Theorem 149 Scaling in Stable Lévy Processes	86
Exercise 11.1 Wiener Process with Constant Drift	86
Exercise 11.2 Perron-Frobenius Operators	86
Exercise 11.3 Continuity of the Wiener Process	86
Exercise 11.4 Independent Increments with Respect to a Filtration	86
Exercise 11.5 Poisson Counting Process	86
Exercise 11.6 Poisson Distribution is Infinitely Divisible	87
Exercise 11.7 Self-Similarity in Lévy Processes	87
Exercise 11.8 Gaussian Stability	87
Exercise 11.9 Poissonian Stability?	87
Exercise 11.10 Lamperti Transformation	87
Chapter 12 Generators of Markov Processes	88
Definition 150 Infinitesimal Generator	89
Definition 151 Limit in the L -norm sense	89
Lemma 152 Generators are Linear	89
Lemma 153 Invariant Distributions of a Semi-group Belong to the Null Space of Its Generator	89
Lemma 154 Invariant Distributions and the Generator of the Time-Evolution Semigroup	89
Lemma 155 Operators in a Semi-group Commute with Its Generator	90
Definition 156 Time Derivative in Function Space	90
Lemma 157 Generators and Derivatives at Zero	90
Theorem 158 The Derivative of a Function Evolved by a Semi-Group	90
Corollary 159 Initial Value Problems in Function Space	91
Corollary 160 Derivative of Conditional Expectations of a Markov Process	91

Definition 161 Resolvents	92
Definition 162 Yosida Approximation of Operators	92
Theorem 163 Hille-Yosida Theorem	93
Corollary 164 Stochastic Approximation of Initial Value Problems	93
Exercise 12.1 Generators are Linear	93
Exercise 12.2 Semi-Groups Commute with Their Generators . . .	94
Exercise 12.3 Generator of the Poisson Counting Process	94
Chapter 13 The Strong Markov Property and Martingale Problems	95
Definition 165 Strongly Markovian at a Random Time	96
Definition 166 Strong Markov Property	96
Example 167 A Markov Process Which Is Not Strongly Markovian	96
Definition 168 Martingale Problem	97
Proposition 169 Cadlag Nature of Functions in Martingale Problems	97
Lemma 170 Alternate Formulation of Martingale Problem	97
Theorem 171 Markov Processes Solve Martingale Problems . . .	97
Theorem 172 Solutions to the Martingale Problem are Strongly Markovian	97
Exercise 13.1 Strongly Markov at Discrete Times	98
Exercise 13.2 Markovian Solutions of the Martingale Problem . .	98
Exercise 13.3 Martingale Solutions are Strongly Markovian . . .	98
Chapter 14 Feller Processes	99
Definition 173 Initial Distribution, Initial State	99
Lemma 174 Kernel from Initial States to Paths	100
Definition 175 Markov Family	100
Definition 176 Mixtures of Path Distributions (Mixed States) . .	100
Definition 177 Feller Process	101
Definition 178 Contraction Operator	101
Definition 179 Strongly Continuous Semigroup	101
Definition 180 Positive Operator	101
Definition 181 Conservative Operator	102
Lemma 182 Continuous semi-groups produce continuous paths in function space	102
Definition 183 Continuous Functions Vanishing at Infinity . . .	102
Definition 184 Feller Semigroup	102
Lemma 185 The First Pair of Feller Properties	103
Lemma 186 The Second Pair of Feller Properties	103
Theorem 187 Feller Processes and Feller Semigroups	103
Theorem 188 Generator of a Feller Semigroup	103
Theorem 189 Feller Semigroups are Strongly Continuous	103
Proposition 190 Cadlag Modifications Implied by a Kind of Modulus of Continuity	104
Lemma 191 Markov Processes Have Cadlag Versions When They Don't Move Too Fast (in Probability)	104

Lemma 192 Markov Processes Have Cadlag Versions If They Don't Move Too Fast (in Expectation)	104
Theorem 193 Feller Implies Cadlag	105
Theorem 194 Feller Processes are Strongly Markovian	105
Theorem 195 Dynkin's Formula	106
Exercise 14.1 Yet Another Interpretation of the Resolvents	106
Exercise 14.2 The First Pair of Feller Properties	106
Exercise 14.3 The Second Pair of Feller Properties	106
Exercise 14.4 Dynkin's Formula	106
Exercise 14.5 Lévy and Feller Processes	106
Chapter 15 Convergence of Feller Processes	107
Definition 196 Convergence in Finite-Dimensional Distribution	107
Lemma 197 Finite and Infinite Dimensional Distributional Con- vergence	108
Definition 198 The Space \mathbf{D}	108
Definition 199 Modified Modulus of Continuity	108
Proposition 200 Weak Convergence in $\mathbf{D}(\mathbb{R}^+, \Xi)$	108
Proposition 201 Sufficient Condition for Weak Convergence	109
Definition 202 Closed and Closable Generators, Closures	109
Definition 203 Core of an Operator	109
Lemma 204 Feller Generators Are Closed	109
Theorem 205 Convergence of Feller Processes	110
Corollary 206 Convergence of Discret-Time Markov Processes on Feller Processes	111
Definition 207 Pure Jump Markov Process	111
Proposition 208 Pure-Jump Markov Processes and ODEs	112
Exercise 15.1 Poisson Counting Process	113
Exercise 15.2 Exponential Holding Times in Pure-Jump Processes	113
Exercise 15.3 Solutions of ODEs are Feller Processes	113
Chapter 16 Convergence of Random Walks	114
Definition 209 Continuous-Time Random Walk (Cadlag)	117
Lemma 210 Increments of Random Walks	117
Lemma 211 Continuous-Time Random Walks are Pseudo-Feller	117
Lemma 212 Evolution Operators of Random Walks	118
Theorem 213 Functional Central Limit Theorem (I)	118
Lemma 214 Convergence of Random Walks in Finite-Dimensional Distribution	119
Theorem 215 Functional Central Limit Theorem (II)	119
Corollary 216 The Invariance Principle	120
Corollary 217 Donsker's Theorem	121
Exercise 16.1 Example 167 Revisited	121
Exercise 16.2 Generator of the d -dimensional Wiener Process	121
Exercise 16.3 Continuous-time random walks are Markovian	121
Exercise 16.4 Donsker's Theorem	121

Exercise 16.5 Diffusion equation	122
Exercise 16.6 Functional CLT for Dependent Variables	122
Chapter 17 Diffusions and the Wiener Process	124
Definition 218 Diffusion	124
Lemma 219 The Wiener Process Is a Martingale	126
Definition 220 Wiener Processes with Respect to Filtrations	127
Definition 221 Gaussian Process	127
Lemma 222 Wiener Process Is Gaussian	127
Theorem 223 Almost All Continuous Curves Are Non-Differentiable	128
Chapter 18 Stochastic Integrals with the Wiener Process	130
Theorem 224 Martingale Characterization of the Wiener Process	131
Chapter 19 Stochastic Differential Equations	133
Definition 225 Progressive Process	133
Definition 226 Non-anticipating filtrations, processes	134
Definition 227 Elementary process	134
Definition 228 Mean square integrable	134
Definition 229 \mathcal{S}_2 norm	134
Proposition 230 $\ \cdot\ _{\mathcal{S}_2}$ is a norm	134
Definition 231 Itô integral of an elementary process	135
Lemma 232 Approximation of Bounded, Continuous Processes by Elementary Processes	135
Lemma 233 Approximation by of Bounded Processes by Bounded, Continuous Processes	136
Lemma 234 Approximation of Square-Integrable Processes by Bounded Processes	136
Lemma 235 Approximation of Square-Integrable Processes by El- ementary Processes	136
Lemma 236 Itô Isometry for Elementary Processes	136
Theorem 237 Itô Integrals of Approximating Elementary Pro- cesses Converge	137
Definition 238 Itô integral	138
Corollary 239 The Itô isometry	138
Definition 240 Itô Process	141
Lemma 241 Itô processes are non-anticipating	141
Theorem 242 Itô's Formula in One Dimension	141
Example 243 Section 19.2.2 summarized	145
Definition 244 Multidimensional Itô Process	145
Theorem 245 Itô's Formula in Multiple Dimensions	146
Theorem 246 Representation of Martingales as Stochastic Inte- grals (Martingale Representation Theorem)	147
Theorem 247 Martingale Characterization of the Wiener Process	147
Definition 248 Stochastic Differential Equation, Solutions	147
Lemma 249 A Quadratic Inequality	148

Definition 250 Maximum Process	148
Definition 251 The Space $\mathcal{QM}(T)$	149
Lemma 252 Completeness of $\mathcal{QM}(T)$	149
Proposition 253 Doob's Martingale Inequalities	149
Lemma 254 A Maximal Inequality for Itô Processes	149
Definition 255 Picard operator	150
Lemma 256 Solutions are fixed points of the Picard operator	150
Lemma 257 A maximal inequality for Picard iterates	150
Lemma 258 Gronwall's Inequality	151
Theorem 259 Existence and Uniqueness of Solutions to SDEs in One Dimension	151
Theorem 260 Existence and Uniqueness for Multidimensional SDEs	153
Exercise 19.1 Basic Properties of the Itô Integral	155
Exercise 19.2 Martingale Properties of the Itô Integral	156
Exercise 19.3 Continuity of the Itô Integral	156
Exercise 19.4 "The square of dW "	156
Exercise 19.5 Itô integrals of elementary processes do not depend on the break-points	156
Exercise 19.6 Itô integrals are Gaussian processes	156
Exercise 19.7 A Solvable SDE	156
Exercise 19.8 Building Martingales from SDEs	157
Exercise 19.9 Brownian Motion and the Ornstein-Uhlenbeck Pro- cess	157
Exercise 19.10 Again with the martingale characterization of the Wiener process	157
Chapter 20 More on Stochastic Differential Equations	158
Theorem 261 Solutions of SDEs are Non-Anticipating and Con- tinuous	158
Theorem 262 Solutions of SDEs are Strongly Markov	159
Theorem 263 SDEs and Feller Diffusions	159
Corollary 264 Convergence of Initial Conditions and of Processes	160
Example 265 Wiener process, heat equation	162
Example 266 Ornstein-Uhlenbeck process	163
Exercise 20.1 Langevin equation with a conservative force	163
Chapter 21 Large Deviations for Small-Noise Stochastic Differen- tial Equations	164
Theorem 267 Small-Noise SDEs Converge in Probability on No- Noise ODEs	166
Lemma 268 A Maximal Inequality for the Wiener Process	167
Lemma 269 A Tail Bound for Maxwell-Boltzmann Distributions	167
Theorem 270 Upper Bound on the Rate of Convergence of Small- Noise SDEs	168

Chapter 22 Spectral Analysis and L_2 Ergodicity	170
Proposition 271 Linearity of White Noise Integrals	172
Proposition 272 White Noise Has Mean Zero	172
Proposition 273 White Noise and Itô Integrals	173
Proposition 274 White Noise is Uncorrelated	173
Proposition 275 White Noise is Gaussian and Stationary	173
Definition 276 Autocovariance Function	174
Lemma 277 Autocovariance and Time Reversal	174
Definition 278 Second-Order Process	174
Definition 279 Spectral Representation, Power Spectrum	174
Lemma 280 Regularity of the Spectral Process	175
Definition 281 Jump of the Spectral Process	175
Proposition 282 Spectral Representations of Weakly Stationary Processes	176
Definition 283 Orthogonal Increments	176
Lemma 284 Orthogonal Spectral Increments and Weak Stationarity	176
Definition 285 Spectral Function, Spectral Density	177
Theorem 286 Weakly Stationary Processes Have Spectral Functions	177
Theorem 287 Existence of Weakly Stationary Processes with Given Spectral Functions	179
Definition 288 Jump of the Spectral Function	179
Lemma 289 Spectral Function Has Non-Negative Jumps	179
Theorem 290 Wiener-Khinchin Theorem	179
Definition 291 Time Averages	181
Definition 292 Integral Time Scale	182
Theorem 293 Mean-Square Ergodic Theorem (Finite Autocovari- ance Time)	182
Corollary 294 Convergence Rate in the Mean-Square Ergodic Theorem	183
Lemma 295 Mean-Square Jumps in the Spectral Process	183
Lemma 296 The Jump in the Spectral Function	183
Lemma 297 Existence of L_2 Limits for Time Averages	183
Lemma 298 The Mean-Square Ergodic Theorem	184
Exercise 22.1 Mean-Square Ergodicity in Discrete Time	184
Exercise 22.2 Mean-Square Ergodicity with Non-Zero Mean	185
Exercise 22.3 Functions of Weakly Stationary Processes	185
Exercise 22.4 Ergodicity of the Ornstein-Uhlenbeck Process?	185
Chapter 23 Ergodic Properties and Ergodic Limits	187
Definition 299 Dynamical System	189
Lemma 300 Dynamical Systems are Markov Processes	189
Definition 301 Observable	189
Definition 302 Invariant Functions, Sets and Measures	189
Lemma 303 Invariant Sets are a σ -Algebra	189
Lemma 304 Invariant Sets and Observables	190
Definition 305 Infinitely Often, i.o.	190

Lemma 306 “Infinitely often” implies invariance	190
Definition 307 Invariance Almost Everywhere	190
Lemma 308 Almost-invariant sets form a σ -algebra	190
Lemma 309 Invariance for simple functions	190
Definition 310 Time Averages	191
Lemma 311 Time averages are observables	191
Definition 312 Ergodic Property	191
Definition 313 Ergodic Limit	191
Lemma 314 Linearity of Ergodic Limits	192
Lemma 315 Non-negativity of ergodic limits	192
Lemma 316 Constant functions have the ergodic property	192
Lemma 317 Ergodic limits are invariantifying	192
Lemma 318 Ergodic limits are invariant functions	193
Lemma 319 Ergodic limits and invariant indicator functions . . .	193
Lemma 320 Ergodic properties of sets and observables	193
Lemma 321 Expectations of ergodic limits	193
Lemma 322 Convergence of Ergodic Limits	194
Lemma 323 Cesàro Mean of Expectations	194
Corollary 324 Replacing Boundedness with Uniform Integrability	194
Definition 325 Asymptotically Mean Stationary	195
Lemma 326 Stationary Implies Asymptotically Mean Stationary	195
Proposition 327 Vitali-Hahn Theorem	195
Theorem 328 Stationary Means are Invariant Measures	195
Lemma 329 Expectations of Almost-Invariant Functions	196
Lemma 330 Limit of Cesàro Means of Expectations	196
Lemma 331 Expectation of the Ergodic Limit is the AMS Expec- tation	196
Corollary 332 Replacing Boundedness with Uniform Integrability	197
Theorem 333 Ergodic Limits of Bounded Observables are Condi- tional Expectations	197
Corollary 334 Ergodic Limits of Integrable Observables are Con- ditional Expectations	197
Exercise 23.1 “Infinitely often” implies invariant	197
Exercise 23.2 Invariant simple functions	197
Exercise 23.3 Ergodic limits of integrable observables	197
Chapter 24 The Almost-Sure Ergodic Theorem	198
Definition 335 Lower and Upper Limiting Time Averages	199
Lemma 336 Lower and Upper Limiting Time Averages are Invariant	199
Lemma 337 “Limits Coincide” is an Invariant Event	199
Lemma 338 Ergodic properties under AMS measures	199
Theorem 339 Birkhoff’s Almost-Sure Ergodic Theorem	199
Corollary 340 Birkhoff’s Ergodic Theorem for Integrable Observ- ables	203

Chapter 25 Ergodicity and Metric Transitivity 204

Definition 341 Ergodic Systems, Processes, Measures and Trans-	
formations	204
Definition 342 Metric Transitivity	205
Lemma 343 Metric transitivity implies ergodicity	205
Lemma 344 Stationary Ergodic Systems are Metrically Transitive	205
Example 345 IID Sequences, Strong Law of Large Numbers . . .	206
Example 346 Markov Chains	206
Example 347 Deterministic Ergodicity: The Logistic Map . . .	206
Example 348 Invertible Ergodicity: Rotations	207
Example 349 Ergodicity when the Distribution Does Not Converge	207
Theorem 350 Ergodicity and the Triviality of Invariant Functions	207
Theorem 351 The Ergodic Theorem for Ergodic Processes . . .	208
Lemma 352 Ergodicity Implies Approach to Independence . . .	208
Theorem 353 Approach to Independence Implies Ergodicity . . .	208
Exercise 25.1 Ergodicity implies an approach to independence . .	209
Exercise 25.2 Approach to independence implies ergodicity . . .	209
Exercise 25.3 Invariant events and tail events	209
Exercise 25.4 Ergodicity of ergodic Markov chains	209

Chapter 26 Decomposition of Stationary Processes into Ergodic Components 210

Proposition 354	211
Proposition 355	211
Proposition 356	211
Definition 357	212
Proposition 358	212
Proposition 359	212
Proposition 360	213
Proposition 361	213
Proposition 362	213
Definition 363	213
Proposition 364	213
Definition 365	214
Proposition 366	214
Proposition 367	214
Proposition 368	215
Lemma 369	215
Lemma 370	215
Lemma 371	215
Theorem 372	216
Definition 373	216
Lemma 374	217
Theorem 375	217
Theorem 376	218
Corollary 377	218

Corollary 378	218
Exercise 26.1	219
Chapter 27 Mixing	220
Definition 379	221
Lemma 380	221
Theorem 381	221
Definition 382	221
Lemma 383	221
Theorem 384	222
Example 385	223
Example 386	223
Example 387	223
Example 388	223
Lemma 389	224
Theorem 390	224
Theorem 391	224
Definition 392	225
Lemma 393	225
Definition 394	226
Definition 395	226
Theorem 396	226
Chapter 28 Shannon Entropy and Kullback-Leibler Divergence	228
Definition 397	229
Lemma 398	230
Definition 399	231
Definition 400	231
Lemma 401	231
Lemma 402	231
Definition 403	231
Lemma 404	232
Lemma 405	232
Definition 406	232
Lemma 407	233
Lemma 408	233
Corollary 409	233
Definition 410	233
Corollary 411	234
Definition 412	234
Proposition 413	234
Proposition 414	234
Definition 415	235
Theorem 416	235

Chapter 29 Entropy Rates and Asymptotic Equipartition	236
Definition 417	236
Definition 418	236
Lemma 419	237
Theorem 420	237
Theorem 421	237
Lemma 422	238
Example 423	238
Example 424	238
Definition 425	238
Definition 426	239
Lemma 427	240
Lemma 428	240
Lemma 429	240
Definition 430	241
Lemma 431	241
Lemma 432	241
Lemma 433	241
Lemma 434	241
Theorem 435	242
Theorem 436	243
Theorem 437	243
Exercise 29.1	244
Exercise 29.2	244
 Chapter 30 General Theory of Large Deviations	 246
Definition 438	246
Definition 439	246
Lemma 440	247
Lemma 441	247
Definition 442	247
Lemma 443	247
Definition 444	248
Lemma 445	248
Lemma 446	248
Definition 447	248
Lemma 448	248
Lemma 449	249
Theorem 450	249
Theorem 451	250
Definition 452	251
Definition 453	251
Definition 454	252
Definition 455 Empirical Process Distribution	252
Corollary 456	252
Corollary 457	253

Definition 458	253
Theorem 459	253
Theorem 460	254
Theorem 461	254
Definition 462 Exponentially Equivalent Random Variables . . .	254
Lemma 463	254

**Chapter 31 Large Deviations for IID Sequences: The Return of
Relative Entropy** **256**

Definition 464	257
Definition 465	257
Lemma 466	258
Definition 467	258
Definition 468	258
Lemma 469	259
Theorem 470	260
Proposition 471	263
Proposition 472	263
Lemma 473	263
Lemma 474	263
Theorem 475	264
Corollary 476	265
Theorem 477	265

Chapter 32 Large Deviations for Markov Sequences **266**

Theorem 478	267
Corollary 479	269
Corollary 480	269
Corollary 481	269
Corollary 482	270
Theorem 483	270
Theorem 484	270

**Chapter 33 Large Deviations for Weakly Dependent Sequences via
the Gärtner-Ellis Theorem** **271**

Definition 485	272
Lemma 486	272
Lemma 487 Upper Bound in the Gärtner-Ellis Theorem for Com- pact Sets	272
Lemma 488 Upper Bound in the Gärtner-Ellis Theorem for Closed Sets	272
Definition 489	273
Lemma 490	273
Lemma 491	273
Definition 492 Exposed Point	273
Definition 493	273

Lemma 494 Lower Bound in the Gärtner-Ellis Theorem for Open Sets	274
Theorem 495 Abstract Gärtner-Ellis Theorem	274
Definition 496 Relative Interior	274
Definition 497 Essentially Smooth	274
Proposition 498	274
Theorem 499 Euclidean Gärtner-Ellis Theorem	274
Exercise 33.1	275
Exercise 33.2	275
Exercise 33.3	275
Exercise 33.4	275
Exercise 33.5	275

Chapter 34 Large Deviations for Stochastic Differential Equations 276

Definition 500 Cameron-Martin Spaces	277
Lemma 501 Cameron-Martin Spaces are Hilbert	277
Definition 502 Effective Wiener Action	278
Proposition 503 Girsanov Formula for Deterministic Drift	278
Lemma 504 Exponential Bound on the Probability of Tubes Around Given Trajectories	278
Lemma 505 Trajectories are Rarely Far from Action Minima	279
Proposition 506 Compact Level Sets of the Wiener Action	280
Theorem 507 Schilder's Theorem on Large Deviations of the Wiener Process on the Unit Interval	280
Corollary 508 Extension of Schilder's Theorem to $[0, T]$	281
Corollary 509 Schilder's Theorem on \mathbb{R}^+	281
Definition 510 SDE with Small State-Independent Noise	282
Definition 511 Effective Action under State-Independent Noise	282
Lemma 512 Continuous Mapping from Wiener Process to SDE Solutions	282
Lemma 513 Mapping Cameron-Martin Spaces Into Themselves	283
Theorem 514 Freidlin-Wentzell Theorem with State-Independent Noise	283
Definition 515 SDE with Small State-Dependent Noise	283
Definition 516 Effective Action under State-Dependent Noise	284
Theorem 517 Freidlin-Wentzell Theorem for State-Dependent Noise	284
Exercise 34.1 Cameron-Martin Spaces are Hilbert Spaces	284
Exercise 34.2 Lower Bound in Schilder's Theorem	284
Exercise 34.3 Mapping Cameron-Martin Spaces Into Themselves	284

Preface: Stochastic Processes in Measure-Theoretic Probability

This is intended to be a *second* course in stochastic processes (at least!); I am going to assume you have all had a first course on stochastic processes, using elementary probability theory, say our 36-703. You might then ask what the added benefit is of taking this course, of re-studying stochastic processes within the framework of measure-theoretic probability. There are a number of reasons to do this.

First, the measure-theoretic framework allows us to greatly generalize the range of processes we can consider. Topics like empirical process theory and stochastic calculus are basically incomprehensible without the measure-theoretic framework. Much of the impetus for developing measure-theoretic probability in the first place came from the impossibility of properly handling continuous random motion, especially the Wiener process, with only the tools of elementary probability.

Second, even topics like Markov processes and ergodic theory, which can be discussed without it, greatly benefit from measure-theoretic probability, because it lets us establish important results which are beyond the reach of elementary methods.

Third, many of the greatest names in twentieth century mathematics have worked in this area, and the theories they have developed are profound, useful and beautiful. Knowing them will make you a better person.

Definitions, lemmas, theorems, corollaries, examples, etc., are all numbered together, consecutively across lectures. Exercises are separately numbered within lectures.

Part I

**Stochastic Processes in
General**

Chapter 1

Basic Definitions: Indexed Collections and Random Functions

Section 1.1 introduces stochastic processes as indexed collections of random variables.

Section 1.2 builds the necessary machinery to consider random functions, especially the product σ -field and the notion of sample paths, and then re-defines stochastic processes as random functions whose sample paths lie in nice sets.

This first chapter begins at the beginning, by defining stochastic processes. Even if you have seen this definition before, it will be useful to review it.

We will flip back and forth between two ways of thinking about stochastic processes: as indexed collections of random variables, and as random functions.

As always, assume we have a nice base probability space (Ω, \mathcal{F}, P) , which is rich enough that all the random variables we need exist.

1.1 So, What Is a Stochastic Process?

Definition 1 (A Stochastic Process Is a Collection of Random Variables) *A stochastic process $\{X_t\}_{t \in T}$ is a collection of random variables X_t , taking values in a common measure space (Ξ, \mathcal{X}) , indexed by a set T .*

That is, for each $t \in T$, $X_t(\omega)$ is an \mathcal{F}/\mathcal{X} -measurable function from Ω to Ξ , which induces a probability measure on Ξ in the usual way.

It's sometimes more convenient to write $X(t)$ in place of X_t . Also, when $S \subset T$, X_s or $X(S)$ refers to that sub-collection of random variables.

Example 2 (Random variables) *Any single random variable is a (trivial) stochastic process. (Take $T = \{1\}$, say.)*

Example 3 (Random vector) Let $T = \{1, 2, \dots, k\}$ and $\Xi = \mathbb{R}$. Then $\{X_t\}_{t \in T}$ is a random vector in \mathbb{R}^k .

Example 4 (One-sided random sequences) Let $T = \{1, 2, \dots\}$ and Ξ be some finite set (or \mathbb{R} or \mathbb{C} or $\mathbb{R}^k \dots$). Then $\{X_t\}_{t \in T}$ is a one-sided discrete (real, complex, vector-valued, ...) random sequence. Most of the stochastic processes you have encountered are probably of this sort: Markov chains, discrete-parameter martingales, etc. Figures 1.2, 1.3, 1.4 and 1.5 illustrate some one-sided random sequences.

Example 5 (Two-sided random sequences) Let $T = \mathbb{Z}$ and Ξ be as in Example 4. Then $\{X_t\}_{t \in T}$ is a two-sided random sequence.

Example 6 (Spatially-discrete random fields) Let $T = \mathbb{Z}^d$ and Ξ be as in Example 4. Then $\{X_t\}_{t \in T}$ is a d -dimensional spatially-discrete random field.

Example 7 (Continuous-time random processes) Let $T = \mathbb{R}$ and $\Xi = \mathbb{R}$. Then $\{X_t\}_{t \in T}$ is a real-valued, continuous-time random process (or random motion or random signal). Figures 1.6 and 1.7 illustrate some of the possibilities.

Vector-valued processes are an obvious generalization.

Example 8 (Random set functions) Let $T = \mathcal{B}$, the Borel field on the reals, and $\Xi = \overline{\mathbb{R}}^+$, the non-negative extended reals. Then $\{X_t\}_{t \in T}$ is a random set function on the reals.

The definition of random set functions on \mathbb{R}^d is entirely parallel. Notice that if we want not just a set function, but a measure or a probability measure, this will imply various forms of dependence among the random variables in the collection, e.g., a measure must respect countable additivity over disjoint sets. We will return to this topic in the next section.

Example 9 (One-sided random sequences of set functions) Let $T = \mathcal{B} \times \mathbb{N}$ and $\Xi = \overline{\mathbb{R}}^+$. Then $\{X_t\}_{t \in T}$ is a one-sided random sequence of set functions.

Example 10 (Empirical distributions) Suppose $Z_i, i = 1, 2, \dots$ are independent, identically-distributed real-valued random variables. (We can see from Example 4 that this is a one-sided real-valued random sequence.) For each Borel set B and each n , define

$$\hat{P}_n(B) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B(Z_i)$$

i.e., the fraction of the samples up to time n which fall into that set. This is the empirical measure. $\hat{P}_n(B)$ is a one-sided random sequence of set functions — in fact, of probability measures. We would like to be able to say something about how it behaves. It would be very reassuring, for instance, to be able to show that it converges to the common distribution of the Z_i (Figure 1.8).

1.2 Random Functions

$X(t, \omega)$ has two arguments, t and ω . For each fixed value of t , $X_t(\omega)$ is straightforward random variable. For each fixed value of ω , however, $X(t)$ is a function from T to Ξ — a *random function*. The advantage of the random function perspective is that it lets us consider the realizations of stochastic processes as single objects, rather than large collections. This isn't just tidier; we will need to talk about *relations* among the variables in the collection or their realizations, rather than just properties of individual variables, and this will help us do so. In Example 10, it's important that we've got random *probability measures*, rather than just random *set functions*, so we need to require that, e.g., $\hat{P}_n(A \cup B) = \hat{P}_n(A) + \hat{P}_n(B)$ when A and B are disjoint Borel sets, and this is a relationship among the three random variables $\hat{P}_n(A)$, $\hat{P}_n(B)$ and $\hat{P}_n(A \cup B)$. Plainly, working out all the dependencies involved here is going to get rather tedious, so we'd like a way to talk about acceptable realizations of the whole stochastic process. This is what the random functions notion will let us do.

We'll make this more precise by defining a random function as a function-valued random variable. To do this, we need a measure space of functions, and a measurable mapping from (Ω, \mathcal{F}, P) to that function space. To get a measure space, we need a carrier set and a σ -field on it. The natural set to use is Ξ^T , the set of all functions from T to Ξ . (We'll see how to restrict this to just the functions we want presently.) Now, how about the σ -field?

Definition 11 (Cylinder Set) *Given an index set T and a collection of σ -fields \mathcal{X}_t on spaces Ξ_t , $t \in T$. Pick any $t \in T$ and any $A_t \in \mathcal{X}_t$. Then $A_t \times \prod_{s \neq t} \Xi_s$ is a one-dimensional cylinder set.*

For any finite k , k -dimensional cylinder sets are defined similarly, and clearly are the intersections of k different one-dimensional cylinder sets. To see why they have this name, notice a cylinder, in Euclidean geometry, consists of all the points where the x and y coordinates fall into a certain set (the base), leaving the z coordinate unconstrained. Similarly, a cylinder set like $A_t \times \prod_{s \neq t} \Xi_s$ consists of all the functions in Ξ^T where $f(t) \in A_t$, and are otherwise unconstrained.

Definition 12 (Product σ -field) *The product σ -field, $\otimes_{t \in T} \mathcal{X}_t$, is the σ -field over Ξ^T generated by all the one-dimensional cylinder sets. If all the \mathcal{X}_t are the same, \mathcal{X} , we write the product σ -field as \mathcal{X}^T .*

The product σ -field is enough to let us define a random function, and is going to prove to be *almost* enough for all our purposes.

Definition 13 (Random Function; Sample Path) *A Ξ -valued random function on T is a map $X : \Omega \mapsto \Xi^T$ which is $\mathcal{F}/\mathcal{X}^T$ -measurable. The realizations of X are functions $x(t)$ taking values in Ξ , called its sample paths.*

N.B., it has become common to apply the term “sample path” or even just “path” even in situations where the geometric analogy it suggests may be somewhat misleading. For instance, for the empirical distributions of Example 10, the “sample path” is the *measure* \hat{P}_n , not the curves shown in Figure 1.8.

Definition 14 (Functional of the Sample Path) *Let E, \mathcal{E} be a measure-space. A functional of the sample path is a mapping $f : \Xi^T \mapsto E$ which is $\mathcal{X}^T/\mathcal{E}$ -measurable.*

Examples of useful and common functionals include maxima, minima, sample averages, etc. Notice that none of these are functions of any one random variable, and in fact their value cannot be determined from any part of the sample path smaller than the whole thing.

Definition 15 (Projection Operator, Coordinate Map) *A projection operator or coordinate map π_t is a map from Ξ^T to Ξ such that $\pi_t X = X(t)$.*

The projection operators are a convenient device for recovering the individual coordinates — the random variables in the collection — from the random function. Obviously, as t ranges over T , $\pi_t X$ gives us a collection of random variables, i.e., a stochastic process in the sense of our first definition. The following lemma lets us go back and forth between the collection-of-variables, coordinate view, and the entire-function, sample-path view.

Theorem 16 (Product σ -field-measurability is equivalent to measurability of all coordinates) *X is $\mathcal{F}/\otimes_{t \in T} \mathcal{X}_t$ -measurable iff $\pi_t X$ is $\mathcal{F}/\mathcal{X}_t$ -measurable for every t .*

PROOF: This follows from the fact that the one-dimensional cylinder sets generate the product σ -field. \square

We have said before that we will want to constrain our stochastic processes to have certain properties — to be probability measures, rather than just set functions, or to be continuous, or twice differentiable, etc. Write the set of all such functions in Ξ^T as U . Notice that U does not have to be an element of the product σ -field, and in general is *not*. (We will consider some of the reasons for this later.) As usual, by $U \cap \mathcal{X}^T$ we will mean the collection of all sets of the form $U \cap C$, where $C \in \mathcal{X}^T$. Notice that $(U, U \cap \mathcal{X}^T)$ is a measure space. What we want is to ensure that the sample path of our random function lies in U .

Definition 17 (A Stochastic Process Is a Random Function) *A Ξ -valued stochastic process on T with paths in U , $U \subseteq \Xi^T$, is a random function $X : \Omega \mapsto U$ which is $\mathcal{F}/U \cap \mathcal{X}^T$ -measurable.*

Corollary 18 (Measurability of constrained sample paths) *A function X from Ω to U is $\mathcal{F}/U \cap \mathcal{X}^T$ -measurable iff X_t is \mathcal{F}/\mathcal{X} -measurable for all t .*

PROOF: Because $X(\omega) \in U$, $\mathcal{F}/U \cap \mathcal{X}^T$ measurability works just the same way as $\mathcal{F}/\mathcal{X}^T$ measurability (Exercise 1.2). So imitate the proof of Theorem 16. \square

Example 19 (Random Measures) *Let $T = \mathcal{B}^d$, the Borel field on \mathbb{R}^d , and let $\Xi = \overline{\mathbb{R}}^+$, the non-negative extended reals. Then Ξ^T is the class of set functions on \mathbb{R}^d . Let M be the class of such set functions which are also measures (i.e., which are countably additive and give zero on the null set). Then a random set function X with realizations in M is a random measure.*

Example 20 (Point Processes) *Let X be a random measure, as in the previous example. If $X(B)$ is a finite integer for every bounded Borel set B , then X is a point process. If in addition $X(r) \leq 1$ for every $r \in \mathbb{R}^d$, then X is simple. The Poisson process is a simple point process. See Figure 1.1.*

Example 21 (Continuous random processes) *Let $T = \mathbb{R}^+$, $\Xi = \mathbb{R}^d$, and $\mathbf{C}(T)$ the class of continuous functions from T to Ξ (in the usual topology). Then a Ξ -valued random process on T with paths in $\mathbf{C}(T)$ is a continuous random process. The Wiener process, or Brownian motion, is an example. We will see that most sample paths in $\mathbf{C}(T)$ are not differentiable.*

1.3 Exercises

Exercise 1.1 (The product σ -field answers countable questions) *Let $\mathcal{D} = \bigcup_S \mathcal{X}^S$, where the union ranges over all countable subsets S of the index set T . For any event $D \in \mathcal{D}$, whether or not a sample path $x \in D$ depends on the value of x_t at only a countable number of indices t .*

1. Show that \mathcal{D} is a σ -field.
2. Show that if $A \in \mathcal{X}^T$, then $A \in \mathcal{X}^S$ for some countable subset S of T .

Exercise 1.2 (The product σ -field constrained to a given set of paths) *Let $U \subset \mathcal{X}^T$ be a set of allowed paths. Show that*

1. $U \cap \mathcal{X}^T$ is a σ -field on U ;
2. $U \cap \mathcal{X}^T$ is generated by sets of the form $U \cap B$, where $B \in \mathcal{X}^S$ for some finite subset S of T .

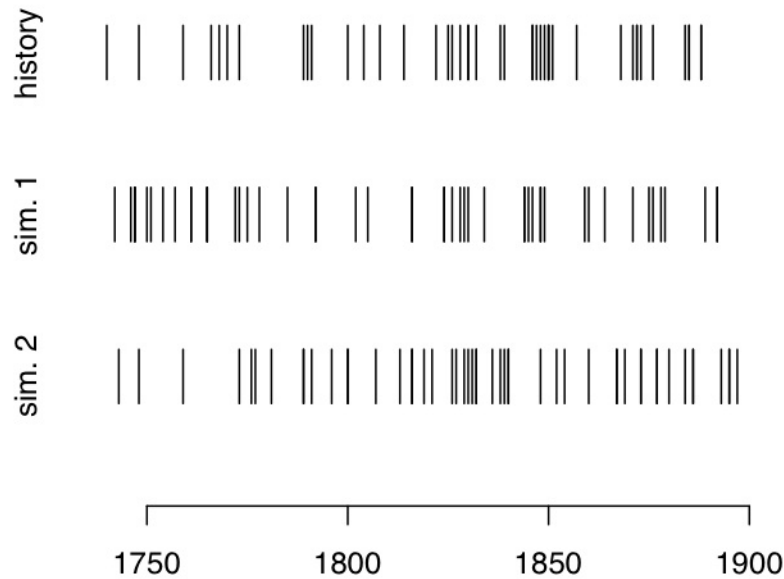


Figure 1.1: Examples of point processes. The top row shows the dates of appearances of 44 genres of English novels (data taken from Moretti (2005)). The bottom two rows show independent realizations of a Poisson process with the same mean time between arrivals as the actual history. The number of tick-marks falling within any measurable set on the horizontal axis determines an integer-valued set function, in fact a measure.

```

AATGAAATAAAAAAAAAACGAAAATAAAAAA
AAGGCCATTAAAGTTAAAATAATGAAAGGA
CAATGATTAGGACAATAACATACAAGTTAT
GGGGTTAATTAATGGTTAGGATGGGTTTTT
CCTTCAAAGTTAATGAAAAGTTAAATTTA
TAAGTATTTGAAGCACAGCAACAACACTAGGT

```

Figure 1.2: Examples of one-sided random sequences ($\Xi = \{A, C, G, T\}$). The top line shows the first thirty bases of the first chromosome of the cellular slime mold *Dictyostelium discoideum* (Eichinger *et al.*, 2005), downloaded from dictybase.org. The lower lines are independent samples from a simple Markov model fitted to the full chromosome.


```

11111111111111110110011011000001
011110110111111011110110110110
01111000110110111111111110000
011011011011111101111000111100
01101111110110110000111110111
000111111111111001101100011011

```

Figure 1.3: Examples of one-sided random sequences. These binary sequences ($\Xi = \{0, 1\}$) are the first thirty steps from samples of a *sofic process*, one with a finite number of underlying states which is nonetheless not a Markov chain of any finite order. Can you figure out the rule specifying which sequences are allowed and which are forbidden? (Hint: these are all samples from the *even* process.)

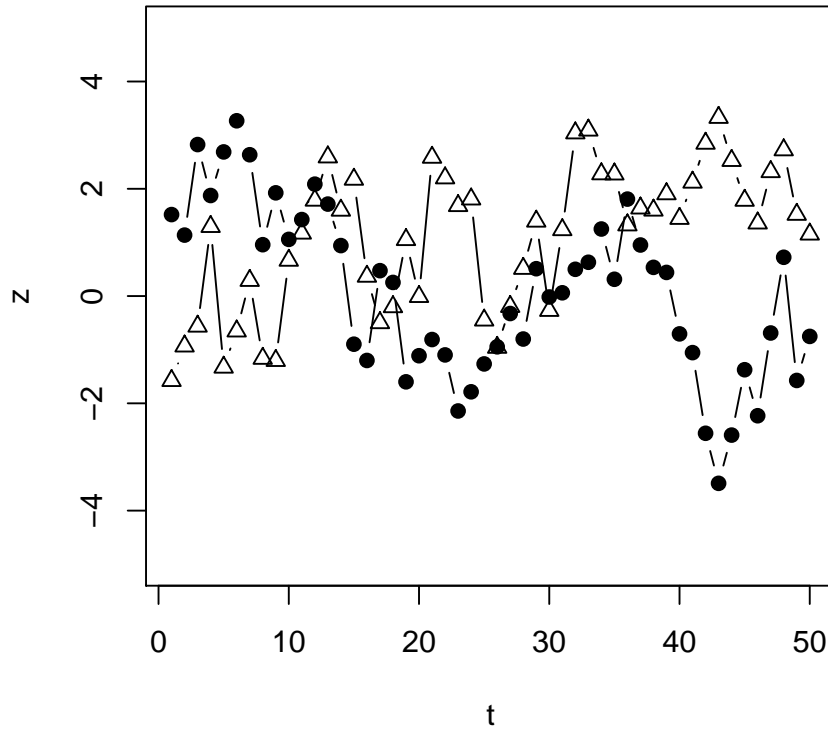


Figure 1.4: Examples of one-sided random sequences. These are linear Gaussian random sequences, $X_{t+1} = 0.8X_t + Z_{t+1}$, where the Z_t are all i.i.d. $\mathcal{N}(0, 1)$, and $X_1 = Z_0$. Different shapes of dots represent different independent samples of this *autoregressive* process. (The line segments are simply guides to the eye.) This is a Markov sequence, but one with a continuous state-space.

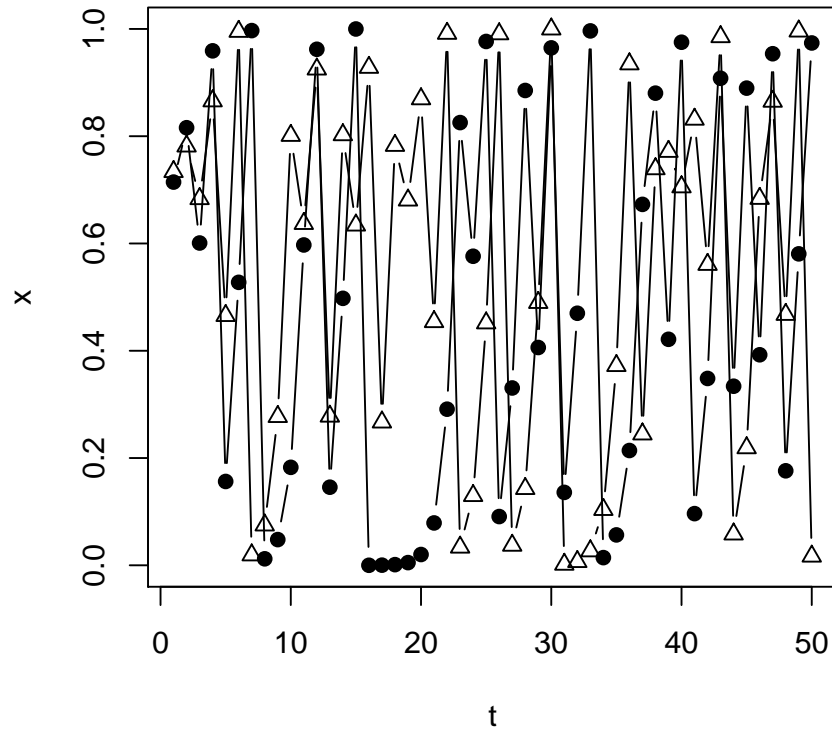


Figure 1.5: Nonlinear, non-Gaussian random sequences. Here $X_1 \sim U(0, 1)$, i.e., uniformly distributed on the unit interval, and $X_{t+1} = 4X_t(1 - X_t)$. Notice that while the two samples begin very close together, they rapidly separate; after a few time-steps their locations are, in fact, effectively independent. We will study both this approach to independence, known as *mixing*, and this example, known as the *logistic map*, in some detail.

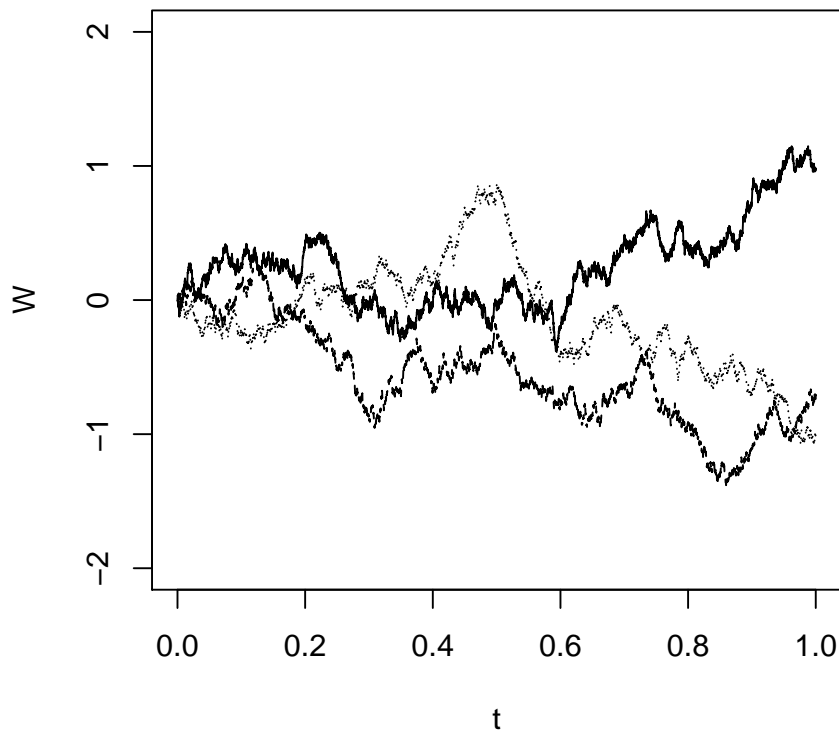


Figure 1.6: Continuous-time random processes. Shown are three samples from the standard Wiener process, also known as Brownian motion, a Gaussian process with independent increments and continuous trajectories. This is a central part of the course, and actually what forced probability to be re-defined in terms of measure theory.

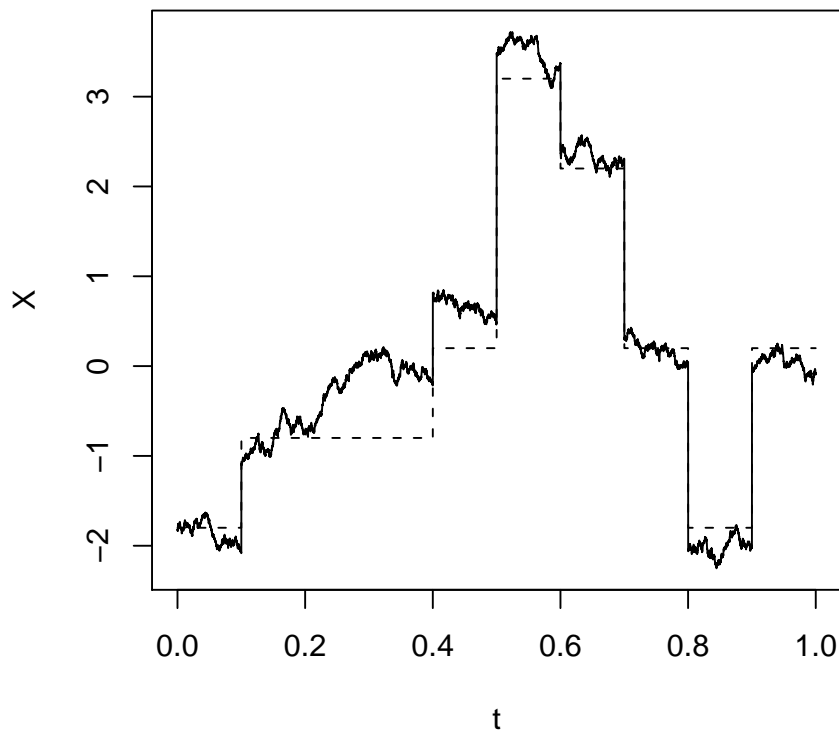


Figure 1.7: Continuous-time random processes can have discontinuous trajectories. Here $X_t = W_t + J_t$, where W_t is a standard Wiener process, and J_t is piecewise constant (shown by the dashed lines), changing at $t = 0.1, 0.2, 0.3, \dots, 1.0$. The trajectory is discontinuous at $t = 0.4$, but continuous from the right there, and there is a limit from the left. In fact, at every point the trajectory is continuous from the right and has a limit from the left. We will see many such *cadlag* processes.

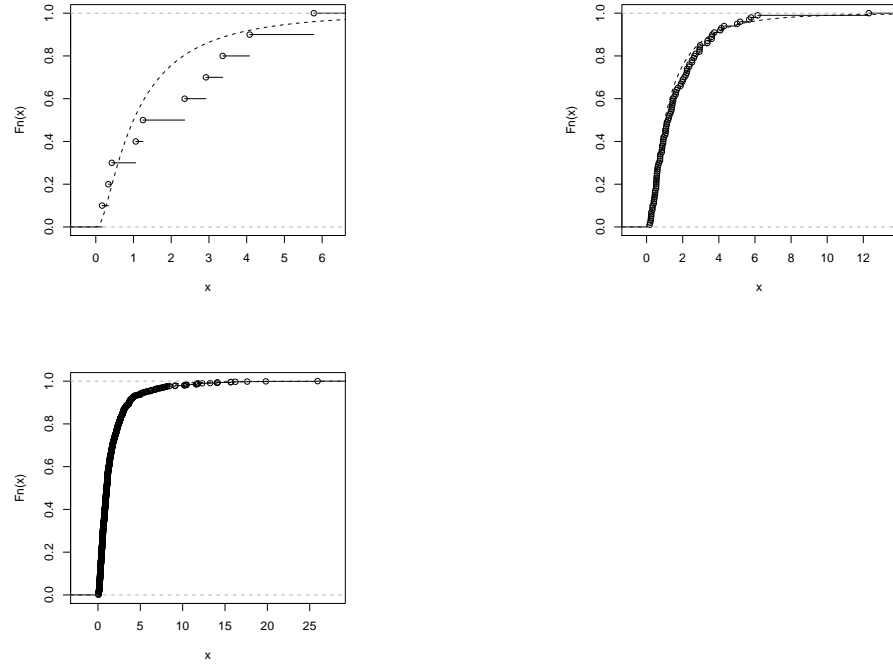


Figure 1.8: Empirical cumulative distribution functions for successively larger samples from the standard log-normal distribution (from left to right, $n = 10, 100, 1000$), with the theoretical CDF as the smooth dashed line. Because the intervals of the form $(-\infty, a]$ are a generating class for the Borel σ -field, the empirical C.D.F. suffices to represent the empirical distribution.

Chapter 2

Building Infinite Processes from Finite-Dimensional Distributions

Section 2.1 introduces the finite-dimensional distributions of a stochastic process, and shows how they determine its infinite-dimensional distribution.

Section 2.2 considers the consistency conditions satisfied by the finite-dimensional distributions of a stochastic process, and the extension theorems (due to Daniell and Kolmogorov) which prove the existence of stochastic processes with specified, consistent finite-dimensional distributions.

2.1 Finite-Dimensional Distributions

So, we now have X , our favorite Ξ -valued stochastic process on T with paths in U . Like any other random variable, it has a probability law or distribution, which is defined over the entire set U . Generally, this is infinite-dimensional. Since it is inconvenient to specify distributions over infinite-dimensional spaces all in a block, we consider the *finite-dimensional distributions*.

Definition 22 (Finite-dimensional distributions) *The finite-dimensional distributions of X are the joint distributions of $X_{t_1}, X_{t_2}, \dots, X_{t_n}$, $t_1, t_2, \dots, t_n \in T$, $n \in \mathbb{N}$.*

You will sometimes see “FDDs” and “fidis” as abbreviations for “finite-dimensional distributions”. Please do not use “fidis”.

We can at least hope to specify the finite-dimensional distributions. But we are going to want to ask a lot of questions about asymptotics, and global properties of sample paths, which go beyond any *finite* dimension, so you might worry

that we'll still need to deal directly with the infinite-dimensional distribution. The next theorem says that this worry is unfounded; the finite-dimensional distributions specify the infinite-dimensional distribution (pretty much) uniquely.

Theorem 23 (Finite-dimensional distributions determine process distributions) *Let X and Y be two Ξ -valued processes on T with paths in U . Then X and Y have the same distribution iff all their finite-dimensional distributions agree.*

PROOF: “Only if”: Since X and Y have the same distribution, applying the any given set of coordinate mappings will result in identically-distributed random vectors, hence all the finite-dimensional distributions will agree.

“If”: We'll use the π - λ theorem.

Let \mathcal{C} be the finite cylinder sets, i.e., all sets of the form

$$C = \{x \in \Xi^T \mid (x_{t_1}, x_{t_2}, \dots, x_{t_n}) \in B\}$$

where $n \in \mathbb{N}$, $B \in \mathcal{X}^n$, $t_1, t_2, \dots, t_n \in T$. Clearly, this is a π -system, since it is closed under intersection.

Now let \mathcal{L} consist of all the sets $L \in \mathcal{X}^T$ where $\mathbb{P}(X \in L) = \mathbb{P}(Y \in L)$. We need to show that this is a λ -system, i.e., that it (i) includes Ξ^T , (ii) is closed under complementation, and (iii) is closed under monotone increasing limits. (i) is clearly true: $\mathbb{P}(X \in \Xi^T) = \mathbb{P}(Y \in \Xi^T) = 1$. (ii) is true because we're looking at a probability: if $L \in \mathcal{L}$, then $\mathbb{P}(X \in L^c) = 1 - \mathbb{P}(X \in L) = 1 - \mathbb{P}(Y \in L) = \mathbb{P}(Y \in L^c)$. To see (iii), let $L_n \uparrow L$ be a monotone-increasing sequence of sets in \mathcal{L} , and recall that, for any measure, $L_n \uparrow L$ implies $\mu L_n \uparrow \mu L$. So $\mathbb{P}(X \in L_n) \uparrow \mathbb{P}(X \in L)$, $\mathbb{P}(Y \in L_n) \uparrow \mathbb{P}(Y \in L)$, and (since $\mathbb{P}(X \in L_n) = \mathbb{P}(Y \in L_n)$), $\mathbb{P}(X \in L_n) \uparrow \mathbb{P}(Y \in L)$ as well. A sequence cannot have two limits, so $\mathbb{P}(X \in L) = \mathbb{P}(Y \in L)$, and $L \in \mathcal{L}$.

Since the finite-dimensional distributions match, $\mathbb{P}(X \in C) = \mathbb{P}(Y \in C)$ for all $C \in \mathcal{C}$, which means that $\mathcal{C} \subset \mathcal{L}$. Also, from the definition of the product σ -field, $\sigma(\mathcal{C}) = \mathcal{X}^T$. Hence, by the π - λ theorem, $\mathcal{X}^T \subseteq \mathcal{L}$. \square

A note of caution is in order here. If X is a Ξ -valued process on T whose paths are constrained to line in U , and Y is a similar process that is not so constrained, it is nonetheless possible that X and Y agree in all their finite-dimensional distributions. The trick comes if U is not, itself, an element of \mathcal{X}^T . The most prominent instance of this is when $\Xi = \mathbb{R}$, $T = \mathbb{R}$, and the constraint is continuity of the sample paths: we will see that $U \notin \mathcal{B}^{\mathbb{R}}$. (This is the point of Exercise 1.1.)

2.2 Consistency and Extension

The finite-dimensional distributions of a given stochastic process are related to one another in the usual way of joint and marginal distributions. Take some collection of indices $t_1, t_2, \dots, t_n \in T$, and corresponding measurable sets

$B_1 \in \mathcal{X}_1, B_2 \in \mathcal{X}_2, \dots, B_n \in \mathcal{X}_n$. Then, for any $m > n$, and any further indices $t_{n+1}, t_{n+2}, \dots, t_m$, it must be the case that

$$\begin{aligned} & \mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2, \dots, X_{t_n} \in B_n) \\ &= \mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2, \dots, X_{t_n} \in B_n, X_{t_{n+1}} \in \Xi, X_{t_{n+2}} \in \Xi, \dots, X_{t_m} \in \Xi) \end{aligned} \quad (2.1)$$

This is going to get really awkward to write over and over, so let's introduce some simplifying notation. $\text{Fin}(T)$ will denote the class of all finite sub-sets of our index set T , and likewise $\text{Denum}(T)$ all denumerable sub-sets. We'll indicate such sub-sets, for the moment, by capital letters like J, K , etc., and extend the definition of coordinate maps (Definition 15) so that π_J maps from Ξ^T to Ξ^J in the obvious way, and π_J^K maps from Ξ^K to Ξ^J , if $J \subset K$. If μ is the measure for the whole process, then the finite-dimensional distributions are $\{\mu_J | J \in \text{Fin}(T)\}$. Clearly, $\mu_J = \mu \circ \pi_J^{-1}$.

Definition 24 (Projective Family of Distributions) *A family of distributions μ_J , $J \in \text{Denum}(T)$, is projective when for every $J, K \in \text{Denum}(T)$, $J \subset K$ implies*

$$\mu_J = \mu_K \circ (\pi_J^K)^{-1} \quad (2.2)$$

Such a family is also said to be consistent or compatible (with one another).

Lemma 25 (FDDs Form Projective Families) *The finite-dimensional distributions of a stochastic process always form a projective family.*

PROOF: This is just the fact that we get marginal distributions by integrating out some variables from the joint distribution. But, to proceed formally: Letting J and K be finite sets of indices, $J \subset K$, we know that $\mu_K = \mu \circ \pi_K^{-1}$, that $\mu_J = \mu \circ \pi_J^{-1}$ and that $\pi_J = \pi_J^K \circ \pi_K$. Hence

$$\mu_J = \mu \circ (\pi_J^K \circ \pi_K)^{-1} \quad (2.3)$$

$$= \mu \circ \pi_K^{-1} \circ (\pi_J^K)^{-1} \quad (2.4)$$

$$= \mu_K \circ (\pi_J^K)^{-1} \quad (2.5)$$

as required. \square

I claimed that the reason to care about finite-dimensional distributions is that if we specify them, we specify the distribution of the whole process. Lemma 25 says that a putative family of finite dimensional distributions must be consistent, if they are to let us specify a stochastic process. Theorem 23 says that there can't be more than one process distribution with all the same finite-dimensional marginals, but it doesn't guarantee that a given collection of consistent finite-dimensional distributions *can* be extended to a process distribution — it gives uniqueness but not existence. Proving the existence of an extension requires some extra assumptions. Either we need to impose topological conditions on Ξ ,

or we need to ensure that all the finite-dimensional distributions can be related through conditional probabilities. The first approach is due to Daniell and Kolmogorov, and will finish this lecture; the second is due to Ionescu-Tulcea, and will begin the next.

We'll start with Daniell's theorem on the existence of random sequences, i.e., where the index set is the natural numbers, which uses mathematical induction to extend the finite-dimensional family. To get *there*, we need a useful proposition about our ability to represent non-trivial random variables as functions of uniform random variables on the unit interval.

Proposition 26 (Randomization, transfer) *Let X and X' be identically-distributed random variables in a measurable space Ξ and Y a random variable in a Borel space Υ . Then there exists a measurable function $f : \Xi \times [0, 1] \mapsto \Upsilon$ such that $\mathcal{L}(X', f(X', Z)) = \mathcal{L}(X, Y)$, when Z is uniformly distributed on the unit interval and independent of X' .*

PROOF: See Kallenberg, Theorem 6.10 (p. 112–113). \square

Basically what this says is that if we have two random variables with a certain joint distribution, we can always represent the pair by a copy of one of the variables (X), and a transformation of an independent random number. It is important that Υ be a Borel space here; the result, while very natural-sounding, does not hold for arbitrary measurable spaces, because the proof relies on having a regular conditional probability.

Theorem 27 (Daniell Extension Theorem) *For each $n \in \mathbb{N}$, let Ξ_n be a Borel space, and μ_n be a probability measure on $\prod_{i=1}^n \Xi_i$. If the μ_n form a projective family, then there exist random variables $X_i : \Omega \mapsto \Xi_i$, $i \in \mathbb{N}$, such that $\mathcal{L}(X_1, X_2, \dots, X_n) = \mu_n$ for all n , and a measure μ on $\prod_{i=1}^{\infty} \Xi_i$ such that μ_n is equal to the projection of μ onto $\prod_{i=1}^n \Xi_i$.*

PROOF: For any fixed n , X_1, X_2, \dots, X_n is just a random vector with distribution μ_n , and we can always construct such an object. The delicate part here is showing that, when we go to $n+1$, we can use the *same* random elements for the first n coordinates. We'll do this by using the representation-by-randomization proposition just introduced, starting with an IID sequence of uniform random variables on the unit interval, and then transforming them to get a sequence of variables in the Ξ_i which have the right joint distribution. (This is like the quantile transform trick for generating random variates.) The proof will go inductively, so first we'll take care of the induction step, and then go back to reassure ourselves about the starting point.

Induction: Assume we already have X_1, X_2, \dots, X_n such that $\mathcal{L}(X_1, X_2, \dots, X_n) = \mu_n$, and that we have a $Z_{n+1} \sim U(0, 1)$ and independent of all the X_i to date. As remarked, we can always get Y_1, Y_2, \dots, Y_{n+1} such that $\mathcal{L}(Y_1, Y_2, \dots, Y_{n+1}) = \mu_{n+1}$. Because the μ_n form a projective family, $\mathcal{L}(Y_1, Y_2, \dots, Y_n) = \mathcal{L}(X_1, X_2, \dots, X_n)$. Hence, by Proposition 26, there is a measurable f such that, if we set $X_{n+1} = f(X_1, X_2, \dots, X_n, Z_{n+1})$, then $\mathcal{L}(X_1, X_2, \dots, X_n, X_{n+1}) = \mu_{n+1}$.

First step: We need there to be an X_1 with distribution μ_1 , and we need a (countably!) unlimited supply of IID variables Z_2, Z_3, \dots all $\sim U(0, 1)$. But the existence of X_1 is just the existence of a random variable with a well-defined distribution, which is unproblematic, and the existence of an infinite sequence of IID uniform random variates is too. (See 36-752, or Lemma 3.21 in Kallenberg, or Exercise 3.1.)

Finally, to convince yourself of the existence of the measure μ on the product space, recall Theorem 16. \square

Remark: Kallenberg, Corollary 6.15, gives a somewhat more abstract version of this theorem.

Daniell's extension theorem works fine for one-sided random sequences, but we often want to work with larger and more interesting index sets. For this we need the full Kolmogorov extension theorem, where the index set T can be completely arbitrary. This in turn needs the Carathéodory Extension Theorem, which I re-state here for convenience.

Proposition 28 (Carathéodory Extension Theorem) *Let μ be a non-negative, finitely additive set function on a field \mathcal{C} of subsets of some space Ω . If μ is also countably additive, then it extends to a measure on $\sigma(\mathcal{C})$, and, if $\mu(\Omega) < \infty$, the extension is unique.*

PROOF: See 36-752 lecture notes (Theorem 50, Exercise 51), or Kallenberg, Theorem 2.5, pp. 26–27. Note that “extension” here means extending from a mere field to a σ -field, not from finite to infinite index sets. \square

Theorem 29 (Kolmogorov Extension Theorem) *Let Ξ_t , $t \in T$, be a collection of Borel spaces, with σ -fields \mathcal{X}_t , and let μ_J , $J \in \text{Fin}(T)$, be a projective family of finite-dimensional distributions on those spaces. Then there exist Ξ_t -valued random variables X_t such that $\mathcal{L}(X_J) = \mu_J$ for all $J \in \text{Fin}(T)$.*

PROOF: This will be easier to follow if we first consider the case where T is countable, which is basically Theorem 27 again, and then the general case, where we need Proposition 28.

Countable T : We can, by definition, put the elements of T in 1–1 correspondence with the elements of \mathbb{N} . This in turn establishes a bijection between the product space $\bigotimes_{t \in T} \Xi_t = \Xi_T$ and the sequence space $\bigotimes_{i=1}^{\infty} \Xi_t$. This bijection also induces a projective family of distributions on finite sequences. The Daniell Extension Theorem (27) gives us a measure on the sequence space, which the bijection takes back to a measure on Ξ_T . To see that this μ does not depend on the order in which we arranged T , notice that any two arrangements must give identical results for any finite set J , and then use Theorem 23.

Uncountable T : For each countable $K \subset T$, the argument of the preceding paragraph gives us a measure μ_K on Ξ_K . And, clearly, these μ_K themselves form a projective family. Now let's define a set function μ on the countable cylinder sets, i.e., on the class \mathcal{D} of sets of the form $A \times \Xi_{T \setminus K}$, for some $K \in \text{Denum}(T)$ and some $A \in \mathcal{X}_K$. Specifically, $\mu : \mathcal{D} \mapsto [0, 1]$, and $\mu(A \times \Xi_{T \setminus K}) = \mu_K(A)$. We would like to use Carathéodory's theorem to extend this set function to

a measure on the product σ -algebra \mathcal{X}_T . First, let's check that the countable cylinder sets form a field: (i) $\Xi_T \in \mathcal{D}$, clearly. (ii) The complement, in Ξ_T , of a countable cylinder $A \times \Xi_{T \setminus K}$ is another countable cylinder, $A^c \times \Xi_{T \setminus K}$. (iii) The union of two countable cylinders $B_1 = A_1 \times \Xi_{T \setminus K_1}$ and $B_2 = A_2 \times \Xi_{T \setminus K_2}$ is another countable cylinder, since we can always write it as $A \times \Xi_{T \setminus K}$ for some $A \in \mathcal{X}_K$, where $K = K_1 \cup K_2$. Clearly, $\mu(\emptyset) = 0$, so we just need to check that μ is countably additive. So consider any sequence of disjoint cylinder sets B_1, B_2, \dots . Because they're cylinder sets, each i , $B_i = A_i \times \Xi_{T \setminus K_i}$, for some $K_i \in \text{Denum}(T)$, and some $A_i \in \mathcal{X}_{K_i}$. Now set $K = \bigcup_i K_i$; this is a countable union of countable sets, and so itself countable. Furthermore, say $C_i = A_i \times \Xi_{K \setminus K_i}$, so we can say that $\bigcup_i B_i = (\bigcup_i C_i) \times \Xi_{T \setminus K}$. With this notation in place,

$$\mu \bigcup_i B_i = \mu_K \bigcup_i C_i \quad (2.6)$$

$$= \sum_i \mu_K C_i \quad (2.7)$$

$$= \sum_i \mu_{K_i} A_i \quad (2.8)$$

$$= \sum_i \mu B_i \quad (2.9)$$

where in the second line we've used the fact that μ_K is a probability measure on Ξ_K , and so countably additive on sets like the C_i . This proves that μ is countably additive, so by Proposition 28 it extends to a measure on $\sigma(\mathcal{D})$, the σ -field generated by the countable cylinder sets. But we know from Definition 12 that this σ -field *is* the product σ -field. Since $\mu(\Xi_T) = 1$, Proposition 28 further tells us that the extension is unique. \square

Borel spaces are good enough for most of the situations we find ourselves modeling, so the Daniell-Kolmogorov Extension Theorem (as it's often known) see a lot of work. Still, some people dislike having to make topological assumptions to solve probabilistic problems; it seems inelegant. The Ionescu-Tulcea Extension Theorem provides a purely probabilistic solution, available if we can write down the FDDs recursively, in terms of regular conditional probability distributions, even if the spaces where the process has its coordinates are not nice and Borel. Doing this properly will involve our revisiting and extending some ideas about conditional probability, which you will have seen in 36-752, so it will be deferred to the next lecture.

Chapter 3

Building Infinite Processes from Regular Conditional Probability Distributions

Section 3.1 introduces the notion of a probability kernel, which is a useful way of systematizing and extending the treatment of conditional probability distributions you will have seen in 36-752.

Section 3.2 gives an extension theorem (due to Ionescu Tulcea) which lets us build infinite-dimensional distributions from a family of finite-dimensional distributions. Rather than assuming topological regularity of the space, as in Section 2.2, we assume that the FDDs can be derived from one another recursively, through applying probability kernels. This is the same as assuming regularity of the appropriate conditional probabilities.

3.1 Probability Kernels

Definition 30 (Probability Kernel) *A measure kernel from a measurable space Ξ, \mathcal{X} to another measurable space Υ, \mathcal{Y} is a function $\kappa : \Xi \times \mathcal{Y} \mapsto \overline{\mathbb{R}}^+$ such that*

1. *for any $Y \in \mathcal{Y}$, $\kappa(x, Y)$ is \mathcal{X} -measurable; and*
2. *for any $x \in \Xi$, $\kappa(x, Y) \equiv \kappa_x(Y)$ is a measure on Υ, \mathcal{Y} . We will write the integral of a function $f : \Upsilon \mapsto \mathbb{R}$, with respect to this measure, as $\int f(y)\kappa(x, dy)$, $\int f(y)\kappa_x(dy)$, or, most compactly, $\kappa f(x)$.*

If, in addition, κ_x is a probability measure on Υ, \mathcal{Y} for all x , then κ is a probability kernel.

(Some authors use “kernel”, unmodified, to mean a probability kernel, and some a measure kernel; read carefully.)

Notice that we can represent any distribution on Υ as a kernel where the first argument is irrelevant: $\kappa(x_1, Y) = \kappa(x_2, Y)$ for all $x_1, x_2 \in \Xi$. The “kernels” in kernel density estimation are probability kernels, as are the stochastic transition matrices of Markov chains. (The kernels in support vector machines, however, generally are not.)

From a previous probability course, like 36-752, you will remember the measure-theoretic definition of the conditional probability of a set A , as the conditional expectation of $\mathbf{1}_A$, i.e., $\mathbb{P}(A|\mathcal{G}) = \mathbf{E}[\mathbf{1}_A(\omega)|\mathcal{G}]$. (I have explicitly included the dependence on ω to make it plain that this is a random set function.) You may also recall a potential bit of trouble, which is that even when these conditional expectations exist for all measurable sets A , it’s not necessarily true that they give us a measure, i.e., that $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i|\mathcal{G}) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|\mathcal{G})$ and all the rest of it. A version of the conditional probabilities for which they do form a measure is said to be a *regular conditional probability*. Clearly, regular conditional probabilities are all probability kernels. The ordinary rules for manipulating conditional probabilities suggest how we can define the composition of kernels.

Definition 31 (Composition of probability kernels) *Let κ_1 be a kernel from Ξ to Υ , and κ_2 a kernel from $\Xi \times \Upsilon$ to Γ . Then we define $\kappa_1 \otimes \kappa_2$ as the kernel from Ξ to $\Upsilon \times \Gamma$ such that*

$$(\kappa_1 \otimes \kappa_2)(x, B) = \int \kappa_1(x, dy) \int \kappa_2(x, y, dz) \mathbf{1}_B(y, z)$$

for every measurable $B \subseteq \Upsilon \times \Gamma$ (where z ranges over the space Γ).

Verbally, κ_1 gives us a distribution on Υ , from any starting point $x \in \Xi$. Given a pair of points $(x, y) \in \Xi \times \Upsilon$, κ_2 gives a distribution on Γ . So their composition says, basically, how to chain together conditional distributions, given a starting point.

3.2 Extension via Recursive Conditioning

With the machinery of probability kernels in place, we are in a position to give an alternative extension theorem, i.e., a different way of proving the existence of stochastic processes with specified finite-dimensional marginal distributions. In Section 2.2, we assumed some topological niceness in the sample spaces, namely that they were Borel spaces. Here, instead, we will assume probabilistic niceness in the FDDs themselves, namely that they can be obtained through composing probability kernels. This is the same as assuming that they can be obtained by chaining together regular conditional probabilities. The general form of this result is attributed in the literature to Ionescu Tulcea.

Just as proving the Kolmogorov Extension Theorem needed a measure-theoretic result, the Carathéodory Extension Theorem, our proof of the Ionescu

Tulcea Extension Theorem will require a different measure-theoretic result, which is not, so far as I know, named after anyone. (It re-appears in the theory of Markov operators, beginning in Section 10.2.)

Proposition 32 (Set functions continuous at \emptyset) *Suppose μ is a finite, non-negative, additive set function on a field \mathcal{A} . If, for any sequence of sets $A_n \in \mathcal{A}$, $A_n \downarrow \emptyset \implies \mu A_n \rightarrow 0$, then*

1. μ is countably additive on \mathcal{A} , and
2. μ extends uniquely to a measure on $\sigma(\mathcal{A})$.

PROOF: Part (1) is a weaker version of Theorem F in Chapter 2, §9 of Halmos (1950, p. 39). (When reading his proof, remember that every field of sets is also a ring of sets.) Part (2) follows from part (1) and the Carathéodory Extension Theorem (Proposition 28). \square

With this preliminary out of the way, let's turn to the main event.

Theorem 33 (Ionescu Tulcea Extension Theorem) *Consider a sequence of measurable spaces Ξ_n, \mathcal{X}_n , $n \in \mathbb{N}$. Suppose that for each n , there exists a probability kernel κ_n from $\prod_{i=1}^{n-1} \Xi_i$ to Ξ_n (taking κ_1 to be a kernel insensitive to its first argument, i.e., a probability measure). Then there exists a sequence of random variables X_n , $n \in \mathbb{N}$, taking values in the corresponding Ξ_n , such that $\mathcal{L}(X_1, X_2, \dots, X_n) = \bigotimes_{i=1}^n \kappa_i$.*

PROOF: As before, we'll be working with the cylinder sets, but now we'll make our life simpler if we consider cylinders where the base set rests in the first n spaces Ξ_1, \dots, Ξ_n . More specifically, set $\mathcal{B}_n = \bigotimes_{i=1}^n \mathcal{X}_i$ (these are the base sets), and $\mathcal{C}_n = \mathcal{B}_n \times \prod_{i=n+1}^{\infty} \Xi_i$ (these are the cylinder sets), and $\mathcal{C} = \bigcup_n \mathcal{C}_n$. \mathcal{C} clearly contains all the finite cylinders, so it generates the product σ -field on infinite sequences. We will use it as the field in Proposition 32. (Checking that \mathcal{C} is a field is entirely parallel to checking that the \mathcal{D} appearing in the proof of Theorem 29 was a field.)

For each base set $A \in \mathcal{B}_n$, let $[A]$ be the corresponding cylinder, $[A] = A \times \prod_{i=n+1}^{\infty} \Xi_i$. Notice that for every set $C \in \mathcal{C}$, there is at least one A , in some \mathcal{B}_n , such that $C = [A]$. Now we define a set function μ on \mathcal{C} .

$$\mu([A]) = \left(\bigotimes_{i=1}^n \kappa_i \right) A \quad (3.1)$$

(Checking that this is well-defined is left as an exercise, 3.2.) Clearly, this is a finite, and finitely-additive, set function defined on a field. So to use Proposition 32, we just need to check continuity from above at \emptyset . Let A_n be any sequence of sets such that $[A_n] \downarrow \emptyset$ and $A_n \in \mathcal{B}_n$. (Any sequence of sets in $\mathcal{C} \downarrow \emptyset$ can be massaged into this form.) We wish to show that $\mu([A_n]) \downarrow 0$. We'll get this to

work by considering functions which are (pretty much) conditional probabilities for these sets:

$$p_{n|k} = \left(\bigotimes_{i=k+1}^n \kappa_i \right) \mathbf{1}_{A_n}, \quad k \leq n \quad (3.2)$$

$$p_{n|n} = \mathbf{1}_{A_n} \quad (3.3)$$

Two facts follow immediately from the definitions:

$$p_{n|0} = \left(\bigotimes_{i=1}^n \kappa_i \right) \mathbf{1}_{A_n} = \mu([A_n]) \quad (3.4)$$

$$p_{n|k} = \kappa_{k+1} p_{n|k+1} \quad (3.5)$$

From the fact that the $[A_n] \downarrow \emptyset$, we know that $p_{n+1|k} \leq p_{n|k}$, for all k . This implies that $\lim_n p_{n|k} = m_k$ exists, for each k , and is approached from above. Applied to $p_{n|0}$, we see from 3.5 that $\mu([A_n]) \rightarrow m_0$. We would like $m_0 = 0$. Assume the contrary, that $m_0 > 0$. From 3.5 and the dominated convergence theorem, we can see that $m_k = \kappa_{k+1} m_{k+1}$. Hence if $m_0 > 0$, $\kappa_1 m_1 > 0$, which means (since that last expression is really an integral) that there is at least one point $x_1 \in \Xi_1$ such that $m_1(s_1) > 0$. Recursing our way down the line, we get a sequence $x = x_1, x_2, \dots \in \Xi^{\mathbb{N}}$ such that $m_n(x_1, \dots, x_n) > 0$ for all n . But now look what we've done: for each n ,

$$0 < m_n(x_1, \dots, x_n) \quad (3.6)$$

$$\leq p_{n|n}(x_1, \dots, x_n) \quad (3.7)$$

$$= \mathbf{1}_{A_n}(x_1, \dots, x_n) \quad (3.8)$$

$$= \mathbf{1}_{[A_n]}(x) \quad (3.9)$$

$$x \in [A_n] \quad (3.10)$$

This is the same as saying that $x \in \bigcap_n [A_n]$. But $[A_n] \downarrow \emptyset$, so there can be no such x . Hence $m_0 = 0$, meaning that $\mu([A_n]) \rightarrow 0$, and μ is continuous at the empty set.

Since μ is finite, finitely-additive, non-negative and continuous at \emptyset , by Proposition 32 it extends uniquely to a measure on the product σ -field. \square

Notes on the proof: It would seem natural that one could show $m_0 = 0$ directly, rather than by contradiction, but I can't think of a way to do it, and every book I've consulted does it in exactly this way.

To appreciate the simplification made possible by the notion of probability kernels, compare this proof to the one given by Fristedt and Gray (1997, §22.1).

Notice that the Daniell, Kolmogorov and Ionescu Tulcea Extension Theorems all give *sufficient* conditions for the existence of stochastic processes, not necessary ones. The necessary and sufficient condition for extending the FDDs to a process probability measure is something called σ -smoothness. (See Pollard (2002) for details.) Generally speaking, we will deal with processes which satisfy both the Kolmogorov and the Ionescu Tulcea type conditions, e.g., real-valued Markov process.

3.3 Exercises

Exercise 3.1 (Łomnick-Ulam Theorem on infinite product measures)

Let T be an uncountable index set, and $(\Xi_t, \mathcal{X}_t, \mu_t)$ a collection of probability spaces. Show that there exist independent random variables X_t in Ξ_t with distributions μ_t . Hint: use the Ionescu Tulcea theorem on countable subsets of T , and then imitate the proof of the Kolmogorov extension theorem.

Exercise 3.2 (Measures of cylinder sets) In the proof of the Ionescu Tulcea Theorem, we employed a set function on the finite cylinder sets, where the measure of an infinite-dimensional cylinder set $[A]$ is taken to be the measure of its finite-dimensional base set A . However, the same cylinder set can be specified by different base sets, so it is necessary to show that Equation 3.1 has a unique value on its right-hand side. In what follows, C is an arbitrary member of the class \mathcal{C} .

1. Show that, when $A, B \in \mathcal{B}_n$, $[A] = [B]$ iff $A = B$. That is, two cylinders generated by bases of equal dimensionality are equal iff their bases are equal.
2. Show that there is a smallest n such that $C = [A]$ for an $A \in \mathcal{B}_n$. Conclude that the right-hand side of Equation 3.1 could be made well-defined if we took n there to be this least possible n .
3. Suppose that $m < n$, $A \in \mathcal{B}_m$, $B \in \mathcal{B}_n$, and $[A] = [B]$. Show that $B = A \times \prod_{i=m+1}^n \Xi_i$.
4. Continuing the situation in (iii), show that

$$\left(\bigotimes_{i=1}^m \kappa_i \right) A = \left(\bigotimes_{i=1}^n \kappa_i \right) B$$

Conclude that the right-hand side of Equation 3.1 is well-defined, as promised.

Part II

One-Parameter Processes in
General

Chapter 4

One-Parameter Processes, Usually Functions of Time

Section 4.1 defines one-parameter processes, and their variations (discrete or continuous parameter, one- or two- sided parameter), including many examples.

Section 4.2 shows how to represent one-parameter processes in terms of “shift” operators.

We’ve been doing a lot of pretty abstract stuff, but the point of this is to establish a common set of tools we can use across many different concrete situations, rather than having to build very similar, specialized tools for each distinct case. Today we’re going to go over some examples of the kind of situation our tools are supposed to let us handle, and begin to see how they let us do so. In particular, the two classic areas of application for stochastic processes are dynamics (systems changing over time) and inference (conclusions changing as we acquire more and more data). Both of these can be treated as “one-parameter” processes, where the parameter is time in the first case and sample size in the second.

4.1 One-Parameter Processes

The index set T isn’t, usually, an amorphous abstract set, but generally something with some kind of topological or geometrical structure. The number of (topological) dimensions of this structure is the number of *parameters* of the process.

Definition 34 (One-Parameter Process) *A process whose index set T has one dimension is a one-parameter process. A process whose index set has more than one dimension is a multi-parameter process. A one-parameter process is*

discrete or continuous depending on whether its index set is countable or uncountable. A one-parameter process where the index set has a minimal element it is one-sided, otherwise it is two-sided.

\mathbb{N} is a one-sided discrete index set, \mathbb{Z} a two-sided discrete index set, \mathbb{R}^+ (including zero!) is a one-sided continuous index set, and \mathbb{R} a two-sided continuous index set.

Most of this course will be concerned with one-parameter processes, which are intensely important in applications. This is because the one-dimensional parameter is usually either time (when we're doing dynamics) or sample size (when we're doing inference), or both at once. There are also some important cases where the single parameter is space.

Example 35 (Bernoulli process) *You all know this one: a one-sided infinite sequence of independent, identically-distributed binary variables, where, for all t , $X_t = 1$ with probability p .*

Example 36 (Markov models) *Markov chains are discrete-parameter stochastic processes. They may be either one-sided or two-sided. So are Markov models of order k , and hidden Markov models. Continuous-time Markov processes are, naturally enough, continuous-parameter stochastic processes, and again may be either one-sided or two-sided.*

Instances of physical processes that may be represented by Markov models include: the positions and velocities of the planets; the positions and velocities of molecules in a gas; the pressure, temperature and volume of the gas (Keizer, 1987); the position and velocity of a tracer particle in a turbulent fluid flow; the three-dimensional velocity field of a turbulent fluid (Frisch, 1995); the gene pool of an evolving population (Fisher, 1958; Haldane, 1932; Gillespie, 1998). Instances of physical processes that may be represented by hidden Markov models include: the spike trains of neurons (Tuckwell, 1989); the sonic wave-forms of human speech (Rabiner, 1989); many economic and social time-series (Durbin and Koopman, 2001); etc.

Example 37 (“White Noise” (Not Really)) *For each $t \in \mathbb{R}^+$, let $X_t \sim \mathcal{N}(0, 1)$, all mutually independent of one another. This is a process with a one-sided continuous parameter.*

It would be character building, at this point, to convince yourself that the process just described exists. (You will need the Kolmogorov Extension Theorem, 29).

Example 38 (Wiener Process) *Here $T = \mathbb{R}^+$ and $\Xi = \mathbb{R}$. The Wiener process is the continuous-parameter random process where*

1. $W(0) = 0$,

2. for any three times, $t_1 < t_2 < t_3$, $W(t_2) - W(t_1)$ and $W(t_3) - W(t_2)$ are independent (the “independent increments” property),
3. $W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1)$ and
4. $W(t, \omega)$ is a continuous function of t for almost all ω .

We will spend a lot of time with the Wiener process, because it turns out to play a role in the theory of stochastic processes analogous to that played by the Gaussian distribution in elementary probability — the easily-manipulated, formally-nice distribution delivered by limit theorems.

When we examine the Wiener process in more detail, we will see that it almost never has a derivative. Nonetheless, in a sense which will be made clearer when we come to stochastic calculus, the Wiener process can be regarded as the integral over time of something *very like* white noise, as described in the preceding example.

Example 39 (Logistic Map) Let $T = \mathbb{N}$, $\Xi = [0, 1]$, $X(0) \sim U(0, 1)$, and $X(t+1) = aX(t)(1 - X(t))$, $a \in [0, 4]$. This is called the logistic map. Notice that all the randomness is in the initial value $X(0)$; given the initial condition, all later values $X(t)$ are fixed. Nonetheless, this is a Markov process, and we will see that, at least for certain values of a , it satisfies versions of the laws of large numbers and the central limit theorem. In fact, large classes of deterministic dynamical systems have such stochastic properties.

Example 40 (Symbolic Dynamics of the Logistic Map) Let $X(t)$ be the logistic map, as in the previous example, and let $S(t) = 0$ if $X(t) \in [0, 0.5]$ and $S(t) = 1$ if $X(t) \in [0.5, 1]$. That is, we partition the state space of the logistic map, and record which cell of the partition the original process finds itself in. $X(t)$ is a Markov process, but these “symbolic” dynamics are not necessarily Markovian. We will want to know when functions of Markov processes are themselves Markov. We will also see that there is a sense in which, Markovian or not, this partition is exactly as informative as the original, continuous state — that it is generating. Finally, when $a = 4$ in the logistic map, the symbol sequence is actually a Bernoulli process, so that a deterministic function of a completely deterministic dynamical system provides a model of IID randomness.

Here are some examples where the parameter is sample size.

Example 41 (IID Samples) Let X_i , $i \in \mathbb{N}$ be samples from an IID distribution, and $Z_n = \frac{1}{n} \sum_{i=1}^n X_i$ be the sample mean. Then Z_n is a one-parameter stochastic process. The point of the ordinary law of large numbers is to reassure us that $Z_n \rightarrow \mathbf{E}[X_n]$ a.s. The point of the central limit theorem is to reassure us that $\sqrt{n}(Z_n - \mathbf{E}[X])$ has constant average size, so that the sampling fluctuation $Z_n - \mathbf{E}[X]$ must be shrinking as \sqrt{n} grows.

If X_i is the indicator of a set, this convergence means that the relative frequency with which the set is occupied will converge on its true probability.

Example 42 (Non-IID Samples) Let X_i be a non-IID one-sided discrete-parameter process, say a Markov chain, and again let Z_n be its sample mean, now often called its “time average”. The usual machinery of the law of large numbers and the central limit theorem are now inapplicable, and someone who has just taken 36-752 has, strictly speaking, no idea as to whether or not their time averages will converge on expectations. Under the heading of ergodic theory, we will see when this will happen. Since this is the situation with all interesting time series, the application of statistical methods to situations where we cannot contrive to randomize depends crucially on ergodic considerations.

Example 43 (Estimating Distributions) Recall Example 10, where we looked at the sequence of empirical distributions \hat{P}_n for samples from an IID data-source. We would like to be able to say that \hat{P}_n converges on P . The usual way to do this, if our samples are of a real-valued random variable, is to consider the empirical cumulative distribution function, F_n . For each n , this may be regarded as a one-parameter random process ($T = \mathbb{R}$, $\Xi = [0, 1]$), and the difficulty is to show that this sequence of random processes converges to F . The usual way is to show that $E_n \equiv \sqrt{n}(F_n - F)$, the empirical process, converges to a fixed process, related to the Wiener process, depending only on the true distribution F . (See Figure 4.1.) Since the magnitude of the limiting process is constant, and \sqrt{n} grows, it follows that $F_n - F$ must shrink. So theorizing even this elementary bit of statistical inference really requires two doses of stochastic process theory, one to get a grip on F_n at each n , and the other to get a grip on what happens to F_n as n grows.

Example 44 (Doob’s Martingale) Let X be a random variable, and \mathcal{F}_i , $i \in \mathbb{N}$, a sequence of increasing σ -algebras (i.e. a filtration). Then $Y_i = \mathbf{E}[X|\mathcal{F}_i]$ is a one-sided discrete-parameter stochastic process, and in fact a martingale. Martingales in general are one-parameter stochastic processes. Note that posterior mean parameter estimates, in Bayesian inference, are instances of Doob’s martingale.

Here are some examples where the one-dimensional parameter is neither time nor sample size.

Example 45 (The One-Dimensional Ising Model) This system serves as a toy model of magnetism in theoretical physics. Atoms sit evenly spaced on the points of a regular, infinite, one-dimensional crystalline lattice. Each atom has a magnetic moment, which is either pointing north (+1) or south (−1). Atoms are more likely to point north if their neighbors point north, and vice-versa. The natural index here is \mathbb{Z} , so the parameter is discrete and two-sided.

Example 46 (Text) Text (at least in most writing systems!) can be represented by a sequence of discrete values at discrete, ordered locations. Since texts can be arbitrarily long, but they all start somewhere, they are discrete-parameter, one-sided processes. Or, more exactly, once we specify a distribution over sequences from the appropriate alphabet, we will have such a process.

Example 47 (Polymer Sequences) *Similarly, DNA, RNA and proteins are all heteropolymers — compounds in which distinct constituent chemicals (the monomers) are joined in a sequence. Position along the sequence (chromosome, protein) provides the index, and the nature of the monomer at that position the value.*

Linguists believe that no Markovian model (with finitely many states) can capture human language (Chomsky, 1957; Pullum, 1991). Whether this is true of DNA sequences is not known. In both cases, hidden Markov models are used extensively Baldi and Brunak (2001); Charniak (1993), even if they can only be approximately true of language.

4.2 Operator Representations of One-Parameter Processes

Consider our favorite discrete-parameter process, say X_t . If we try to relate X_t to its *history*, i.e., to the preceding values from the process, we will often get a horribly complicated probabilistic expression. There is however an alternative, which represents the dynamical part of any process as a remarkably simple semi-group of operators.

Definition 48 (Shift Operators) *Consider Ξ^T , $T = \mathbb{N}$, $= \mathbb{Z}$, $= \mathbb{R}^+$ or $= \mathbb{R}$. The shift-by- τ operator Σ_τ , $\tau \geq 0$, maps Ξ^T into itself by shifting forward in time: $(\Sigma_\tau)x(t) = x(t + \tau)$. The collection of all shift operators is the shift semi-group or time-evolution semi-group.*

(A semi-group does not need to have an identity element, and one which does is technically called a “monoid”. No one talks about the shift monoid or time-evolution monoid, however.)

Before we had a Ξ -valued stochastic process X on T , i.e., our process was a random function from T to Ξ . To extract individual random variables, we used the projection operators π_t , which took X to X_t . With the shift operators, we simply have $\pi_t = \pi_0 \circ \Sigma_t$. To represent the passage of time, then, we just apply elements of this semi-group to the function space. Rather than having complicated dynamics which gets us from one value to the next, by working with shifts on function space, all of the complexity is shifted to the initial distribution. This will prove to be extremely useful when we consider stationary processes in the next lecture, and even more useful when, later on, we want to extend the limit theorems from IID sequences to dependent processes.

4.3 Exercises

Exercise 4.1 (Existence of proto-Wiener processes) *Use Theorem 29 and the properties of Gaussian distributions to show that processes exist which satisfy*

points (1)–(3) of Example 38 (but not necessarily continuity). You will want to begin by finding a way to write down the FDDs recursively.

Exercise 4.2 (Time-Evolution Semi-Group) *These are all very easy, but worth the practice.*

1. Verify that the time-evolution semi-group, as described, is a monoid, i.e., that it is closed under composition, that composition is associative, and that there is an identity element. What, in fact, is the identity?
2. Can a one-sided process have a shift group, rather than just a semi-group?
3. Verify that $\pi_\tau = \pi_0 \circ \Sigma_\tau$.
4. Verify that, for a discrete-parameter process, $\Sigma_t = (\Sigma_1)^t$, and so Σ_1 generates the semi-group. (For this reason it is often abbreviated to Σ .)

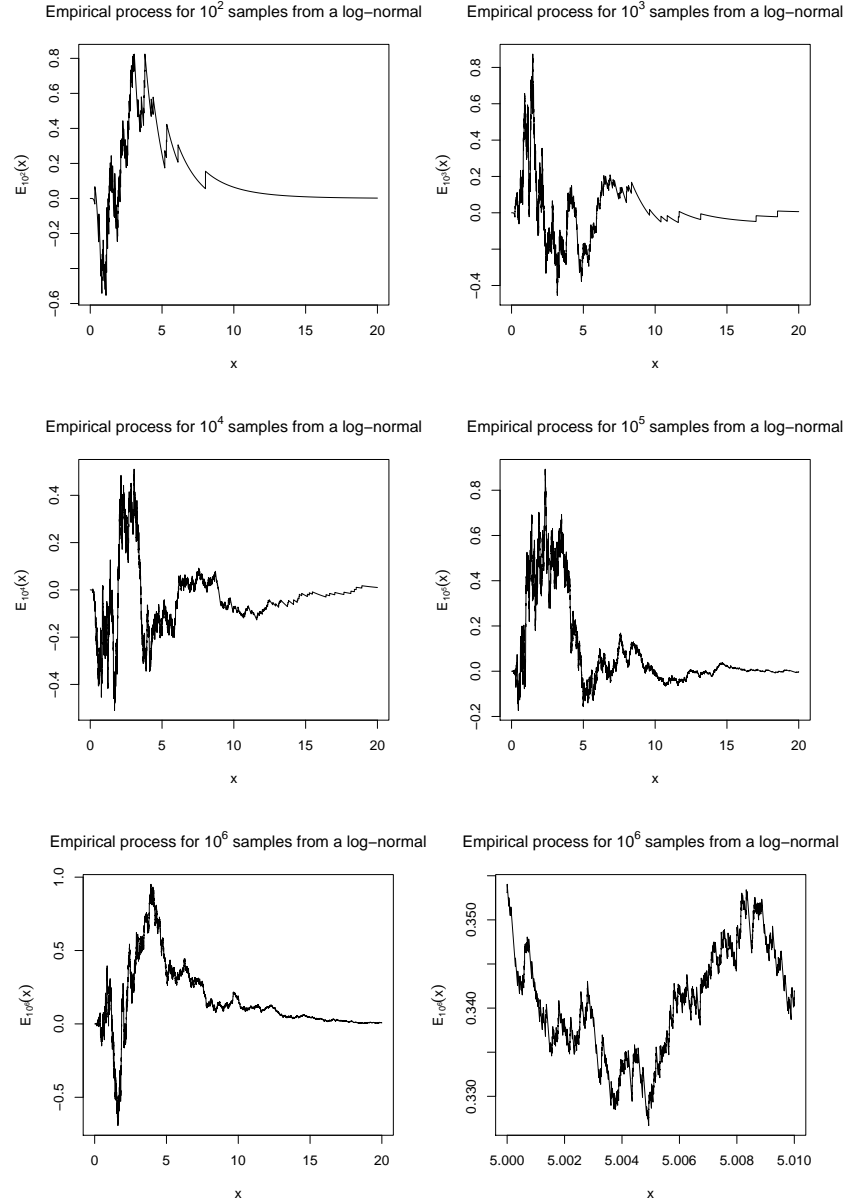


Figure 4.1: Empirical process for successively larger samples from a log-normal distribution, mean of the log = 0.66, standard deviation of the log = 0.66. The samples were formed by taking the first n variates from an over-all sample of 1,000,000 variates. The bottom right figure magnifies a small part of E_{10^6} to show how spiky the paths are.

Chapter 5

Stationary One-Parameter Processes

Section 5.1 describes the three main kinds of stationarity: strong, weak, and conditional.

Section 5.2 relates stationary processes to the shift operators introduced in the last chapter, and to measure-preserving transformations more generally.

5.1 Kinds of Stationarity

Stationary processes are those which are, in some sense, the same at different times — slightly more formally, which are invariant under translation in time. There are three particularly important forms of stationarity: strong or strict, weak, and conditional.

Definition 49 (Strong Stationarity) *A one-parameter process is strongly stationary or strictly stationary when all its finite-dimensional distributions are invariant under translation of the indices. That is, for all $\tau \in T$, and all $J \in \text{Fin}(T)$,*

$$\mathcal{L}(X_J) = \mathcal{L}(X_{J+\tau}) \quad (5.1)$$

Notice that when the parameter is discrete, we can get away with just checking the distributions of blocks of consecutive indices.

Definition 50 (Weak Stationarity) *A one-parameter process is weakly stationary or second-order stationary when, for all $t \in T$,*

$$\mathbf{E}[X_t] = \mathbf{E}[X_0] \quad (5.2)$$

and for all $t, \tau \in T$,

$$\mathbf{E}[X_\tau X_{\tau+t}] = \mathbf{E}[X_0 X_t] \quad (5.3)$$

At this point, you should check that a weakly stationary process has time-invariant correlations. (We will say much more about this later.) You should also check that strong stationarity implies weak stationarity. It will turn out that weak and strong stationarity coincide for Gaussian processes, but not in general.

Definition 51 (Conditional (Strong) Stationarity) *A one-parameter process is conditionally stationary if its conditional distributions are invariant under time-translation: $\forall n \in \mathbb{N}$, for every set of $n + 1$ indices $t_1, \dots, t_{n+1} \in T$, $t_i < t_{i+1}$, and every shift τ ,*

$$\mathcal{L}(X_{t_{n+1}} | X_{t_1}, X_{t_2} \dots X_{t_n}) = \mathcal{L}(X_{t_{n+1}+\tau} | X_{t_1+\tau}, X_{t_2+\tau} \dots X_{t_n+\tau}) \quad (5.4)$$

(a.s.).

Strict stationarity implies conditional stationarity, but the converse is not true, in general. (Homogeneous Markov processes, for instance, are all conditionally stationary, but most are not stationary.) Many methods which are normally presented using strong stationarity can be adapted to processes which are merely conditionally stationary.¹

Strong stationarity will play an important role in what follows, because it is the natural generaliation of the IID assumption to situations with dependent variables — we allow for dependence, but the probabilistic set-up remains, in a sense, unchanging. This will turn out to be enough to let us learn a great deal about the process from observation, just as in the IID case.

5.2 Strictly Stationary Processes and Measure-Preserving Transformations

The shift-operator representation of Section 4.2 is particularly useful for strongly stationary processes.

Theorem 52 (Stationarity is Shift-Invariance) *A process X with measure μ is strongly stationary if and only if μ is shift-invariant, i.e., $\mu = \mu \circ \Sigma_\tau^{-1}$ for all Σ_τ in the time-evolution semi-group.*

PROOF: “If” (invariant distributions imply stationarity): For any finite collection of indices J , $\mathcal{L}(X_J) = \mu \circ \pi_J^{-1}$ (Lemma 25), and similarly $\mathcal{L}(X_{J+\tau}) = \mu \circ \pi_{J+\tau}^{-1}$.

$$\pi_{J+\tau} = \pi_J \circ \Sigma_\tau \quad (5.5)$$

$$\pi_{J+\tau}^{-1} = \Sigma_\tau^{-1} \circ \pi_J^{-1} \quad (5.6)$$

$$\mu \circ \pi_{J+\tau}^{-1} = \mu \circ \Sigma_\tau^{-1} \circ \pi_J^{-1} \quad (5.7)$$

$$\mathcal{L}(X_{J+\tau}) = \mu \circ \pi_{J+\tau}^{-1} \quad (5.8)$$

$$= \mathcal{L}(X_J) \quad (5.9)$$

¹For more on conditional stationarity, see Caires and Ferreira (2005).

“Only if”: The statement that $\mu = \mu \circ \Sigma_\tau^{-1}$ really means that, for any set $A \in \mathcal{X}^T$, $\mu(A) = \mu(\Sigma_\tau^{-1}A)$. Suppose A is a finite-dimensional cylinder set. Then the equality holds, because all the finite-dimensional distributions agree (by hypothesis). But this means that X and $\Sigma_\tau X$ are two processes with the same finite-dimensional distributions, and so their infinite-dimensional distributions agree (Theorem 23), and the equality holds on all measurable sets A . \square

This can be generalized somewhat.

Definition 53 (Measure-Preserving Transformation) *A measurable mapping F from a measurable space Ξ, \mathcal{X} into itself preserves measure μ iff, $\forall A \in \mathcal{X}$, $\mu(A) = \mu(F^{-1}A)$, i.e., iff $\mu = \mu \circ F^{-1}$. This is true just when $F(X) \stackrel{d}{=} X$, when X is a Ξ -valued random variable with distribution μ . We will often say that F is measure-preserving, without qualification, when the context makes it clear which measure is meant.*

Remark on the definition. It is natural to wonder why we write the defining property as $\mu = \mu \circ F^{-1}$, rather than $\mu = \mu \circ F$. There is actually a subtle difference, and the former is stronger than the latter. To see this, unpack the statements, yielding respectively

$$\forall A \in \mathcal{X}, \mu(A) = \mu(F^{-1}(A)) \quad (5.10)$$

$$\forall A \in \mathcal{X}, \mu(A) = \mu(F(A)) \quad (5.11)$$

To see that Eq. 5.10 implies Eq. 5.11, pick any measurable set B , and then apply 5.10 to $F(B)$ (which is $\in \mathcal{X}$, because F is measurable). To go the other way, from 5.11 to 5.10, it would have to be the case that, $\forall A \in \mathcal{X}$, $\exists B \in \mathcal{X}$ such that $A = F(B)$, i.e., every measurable set would have to be the image, under F , of another measurable set. This is not necessarily the case; it would require, for starters, that F be onto (surjective).

Theorem 52 says that every stationary process can be represented by a measure-preserving transformation, namely the shift. Since measure-preserving transformations arise in many other ways, however, it is useful to know about the processes they generate.

Corollary 54 (Measure-preservation implies stationarity) *If F is a measure-preserving transformation on Ξ with invariant measure μ , and X is a Ξ -valued random variable, $\mathcal{L}(X) = \mu$, then the sequence $F^n(X)$, $n \in \mathbb{N}$ is strongly stationary.*

PROOF: Consider shifting the sequence $F^n(X)$ by one: the n^{th} term in the shifted sequence is $F^{n+1}(X) = F^n(F(X))$. But since $\mathcal{L}(F(X)) = \mathcal{L}(X)$, by hypothesis, $\mathcal{L}(F^{n+1}(X)) = \mathcal{L}(F^n(X))$, and the measure is shift-invariant. So, by Theorem 52, the process $F^n(X)$ is stationary.

5.3 Exercises

Exercise 5.1 (Functions of Stationary Processes) *Use Corollary 54 to show that if g is any measurable function on Ξ , then the sequence $g(F^n(X))$ is also stationary.*

Exercise 5.2 (Continuous Measure-Preserving Families of Transformations) *Let F_t , $t \in \mathbb{R}^+$, be a semi-group of measure-preserving transformations, with F_0 being the identity. Prove the analog of Corollary 54, i.e., that $F_t(X)$, $t \in \mathbb{R}^+$, is a stationary process.*

Exercise 5.3 (The Logistic Map as a Measure-Preserving Transformation) *The logistic map with $a = 4$ is a measure-preserving transformation, and the measure it preserves has the density $1/\pi\sqrt{x(1-x)}$ (on the unit interval).*

1. *Verify that this density is invariant under the action of the logistic map.*
2. *Simulate the logistic map with uniformly distributed X_0 . What happens to the density of X_t as $t \rightarrow \infty$?*

Chapter 6

Random Times and Their Properties

Section 6.1 recalls the definition of a filtration (a growing collection of σ -fields) and of “stopping times” (basically, measurable random times).

Section 6.2 defines various sort of “waiting” times, including hitting, first-passage, and return or recurrence times.

Section 6.3 proves the Kac recurrence theorem, which relates the finite-dimensional distributions of a stationary process to its mean recurrence times.

6.1 Reminders about Filtrations and Stopping Times

You will have seen these in 36-752 as part of martingale theory, though their application is more general, as we’ll see.

Definition 55 (Filtration) *Let T be an ordered index set. A collection \mathcal{F}_t , $t \in T$ of σ -algebras is a filtration (with respect to this order) if it is non-decreasing, i.e., $f \in \mathcal{F}_t$ implies $f \in \mathcal{F}_s$ for all $s > t$. We generally abbreviate this filtration by $\{\mathcal{F}_t\}$. Define $\{\mathcal{F}\}_{t+}$ as $\bigcap_{s>t} \mathcal{F}_s$. If $\{\mathcal{F}\}_{t+} = \{\mathcal{F}_t\}$, then $\{\mathcal{F}_t\}$ is right-continuous.*

Recall that we generally think of a σ -algebra as representing available information — for any event $f \in \mathcal{F}$, we can answer the question “did f happen?” A filtration is a way of representing our information about a system growing over time. To see what right-continuity is about, imagine it failed, which would mean $\mathcal{F}_t \subset \bigcap_{s>t} \mathcal{F}_s$. Then there would have to be events which were detectable at *all* times after t , but not at t itself, i.e., some sudden jump in our information right after t . This is what right-continuity rules out.

Definition 56 (Adapted Process) A stochastic process X on T is adapted to a filtration $\{\mathcal{F}_t\}$ if $\forall t$, X_t is \mathcal{F}_t -measurable. Any process is adapted to the filtration it induces, $\sigma\{X_s : s \leq t\}$. This natural filtration is written $\{\mathcal{F}_t^X\}$.

A process being adapted to a filtration just means that, at every time, the filtration gives us enough information to find the value of the process.

Definition 57 (Stopping Time, Optional Time) An optional time or a stopping time, with respect to a filtration $\{\mathcal{F}_t\}$, is a T -valued random variable τ such that, for all t ,

$$\{\omega \in \Omega : \tau(\omega) \leq t\} \in \mathcal{F}_t \quad (6.1)$$

If Eq. 6.1 holds with $<$ instead of \leq , then τ is weakly optional or a weak stopping time.

Basically, all we're doing here is defining what we mean by "a random time at which something detectable happens". That being the case, it is natural to ask what information we have when that detectable thing happens.

Definition 58 (\mathcal{F}_τ for a Stopping Time τ) If τ is a $\{\mathcal{F}_t\}$ stopping time, then the σ -algebra \mathcal{F}_τ is given by

$$\mathcal{F}_\tau \equiv \{A \in \mathcal{F} : \forall t, A \cap \{\omega : \tau(\omega) \leq t\} \in \mathcal{F}_t\} \quad (6.2)$$

I admit that the definition of \mathcal{F}_τ looks bizarre, and I won't blame you if you have to read it a few times to convince yourself it isn't circular. Here is a simple case where it makes sense. Let X be a one-sided process, and τ a discrete $\{\mathcal{F}_t^X\}$ stopping time. Then

$$\mathcal{F}_\tau^X = \sigma(X(t \wedge \tau) : t \geq 0) \quad (6.3)$$

That is, \mathcal{F}_τ^X is everything we know from observing X up to time τ . (This is Exercise 6.2.) The convoluted-looking definition of \mathcal{F}_τ carries this idea over to the more general situation where τ is continuous and we don't necessarily have a single variable generating the filtration. A filtration lets us tell whether some event A happened by the random time τ if simultaneously gives us enough information to notice τ and A .

The process $Y(t) = X(t \wedge \tau)$ follows along with X up until τ , at which point it becomes fixed in place. It is accordingly called an *arrested*, *halted* or *stopped* version of the process. This seems to be the origin of the name "stopping time".

6.2 Waiting Times

"Waiting times" are particular kinds of optional kinds: how much time must elapse before a given event happens, either from a particular starting point, or averaging over all trajectories? Often, these are of particular interest in themselves, and some of them can be related to other quantities of interest.

Definition 59 (Hitting Time) *Given a one-sided Ξ -valued process X , the hitting time τ_B of a measurable set $B \subset \Xi$ is the first time at which $X(t) \in B$;*

$$\tau_B = \inf \{t > 0 : X_t \in B\} \quad (6.4)$$

Example 60 (Fixation through Genetic Drift) *Consider the variation in a given locus (roughly, gene) in an evolving population. If there are k different versions of the gene (“alleles”), the state of the population can be represented by a vector $X(t) \in \mathbb{R}^k$, where at each time $X_i(t) \geq 0$ and $\sum_i X_i(t) = 1$. This set is known as the k -dimensional probability simplex S_k . We say that a certain allele has been fixed in the population or gone to fixation at t if $X_i(t) = 1$ for some i , meaning that all members of the population have that version of the gene. Fixation corresponds to $X(t) \in V$, where V consists of the vertices of the simplex. An important question in evolutionary theory is how long it takes the population to go to fixation. By comparing the actual rate of fixation to that expected under a model of adaptively-neutral genetic drift, it is possible to establish that some genes are under the influence of natural selection.*

Gillespie (1998) is a nice introduction to population genetics, including this problem among many others, using only elementary probability. More sophisticated models treat populations as measure-valued stochastic processes.

Example 61 (Stock Options) *A stock option¹ is a legal instrument giving the holder the right to buy a stock at a specified price (the strike price, c) before a certain expiration date t_e . The point of the option is that, if you exercise it at a time t when the price of the stock $p(t)$ is above c , you can turn around and sell the stock to someone else, making a profit of $p(t) - c$. When $p(t) > c$, the option is said to be in money or above water. Options can themselves be sold, and the value of an option depends on how much money it could be used to make, which in turn depends on the probability that it will be “in money” before time t_e . An important part of mathematical finance thus consists of problems of the form “assuming prices $p(t)$ follow a process distribution μ , what is the distribution of hitting times of the set $p(t) > c$?”*

While the financial industry is a major consumer of stochastics, and it has a legitimate role to play in capitalist society, I do hope you will find something more interesting to do with your new-found mastery of random processes, so I will not give many examples of this sort. If you want much, much more, read Shiryaev (1999).

Definition 62 (First Passage Time) *When $\Xi = \mathbb{R}$ or \mathbb{Z} , we call the hitting time of the origin the time of first passage through the origin, and similarly for other points.*

¹ Actually, this is just one variety of option (an “American call”), out of a huge variety. I will not go into details.

Definition 63 (Return Time, Recurrence Time) Fix a set $B \in \Xi$. Suppose that $X(t_0) \in B$. Then the return time or first return time of B is recurrence time of B is $\inf\{t > t_0 : X(t) \in B\}$, and the recurrence time θ_B is the difference between the first return time and t_0 .

Note 1: If I'm to be honest with you, I should admit that “return time” and “recurrence time” are used more or less interchangeably in the literature to refer to either the time *coordinate* of the first return (what I'm calling the return time) or the time *interval* which elapses before that return (what I'm calling the recurrence time). I will try to keep these straight here. Check definitions carefully when reading papers!

Note 2: Observe that if we have a discrete-parameter process, and are interested in recurrences of a finite-length sequence of observations $w \in \Xi^k$, we can handle this situation by the device of working with the shift operator in sequence space.

The question of whether any of these waiting times is optional (i.e., measurable) must, sadly, be raised. The following result is generally enough for our purposes.

Proposition 64 (Some Sufficient Conditions for Waiting Times to be Weakly Optional) Let X be a Ξ -valued process on a one-sided parameter T , adapted to a filtration $\{\mathcal{F}_t\}$, and let B be an arbitrary measurable set in Ξ . Then τ_B is weakly $\{\mathcal{F}_t\}$ -optional under any of the following (sufficient) conditions, and $\{\mathcal{F}_t\}$ -optional under the first two:

1. T is discrete.
2. T is \mathbb{R}^+ , Ξ is a metric space, B is closed, and $X(t)$ is a continuous function of t .
3. T is \mathbb{R}^+ , Ξ is a topological space, B is open, and $X(t)$ is right-continuous as a function of t .

PROOF: See, for instance, Kallenberg, Lemma 7.6, p. 123. \square

6.3 Kac's Recurrence Theorem

For strictly stationary, discrete-parameter sequences, a very pretty theorem, due to Mark Kac (1947), relates the probability of seeing a particular event to the mean time between recurrences of the event. Throughout, we consider an arbitrary Ξ -valued process X , subject only to the requirements of stationarity and a discrete parameter.

Fix an arbitrary measurable set $A \in \Xi$ with $\mathbb{P}(X_1 \in A) > 0$, and consider a new process $Y(t)$, where $Y_t = 1$ if $X_t \in A$ and $Y_t = 0$ otherwise. By Exercise 5.1, Y_t is also stationary. Thus $\mathbb{P}(X_1 \in A, X_2 \notin A) = \mathbb{P}(Y_1 = 1, Y_2 = 0)$. Let us abbreviate $\mathbb{P}(Y_1 = 0, Y_2 = 0, \dots, Y_{n_1} = 0, Y_n = 0)$ as w_n ; this is the probability of making n consecutive observations, none of which belong to the event

A. Clearly, $w_n \geq w_{n+1}$. Similarly, let $e_n = \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_n = 0)$ and $r_n = \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_n = 1)$ — these are, respectively, the probabilities of starting in A and not returning within $n - 1$ steps, and of starting in A and returning for the first time after $n - 2$ steps. (Set e_1 to $\mathbb{P}(Y_1 = 1)$, and $w_0 = e_0 = 1$.)

Lemma 65 (Some Recurrence Relations for Kac's Theorem) *The following recurrence relations hold among the probabilities w_n , e_n and r_n :*

$$e_n = w_{n-1} - w_n, \quad n \geq 1 \quad (6.5)$$

$$r_n = e_{n-1} - e_n, \quad n \geq 2 \quad (6.6)$$

$$r_n = w_{n-2} - 2w_{n-1} + w_n, \quad n \geq 2 \quad (6.7)$$

PROOF: To see the first equality, notice that

$$\begin{aligned} \mathbb{P}(Y_1 = 0, Y_2 = 0, \dots, Y_{n-1} = 0) & \quad (6.8) \\ &= \mathbb{P}(Y_2 = 0, Y_3 = 0, \dots, Y_n = 0) \\ &= \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_n = 0) + \mathbb{P}(Y_1 = 0, Y_2 = 0, \dots, Y_n = 0) \end{aligned} \quad (6.9)$$

using first stationarity and then total probability. To see the second equality, notice that, by total probability,

$$\begin{aligned} \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{n-1} = 0) & \quad (6.10) \\ &= \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{n-1} = 0, Y_n = 0) + \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{n-1} = 0, Y_n = 1) \end{aligned}$$

The third relationship follows from the first two. \square

Theorem 66 (Recurrence in Stationary Processes) *Let X be a Ξ -valued discrete-parameter stationary process. For any set A with $\mathbb{P}(X_1 \in A) > 0$, for almost all ω such that $X_1(\omega) \in A$, there exists a τ for which $X_\tau(\omega) \in A$.*

$$\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) = 1 \quad (6.11)$$

PROOF: The event $\{\theta_A = k, X_1 \in A\}$ is the same as the event $\{Y_1 = 1, Y_2 = 0, \dots, Y_{k+1} = 1\}$. Since $\mathbb{P}(X_1 \in A) > 0$, we can handle the conditional probabilities in an elementary fashion:

$$\mathbb{P}(\theta_A = k | X_1 \in A) = \frac{\mathbb{P}(\theta_A = k, X_1 \in A)}{\mathbb{P}(X_1 \in A)} \quad (6.12)$$

$$= \frac{\mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} \quad (6.13)$$

$$\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) = \frac{\sum_{k=1}^{\infty} \mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} \quad (6.14)$$

$$= \frac{\sum_{k=2}^{\infty} r_k}{e_1} \quad (6.15)$$

Now consider the finite sums, and apply Eq. 6.7.

$$\sum_{k=2}^n r_k = \sum_{k=2}^n w_{k-2} - 2w_{k-1} + w_k \quad (6.16)$$

$$= \sum_{k=0}^{n-2} w_k + \sum_{k=2}^n w_k - 2 \sum_{k=1}^{n-1} w_k \quad (6.17)$$

$$= w_0 + w_n - w_1 - w_{n-1} \quad (6.18)$$

$$= (w_0 - w_1) - (w_{n-1} - w_n) \quad (6.19)$$

$$= e_1 - (w_{n-1} - w_n) \quad (6.20)$$

where the last line uses Eq. 6.6. Since $w_{n-1} \geq w_n$, there exists a $\lim_n w_n$, which is ≥ 0 since every individual w_n is. Hence $\lim_n w_{n-1} - w_n = 0$.

$$\sum_{k=1}^{\infty} \mathbb{P}(\theta_A = k | X_1 \in A) = \frac{\sum_{k=2}^{\infty} r_k}{e_1} \quad (6.21)$$

$$= \lim_{n \rightarrow \infty} \frac{e_1 - (w_{n-1} - w_n)}{e_1} \quad (6.22)$$

$$= \frac{e_1}{e_1} \quad (6.23)$$

$$= 1 \quad (6.24)$$

which was to be shown. \square

Corollary 67 (Poincaré Recurrence Theorem) *Let F be a transformation which preserves measure μ . Then for any set A of positive μ measure, for μ -almost-all $x \in A$, $\exists n \geq 1$ such that $F^n(x) \in A$.*

PROOF: A direct application of the theorem, given the relationship between stationary processes and measure-preserving transformations we established by Corollary 54. \square

Corollary 68 (“Nietzsche”) *In the set-up of the previous theorem, if $X_1(\omega) \in A$, then $X_t \in A$ for infinitely many t (a.s.).*

PROOF: Repeated application of the theorem yields an infinite sequence of times $\tau_1, \tau_2, \tau_3, \dots$ such that $X_{\tau_i}(\omega) \in A$, for almost all ω such that $X_1(\omega) \in A$ in the first place. \square

Now that we’ve established that once something happens, it will happen again and again, we would like to know how long we have to wait between recurrences.

Theorem 69 (Kac’s Recurrence Theorem) *Continuing the previous notation, $\mathbf{E}[\theta_A | X_1 \in A] = 1/\mathbb{P}(X_1 \in A)$ if and only if $\lim_n w_n = 0$.*

PROOF: “If”: Unpack the expectation:

$$\mathbf{E}[\theta_A | X_1 \in A] = \sum_{k=1}^{\infty} k \frac{\mathbb{P}(Y_1 = 1, Y_2 = 0, \dots, Y_{k+1} = 1)}{\mathbb{P}(Y_1 = 1)} \quad (6.25)$$

$$= \frac{1}{\mathbb{P}(X_1 \in A)} \sum_{k=1}^{\infty} k r_{k+1} \quad (6.26)$$

so we just need to show that the last series above sums to 1. Using Eq. 6.7 again,

$$\sum_{k=1}^n k r_{k+1} = \sum_{k=1}^n k(w_{k-1} - 2w_k + w_{k+1}) \quad (6.27)$$

$$= \sum_{k=1}^n k w_{k-1} + \sum_{k=1}^n k w_{k+1} - 2 \sum_{k=1}^n k w_k \quad (6.28)$$

$$= \sum_{k=0}^{n-1} (k+1) w_k + \sum_{k=2}^{n+1} (k-1) w_k - 2 \sum_{k=1}^n k w_k \quad (6.29)$$

$$= w_0 + n w_{n+1} - (n+1) w_n \quad (6.30)$$

$$= 1 - w_n - n(w_n - w_{n+1}) \quad (6.31)$$

We therefore wish to show that $\lim_n w_n = 0$ implies $\lim_n w_n + n(w_n - w_{n+1}) = 0$. By hypothesis, it is enough to show that $\lim_n n(w_n - w_{n+1}) = 0$. The partial sums on the left-hand side of Eq. 6.27 are non-decreasing, so $w_n + n(w_n - w_{n+1})$ is non-increasing. Since it is also ≥ 0 , the limit $\lim_n w_n + n(w_n - w_{n+1})$ exists. Since $w_n \rightarrow 0$, $w_n - w_{n+1} \leq w_n$ must also go to zero; the only question is whether it goes to zero fast enough. So consider

$$\lim_n \sum_{k=1}^n w_k - \sum_{k=1}^n w_{k+1} \quad (6.32)$$

Telescoping the sums again, this is $\lim_n w_1 - w_{n+1}$. Since $\lim_{n+1} w_{n+1} = \lim_n w_n = 0$, the limit exists. But we can equally re-write Eq. 6.32 as

$$\lim_n \sum_{k=1}^n w_k - w_{k+1} = \sum_{n=1}^{\infty} w_n - w_{n+1} \quad (6.33)$$

Since the sum converges, the individual terms $w_n - w_{n+1}$ must be $o(n^{-1})$. Hence $\lim_n n(w_n - w_{n+1}) = 0$, as was to be shown.

“Only if”: From Eq. 6.31 in the “if” part, we see that the hypothesis is equivalent to

$$1 = \lim_n (1 - w_n - n(w_n - w_{n+1})) \quad (6.34)$$

Since $w_n \geq w_{n+1}$, $1 - w_n - n(w_n - w_{n+1}) \leq 1 - w_n$. We know from the proof of Theorem 66 that $\lim_n w_n$ exists, whether or not it is zero. If it is not zero, then $\lim_n (1 - w_n - n(w_n - w_{n+1})) \leq 1 - \lim_n w_n < 1$. Hence $w_n \rightarrow 0$ is a necessary condition. \square

Example 70 (Counter-example for Kac's Recurrence Theorem) *One might imagine that the condition $w_n \rightarrow 0$ in Kac's Theorem is redundant, given the assumption of stationarity. Here is a counter-example. Consider a homogeneous Markov chain on a finite space Ξ , which is partitioned into two non-communicating components, Ξ_1 and Ξ_2 . Each component is, internally, irreducible and aperiodic, so there will be an invariant measure μ_1 supported on Ξ_1 , and another invariant measure μ_2 supported on Ξ_2 . But then, for any $s \in [0, 1]$, $s\mu_1 + (1-s)\mu_2$ will also be invariant. (Why?) Picking $A \subset \Xi_2$ gives $\lim_n w_n = s$, the probability that the chain begins in the wrong component to ever reach A .*

Kac's Theorem turns out to be the foundation for a fascinating class of methods for learning the distributions of stationary processes, and for “universal” prediction and data compression. There is also an interesting interaction with large deviations theory. This subject is one possibility for further discussion at the end of the course. Whether or not we get there, let me recommend some papers in a footnote.²

6.4 Exercises

Exercise 6.1 (Weakly Optional Times and Right-Continuous Filtrations) *Show that a random time τ is weakly $\{\mathcal{F}_t\}$ -optional iff it is $\{\mathcal{F}\}_{t+}$ -optional.*

Exercise 6.2 (Discrete Stopping Times and Their σ -Algebras) *Prove Eq. 6.3. Does the corresponding statement hold for two-sided processes?*

Exercise 6.3 (Kac's Theorem for the Logistic Map) *First, do Exercise 5.3. Then, using the same code, suitably modified, numerically check Kac's Theorem for the logistic map with $a = 4$. Pick any interval $I \subset [0, 1]$ you like, but be sure not to make it too small.*

1. *Generate n initial points in I , according to the invariant measure $\frac{1}{\pi\sqrt{x(1-x)}}$. For each point x_i , find the first t such that $F^t(x_i) \in I$, and take the mean over the sample. What happens to this space average as n grows?*
2. *Generate a single point x_0 in I , according to the invariant measure. Iterate it T times. Record the successive times t_1, t_2, \dots at which $F^t(x_0) \in I$, and find the mean of $t_i - t_{i-1}$ (taking $t_0 = 0$). What happens to this time average as T grows?*

² “How Sampling Reveals a Process” (Ornstein and Weiss, 1990); Algoet (1992); Kontoyiannis *et al.* (1998).

Chapter 7

Continuity of Stochastic Processes

Section 7.1 describes the leading kinds of continuity for stochastic processes, which derive from the modes of convergence of random variables. It also defines the idea of versions of a stochastic process.

Section 7.2 explains why continuity of sample paths is often problematic, and why we need the whole “paths in U ” song-and-dance. As an illustration, we consider a Gaussian process which is close to the Wiener process, except that it’s got a nasty non-measurability.

Section 7.3 introduces separable random functions.

7.1 Kinds of Continuity for Processes

Continuity is a convergence property: a continuous function is one where convergence of the inputs implies convergence of the outputs. But we have several kinds of convergence for random variables, so we may expect to encounter several kinds of continuity for random processes. Note that the following definitions are stated broadly enough that the index set T does not have to be one-dimensional.

Definition 71 (Continuity in Mean) *A stochastic process X is continuous in the mean at t_0 if $t \rightarrow t_0$ implies $\mathbf{E} [|X(t) - X(t_0)|^2] \rightarrow 0$. X is continuous in the mean if this holds for all t_0 .*

It would, of course, be more natural to refer to this as “continuity in mean square”, or even “continuity in L_2 ”, and one can define continuity in L_p for arbitrary p .

Definition 72 (Continuity in Probability, Stochastic Continuity) *X is continuous in probability at t_0 if $t \rightarrow t_0$ implies $X(t) \xrightarrow{P} X(t_0)$. X is continuous in probability or stochastically continuous if this holds for all t_0 .*

Note that neither L_p -continuity nor stochastic continuity says that the individual sample paths, themselves, are continuous.

Definition 73 (Continuous Sample Paths) *A process X is continuous at t_0 if, for almost all ω , $t \rightarrow t_0$ implies $X(t, \omega) \rightarrow X(t_0, \omega)$. A process is continuous if, for almost all ω , $X(\cdot, \omega)$ is a continuous function.*

Obviously, continuity of sample paths implies stochastic continuity and L_p -continuity.

A weaker pathwise property than strict continuity, frequently used in practice, is the combination of continuity from the right with limits from the left. This is usually known by the term “cadlag”, abbreviating the French phrase “continues à droite, limites à gauche”; “rcll” is an unpronounceable synonym.

Definition 74 (Cadlag) *A sample function x on a well-ordered set T is cadlag if it is continuous from the right and limited from the left at every point. That is, for every $t_0 \in T$, $t \downarrow t_0$ implies $x(t) \rightarrow x(t_0)$, and for $t \uparrow t_0$, $\lim_{t \uparrow t_0} x(t)$ exists, but need not be $x(t_0)$. A stochastic process X is cadlag if almost all its sample paths are cadlag.*

As we will see, it will not be easy to show that our favorite random processes have any of these desirable properties. What will be easy will be to show that they are, in some sense, easily modified into ones which do have good regularity properties, without loss of probabilistic content. This is made more precise by the notion of *versions* of a stochastic process, related to that of versions of conditional probabilities.

Definition 75 (Versions of a Stochastic Process, Stochastically Equivalent Processes) *Two stochastic processes X and Y with a common index set T are called versions of one another if*

$$\forall t \in T, \mathbb{P}(\omega : X(t, \omega) = Y(t, \omega)) = 1$$

Such processes are also said to be stochastically equivalent.

Lemma 76 (Versions Have Identical FDDs) *If X and Y are versions of one another, they have the same finite-dimensional distributions.*

PROOF: Clearly it will be enough to show that $\mathbb{P}(X_J = Y_J) = 1$ for arbitrary finite collections of indices J . Pick any such collection $J = \{t_1, t_2, \dots, t_j\}$. Then

$$\mathbb{P}(X_J = Y_J) = \mathbb{P}(X_{t_1} = Y_{t_1}, \dots, X_{t_j} = Y_{t_j}) \quad (7.1)$$

$$= 1 - \mathbb{P}\left(\bigcup_{t_i \in J} X_{t_i} \neq Y_{t_i}\right) \quad (7.2)$$

$$\geq 1 - \sum_{t_i \in J} \mathbb{P}(X_{t_i} \neq Y_{t_i}) \quad (7.3)$$

$$= 1 \quad (7.4)$$

using only *finite* sub-additivity. \square

There is a stronger notion of similarity between processes than that of versions, which will sometimes be useful.

Definition 77 (Indistinguishable Processes) *Two stochastic processes X and Y are indistinguishable, or equivalent up to evanescence, when*

$$\mathbb{P}(\omega : \forall t, X(t, \omega) = Y(t, \omega)) = 1$$

Notice that saying X and Y are indistinguishable means that their sample paths are equal almost surely, while saying they are versions of one another means that, at any time, they are almost surely equal. Indistinguishable processes are versions of one another, but not necessarily the reverse. (Look at where the quantifier and the probability statements go.) However, if $T = \mathbb{R}^d$, then any two right-continuous versions of the same process are indistinguishable (Exercise 7.2).

7.2 Why Continuity Is an Issue

In many situations, we want to use stochastic processes to model dynamical systems, where we know that the dynamics are continuous in time (i.e. the index set is \mathbb{R} , or maybe \mathbb{R}^+ or $[0, T]$ for some real T).¹ This means that we ought to restrict the sample paths to be continuous functions; in some cases we'd even want them to be differentiable, or yet more smooth. As well as being a matter of physical plausibility or realism, it is also a considerable mathematical convenience, as the following shows.

Proposition 78 (Continuous Sample Paths Have Measurable Extrema)

Let $X(t, \omega)$ be a real-valued continuous-parameter process with continuous sample paths. Then on any finite interval I , $M(\omega) \equiv \sup_{t \in I} X(t, \omega)$ and $m(\omega) \equiv \inf_{t \in I} X(t, \omega)$ are measurable random variables.

PROOF: It'll be enough to prove this for the supremum function M ; the proof for m is entirely parallel. First, notice that $M(\omega)$ must be finite, because the sample paths $X(\cdot, \omega)$ are continuous functions, and continuous functions are bounded on bounded intervals. Next, notice that $M(\omega) > a$ if and only if $X(t, \omega) > a$ for some $t \in I$. But then, by continuity, there will be some rational $t' \in I \cap \mathbb{Q}$ such that $X(t', \omega) > a$; countably many, in fact.² Hence

$$\{\omega : M(\omega) > a\} = \bigcup_{t \in I \cap \mathbb{Q}} \{\omega : X(t, \omega) > a\}$$

¹Strictly speaking, we don't *really* know that space-time is a continuum, but the discretization, if there is one, is so fine that it might as well be.

²Continuity means that we can pick a δ such that, for all t' within δ of t , $X(t', \omega)$ is within $\frac{1}{2}(X(t, \omega) - a)$ of $X(t, \omega)$. And there are countably many rational numbers within any real interval. — There is nothing special about the rational numbers here; any countable, dense subset of the real numbers would work as well.

Since, for each t , $X(t, \omega)$ is a random variable, the sets in the union on the right-hand side are all measurable, and the union of a countable collection of measurable sets is itself measurable. Since intervals of the form (a, ∞) generate the Borel σ -field on the reals, we have shown that $M(\omega)$ is a measurable function from Ω to the reals, i.e., a random variable. \square

Continuity raises some very tricky technical issues. The product σ -field is the usual way of formalizing the notion that what we know about a stochastic process are values observed at certain particular times. What we saw in Exercise 1.1 is that “the product σ -field answers countable questions”: for any measurable set A , whether $x(\cdot, \omega) \in A$ depends only on the value of $x(t, \omega)$ at countably many indices t . It follows that the class of all continuous sample paths is *not* product- σ -field measurable, because $x(\cdot, \omega)$ is continuous at t iff $x(t_n, \omega) \rightarrow x(t, \omega)$ along *every* sequence $t_n \rightarrow t$, and this involves the value of the function at uncountably many coordinates. It is further true that the class of differentiable functions is not product σ -field measurable. For that matter, neither is the class of piecewise linear functions! (See Exercise 7.1.)

You might think that, on the basis of Theorem 23, this should not *really* be much of an issue: that even if the class of continuous sample paths (say) isn’t strictly measurable, it could be well-approximated by measurable sets, and so getting the finite-dimensional distributions right is all that matters. This would make the theory of stochastic processes in continuous time much simpler, but unfortunately it’s not quite the case. Here is an example to show just how bad things can get, even when all the finite-dimensional distributions agree.³

Example 79 (A Horrible Version of the proto-Wiener Process) *Example 38 defined the Wiener process by four requirements: starting at the origin, independent increments, a Gaussian distribution of increments, and continuity of sample paths. Take a Wiener process $W(t, \omega)$ and consider $M(\omega) \equiv \sup_{t \in [0, 1]} W(t, \omega)$, its supremum over the unit interval. By the preceding proposition, we know that M is a measurable random variable. But we can construct a version of W for which the supremum is not measurable.*

For starters, assume that Ω can be partitioned into an uncountable collection of disjoint measurable sets, one for each $t \in [0, 1]$. (This can be shown as an exercise in real analysis.) Select any non-measurable real-valued function $B(\omega)$, so long as $B(\omega) > M(\omega)$ for all ω . (There are uncountably many suitable functions.) Set $W^(t, \omega) = W(t, \omega)$ if $\omega \notin \Omega_t$, and $= B(\omega)$ if $\omega \in \Omega_t$. Now, at every t , $\mathbb{P}(W(t, \omega) = W^*(t, \omega)) = 1$. W^* is a version of W , and all their finite-dimensional distributions agree. But, for every ω , there is a t such that $W^*(t, \omega) = B(\omega) > \sup_t W(t, \omega)$, so $\sup_t W^*(t, \omega) = B(\omega)$, which by design is non-measurable.*

Why care about this example? Two reasons. First, and technically, we’re going to want to take suprema of the Wiener process a lot when we deal with large deviations theory, and with the approximation of discrete processes by continuous ones. Second, nothing in the example *really* depended on starting

³I stole this example from Pollard (2002, p. 214).

from the Wiener process; any process with continuous sample paths would have worked as well. So controlling the finite-dimensional distributions is not enough to guarantee that a process has measurable extrema, which would lead to all kinds of embarrassments. We could set up what would look like a reasonable stochastic model, and then be unable to say things like “with probability p , the { temperature/ demand for electricity/ tracking error/ interest rate } will not exceed r over the course of the year”.

Fundamentally, the issues with continuity are symptoms of a deeper problem. The reason the supremum function is non-measurable in the example is that it involves uncountably many indices. A countable collection of ill-behaved sets of measure zero is a set of measure zero, and may be ignored, but an uncountable collection of them can have probability 1. Fortunately, there are standard ways of evading these measure-theoretic issues, by showing that one can always find random functions which not only have prescribed finite-dimensional distributions (what we did in Lectures 2 and 3), but also are regular enough that we can take suprema, or integrate over time, or force them to be continuous. This hinges on the notion of *separability* for random functions.

7.3 Separable Random Functions

The basic idea of a *separable* random function is one whose properties can be handled by dealing only with a countable, dense subset, just as, in the proof of Proposition 78, we were able to get away with only looking at $X(t)$ at only rational values of t . Because a space with a countable, dense subset is called a “separable” space, we will call such functions “separable functions”.

Definition 80 (Separable Function) *Let Ξ and T be metric spaces, and D be a countable, dense subset of T . A function $x : T \mapsto \Xi$ is D -separable or separable with respect to D if, for all $t \in T$, there exists a sequence $t_i \in D$ such that $t_i \rightarrow t$ and $x(t_i) \rightarrow x(t)$.*

Lemma 81 (Some Sufficient Conditions for Separability) *The following conditions are sufficient for separability:*

1. T is countable.
2. x is continuous.
3. T is well-ordered and x is right-continuous.

PROOF: (1) Take the separating set to be T itself. (2) Pick any countable dense D . By density, for every t there will be a sequence $t_i \in D$ such that $t_i \rightarrow t$. By continuity, along any sequence converging to t , $x(t_i) \rightarrow x(t)$. (3) Just like (2), only be sure to pick the $t_i > t$. (You can do this, again, for any countable dense D .) \square

Definition 82 (Separable Process) A Ξ -valued process X on T is separable with respect to D if D is a countable, dense subset of T , and there is a measure-zero set $N \subset \Omega$ such that for every $\omega \notin N$, $X(\cdot, \omega)$ is D -separable. That is, $X(\cdot, \omega)$ is almost surely D -separable.

We cannot easily guarantee that a process is separable. What we can easily do is go from one process, which may or may not be separable, to a separable process with the same finite-dimensional distributions. This is known as a *separable modification* of the original process. Combined with the extension theorems (Theorems 27, 29 and 33), this tells that we can always construct a separable process with desired finite-dimensional distributions. We shall therefore feel entitled to assume that our processes *are* separable, without further ado. The proofs of the existence of separable and continuous versions of general processes are, however, somewhat involved, and so postponed to the next lecture.

7.4 Exercises

Exercise 7.1 (Piecewise Linear Paths Not in the Product σ -Field)

Consider real-valued functions on the unit interval (i.e., $\Xi = \mathbb{R}$, $T = [0, 1]$, $\mathcal{X} = \mathcal{B}$). The product σ -field is thus $\mathcal{B}^{[0,1]}$. In many circumstances, it would be useful to constrain sample paths to be piece-wise linear functions of the index. Let $\mathbf{PL}([0, 1])$ denote this class of functions. Use the argument of Exercise 1.1 to show that $\mathbf{PL}([0, 1]) \notin \mathcal{B}^{[0,1]}$.

Exercise 7.2 (Indistinguishability of Right-Continuous Versions)

Show that, if X and Y are versions of one another, with index set \mathbb{R}^d , and both are right-continuous, then they are indistinguishable.

Chapter 8

More on Continuity

Section 8.1 constructs separable modifications of reasonable but non-separable random functions, and explains how separability relates to non-denumerable properties like continuity.

Section 8.2 constructs versions of our favorite one-parameter processes where the sample paths are measurable functions of the parameter.

Section 8.3 gives conditions for the existence of cadlag versions.

Section 8.4 gives some criteria for continuity, and for the existence of “continuous modifications” of discontinuous processes.

Recall the story so far: last time we saw that the existence of processes with given finite-dimensional distributions does not guarantee that they have desirable and natural properties, like continuity, and in fact that one can construct discontinuous versions of processes which *ought* to be continuous. We therefore need extra theorems to guarantee the existence of *continuous* versions of processes with specified FDDs. To get there, we will first prove the existence of *separable* versions. This will require various topological conditions on both the index set T and the value space Ξ .

In the interest of space (or is it time?), Section 8.1 will provide complete and detailed proofs. The other sections will simply state results, and refer proofs to standard sources, mostly Gikhman and Skorokhod (1965/1969). (They in turn follow Doob (1953), but are explicit about what he regarded as obvious generalizations and extensions, *and* they cost about \$20, whereas Doob costs \$120 in paperback.)

8.1 Separable Versions

We can show that separable versions of our favorite stochastic processes exist under quite general conditions, but first we will need some preliminary results, living at the border between topology and measure theory. This starts by recalling some facts about compact spaces.

Definition 83 (Compactness, Compactification) *A set A in a topological space Ξ is compact if every covering of A by open sets contains a finite sub-cover. Ξ is a compact space if it is itself a compact set. If Ξ is not compact, but is a sub-space of a compact topological space $\tilde{\Xi}$, the super-space is a compactification of Ξ .*

Proposition 84 (Compact Spaces are Separable) *Every compact space is separable.*

PROOF: A standard result of analysis. Interestingly, it relies on the axiom of choice. \square

Proposition 85 (Compactification is Possible) *Every non-compact topological space Ξ is a sub-space of some compact topological space $\tilde{\Xi}$, i.e., every non-compact topological space can be compactified.*

Example 86 (Compactifying the Reals) *The real numbers \mathbb{R} are not compact: they have no finite covering by open intervals (or other open sets). The extended reals, $\bar{\mathbb{R}} \equiv \mathbb{R} \cup +\infty \cup -\infty$, are compact, since intervals of the form $(a, \infty]$ and $[-\infty, a)$ are open. This is a two-point compactification of the reals. There is also a one-point compactification, with a single point at $\pm\infty$, but this has the undesirable property of making big negative and positive numbers close to each other.*

Recall that a random function is separable if its value at any arbitrary index can be determined almost surely by examining its values on some fixed, countable collection of indices. The next lemma states an alternative characterization of separability. The lemma after that gives conditions under which a weaker property holds — the almost-sure determination of whether $X(t, \omega) \in B$, for a specific t and set B , by the behavior of $X(t_n, \omega)$ at countably many t_n . The final lemma extends this to large collections of sets, and then the proof of the theorem puts all the parts together.

Lemma 87 (Alternative Characterization of Separability) *Let T be a separable set, Ξ a compact metric space, and D a countable dense subset of T . Define V as the class of all open balls in T centered at points in D and with rational radii. For any $G \subset T$, let*

$$R(G, \omega) \equiv \text{closure} \left(\bigcup_{t \in G \cap D} X(t, \omega) \right) \quad (8.1)$$

$$R(t, \omega) \equiv \bigcap_{S: S \in V, t \in S} R(S, \omega) \quad (8.2)$$

Then $X(t, \omega)$ is D -separable if and only if there exists a set $N \subset \Omega$ such that

$$\omega \notin N \Rightarrow \forall t, X(t, \omega) \in R(t, \omega) \quad (8.3)$$

and $\mathbb{P}(N) = 0$.

PROOF: Roughly speaking, $R(t, \omega)$ is what we'd think the range of the function would be, in the vicinity of t , if it went just by what it did at points in the separating set D . The actual value of the function falling into this range (almost surely) is necessary and sufficient for the function to be separable. But let's speak less roughly.

"Only if": Since $X(t, \omega)$ is D -separable, for almost all ω , for any t there is some sequence $t_n \in D$ such that $t_n \rightarrow t$ and $X(t_n, \omega) \rightarrow X(t, \omega)$. For any ball S centered at t , there is some N such that $t_n \in S$ if $n \geq N$. Hence the values of $x(t_n)$ are eventually confined to the set $\bigcup_{t \in S \cap D} X(t, \omega)$. Recall that the closure of a set A consists of the points x such that, for some sequence $x_n \in A$, $x_n \rightarrow x$. As $X(t_n, \omega) \rightarrow X(t, \omega)$, it must be the case that $X(t, \omega) \in \text{closure}(\bigcup_{t \in S \cap D} X(t, \omega))$. Since this applies to all S , $X(t, \omega)$ must be in the intersection of all those closures, hence $X(t, \omega) \in R(t, \omega)$ — unless we are on one of the probability-zero bad sample paths, i.e., unless $\omega \in N$.

"If": Assume that, with probability 1, $X(t, \omega) \in R(t, \omega)$. Thus, for any $S \in V$, we know that there exists a sequence of points $t_n \in S \cap D$ such that $X(t_n, \omega) \rightarrow X(t, \omega)$. However, this doesn't say that $t_n \rightarrow t$, which is what we need for separability. We will now build such a sequence. Consider a series of spheres $S_k \in V$ such that (i) every point in S_k is within a distance 2^{-k} of t and (ii) $S_{k+1} \subset S_k$. For each S_k , there is a sequence $t_n^{(k)} \in S_k$ such that $X(t_n^{(k)}, \omega) \rightarrow X(t, \omega)$. In fact, for any $m > 0$, $|X(t_n^{(k)}, \omega) - X(t, \omega)| < 2^{-m}$ if $n \geq N(k, m)$, for some $N(k, m)$. Our final sequence of indices t_i then consists of the following points: $t_n^{(1)}$ for n from $N(1, 1)$ to $N(1, 2)$; $t_n^{(2)}$ for n from $N(2, 2)$ to $N(2, 3)$; and in general $t_n^{(k)}$ for n from $N(k, k)$ to $N(k, k+1)$. Clearly, $t_i \rightarrow t$, and $X(t_i, \omega) \rightarrow X(t, \omega)$. Since every $t_i \in D$, we have shown that $X(t, \omega)$ is D -separable. \square

Lemma 88 (Confining Bad Behavior to a Measure Zero Event) *Let T be a separable index set, Ξ a compact space, X a random function from T to Ξ , and B be an arbitrary Borel set of Ξ . Then there exists a denumerable set of points $t_n \in T$ such that, for any $t \in T$, the set*

$$N(t, B) \equiv \{\omega : X(t, \omega) \notin B\} \cap \left(\bigcap_{n=1}^{\infty} \{\omega : X(t_n, \omega) \in B\} \right) \quad (8.4)$$

has probability 0.

PROOF: We proceed recursively. The first point, t_1 , can be whatever we like. Suppose t_1, t_2, \dots, t_n are already found, and define the following:

$$M_n \equiv \bigcap_{k=1}^n \{\omega : X(t_k, \omega) \in B\} \quad (8.5)$$

$$L_n(t) \equiv M_n \cap \{\omega : X(t, \omega) \notin B\} \quad (8.6)$$

$$p_n \equiv \sup_t \mathbb{P}(L_n(t)) \quad (8.7)$$

M_n is the set where the random function, evaluated at the first n indices, gives a value in our favorite set; it's clearly measurable. $L_n(t)$, also clearly measurable, gives the collection of points in Ω where, if we chose t for the next point in the collection, this will break down. p_n is the worst-case probability of this happening. For each t , $L_{n+1}(t) \subseteq L_n(t)$, so $p_{n+1} \leq p_n$. Suppose $p_n = 0$; then we've found the promised denumerable sequence, and we're done. Suppose instead that $p_n > 0$. Pick any t such that $\mathbb{P}(L_n(t)) \geq \frac{1}{2}p_n$, and call it t_{n+1} . (There has to be such a point, or else p_n wouldn't be the supremum.) Now notice that $L_1(t_2), L_2(t_3), \dots, L_n(t_{n+1})$ are all mutually exclusive, but not necessarily jointly exhaustive. So

$$1 = \mathbb{P}(\Omega) \quad (8.8)$$

$$\geq \mathbb{P}\left(\bigcup_n L_n(t_{n+1})\right) \quad (8.9)$$

$$= \sum_n \mathbb{P}(L_n(t_{n+1})) \quad (8.10)$$

$$\geq \sum_n \frac{1}{2}p_n > 0 \quad (8.11)$$

so $p_n \rightarrow 0$ as $n \rightarrow \infty$.

We saw that $L_n(t)$ is a monotone-decreasing sequence of sets, for each t , so a limiting set exists, and in fact $\lim_n L_n(t) = N(t, B)$. So, by monotone convergence,

$$\mathbb{P}(N(t, B)) = \mathbb{P}\left(\lim_n L_n(t)\right) \quad (8.12)$$

$$= \lim_n \mathbb{P}(L_n(t)) \quad (8.13)$$

$$\leq \lim_n p_n \quad (8.14)$$

$$= 0 \quad (8.15)$$

as was to be shown. \square

Lemma 89 (A Null Exceptional Set) *Let \mathcal{B}_0 be any countable class of Borel sets in Ξ , and \mathcal{B} the closure of \mathcal{B}_0 under countable intersection. Under the hypotheses of the previous lemma, there is a denumerable sequence t_n such that, for every $t \in T$, there exists a set $N(t) \subset \Omega$ with $\mathbb{P}(N(t)) = 0$, and, for all $B \in \mathcal{B}$,*

$$\{\omega : X(t, \omega) \notin A\} \cap \left(\bigcap_{n=1}^{\infty} \{\omega : X(t_n, \omega) \in A\} \right) \subseteq N(t) \quad (8.16)$$

PROOF: For each $B \in \mathcal{B}_0$, construct the sequence of indices as in the previous lemma. Since there are only countably many sets in \mathcal{B} , if we take the union of all of these sequences, we will get another countable sequence, call it t_n . Then we have that, $\forall B \in \mathcal{B}_0$, $\forall t \in T$, $\mathbb{P}(X(t_n, \omega) \in B, n \geq 1, X(t, \omega) \notin B) = 0$. Take this set

to be $N(t, B)$, and define $N(t) \equiv \bigcup_{B \in \mathcal{B}_0} N(t, B)$. Since $N(t)$ is a countable union of probability-zero events, it is itself a probability-zero event. Now, take any $B \in \mathcal{B}$, and any $B_0 \in \mathcal{B}_0$ such that $B \subseteq B_0$. Then

$$\{X(t, \omega) \notin B_0\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n, \omega) \in B\} \right) \quad (8.17)$$

$$\begin{aligned} &\subseteq \{X(t, \omega) \notin B_0\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n, \omega) \in B_0\} \right) \\ &\subseteq N(t) \end{aligned} \quad (8.18)$$

Since $B = \bigcap_k B_0^{(k)}$ for some sequence of sets $B_0^{(k)} \in \mathcal{B}_0$, it follows (via De Morgan's laws and the distributive law) that

$$\{X(t, \omega) \notin B\} = \bigcup_{k=1}^{\infty} \{X(t, \omega) \notin B_0^{(k)}\} \quad (8.19)$$

$$\begin{aligned} &\{X(t, \omega) \notin B\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n, \omega) \in B\} \right) \\ &= \bigcup_{k=1}^{\infty} \{X(t, \omega) \notin B_0^{(k)}\} \cap \left(\bigcap_{n=1}^{\infty} \{X(t_n, \omega) \in B\} \right) \end{aligned} \quad (8.20)$$

$$\subseteq \bigcup_{n=1}^{\infty} N(t) \quad (8.21)$$

$$= N(t) \quad (8.22)$$

which was to be shown. \square

Theorem 90 (Separable Versions, Separable Modification) *Suppose that Ξ is a compact metric space and T is a separable metric space. Then, for any Ξ -valued stochastic process X on T , there exists a separable version \tilde{X} . This is called a separable modification of X .*

PROOF: Let D be a countable dense subset of T , and V the class of open spheres of rational radius centered at points in D . Any open subset of T is a union of countably many sets from V , which is itself countable. Similarly, let C be a countable dense subset of Ξ , and let \mathcal{B}_0 consist of the complements of spheres centers at points in D with rational radii, and (as in the previous lemma) let \mathcal{B} be the closure of \mathcal{B}_0 under countable intersection. Every closed set in Ξ belongs to \mathcal{B} .¹ For every $S \in V$, consider the restriction of $X(t, \omega)$ to $t \in S$, and apply Lemma 89 to the random function $X(t, \omega)$ to get a sequence of indices $I(S) \subset T$, and, for every $t \in S$, a measure-zero set $N_S(t) \subset \Omega$ where

¹ You show this.

things can go wrong. Set $I = \bigcup_{S \in V} I(S)$ and $N(t) = \bigcup_{S \in V} N_S(t)$. Because V is countable, I is still a countable set of indices, and $N(t)$ is still of measure zero. I is going to be our separating set, and we're going to show that we have uncountably many sets $N(t)$ won't be a problem.

Define $\tilde{X}(t, \omega) = X(t, \omega)$ if $t \in I$ or $\omega \notin N(t)$ — if we're at a time in the separating set, or we're at some other time but have avoided the bad set, we'll just copy our original random function. What to do otherwise, when $t \notin I$ and $\omega \in N(t)$? Construct $R(t, \omega)$, as in the proof of Lemma 87, and let $\tilde{X}(t, \omega)$ take *any* value in this set. Since $R(t, \omega)$ depends only on the value of the function at indices in the separating set, it doesn't matter whether we build it from X or from \tilde{X} . In fact, for all t and ω , $\tilde{X}(t, \omega) \in R(t, \omega)$, so, by Lemma 87, $\tilde{X}(t, \omega)$ is separable. Finally, for every t , $\{\tilde{X}(t, \omega) = X(t, \omega)\} \subseteq N(t)$, so $\forall t$, $\mathbb{P}(\tilde{X}(t) = X(t))$, and \tilde{X} is a version of X (Definition 75). \square

Corollary 91 (Separable Modifications in Compactified Spaces) *If the situation is as in the previous theorem, but Ξ is not compact, there exists a separable version of X in any compactification $\tilde{\Xi}$ of Ξ .*

PROOF: Because Ξ is a sub-space of any of its compactifications $\tilde{\Xi}$, X is also a process with values in $\tilde{\Xi}$.² Since $\tilde{\Xi}$ is compact, X has a separable modification \tilde{X} with values in $\tilde{\Xi}$, but (with probability 1) $\tilde{X}(t) \in \Xi$. \square

Corollary 92 (Separable Versions with Prescribed Distributions Exist) *Let Ξ be a complete, compact metric space, T a separable index set, and μ_J , $J \in \text{Fin}(T)$ a projective family of probability distributions. Then there is a separable stochastic process with finite-dimensional distributions given by μ_J .*

PROOF: Combine Theorem 90 with the Kolmogorov Extension Theorem 29. (By Proposition 84, a complete, compact metric space is a complete, separable metric space.) \square

8.2 Measurable Versions

It would be nice for us if $X(t)$ is a measurable function *of* t , because we are going to want to write down things like

$$\int_{t=a}^{t=b} X(t) dt$$

and have them mean something. Irritatingly, this will require another modification.

²If you want to be really picky, define a 1-1 function $h : \Xi \mapsto \tilde{\Xi}$ taking points to their counterparts. Then X and $h^{-1}(X)$ are indistinguishable. Do I need to go on?

Definition 93 (Measurable random function) *Let T, \mathcal{T}, τ be a measurable space, its σ -field and a measure defined thereon. A random function X on T with values in Ξ, \mathcal{X} is measurable if $X : T \times \Omega \mapsto \Xi$ is $\widetilde{\mathcal{T} \times \mathcal{F}}/\mathcal{X}$ measurable, where $\mathcal{T} \times \mathcal{F}$ is the product σ -field on $T \times \Omega$, and $\widetilde{\mathcal{T} \times \mathcal{F}}$ its completion by the null sets of the product measure $\tau \times \mathbb{P}$.*

It would seem more natural to simply define measurable processes by saying that the sample paths are measurable, that $X(\cdot, \omega)$ is a \mathcal{T} -measurable function of t for \mathbb{P} -almost-all ω . This would be like our definition of almost-sure continuity. However, Definition 93 implies this version, via Fubini's Theorem, and facilitates the proofs of the two following theorems.

Theorem 94 (Exchanging Expectations and Time Integrals) *If $X(t)$ is measurable, and $\mathbf{E}[X(t)]$ is integrable (with respect to the measure τ on T), then for any set $I \in \mathcal{T}$,*

$$\int_I \mathbf{E}[X(t)] \tau(dt) = \mathbf{E} \left[\int_I X(t) \tau(dt) \right] \quad (8.23)$$

PROOF: This is just Fubini's Theorem! \square

Theorem 95 (Measurable Separable Modifications) *Suppose that T and Ξ are both compact. If $X(t, \omega)$ is continuous in probability at τ -almost-all t , then it has a version which is both separable and measurable, its measurable separable modification.*

PROOF: See Gikhman and Skorokhod (1965/1969, ch. IV, sec. 3, thm. 1, p. 157). \square

8.3 Cadlag Versions

Theorem 96 (Cadlag Versions) *Let X be a separable random process with $T = [a, b] \subseteq \mathbb{R}$, and Ξ a complete metric space with metric ρ . Suppose that $X(t)$ is continuous in probability on T , and there are real constants $p, q, C \geq 0, r > 1$ such that, for any three indices $t_1 < t_2 < t_3 \in T$,*

$$\mathbf{E}[\rho^p(X(t_1), X(t_2))\rho^q(X(t_2), X(t_3))] \leq C|t_3 - t_1|^r \quad (8.24)$$

Then there is a version of X whose sample paths are cadlag (a.s.).

PROOF: Combine Theorem 1 and Theorem 3 of Gikhman and Skorokhod (1965/1969, ch. IV, sec. 4, pp. 159–169). \square

8.4 Continuous Modifications

Theorem 97 (Continuous Versions) *Let X be a separable stochastic process with $T = [a, b] \subseteq \mathbb{R}$, and Ξ a complete metric space with metric ρ . Suppose that there are constants $C, p > 0, r > 1$ such that, for any $t_1 < t_2 \in T$,*

$$\mathbf{E}[\rho^p(X(t_1), X(t_2))] \leq C|t_2 - t_1|^r \quad (8.25)$$

Then $X(t)$ has a continuous version.

PROOF: See Gikhman and Skorokhod (1965/1969, ch. IV, sec. 5, thm. 2, p. 170), and the first remark following the theorem. \square

A slightly more refined result requires two preliminary definitions.

Definition 98 (Modulus of continuity) *For any function x from a metric space T, d to a metric space Ξ, ρ , the modulus of continuity is the function $m_x(r) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ given by*

$$m_x(r) = \sup \{ \rho(x(s), x(t)) : s, t \in T, d(s, t) \leq r \} \quad (8.26)$$

Lemma 99 (Modulus of continuity and uniform continuity) *x is uniformly continuous if and only if its modulus of continuity $\rightarrow 0$ as $r \rightarrow 0$.*

PROOF: Obvious from Definition 98 and the definition of uniform continuity.

Definition 100 (Hölder continuity) *Continuing the notation of Definition 98, we say that x is Hölder-continuous with exponent c if there are positive constants c, γ such that $m_x(r) \leq \gamma r^c$ for all sufficiently small r ; i.e., $m_x(r) = O(r^c)$. If this holds on every bounded subset of T , then the function is locally Hölder-continuous.*

Theorem 101 (Hölder-continuous versions) *Let T be \mathbb{R}^d and Ξ a complete metric space with metric ρ . If there are constants $p, q, \gamma > 0$, such that, for any $t_1, t_2 \in T$,*

$$\mathbf{E}[\rho^p(X(t_1), X(t_2))] \leq \gamma |t_1 - t_2|^{d+q} \quad (8.27)$$

then X has a continuous version \tilde{X} , and almost all sample paths of \tilde{X} are locally Hölder-continuous for any exponent between 0 and q/p exclusive.

PROOF: See Kallenberg, theorem 3.23 (pp. 57–58). Note that part of Kallenberg's proof is a restricted case of what we've already done in proving the existence of a separable version! \square

This lecture, the last, and even a lot of the one before have all been pretty hard and abstract. As a reward for our labor, however, we now have a collection of very important tools — operator representations, filtrations and optional times, recurrence times, and finally existence theorems for continuous processes. These are the devices which will let us take the familiar theory of elementary Markov chains, with finitely many states in discrete time, and produce the general theory of Markov processes with continuously-many states and/or continuous time. The next lecture will begin this work, starting with the operators.

Part III

Markov Processes

Chapter 9

Markov Processes

This chapter begins our study of Markov processes.

Section 9.1 is mainly “ideological”: it formally defines the Markov property for one-parameter processes, and explains why it is a natural generalization of both complete determinism and complete statistical independence.

Section 9.2 introduces the description of Markov processes in terms of their transition probabilities and proves the existence of such processes.

Section 9.3 deals with the question of when being Markovian relative to one filtration implies being Markov relative to another.

9.1 The Correct Line on the Markov Property

The Markov property is the independence of the future from the past, given the present. Let us be more formal.

Definition 102 (Markov Property) *A one-parameter process X is a Markov process with respect to a filtration $\{\mathcal{F}_t\}$ when X_t is adapted to the filtration, and, for any $s > t$, X_s is independent of \mathcal{F}_t given X_t , $X_s \perp\!\!\!\perp \mathcal{F}_t | X_t$. If no filtration is mentioned, it may be assumed to be the natural one generated by X . If X is also conditionally stationary, then it is a time-homogeneous (or just homogeneous) Markov process.*

Lemma 103 (The Markov Property Extends to the Whole Future) *Let X_t^+ stand for the collection of X_u , $u > t$. If X is Markov, then $X_t^+ \perp\!\!\!\perp \mathcal{F}_t | X_t$.*

PROOF: See Exercise 9.1. \square

There are two routes to the Markov property. One is the path followed by Markov himself, of desiring to weaken the assumption of strict statistical independence between variables to mere conditional independence. In fact, Markov

specifically wanted to show that independence was *not* a necessary condition for the law of large numbers to hold, because his arch-enemy claimed that it was, and used that as grounds for believing in free will and Christianity.¹ It turns out that all the key limit theorems of probability — the weak and strong laws of large numbers, the central limit theorem, etc. — work perfectly well for Markov processes, as well as for IID variables.

The other route to the Markov property begins with completely deterministic systems in physics and dynamics. The *state* of a deterministic dynamical system is some variable which fixes the value of all present and future observables. As a consequence, the present state determines the state at all future times. However, strictly deterministic systems are rather thin on the ground, so a natural generalization is to say that the present state determines the *distribution* of future states. This is precisely the Markov property.

Remarkably enough, it is possible to represent any one-parameter stochastic process X as a noisy function of a Markov process Z . The shift operators give a trivial way of doing this, where the Z process is not just homogeneous but actually fully deterministic. An equally trivial, but slightly more probabilistic, approach is to set $Z_t = X_t^-$, the complete past up to and including time t . (This is not necessarily homogeneous.) It turns out that, subject to mild topological conditions on the space X lives in, there is a unique *non-trivial* representation where $Z_t = \epsilon(X_t^-)$ for some function ϵ , Z_t is a homogeneous Markov process, and $X_u \perp\!\!\!\perp \sigma(\{X_t, t \leq u\}) | Z_t$. (See Knight (1975, 1992); Shalizi and Crutchfield (2001).) We may explore such *predictive Markovian representations* at the end of the course, if time permits.

9.2 Transition Probability Kernels

The most obvious way to specify a Markov process is to say what its transition probabilities are. That is, we want to know $\mathbb{P}(X_s \in B | X_t = x)$ for every $s > t$, $x \in \Xi$, and $B \in \mathcal{X}$. Probability kernels (Definition 30) were invented to let us do just this. We have already seen how to compose such kernels; we also need to know how to take their product.

Definition 104 (Product of Probability Kernels) *Let μ and ν be two probability kernels from Ξ to Ξ . Then their product $\mu\nu$ is a kernel from Ξ to Ξ , defined by*

$$(\mu\nu)(x, B) \equiv \int \mu(x, dy) \nu(y, B) \quad (9.1)$$

$$= (\mu \otimes \nu)(x, \Xi \times B) \quad (9.2)$$

¹I am not making this up. See Basharin *et al.* (2004) for a nice discussion of the origin of Markov chains and of Markov's original, highly elegant, work on them. There is a translation of Markov's original paper in an appendix to Howard (1971), and I dare say other places as well.

Intuitively, all the product does is say that the probability of starting at the point x and landing in the set B is equal the probability of first going to y and then ending in B , integrated over all intermediate points y . (Strictly speaking, there is an abuse of notation in Eq. 9.2, since the second kernel in a composition \otimes should be defined over a product space, here $\Xi \times \Xi$. So suppose we have such a kernel ν' , only $\nu'((x, y), B) = \nu(y, B)$.) Finally, observe that if $\mu(x, \cdot) = \delta_x$, the delta function at x , then $(\mu\nu)(x, B) = \nu(x, B)$, and similarly that $(\nu\mu)(x, B) = \nu(x, B)$.

Definition 105 (Transition Semi-Group) *For every $(t, s) \in T \times T$, $s \geq t$, let $\mu_{t,s}$ be a probability kernel from Ξ to Ξ . These probability kernels form a transition semi-group when*

1. For all t , $\mu_{t,t}(x, \cdot) = \delta_x$.
2. For any $t \leq s \leq u \in T$, $\mu_{t,u} = \mu_{t,s}\mu_{s,u}$.

A transition semi-group for which $\forall t \leq s \in T$, $\mu_{t,s} = \mu_{0,s-t} \equiv \mu_{s-t}$ is homogeneous.

As with the shift semi-group, this is really a monoid (because $\mu_{t,t}$ acts as the identity).

The major theorem is the existence of Markov processes with specified transition kernels.

Theorem 106 (Existence of Markov Process with Given Transition Kernels) *Let $\mu_{t,s}$ be a transition semi-group and ν_t a collection of distributions on a Borel space Ξ . If*

$$\nu_s = \nu_t \mu_{t,s} \quad (9.3)$$

then there exists a Markov process X such that $\forall t$,

$$\mathcal{L}(X_t) = \nu_t \quad (9.4)$$

and $\forall t_1 \leq t_2 \leq \dots \leq t_n$,

$$\mathcal{L}(X_{t_1}, X_{t_2} \dots X_{t_n}) = \nu_{t_1} \otimes \mu_{t_1, t_2} \otimes \dots \otimes \mu_{t_{n-1}, t_n} \quad (9.5)$$

Conversely, if X is a Markov process with values in Ξ , then there exist distributions ν_t and a transition kernel semi-group $\mu_{t,s}$ such that Equations 9.4 and 9.3 hold, and

$$\mathbb{P}(X_s \in B | \mathcal{F}_t) = \mu_{t,s} \text{ a.s.} \quad (9.6)$$

PROOF: (From transition kernels to a Markov process.) For any finite set of times $J = \{t_1, \dots, t_n\}$ (in ascending order), define a distribution on Ξ_J as

$$\nu_J \equiv \nu_{t_1} \otimes \mu_{t_1, t_2} \otimes \dots \otimes \mu_{t_{n-1}, t_n} \quad (9.7)$$

It is easily checked, using point (2) in the definition of a transition kernel semi-group (Definition 105), that the ν_J form a projective family of distributions. Thus, by the Kolmogorov Extension Theorem (Theorem 29), there exists a stochastic process whose finite-dimensional distributions are the ν_J . Now pick a J of size n , and two sets, $B \in \mathcal{X}^{n-1}$ and $C \in \mathcal{X}$.

$$\mathbb{P}(X_J \in B \times C) = \nu_J(B \times C) \quad (9.8)$$

$$= \mathbf{E}[\mathbf{1}_{B \times C}(X_J)] \quad (9.9)$$

$$= \mathbf{E}[\mathbf{1}_B(X_{J \setminus t_n})\mu_{t_{n-1}, t_n}(X_{t_{n-1}}, C)] \quad (9.10)$$

Set $\{\mathcal{F}_t\}$ to be the natural filtration, $\sigma(\{X_u, u \leq s\})$. If $A \in \mathcal{F}_s$ for some $s \leq t$, then by the usual generating class arguments we have

$$\mathbb{P}(X_t \in C, X_s^- \in A) = \mathbf{E}[\mathbf{1}_A \mu_{s,t}(X_s, C)] \quad (9.11)$$

$$\mathbb{P}(X_t \in C | \mathcal{F}_s) = \mu_{s,t}(X_s, C) \quad (9.12)$$

i.e., $X_t \perp\!\!\!\perp \mathcal{F}_s | X_s$, as was to be shown.

(From the Markov property to the transition kernels.) From the Markov property, for any measurable set $C \in \mathcal{X}$, $\mathbb{P}(X_t \in C | \mathcal{F}_s)$ is a function of X_s alone. So define the kernel $\mu_{s,t}$ by $\mu_{s,t}(x, C) = \mathbb{P}(X_t \in C | X_s = x)$, with a possible measure-0 exceptional set from (ultimately) the Radon-Nikodym theorem. (The fact that Ξ is Borel guarantees the existence of a regular version of this conditional probability.) We get the semi-group property for these kernels thus: pick any three times $t \leq s \leq u$, and a measurable set $C \subseteq \Xi$. Then

$$\mu_{t,u}(X_t, C) = \mathbb{P}(X_u \in C | \mathcal{F}_t) \quad (9.13)$$

$$= \mathbb{P}(X_u \in C, X_s \in \Xi | \mathcal{F}_t) \quad (9.14)$$

$$= (\mu_{t,s} \otimes \mu_{s,u})(X_t, \Xi \times C) \quad (9.15)$$

$$= (\mu_{t,s} \mu_{s,u})(X_t, C) \quad (9.16)$$

The argument to get Eq. 9.3 is similar. \square

Note: For one-sided discrete-parameter processes, we could use the Ionescu-Tulcea Extension Theorem 33 to go from a transition kernel semi-group to a Markov process, even if Ξ is not a Borel space.

Definition 107 (Invariant Distribution) Let X be a homogeneous Markov process with transition kernels μ_t . A distribution ν on Ξ is invariant when, $\forall t$, $\nu = \nu \mu_t$, i.e.,

$$(\nu \mu_t)(B) \equiv \int \nu(dx) \mu_t(x, B) \quad (9.17)$$

$$= \nu(B) \quad (9.18)$$

ν is also called an equilibrium distribution.

The term “equilibrium” comes from statistical physics, where however its meaning is a bit more strict, in that “detailed balance” must also be satisfied: for any two sets $A, B \in \mathcal{X}$,

$$\int \nu(dx) \mathbf{1}_A \mu_t(x, B) = \int \nu(dx) \mathbf{1}_B \mu_t(x, A) \quad (9.19)$$

i.e., the flow of probability from A to B must equal the flow in the opposite direction. Much confusion has resulted from neglecting the distinction between equilibrium in the strict sense of detailed balance and equilibrium in the weaker sense of invariance.

Theorem 108 (Stationarity and Invariance for Homogeneous Markov Processes) *Suppose X is homogeneous, and $\mathcal{L}(X_t) = \nu$, where ν is an invariant distribution. Then the process X_t^+ is stationary.*

PROOF: Exercise 9.4. \square

9.3 The Markov Property Under Multiple Filtrations

Definition 102 specifies what it is for a process to be Markovian relative to a given filtration $\{\mathcal{F}_t\}$. The question arises of when knowing that X is Markov with respect to one filtration $\{\mathcal{F}_t\}$ will allow us to deduce that it is Markov with respect to another, say $\{\mathcal{G}_t\}$.

To begin with, let’s introduce a little notation.

Definition 109 (Natural Filtration) *The natural filtration for a stochastic process X is $\{\mathcal{F}_t^X\} \equiv \sigma(\{X_u, u \leq t\})$. Every process X is adapted to its natural filtration.*

Definition 110 (Comparison of Filtrations) *A filtration $\{\mathcal{G}_t\}$ is finer than or more refined than or a refinement of $\{\mathcal{F}_t\}$, $\{\mathcal{F}_t\} \prec \{\mathcal{G}_t\}$, if, for all t , $\mathcal{F}_t \subseteq \mathcal{G}_t$, and at least sometimes the inequality is strict. $\{\mathcal{F}_t\}$ is coarser or less fine than $\{\mathcal{G}_t\}$. If $\{\mathcal{F}_t\} \prec \{\mathcal{G}_t\}$ or $\{\mathcal{F}_t\} = \{\mathcal{G}_t\}$, we write $\{\mathcal{F}_t\} \preceq \{\mathcal{G}_t\}$.*

Lemma 111 (The Natural Filtration Is the Coarsest One to Which a Process Is Adapted) *If X is adapted to $\{\mathcal{G}_t\}$, then $\{\mathcal{F}_t^X\} \preceq \{\mathcal{G}_t\}$.*

PROOF: For each t , X_t is \mathcal{G}_t measurable. But \mathcal{F}_t^X is, by construction, the smallest σ -algebra with respect to which X_t is measurable, so, for every t , $\mathcal{F}_t^X \subseteq \mathcal{G}_t$, and the result follows. \square

Theorem 112 (Markovianity Is Preserved Under Coarsening) *If X is Markovian with respect to $\{\mathcal{G}_t\}$, then it is Markovian with respect to any coarser filtration to which it is adapted, and in particular with respect to its natural filtration.*

PROOF: Use the smoothing property of conditional expectations: For any two σ -fields $\mathcal{H} \subset \mathcal{K}$ and random variable Y , $\mathbf{E}[Y|\mathcal{H}] = \mathbf{E}[\mathbf{E}[Y|\mathcal{K}]|\mathcal{H}]$ a.s. So, if $\{\mathcal{F}_t\}$ is coarser than $\{\mathcal{G}_t\}$, and X is Markovian with respect to the latter, for any function $f \in L_1$ and time $s > t$,

$$\mathbf{E}[f(X_s)|\mathcal{F}_t] = \mathbf{E}[\mathbf{E}[f(X_s)|\mathcal{G}_t]|\mathcal{F}_t] \text{ a.s.} \quad (9.20)$$

$$= \mathbf{E}[\mathbf{E}[f(X_s)|X_t]|\mathcal{F}_t] \quad (9.21)$$

$$= \mathbf{E}[f(X_s)|X_t] \quad (9.22)$$

The next-to-last line uses the fact that $X_s \perp\!\!\!\perp \mathcal{G}_t | X_t$, because X is Markovian with respect to $\{\mathcal{G}_t\}$, and this in turn implies that conditioning X_s , or any function thereof, on \mathcal{G}_t is equivalent to conditioning on X_t alone. (Recall that X_t is \mathcal{G}_t -measurable.) The last line uses the facts that (i) $\mathbf{E}[f(X_s)|X_t]$ is a function X_t , (ii) X is adapted to $\{\mathcal{F}_t\}$, so X_t is \mathcal{F}_t -measurable, and (iii) if Y is \mathcal{F} -measurable, then $\mathbf{E}[Y|\mathcal{F}] = Y$. Since this holds for all $f \in L_1$, it holds in particular for $\mathbf{1}_A$, where A is any measurable set, and this established the conditional independence which constitutes the Markov property. Since (Lemma 111) the natural filtration is the coarsest filtration to which X is adapted, the remainder of the theorem follows. \square

The converse is false, as the following example shows.

Example 113 (The Logistic Map Shows That Markovianity Is Not Preserved Under Refinement) *We revert to the symbolic dynamics of the logistic map, Examples 39 and 40. Let S_1 be distributed on the unit interval with density $1/\pi\sqrt{s(1-s)}$, and let $S_n = 4S_{n-1}(1 - S_{n-1})$. Finally, let $X_n = \mathbf{1}_{[0.5, 1.0]}(S_n)$. It can be shown that the X_n are a Markov process with respect to their natural filtration; in fact, with respect to that filtration, they are independent and identically distributed Bernoulli variables with probability of success $1/2$. However, $\mathbb{P}(X_{n+1}|\mathcal{F}_n^S, X_n) \neq \mathbb{P}(X_{n+1}|X_n)$, since X_{n+1} is a deterministic function of S_n . But, clearly, $\mathcal{F}_n^X \subset \mathcal{F}_n^S$ for each n , so $\{\mathcal{F}_t^X\} \prec \{\mathcal{F}_t^S\}$.*

The issue can be illustrated with graphical models (Spirtes *et al.*, 2001; Pearl, 1988). A discrete-time Markov process looks like Figure 9.1a. X_n blocks all the paths from the past to the future (in the diagram, from left to right), so it produces the desired conditional independence. Now let's add another variable which actually drives the X_n (Figure 9.1b). If we can't measure the S_n variables, just the X_n ones, then it can still be the case that we've got the conditional independence among what we can see. But if we can see X_n as well as S_n — which is what refining the filtration amounts to — then simply conditioning on X_n does not block all the paths from the past of X to its future, and, generally speaking, we will lose the Markov property. Note that knowing S_n *does* block all paths from past to future — so this remains a *hidden* Markov model. Markovian representation theory is about finding conditions under which we can get things to look like Figure 9.1b, even if we can't get them to look like Figure 9.1a.

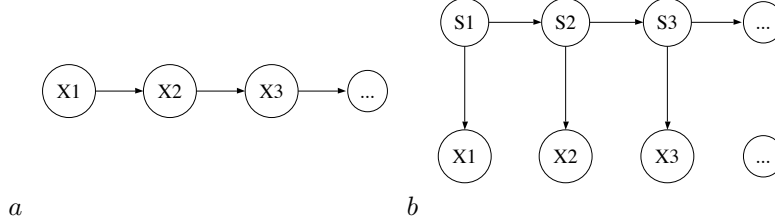


Figure 9.1: (a) Graphical model for a Markov chain. (b) Refining the filtration, say by conditioning on an additional random variable, can lead to a failure of the Markov property.

9.4 Exercises

Exercise 9.1 (Extension of the Markov Property to the Whole Future) *Prove Lemma 103.*

Exercise 9.2 (Futures of Markov Processes Are One-Sided Markov Processes) *Show that if X is a Markov process, then, for any $t \in T$, X_t^+ is a one-sided Markov process.*

Exercise 9.3 (Discrete-Time Sampling of Continuous-Time Markov Processes) *Let X be a continuous-parameter Markov process, and t_n a countable set of strictly increasing indices. Set $Y_n = X_{t_n}$. Is Y_n a Markov process? If X is homogeneous, is Y also homogeneous? Does either answer change if $t_n = nt$ for some constant interval $t > 0$?*

Exercise 9.4 (Stationarity and Invariance for Homogeneous Markov Processes) *Prove Theorem 108.*

Exercise 9.5 (Rudiments of Likelihood-Based Inference for Markov Chains) *(This exercise presumes some knowledge of sufficient statistics and exponential families from theoretical statistics.)*

Assume $T = \mathbb{N}$, Ξ is a finite set, and X a homogeneous Markov Ξ -valued Markov chain. Further assume that X_1 is constant, $= x_1$, with probability 1. (This last is not an essential restriction.) Let $p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$.

1. *Show that p_{ij} fixes the transition kernel, and vice versa.*
2. *Write the probability of a sequence $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, for short $X_1^n = x_1^n$, as a function of p_{ij} .*
3. *Write the log-likelihood ℓ as a function of p_{ij} and x_1^n .*

4. Define n_{ij} to be the number of times t such that $x_t = i$ and $x_{t+1} = j$. Similarly define n_i as the number of times t such that $x_t = i$. Write the log-likelihood as a function of p_{ij} and n_{ij} .
5. Show that the n_{ij} are sufficient statistics for the parameters p_{ij} .
6. Show that the distribution has the form of a canonical exponential family, with sufficient statistics n_{ij} , by finding the natural parameters. (Hint: the natural parameters are transformations of the p_{ij} .) Is the family of full rank?
7. Find the maximum likelihood estimators, for either the p_{ij} parameters or for the natural parameters. (Hint: Use Lagrange multipliers to enforce the constraint $\sum_j n_{ij} = n_i$.)

The classic book by Billingsley (1961) remains an excellent source on statistical inference for Markov chains and related processes.

Exercise 9.6 (Implementing the MLE for a Simple Markov Chain)

(This exercise continues the previous one.)

Set $\Xi = \{0, 1\}$, $x_1 = 0$, and

$$\mathbf{p}_0 = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$$

1. Write a program to simulate sample paths from this Markov chain.
2. Write a program to calculate the maximum likelihood estimate $\hat{\mathbf{p}}$ of the transition matrix from a sample path.
3. Calculate $2(\ell(\hat{\mathbf{p}}) - \ell(\mathbf{p}_0))$ for many independent sample paths of length n . What happens to the distribution as $n \rightarrow \infty$? (Hint: see Billingsley (1961), Theorem 2.2.)

Exercise 9.7 (The Markov Property and Conditional Independence from the Immediate Past) Let X be a Ξ -valued discrete-parameter random process. Suppose that, for all t , $X_{t-1} \perp\!\!\!\perp X_{t+1} | X_t$. Either prove that X is a Markov process, or provide a counter-example. You may assume that Ξ is a Borel space if you find that helpful.

Exercise 9.8 (Higher-Order Markov Processes) A discrete-time process X is a k^{th} order Markov process with respect to a filtration $\{\mathcal{F}_t\}$ when

$$X_{t+1} \perp\!\!\!\perp \mathcal{F}_t | \sigma(X_t, X_{t-1}, \dots, X_{t-k+1}) \quad (9.23)$$

for some finite integer k . For any Ξ -valued discrete-time process X , define the k -block process $Y^{(k)}$ as the Ξ^k -valued process where $Y_t^{(k)} = (X_t, X_{t+1}, \dots, X_{t+k-1})$.

1. Prove that if X is Markovian to order k , it is Markovian to any order $l > k$. (For this reason, saying that X is a k^{th} order Markov process conventionally means that k is the smallest order at which Eq. 9.23 holds.)
2. Prove that X is k^{th} -order Markovian if and only if $Y^{(k)}$ is Markovian.

The second result shows that studying on the theory of first-order Markov processes involves no essential loss of generality. For a test of the hypothesis that X is Markovian of order k against the alternative that is Markovian of order $l > k$, see Billingsley (1961). For recent work on estimating the order of a Markov process, assuming it is Markovian to some finite order, see the elegant paper by Peres and Shields (2005).

Exercise 9.9 (AR(1) Models) A first-order autoregressive model, or $AR(1)$ model, is a real-valued discrete-time process defined by an evolution equation of the form

$$X(t) = a_0 + a_1 X(t-1) + \epsilon(t)$$

where the innovations $\epsilon(t)$ are independent and identically distributed, and independent of $X(0)$. A p^{th} -order autoregressive model, or $AR(p)$ model, is correspondingly defined by

$$X(t) = a_0 + \sum_{i=1}^p a_i X(t-i) + \epsilon(t)$$

Finally an $AR(p)$ model in state-space form is an \mathbb{R}^p -valued process defined by

$$\vec{Y}(t) = a_0 \vec{e}_1 + \begin{bmatrix} a_1 & a_2 & \dots & a_{p-1} & a_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \vec{Y}(t-1) + \epsilon(t) \vec{e}_1$$

where \vec{e}_1 is the unit vector along the first coordinate axis.

1. Prove that $AR(1)$ models are Markov for all choices of a_0 and a_1 , and all distributions of $X(0)$ and ϵ .
2. Give an explicit form for the transition kernel of an $AR(1)$ in terms of the distribution of ϵ .
3. Are $AR(p)$ models Markovian when $p > 1$? Prove or give a counterexample.
4. Prove that \vec{Y} is a Markov process, without using Exercise 9.8. (Hint: What is the relationship between $Y_i(t)$ and $X(t-i+1)$?)

Chapter 10

Alternative Characterizations of Markov Processes

This lecture introduces two ways of characterizing Markov processes other than through their transition probabilities.

Section 10.1 describes discrete-parameter Markov processes as transformations of sequences of IID uniform variables.

Section 10.2 describes Markov processes in terms of measure-preserving transformations (Markov operators), and shows this is equivalent to the transition-probability view.

10.1 Markov Sequences as Transduced Noise

A key theorem says that discrete-time Markov processes can be viewed as the result of applying a certain kind of filter to pure noise.

Theorem 114 (Markov Sequences as Transduced Noise) *Let X be a one-sided discrete-parameter process taking values in a Borel space Ξ . X is Markov iff there are measurable functions $f_n : \Xi \times [0, 1] \mapsto \Xi$ such that, for IID random variables $Z_n \sim U(0, 1)$, all independent of X_1 , $X_{n+1} = f_n(X_n, Z_n)$ almost surely. X is homogeneous iff $f_n = f$ for all n .*

PROOF: Kallenberg, Proposition 8.6, p. 145. Notice that, in order to get the “only if” direction to work, Kallenberg invokes what we have as Proposition 26, which is where the assumptions that Ξ is a Borel space comes in. You should verify that the “if” direction does not require this assumption. \square

Let us stick to the homogeneous case, and consider the function f in somewhat more detail.

In engineering or computer science, a *transducer* is an apparatus — really, a function — which takes a stream of inputs of one kind and produces a stream of outputs of another kind.

Definition 115 (Transducer) *A (deterministic) transducer is a sextuple $\langle \Sigma, \Upsilon, \Xi, f, h, s_0 \rangle$ where Σ , Υ and Ξ are, respectively, the state, input and output spaces, $f : \Sigma \times \Xi \mapsto \Sigma$ is the state update function or state transition function, $h : \Sigma \times \Upsilon \mapsto \Xi$ is the measurement or observation function, and $s_0 \in \Sigma$ is the starting state. (We shall assume both f and h are always measurable.) If h does not depend on its state argument, the transducer is memoryless. If f does not depend on its state argument, the transducer is without after-effect.*

It should be clear that if a memoryless transducer is presented with IID inputs, its output will be IID as well. What Theorem 114 says is that, if we have a transducer with memory (so that h depends on the state) but is without after-effect (so that f does not depend on the state), IID inputs will produce Markovian outputs, and conversely any reasonable Markov process can be represented in this way. Notice that if a transducer is without memory, we can replace it with an equivalent with a single state, and if it is without after-effect, we can identify Σ and Ξ .

Notice also that the two functions f and h determine a transition function where we use the input to update the state: $g : \Sigma \times \Upsilon \mapsto \Sigma$, where $g(s, y) = f(s, h(s, y))$. Thus, if the inputs are IID and uniformly distributed, then (Theorem 114) the successive states of the transducer are always Markovian. The question of which processes can be produced by noise-driven transducers is this intimately bound up with the question of Markovian representations. While, as mentioned, quite general stochastic processes *can* be put in this form (Knight, 1975, 1992), it is not necessarily possible to do this with a finite internal state space Σ , even when Ξ is finite. The distinction between finite and infinite Σ is crucial to theoretical computer science, and we might come back to it later, but

Two issues suggest themselves in connection with this material. One is whether, given a *two-sided* process, we can pull the same trick, and represent a Markovian X as a transformation of an IID sequence extending into the infinite past. (Remember that the theorem is for one-sided processes, and starts with an initial X_1 .) This is more subtle than it seems at first glance, or even than it seemed to Norbert Wiener when he first posed the question (Wiener, 1958); for a detailed discussion, see Rosenblatt (1971), and, for recent set of applications, Wu (2005). The other question is whether the same trick can be pulled in continuous time; here much less is known.

10.2 Time-Evolution (Markov) Operators

Let's look again at the evolution of the one-dimensional distributions for a Markov process:

$$\nu_s = \nu_t \mu_{t,s} \quad (10.1)$$

$$\nu_s(B) = \int \nu_t(dx) \mu_{t,s}(x, B) \quad (10.2)$$

The transition kernels define linear operators taking probability measures on Ξ to probability measures on Ξ . This can be abstracted.

Definition 116 (Markov Operator on Measures) *Take any measurable space Ξ, \mathcal{X} . A Markov operator on measures is an operator M which takes finite measures on this space to other finite measures on this space such that*

1. M is linear, i.e., for any $a_1, a_2 \in [0, 1]$ and any two measures μ_1, μ_2 ,

$$M(a_1\mu_1 + a_2\mu_2) = a_1M\mu_1 + a_2M\mu_2 \quad (10.3)$$

2. M is norm-preserving, i.e., $M\mu(\Xi) = \mu(\Xi)$.

(In particular P must take probability measures to probability measures.)

Definition 117 (Markov Operator on Densities) *Take any probability space Ξ, \mathcal{X}, μ , and let L_1 be as usual the class of all μ -integrable generalized functions on Ξ . A linear operator $P : L_1 \mapsto L_1$ is a Markov operator on densities when:*

1. If $f \geq 0$ (a.e. μ), $Pf \geq 0$ (a.e. μ).
2. If $f \geq 0$ (a.e. μ), $\|Pf\| = \|f\|$.

By “a Markov operator” I will often mean a Markov operator on densities, with the reference measure μ being some suitable uniform distribution on Ξ . However, the general theory applies to operators on measures.

Lemma 118 (Markov Operators on Measures Induce Those on Densities) *Let M be a Markov operator on measures. If M takes measures absolutely continuous with respect to μ to measures absolutely continuous with respect to μ , i.e., it preserves domination by μ , then it induces an almost-unique Markov operator P on densities with respect to μ .*

PROOF: Let f be a function which is in $L_1(\mu)$ and is non-negative (μ -a.e.). If $f \geq 0$ μ a.e., the set function $\nu_f(A) = \int_A f(x) d\mu$ is a finite measure which is absolutely continuous with respect to μ . (Why is it finite?) By hypothesis, then, $M\nu_f$ is another finite measure which is absolutely continuous with respect to μ , and $\nu_f(\Xi) = M\nu_f(\Xi)$. Hence, by the Radon-Nikodym theorem, there is an $L_1(\mu)$ function, call it Pf , such that $M\nu_f(A) = \int_A Pf(x) d\mu$. (“Almost unique”

refers to the possibility of replacing Pf with another version of $dM\nu_f/d\mu$.) In particular, $Pf(x) \geq 0$ for μ -almost-all x , and so $\|Pf\| = \int_{\Xi} |Pf(x)|d\mu = M\nu_f(\Xi) = \nu_f(\Xi) = \int_{\Xi} |f(x)|d\mu = \|f\|$. Finally, the linearity of the operator on densities follows directly from the linearity of the operator on measures and the linearity of integration. If f is sometimes negative, apply the reasoning above to f^+ and f^- , its positive and negative parts, and then use linearity again. \square

Recall from Definition 30 that, for an arbitrary kernel κ , $\kappa f(x)$ is defined as $\int f(y)\kappa(x, dy)$. Applied to our transition kernels, this suggests another kind of operator.

Definition 119 (Transition Operators) *Take any measurable space Ξ, \mathcal{X} , and let $B(\Xi)$ be the class of bounded measurable functions. An operator $K : B(\Xi) \mapsto B(\Xi)$ is a transition operator when:*

1. K is linear
2. If $f \geq 0$ (a.e. μ), $Kf \geq 0$ (a.e. μ)
3. $K1_{\Xi} = 1_{\Xi}$.
4. If $f_n \downarrow 0$, then $Kf_n \downarrow 0$.

Definition 120 (L_{∞} Transition Operators) *For a probability space Ξ, \mathcal{X}, μ , an L_{∞} -transition operator is an operator on $L_{\infty}(\mu)$ satisfying points (1)–(4) of Definition 119.*

Note that every function in $B(\Xi)$ is in $L_{\infty}(\mu)$ for each μ , so the definition of an L_{∞} transition operator is actually stricter than that of a plain transition operator. This is unlike the case with Markov operators, where the L_1 version is weaker than the unrestricted version.

Lemma 121 (Kernels and Operators) *Every probability kernel κ from Ξ to Ξ induces a Markov operator M on measures,*

$$M\nu = \nu\kappa \quad (10.4)$$

and every Markov operator M on measures induces a probability kernel κ ,

$$\kappa(x, B) = M\delta_x(B) \quad (10.5)$$

Similarly, every transition probability kernel induces a transition operator K on functions,

$$Kf(x) = \kappa f(x) \quad (10.6)$$

and every transition operator K induces a transition probability kernel,

$$\kappa(x, B) = K1_B(x) \quad (10.7)$$

PROOF: Exercise 10.1. \square

Now we need some basic concepts from functional analysis; see, e.g., Kolmogorov and Fomin (1975) for background.

Definition 122 (Functional) A functional is a map $g : V \mapsto \mathbb{R}$, that is, a real-valued function of a function. A functional g is

- linear when $g(af_1 + bf_2) = ag(f_1) + bg(f_2)$;
- continuous when $f_n \rightarrow f$ implies $g(f_n) \rightarrow g(f)$;
- bounded by M when $|g(f)| \leq M$ for all $f \in V$;
- bounded when it is bounded by M for some M ;
- non-negative when $g(f) \geq 0$ for all f ;

etc.

Definition 123 (Conjugate or Adjoint Space) The conjugate space or adjoint space of a vector space V is the space V^\dagger of its continuous linear functionals. For $f \in V$ and $g \in V^\dagger$, $\langle f, g \rangle$ denotes $g(f)$. This is sometimes called the inner product.

Proposition 124 (Conjugate Spaces are Vector Spaces) For every V , V^\dagger is also a vector space.

Proposition 125 (Inner Product is Bilinear) For any $a, b, c, d \in \mathbb{R}$, any $f_1, f_2 \in V$ and any $g_1, g_2 \in V^\dagger$,

$$\langle af_1 + bf_2, cg_1 + dg_2 \rangle = ac\langle f_1, g_1 \rangle + ad\langle f_1, g_2 \rangle + bc\langle f_2, g_1 \rangle + bd\langle f_2, g_2 \rangle \quad (10.8)$$

PROOF: Follows from the fact that V^\dagger is a vector space, and each g_i is a linear operator. \square

You are already familiar with an example of a conjugate space.

Example 126 (Vectors in \mathbb{R}^n) The vector space \mathbb{R}^n is self-conjugate. If $g(\vec{x})$ is a continuous linear function of \vec{x} , then $g(\vec{x}) = \sum_{i=1}^n y_i x_i$ for some real constants y_i , which means $g(\vec{x}) = \vec{y} \cdot \vec{x}$.

Here is the simplest example where the conjugate space is not equal to the original space.

Example 127 (Row and Column Vectors) The space of row vectors is conjugate to the space of column vectors, since every continuous linear functional of a column vector x takes the form of $y^T x$ for some other column vector y .

Example 128 (L_p spaces) *The function spaces $L_p(\mu)$ and $L_q(\mu)$ are conjugate to each other, when $1/p + 1/q = 1$, and the inner product is defined through*

$$\langle f, g \rangle \equiv \int f g d\mu \quad (10.9)$$

In particular, L_1 and L_∞ are conjugates.

Example 129 (Measures and Functions) *The space of $C_b(\Xi)$ of bounded, continuous functions on Ξ and the spaces $\mathcal{M}(\Xi, \mathcal{X})$ of finite measures on Ξ are conjugates, with inner product*

$$\langle \mu, f \rangle = \int f d\mu \quad (10.10)$$

Definition 130 (Adjoint Operator) *For conjugate spaces V and V^\dagger , the adjoint operator, O^\dagger , to an operator O on V is an operator on V^\dagger such that*

$$\langle Of, g \rangle = \langle f, O^\dagger g \rangle \quad (10.11)$$

for all $f \in V, g \in V^\dagger$.

Proposition 131 (Adjoint of a Linear Operator) *If O is a continuous linear operator on V , then its adjoint O^\dagger exists and is linear.*

Lemma 132 (Markov Operators on Densities and L_∞ Transition Operators) *Every Markov operator P on densities induces an L_∞ transition operator U on essentially-bounded functions, and vice versa.*

PROOF: Exercise 10.2. \square

Clearly, if κ is part of a transition kernel semi-group, then the collection of induced Markov operators and transition operators also form semi-groups.

Theorem 133 (Transition operator semi-groups and Markov processes)

Let X be a Markov process with transition kernels $\mu_{t,s}$, and let $K_{t,s}$ be the corresponding semi-group of transition operators. Then for any $f \in B(\Xi)$,

$$\mathbf{E}[f(X_s) | \mathcal{F}_t] = (K_{t,s} f)(X_t) \quad (10.12)$$

Conversely, let X be any stochastic process, and let $K_{t,s}$ be a semi-group of transition operators such that Equation 10.12 is valid (a.s.). Then X is a Markov process.

PROOF: Exercise 10.3. \square

Remark. The proof works because the expectations of all $B(\Xi)$ functions together determine a probability measure. (Recall that $\mathbb{P}(B) = \mathbf{E}[\mathbf{1}_B]$, and indicators are bounded everywhere.) If we knew of another collection of functions which also sufficed to determine a measure, then linear operators on that collection would work just as well, in the theorem, as do the transition operators, which by definition apply to all of $B(\Xi)$. In particular, it is sometimes possible to define operators only on much smaller, more restricted collections of functions, which can have technical advantages. See Ethier and Kurtz (1986, ch. 4, sec. 1) for details.

The next two lemmas are useful in establishing asymptotic results.

Lemma 134 (Markov Operators are Contractions) *For any Markov operator P and $f \in L_1$,*

$$\|Pf\| \leq \|f\| \quad (10.13)$$

PROOF (after Lasota and Mackey (1994, prop. 3.1.1, pp. 38–39)): First, notice that $(Pf(x))^+ \leq Pf^+(x)$, because

$$(Pf(x))^+ = (Pf^+ - Pf^-)^+ = \max(0, Pf^+ - Pf^-) \leq \max(0, Pf^+) = Pf^+$$

Similarly $(Pf(x))^- \leq Pf^-(x)$. Therefore $|Pf| \leq P|f|$, and then the statement follows by integration. \square

Lemma 135 (Markov Operators Bring Distributions Closer) *For any Markov operator, and any $f, g \in L_1$, $\|P^n f - P^n g\|$ is non-increasing.*

PROOF: By linearity, $\|P^n f - P^n g\| = \|P^n(f - g)\|$. By the definition of P^n , $\|P^n(f - g)\| = \|PP^{n-1}(f - g)\|$. By the contraction property (Lemma 134), $\|PP^{n-1}(f - g)\| \leq \|P^{n-1}(f - g)\| = \|P^{n-1}f - P^{n-1}g\|$ (by linearity again). \square

Theorem 136 (Invariant Measures Are Fixed Points) *A probability measure ν is invariant for a homogeneous Markov process iff it is a fixed point of all the Markov operators, $M_t\nu = \nu$.*

PROOF: Clear from the definitions! \square

10.3 Exercises

Exercise 10.1 (Kernels and Operators) *Prove Lemma 121. Hints: 1. You will want to use the fact that $\mathbf{1}_A \in B(\Xi)$ for all measurable sets A . 2. In going back and forth between transition kernels and transition operators, you may find Proposition 32 helpful.*

Exercise 10.2 (L_1 and L_∞) *Prove Lemma 132.*

Exercise 10.3 (Operators and Expectations) *Prove Theorem 133. Hint: in showing that a collection of operators determines a Markov process, try using mathematical induction on the finite-dimensional distributions.*

Exercise 10.4 (Bayesian Updating as a Markov Process) *Consider a simple version of Bayesian learning, where the set of hypotheses Θ is finite, and, for each $\theta \in \Theta$, $f_\theta(x)$ is a probability density on Ξ with respect to a common dominating measure, μ say, and the Ξ -valued data X_1, X_2, \dots are all IID, both under the hypotheses and in reality. Given a prior probability vector π_0 on Θ , the posterior π_n is defined via Bayes's rule:*

$$\pi_n(\theta) = \frac{\pi_0(\theta) \prod_{i=1}^n f_\theta(X_i)}{\sum_{\theta \in \Theta} \pi_0(\theta) \prod_{i=1}^n f_\theta(X_i)}$$

1. *Prove that the random sequence π_1, π_2, \dots is adapted to $\{\mathcal{F}_t\}$ if X is adapted.*
2. *Prove that the sequence of posterior distributions is Markovian with respect to its natural filtration.*
3. *Is this still a Markov process if X is not IID? If the hypotheses θ do not model X as IID?*
4. *When, if ever, is the Markov process homogeneous? (If it is sometimes homogeneous, you may give either necessary or sufficient conditions, as you find easier.)*

Exercise 10.5 (More on Bayesian Updating) *Consider a more complicated version of Bayesian updating. Let T be one-sided, H be a Θ -valued random variable, and $\{\mathcal{G}_t\}$ be any filtration. Assume that $\pi_t = \mathcal{L}(H|\mathcal{G}_t)$ is a regular conditional probability distribution on Θ for all t and all ω . As before, π_0 is the prior. Prove that π_t is Markovian with respect to $\{\mathcal{G}_t\}$. (Hint: $\mathbf{E}[\mathbf{E}[Y|\mathcal{G}_t]|\mathcal{G}_s] = \mathbf{E}[Y|\mathcal{G}_s]$ a.s., when $s \leq t$ and $Y \in L_1$ so the expectations exist.)*

Chapter 11

Examples of Markov Processes

Section 11.1 looks at the evolution of densities under the action of the logistic map; this shows how deterministic dynamical systems can be brought under the sway of the theory we've developed for Markov processes.

Section 11.2 finds the transition kernels for the Wiener process, as an example of how to manipulate such things.

Section 11.3 generalizes the Wiener process example to other processes with stationary and independent increments, and in doing so uncovers connections to limits of sums of IID random variables and to self-similarity.

11.1 Probability Densities in the Logistic Map

Let's revisit the first part of Exercise 5.3, from the point of view of what we now know about Markov processes. The exercise asks us to show that the density $\frac{1}{\pi\sqrt{x(1-x)}}$ is invariant under the action of the logistic map with $a = 4$.

Let's write the mapping as $F(x) = 4x(1-x)$. Solving a simple quadratic equation gives us the fact that $F^{-1}(x)$ is the set $\{\frac{1}{2}(1 - \sqrt{1-x}), \frac{1}{2}(1 + \sqrt{1-x})\}$. Notice, for later use, that the two solutions add up to 1. Notice also that $F^{-1}([0, x]) = [0, \frac{1}{2}(1 - \sqrt{1-x})] \cup [\frac{1}{2}(1 + \sqrt{1-x}), 1]$. Now we consider the

cumulative distribution function of X_{n+1} , $\mathbb{P}(X_{n+1} \leq x)$.

$$\begin{aligned} \mathbb{P}(X_{n+1} \leq x) &= \mathbb{P}(X_{n+1} \in [0, x]) \end{aligned} \quad (11.1)$$

$$= \mathbb{P}(X_n \in F^{-1}([0, x])) \quad (11.2)$$

$$= \mathbb{P}\left(X_n \in \left[0, \frac{1}{2}(1 - \sqrt{1-x})\right] \cup \left[\frac{1}{2}(1 + \sqrt{1-x}), 1\right]\right) \quad (11.3)$$

$$= \int_0^{\frac{1}{2}(1-\sqrt{1-x})} \rho_n(y) dy + \int_{\frac{1}{2}(1+\sqrt{1-x})}^1 \rho_n(y) dy \quad (11.4)$$

where ρ_n is the density of X_n . So we have an integral equation for the evolution of the density,

$$\int_0^x \rho_{n+1}(y) dy = \int_0^{\frac{1}{2}(1-\sqrt{1-x})} \rho_n(y) dy + \int_{\frac{1}{2}(1+\sqrt{1-x})}^1 \rho_n(y) dy \quad (11.5)$$

This sort of integral equation is complicated to solve directly. Instead, take the derivative of both sides with respect to x ; we can do this through the fundamental theorem of calculus. On the left hand side, this will just give $\rho_{n+1}(x)$, the density we want.

$$\rho_{n+1}(x) \quad (11.6)$$

$$\begin{aligned} &= \frac{d}{dx} \int_0^{\frac{1}{2}(1-\sqrt{1-x})} \rho_n(y) dy + \frac{d}{dx} \int_{\frac{1}{2}(1+\sqrt{1-x})}^1 \rho_n(y) dy \\ &= \rho_n\left(\frac{1}{2}(1 - \sqrt{1-x})\right) \frac{d}{dx} \left(\frac{1}{2}(1 - \sqrt{1-x})\right) \end{aligned} \quad (11.7)$$

$$\begin{aligned} &\quad - \rho_n\left(\frac{1}{2}(1 + \sqrt{1-x})\right) \frac{d}{dx} \left(\frac{1}{2}(1 + \sqrt{1-x})\right) \\ &= \frac{1}{4\sqrt{1-x}} \left(\rho_n\left(\frac{1}{2}(1 - \sqrt{1-x})\right) + \rho_n\left(\frac{1}{2}(1 + \sqrt{1-x})\right) \right) \end{aligned} \quad (11.8)$$

Notice that this defines a linear operator taking densities to densities. (You should verify the linearity.) In fact, this is a Markov operator, by the terms of Definition 117. Markov operators of this sort, derived from deterministic maps, are called *Perron-Frobenius* or *Frobenius-Perron* operators, and accordingly denoted by P . Thus an invariant density is a ρ^* such that $\rho^* = P\rho^*$. All the

problem asks us to do is to verify that $\frac{1}{\pi\sqrt{x(1-x)}}$ is such a solution.

$$\rho^*\left(\frac{1}{2}(1 - \sqrt{1-x})\right) \quad (11.9)$$

$$= \frac{1}{\pi} \left(\frac{1}{2}(1 - \sqrt{1-x}) \left(1 - \left(\frac{1}{2}(1 - \sqrt{1-x}) \right) \right) \right)^{-1/2}$$

$$= \frac{1}{\pi} \left(\frac{1}{2}(1 - \sqrt{1-x}) \frac{1}{2}(1 + \sqrt{1-x}) \right)^{-1/2} \quad (11.10)$$

$$= \frac{2}{\pi\sqrt{x}} \quad (11.11)$$

Since $\rho^*(x) = \rho^*(1-x)$, it follows that

$$P\rho^* = 2\frac{1}{4\sqrt{1-x}}\rho^*\left(\frac{1}{2}(1 - \sqrt{1-x})\right) \quad (11.12)$$

$$= \frac{1}{\pi\sqrt{x(1-x)}} \quad (11.13)$$

$$= \rho^* \quad (11.14)$$

as desired.

By Lemma 135, for any distribution ρ , $\|P^n\rho - P^n\rho^*\|$ is a non-increasing function of n . However, $P^n\rho^* = \rho^*$, so the iterates of *any* distribution, under the map, approach the invariant distribution monotonically. It would be very handy if we could show that any initial distribution ρ eventually converged on ρ^* , i.e. that $\|P^n\rho - \rho^*\| \rightarrow 0$. When we come to ergodic theory, we will see conditions under which such distributional convergence holds, as it does for the logistic map, and learn how such convergence in distribution is connected to both pathwise convergence properties, and to the decay of correlations.

11.2 Transition Kernels and Evolution Operators for the Wiener Process

We have previously defined the Wiener process (Examples 38 and 79) as the real-valued process on \mathbb{R}^+ with the following properties:

1. $W(0) = 0$;
2. For any three times $t_1 \leq t_2 \leq t_3$, $W(t_3) - W(t_2) \perp W(t_2) - W(t_1)$ (independent increments);
3. For any two times $t_1 \leq t_2$, $W(t_2) - W(t_1) \sim \mathcal{N}(0, t_2 - t_1)$ (Gaussian increments);
4. Continuous sample paths (in the sense of Definition 73).

In Exercise 4.1, you showed that a process satisfying points (1)–(3) above exists. To show that any such process has a version with continuous sample paths, invoke Theorem 97 (Exercise 11.3). Here, however, we will show that it is a homogeneous Markov process, and find both the transition kernels and the evolution operators. Markovianity follows from the independent increments property (2), while homogeneity and the form of the transition operators comes from the Gaussian assumption (3).

First, here's how the Gaussian increments property gives us the transition probabilities:

$$\mathbb{P}(W(t_2) \in B | W(t_1) = w_1) = \mathbb{P}(W(t_2) - W(t_1) \in B - w_1) \quad (11.15)$$

$$= \int_{B-w_1} du \frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{u^2}{2(t_2-t_1)}} \quad (11.16)$$

$$= \int_B dw_2 \frac{1}{\sqrt{2\pi(t_2-t_1)}} e^{-\frac{(w_2-w_1)^2}{2(t_2-t_1)}} \quad (11.17)$$

$$\equiv \mu_{t_1, t_2}(w_1, B) \quad (11.18)$$

Since this evidently depends only on $t_2 - t_1$, and not the individual times, the process is homogeneous. Now, assuming homogeneity, we can find the time-evolution operators for well-behaved observables:

$$K_t f(w) = \mathbf{E}[f(W_t + s) | W_s = w] \quad (11.19)$$

$$= \int f(u) \mu_t(w, du) \quad (11.20)$$

$$= \int f(u) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(u-w)^2}{2t}} du \quad (11.21)$$

$$= \mathbf{E}[f(w + \sqrt{t}Z)] \quad (11.22)$$

where Z is a standard Gaussian random variable independent of W_s .

To show that $W(t)$ is a Markov process, we must show that, for any finite collection of times $t_1 \leq t_2 \leq \dots \leq t_k$,

$$\mathcal{L}(W_{t_1}, W_{t_2}, \dots, W_{t_k}) = \mu_{t_1, t_2} \mu_{t_2, t_3} \dots \mu_{t_{k-1}, t_k} \quad (11.23)$$

Let's just go through the $k = 3$ case, as the others are fundamentally similar, but with more notation. Notice that $W(t_3) - W(t_1) = (W(t_3) - W(t_2)) + (W(t_2) - W(t_1))$. Because increments are independent, then, $W(t_3) - W(t_1)$ is the sum of two independent random variables, $W(t_3) - W(t_2)$ and $W(t_2) - W(t_1)$. The distribution of $W(t_3) - W(t_1)$ is then the convolution of distributions of $W(t_3) - W(t_2)$ and $W(t_2) - W(t_1)$. Those are $\mathcal{N}(0, t_3 - t_2)$ and $\mathcal{N}(0, t_2 - t_1)$ respectively. The convolution of two Gaussian distributions is a third Gaussian, summing their parameters, so according to this argument, we must have $W(t_3) - W(t_1) \sim \mathcal{N}(0, t_3 - t_1)$. But this is precisely what we should have, by the Gaussian-increments property. Since the trick we used above to get the transition kernel from the increment distribution can be applied again, we conclude that $\mu_{t_1, t_2} \mu_{t_2, t_3} = \mu_{t_1, t_3}$. The same trick applies when $k > 3$. Therefore (Theorem 106), $W(t)$ is a Markov process, with respect to its natural filtration.

11.3 Lévy Processes and Limit Laws

Let's think a bit more about the trick we've just pulled with the Wiener process. We wrote the time evolution operators in terms of a random variable,

$$K_t f(w) = \mathbf{E} \left[f(w + \sqrt{t}Z) \right] \quad (11.24)$$

Notice that $\sqrt{t}Z \sim \mathcal{N}(0, t)$, i.e., it is the increment of the process over an interval of time t . This can be generalized to other processes of a similar sort. These are one natural way of generalizing the idea of a sum of IID random variables.

Definition 137 (Processes with Stationary and Independent Increments) *A stochastic process X is a stationary, independent-increments process, or has stationary, independent increments when*

1. *The increments are independent; for any collection of indices $t_1 \leq t_2 \leq \dots \leq t_k$, with k finite, the increments $X(t_i) - X(t_{i-1})$ are all jointly independent.*
2. *The increments are stationary; for any $\tau > 0$,*

$$\begin{aligned} & \mathcal{L}(X(t_2) - X(t_1), X(t_3) - X(t_2), \dots, X(t_k) - X(t_{k-1})) \\ &= \mathcal{L}(X(t_2 + \tau) - X(t_1 + \tau), X(t_3 + \tau) - X(t_2 + \tau), \dots, X(t_k + \tau) - X(t_{k-1} + \tau)) \end{aligned} \quad (11.25)$$

Definition 138 (Lévy Processes) *A Lévy process is a process with stationary and independent increments and cadlag sample paths, which start at zero, $X(0) = 0$.*

Example 139 (Wiener Process is Lévy) *The Wiener process is a Lévy process (since it has not just cadlag but continuous sample paths).*

Example 140 (Poisson Counting Process) *The Poisson counting process with intensity λ is the integer-valued process N on \mathbb{R}^+ where $N(0) = 0$, $N(t) \sim \text{Poisson}(\lambda t)$, and independent increments. It defines a Poisson point process (in the sense of Example 20) by assigning the interval $[0, t]$ the measure $N(t)$, the measure extending to other Borel sets in the usual way. Conversely, any Poisson process defines a counting process in this sense. N can be shown to be continuous in probability, and so, via Theorem 96, to have a cadlag modification. (Exercise 11.5.) This is a Lévy process.*

It is not uncommon to see people writing just “processes with independent increments” when they mean “stationary and independent increments”.

Theorem 141 (Processes with Stationary Independent Increments are Markovian) *Any process with stationary, independent increments is Markovian with respect to its natural filtration, and homogeneous in time.*

PROOF: The joint distribution for any finite collection of times factors into a product of incremental distributions:

$$\begin{aligned} \mathcal{L}(X(t_1), X(t_2), \dots, X(t_k)) \\ = \mathcal{L}(X(t_1)) \mathcal{L}(X(t_2) - X(t_1)) \dots \mathcal{L}(X(t_k) - X(t_{k-1})) \end{aligned} \quad (11.26)$$

Define $\mu_t(x, B) = \mathbb{P}(X(t_1 + t) - X(t_1) + x \in B)$. This is a probability kernel. Moreover, it satisfies the semi-group property, since $X(t_3) - X(t_1) = X(t_3) - X(t_2) + X(t_2) - X(t_1)$ (and similarly for more indices). Thus the μ_t , so defined, are transition probability kernels (Definition 105). By Theorem 106, X is Markovian with respect to its natural filtration. \square

Remark: The assumption that the increments are stationary is not necessary to prove Markovianity. What property of X does it then deliver?

Theorem 142 (Time-Evolution Operators of Processes with Stationary, Independent Increments) *If X has stationary and independent increments, then its time-evolution operators K_t are given by*

$$K_t f(x) = \mathbf{E}[f(x + Z_t)] \quad (11.27)$$

where $\mathcal{L}(Z_t) = \mathcal{L}(X(t) - X(0))$, the increment from 0 to t .

PROOF: Exactly parallel to the Wiener process case.

$$K_t f(x) = \mathbf{E}[f(X_t) | \mathcal{F}_0] \quad (11.28)$$

$$= \mathbf{E}[f(X_t) | X_0 = x] \quad (11.29)$$

$$= \int f(y) \mu_t(x, dy) \quad (11.30)$$

Since $\mu_t(x, B) = \mathbb{P}(X(t_1 + t) - X(t_1) + x \in B)$, the theorem follows. \square

Notice that so far everything has applied equally to discrete or to continuous time. For discrete time, we can chose the distribution of increments over a single time-step, $\mathcal{L}(X(1) - X(0))$, to be essentially whatever we like; stationarity and independence then determine the distribution of all other increments. For continuous time, however, our choices are more constrained, and in an interesting way.

By the semi-group property of the time-evolution operators, we must have $K_s K_t = K_{t+s}$ for all times s and t . Applying Theorem 142, it must be the case that

$$K_{t+s} f(x) = \mathbf{E}[f(x + Z_{t+s})] \quad (11.31)$$

$$= K_s \mathbf{E}[f(x + Z_t)] \quad (11.32)$$

$$= \mathbf{E}[f(x + Z_s + Z_t)] \quad (11.33)$$

where $Z_s \perp\!\!\!\perp Z_t$. That is, the distribution of Z_{t+s} must be the convolution of the distributions of Z_t and Z_s . Let us in particular pick $s = t$; this implies $\mathcal{L}(Z_{2t}) = \mathcal{L}(Z_t) \star \mathcal{L}(Z_t)$, where \star indicates convolution. Writing ν_t for $\mathcal{L}(Z_t)$,

another way to phrase our conclusion is that $\nu_t = \nu_{t/2} \star \nu_{t/2}$. Clearly, this argument could be generalized to relate ν_t to $\nu_{t/n}$ for any n :

$$\nu_t = \nu_{t/n}^{\star n} \quad (11.34)$$

where the superscript $\star n$ indicates the n -fold convolution of the distribution with itself. Equivalently, for each t , and for all n ,

$$Z_t \stackrel{d}{=} \sum_{i=1}^n D_i \quad (11.35)$$

where the IID $D_i \sim \nu_{t/n}$. This is a very curious-looking property with a name.

Definition 143 (Infinitely-Divisible Distributions and Random Variables) *A probability distribution ν is infinitely divisible when, for each n , there exists a ν_n such that $\nu = \nu_n^{\star n}$. A random variable Z is infinitely divisible when, for each n , there are n IID random variables $D_i^{(n)}$ such that $Z \stackrel{d}{=} \sum_i D_i^{(n)}$, i.e., when its distribution is infinitely divisible.*

Clearly, if ν is an infinitely divisible distribution, then it can be obtained as the limiting distribution of a sum of IID random variables $D_i^{(n)}$. (To converge, the individual terms must be going to zero in probability as n grows.) More remarkably, the converse is also true:

Proposition 144 (Limiting Distributions Are Infinitely Divisible) *If ν is the limiting distribution of a sequence of IID sums, then it is an infinitely-divisible distribution.*

PROOF: See, for instance, Kallenberg (2002, Theorem 15.12). \square

Remark: This should *not* be obvious. If we take larger and larger sums of (centered, standardized) IID Bernoulli variables, we obtain a Gaussian as the limiting distribution, but at no time is the Gaussian the distribution of any of the finite sums. That is, the IID sums in the definition of infinite divisibility are not necessarily at all the same as those in the proposition.

Theorem 145 (Infinitely Divisible Distributions and Stationary Independent Increments) *If X is a process with stationary and independent increments, then for all t , $X(t) - X(0)$ is infinitely divisible.*

PROOF: See the remarks before the definition of infinite divisibility. \square

Corollary 146 (Infinitely Divisible Distributions and Lévy Processes) *If ν is an infinitely divisible distribution, there exists a Lévy process where $X(1) \sim \nu$, and this determines the other distributions.*

PROOF: If $X(1) \sim \nu$, then by Eq. 11.34, $X(n) \sim \nu^{\star n}$ for all integer n . Conversely, the distribution of $X(1/n)$ is also fixed, since $\nu = \mathcal{L}(X(1/n))^{\star n}$, which means that the characteristic function of ν is equal to the n^{th} power of

that of $X(1/n)$; inverting the latter then gives the desired distribution. Because increments are independent, $X(n+1/m) \stackrel{d}{=} X(n) + X(1/m)$, hence $\mathcal{L}(X(1))$ fixes the increment distribution for all rational time-intervals. By continuity, however, this also fixes it for time intervals of any length. Since the distribution of increments and the condition $X(0) = 0$ fix the finite-dimensional distributions, it remains only to show that the process is cadlag, which can be done by observing that it is continuous in probability, and then using Theorem 96. \square

We have established a correspondence between the limiting distributions of IID sums, and processes with stationary independent increments. There are many possible applications of this correspondence; one is to reduce problems about limits of IID sums to problems about certain sorts of Markov process, and vice versa. Another is to relate discrete and continuous time processes. For $0 \leq t \leq 1$, set

$$X_t^{(n)} = \frac{1}{\sqrt{n}} \sum_{i=0}^{\lfloor nt \rfloor} Y_i \quad (11.36)$$

where $Y_0 = 0$ but otherwise the Y_i are IID with mean 0 and variance 1. Then each $X^{(n)}$ is a Markov process in discrete time, with stationary, independent increments, and its time-evolution operator is

$$K_n f(x) = \mathbf{E} \left[f \left(x + \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} Y_i \right) \right] \quad (11.37)$$

As n grows, the normalized sum approaches $\sqrt{t}Z$, with Z a standard Gaussian random variable. (Where does the \sqrt{t} come from?) Since the time-evolution operators determine the finite-dimensional distributions of a Markov process (Theorem 133), this suggests that in some sense $X^{(n)}$ should be converging in distribution on the Wiener process W . This is in fact true, but we will have to go deeper into the structure of the operators concerned before we can make it precise, as a first form of the *functional central limit theorem*. Just as the infinitely divisible distributions are the limits of IID sums, processes with stationary and independent increments are the limits of the corresponding random walks.

There is a yet further sense which the Gaussian distribution is a special kind of limiting distribution, which is reflected in another curious property of the Wiener process. Gaussian random variables are not only infinitely divisible, but they can be additively decomposed into more Gaussians. Distributions which can be infinitely divided into others of the same kind will be called “stable” (under convolution or averaging). “Of the same kind” isn’t very mathematical; here is a more precise expression of the idea.

Definition 147 (Self-similarity) *A process is self-similar if, for all t , there exists a measurable $h(t)$, the scaling function, such that $h(t)X(1) \stackrel{d}{=} X(t)$.*

Definition 148 (Stable Distributions) A distribution ν is stable if, for any Lévy process X where $X(1) \sim \nu$, X is self-similar.

Theorem 149 (Scaling in Stable Lévy Processes) In a self-similar Lévy process, the scaling function $h(t)$ must take the form t^α for some real α , the index of stability.

PROOF: Exercise 11.7. \square

It turns out that there are analogs to the functional central limit theorem for a broad range of self-similar processes, especially ones with stable increment distributions; Embrechts and Maejima (2002) is a good short introduction to this topic, and some of its statistical implications.

11.4 Exercises

Exercise 11.1 (Wiener Process with Constant Drift) Consider a process $X(0)$ which, like the Wiener process, has $X(0) = 0$ and independent increments, but where $X(t_2) - X(t_1) \sim \mathcal{N}(a(t_2 - t_1), \sigma^2(t_2 - t_1))$. a is called the drift rate and σ^2 the diffusion constant. Show that $X(t)$ is a Markov process, following the argument for the standard Wiener process ($a = 0$, $\sigma^2 = 1$) above. Do such processes have continuous modifications for all (finite) choices of a and σ^2 ? If so, prove it; if not, give at least one counter-example.

Exercise 11.2 (Perron-Frobenius Operators) Verify that P defined in the section on the logistic map above is a Markov operator.

Exercise 11.3 (Continuity of the Wiener Process) Show that the Wiener process has continuous sample paths, using its finite-dimensional distributions and Theorem 97.

Exercise 11.4 (Independent Increments with Respect to a Filtration) Let X be adapted to a filtration $\{\mathcal{F}_t\}$. Then X has independent increments with respect to $\{\mathcal{F}_t\}$ when $X_t - X_s$ is independent of \mathcal{F}_s for all $s \leq t$. Show that X is Markovian with respect to $\{\mathcal{F}_t\}$, by analogy with Theorem 141.

Exercise 11.5 (Poisson Counting Process) Consider the Poisson counting process N of Example 140.

1. Prove that N is continuous in probability.
2. Prove that N has a cadlag modification. Hint: Theorem 96 is one way, but there are others.

Exercise 11.6 (Poisson Distribution is Infinitely Divisible) *Prove, directly from the definition, that every Poisson distribution is infinitely divisible. What is the corresponding stationary, independent-increment process?*

Exercise 11.7 (Self-Similarity in Lévy Processes) *Prove Theorem 149.*

Exercise 11.8 (Gaussian Stability) *Find the index of stability a standard Gaussian distribution.*

Exercise 11.9 (Poissonian Stability?) *Is the standard Poisson distribution stable? If so, prove it, and find the index of stability. If not, prove that it is not.*

Exercise 11.10 (Lamperti Transformation) *Prove Lamperti's Theorem: If Y is a strictly stationary process and $\alpha > 0$, then $X(t) = t^\alpha Y(\log t)$ is self-similar, and if X is self-similar with index of stability α , then $Y(t) = e^{-\alpha t} X(e^t)$ is strictly stationary. These operations are called the Lamperti transformations, after Lamperti (1962).*

Chapter 12

Generators of Markov Processes

This lecture is concerned with the infinitesimal generator of a Markov process, and the sense in which we are able to write the evolution operators of a homogeneous Markov process as exponentials of their generator.

Take our favorite continuous-time homogeneous Markov process, and consider its semi-group of time-evolution operators K_t . They obey the relationship $K_{t+s} = K_t K_s$. That is, composition of the operators corresponds to addition of their parameters, and vice versa. This is reminiscent of the exponential functions on the reals, where, for any $k \in \mathbb{R}$, $k^{(t+s)} = k^t k^s$. In the discrete-parameter case, in fact, $K_t = (K_1)^t$, where integer powers of operators are defined in the obvious way, through iterated composition, i.e., $K^2 f = K \circ (K f)$. It would be nice if we could extend this analogy to continuous-parameter Markov processes. One approach which suggests itself is to notice that, for any k , there's another real number g such that $k^t = e^{tg}$, and that e^{tg} has a nice representation involving integer powers of g :

$$e^{tg} = \sum_{i=0}^{\infty} \frac{(tg)^i}{i!}$$

The strategy this suggests is to look for some other operator G such that

$$K_t = e^{tG} \equiv \sum_{i=0}^{\infty} \frac{t^i G^i}{i!}$$

Such an operator G is called the *generator* of the process, and the purpose of this chapter is to work out the conditions under which this analogy can be carried through.

In the exponential function case, we notice that g can be extracted by taking the derivative at zero: $\frac{d}{dt} e^{tg} \big|_{t=0} = g$. This suggests the following definition.

Definition 150 (Infinitesimal Generator) Let K_t be a continuous-parameter semi-group of linear operators on L , where L is a normed linear vector space, e.g., L_p for some $1 \leq p \leq \infty$. Say that a function $f \in L$ belongs to $\text{Dom}(G)$ if the limit

$$\lim_{h \downarrow 0} \frac{K_h f - K_0 f}{h} \equiv Gf \quad (12.1)$$

exists in an L -norm sense (Definition 151). The operator G defined through Eq. 12.1 is called the infinitesimal generator of the semi-group K_t .

Definition 151 (Limit in the L -norm sense) Let L be a normed vector space. We say that a sequence of elements $f_n \in L$ has a limit f in L when

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0 \quad (12.2)$$

This definition extends in the natural way to continuously-indexed collections of elements.

Lemma 152 (Generators are Linear) For every semi-group of homogeneous transition operators K_t , the generator G is a linear operator.

PROOF: Exercise 12.1. \square

Lemma 153 (Invariant Distributions of a Semi-group Belong to the Null Space of Its Generator) If ν is an invariant distribution of a semi-group of Markov operators M_t with generator G , then $G\nu = 0$.

PROOF: Since ν is invariant, $M_t\nu = \nu$ for all t (Theorem 136). Hence $M_t\nu - \nu = M_t\nu - M_0\nu = 0$, so, applying the definition of the generator (Eq. 12.1), $G\nu = 0$. \square

Remark: The converse assertion, that $G\nu = 0$ implies ν is invariant under M_t , requires extra conditions.

There is a conjugate version of this lemma.

Lemma 154 (Invariant Distributions and the Generator of the Time-Evolution Semigroup) If ν is an invariant distribution of a Markov process, and the time-evolution semi-group K_t is generated by G , then, $\forall f \in \text{Dom}(G)$, $\nu Gf = 0$.

PROOF: Since ν is invariant, $\nu K_t = \nu$ for all t , hence $\nu K_t f = \nu f$ for all $t \geq 0$ and all f . Since taking expectations with respect to a measure is a linear operator, $\nu(K_t f - f) = 0$, and obviously then (Eq. 12.1) $\nu Gf = 0$. \square

Remark: Once again, $\nu Gf = 0$ for all f is not enough, in itself, to show that ν is an invariant measure.

You will usually see the definition of the generator written with f instead of $K_0 f$, but I chose this way of doing it to emphasize that G is, basically,

the derivative at zero, that $G = dK/dt|_{t=0}$. Recall, from calculus, that the exponential function can k^t be defined by the fact that $\frac{d}{dt}k^t \propto k^t$ (and e can be defined as the k such that the constant of proportionality is 1). As part of our program, we will want to extend this differential point of view. The next lemma builds towards it, by showing that if $f \in \text{Dom}(G)$, then $K_t f$ is too.

Lemma 155 (Operators in a Semi-group Commute with Its Generator) *If G is the generator of the semi-group K_t , and f is in the domain of G , then K_t and G commute, for all t :*

$$K_t G f = \lim_{t' \rightarrow t} \frac{K_{t'} f - K_t f}{t' - t} \quad (12.3)$$

$$= G K_t f \quad (12.4)$$

PROOF: Exercise 12.2. \square

Definition 156 (Time Derivative in Function Space) *For every $t \in T$, let $u(t, x)$ be a function in L . When the limit*

$$u'(t_0, x) = \lim_{t \rightarrow t_0} \frac{u(t, x) - u(t_0, x)}{t - t_0} \quad (12.5)$$

exists in the L sense, then we say that $u'(t_0)$ is the time derivative or strong derivative of $u(t)$ at t_0 .

Lemma 157 (Generators and Derivatives at Zero) *Let K_t be a homogeneous semi-group of operators with generator G . Let $u(t) = K_t f$ for some $f \in \text{Dom}(G)$. Then $u(t)$ is differentiable at $t = 0$, and its derivative there is Gf .*

PROOF: Obvious from the definitions. \square

Theorem 158 (The Derivative of a Function Evolved by a Semi-Group) *Let K_t be a homogeneous semi-group of operators with generator G , and let $u(t, x) = (K_t f)(x)$, for fixed $f \in \text{Dom}(G)$. Then $u'(t)$ exists for all t , and is equal to $G u(t)$.*

PROOF: Since $f \in \text{Dom}(G)$, $K_t G f$ exists, but then, by Lemma 155, $K_t G f = G K_t f = G u(t)$, so $u(t) \in \text{Dom}(G)$ for all t . Now let's consider the time derivative of $u(t)$ at some arbitrary t_0 , working from above:

$$\frac{(u(t) - u(t_0))}{t - t_0} = \frac{K_{t-t_0} u(t_0) - u(t_0)}{t - t_0} \quad (12.6)$$

$$= \frac{K_h u(t_0) - u(t_0)}{h} \quad (12.7)$$

Taking the limit as $h \downarrow 0$, we get that $u'(t_0) = G u(t_0)$, which exists, because $u(t_0) \in \text{Dom}(G)$. \square

Corollary 159 (Initial Value Problems in Function Space) $u(t) = K_t f$, $f \in \text{Dom}(G)$, solves the initial value problem $u(0) = f$, $u'(t) = Gu(t)$.

PROOF: Immediate from the theorem. \square

Remark: Such initial value problems are sometimes called *Cauchy problems*, especially when G takes the form of a differential operator.

Corollary 160 (Derivative of Conditional Expectations of a Markov Process) Let X be a homogeneous Markov process whose time-evolution operators are K_t , with generator G . If $f \in \text{Dom}(G)$, then its condition expectation $\mathbf{E}[f(X_s)|X_t]$ has strong derivative $Gu(t)$.

PROOF: An immediate application of the theorem.

We are now almost ready to state the sense in which K_t is the result of exponentiating G . This is given by the remarkable Hille-Yosida theorem, which in turn involves a family of operators related to the time-evolution operators, the “resolvents”, again built by analogy to the exponential functions, and to Laplace transforms.

Recall that the *Laplace transform* of a function $f : \mathbb{R} \mapsto \mathbb{R}$ is another function, \tilde{f} , defined by

$$\tilde{f}(\lambda) \equiv \int_0^\infty e^{-\lambda t} f(t) dt$$

for positive λ . Laplace transforms arise in many contexts (linear systems theory, integral equations, etc.), one of which is *moment-generating functions* in basic probability theory. If Y is a real-valued random variable with probability law P , then the moment-generating function is

$$M_Y(\lambda) \equiv \mathbf{E}[e^{\lambda Y}] = \int e^{\lambda y} dP = \int e^{\lambda y} p(y) dy$$

when the density in the last expression exists. You may recall, from this context, that the distributions of well-behaved random variables are completely specified by their moment-generating functions; this is actually a special case of a more general result about when functions are uniquely described by their Laplace transforms, i.e., when f can be expressed uniquely in terms of \tilde{f} . This is important to us, because it turns out that the Laplace transform, so to speak, of a semi-group of operators is better-behaved than the semi-group itself, and we’ll want to say when we can use the Laplace transform to recover the semi-group.

The analogy with exponential functions, again, is a key. Notice that, for any positive constant λ ,

$$\int_{t=0}^\infty e^{-\lambda t} e^{tg} dt = \frac{1}{\lambda - g} \quad (12.8)$$

from which we could recover g , as that value of λ for which the Laplace transform is singular. In our analogy, we will want to take the Laplace transform of the semi-group. Just as $\tilde{f}(\lambda)$ is another real number, the Laplace transform of a semi-group of operators is going to be another operator. We will want that to be the inverse operator to $\lambda - G$.

Definition 161 (Resolvents) *Given a continuous-parameter homogeneous semi-group K_t , for each $\lambda > 0$, the resolvent operator or resolvent R_λ is the Laplace transform of K_t : for every $f \in L$,*

$$(R_\lambda f)(x) \equiv \int_{t=0}^{\infty} e^{-\lambda t} (K_t f)(x) dt \quad (12.9)$$

Remark 1: Think of K_t as a function from the real numbers to the linear operators on L . Symbolically, its Laplace transform would be $\int_0^\infty e^{-\lambda t} K_t dt$. The equation above just fills in the content of that symbolic expression.

Remark 2: The name “resolvent”, like some of the other ideas in terminology of operator semi-groups, comes from the theory of integral equations; invariant densities (when they exist) are solutions of homogeneous linear Fredholm integral equations of the second kind. Rather than pursue this connection, or even explain what that phrase means, I will refer you to the classic treatment of integral equations by Courant and Hilbert (1953, ch. 3), which everyone else seems to follow *very closely*.

Remark 3: When the function f is a value (loss, benefit, utility, ...) function, $(K_t f)(x)$ is the expected value at time t when starting the process in state x . $(R_\lambda f)(x)$ can be thought of as the *net present expected value* when starting at x and applying a discount rate λ .

Definition 162 (Yosida Approximation of Operators) *The Yosida approximation to a semi-group K_t with generator G is given by*

$$K_t^{(\lambda)} \equiv e^{tG^{(\lambda)}} \quad (12.10)$$

$$G^{(\lambda)} \equiv \lambda(\lambda R_\lambda - I) = \lambda G R_\lambda \quad (12.11)$$

The domain of $G^{(\lambda)}$ is all of L , not just $\text{Dom}(G)$.

Theorem 163 (Hille-Yosida Theorem) *Let G be a linear operator on some linear subspace \mathcal{D} of L . G is the generator of a continuous semi-group of contractions K_t if and only if*

1. \mathcal{D} is dense in L ;
2. For every $f \in L$ and $\lambda > 0$, there exists a unique $g \in \mathcal{D}$ such that $\lambda g - Gg = f$;
3. For every $g \in \mathcal{D}$ and positive λ , $\|\lambda g - Gg\| \geq \lambda \|g\|$.

Under these conditions, the resolvents of K_t are given by $R_\lambda = (\lambda I - G)^{-1}$, and K_t is the limit of the Yosida approximations as $\lambda \rightarrow \infty$:

$$K_t f = \lim_{\lambda \rightarrow \infty} K_t^\lambda f, \quad \forall f \in L \quad (12.12)$$

PROOF: See Kallenberg, Theorem 19.11. \square

Remark 1: Other good sources on the Hille-Yosida theorem include Ethier and Kurtz (1986, sec. 1.2), and of course Hille's own book (Hille, 1948, ch. XII). I have not read Yosida's original work.

Remark 2: The point of condition (1) in the theorem is that for any $f \in L$, we can choose $f_n \in \mathcal{D}$ such that $f_n \rightarrow f$. Then, even if Gf is not directly defined, Gf_n exists for each n . Because G is linear and therefore continuous, Gf_n goes to a limit, which we can choose to write as Gf . Similarly for G^2 , G^3 , etc. Thus we can define $e^{tG}f$ as

$$\lim_{n \rightarrow \infty} \sum_{j=0}^{\infty} \frac{t^j G^j f_n}{j!}$$

assuming all of the sums inside the limit converge, which is yet to be shown.

Remark 3: The point of condition (2) in the theorem is that, when it holds, $(\lambda I - G)^{-1}$ is well-defined, i.e., there is an inverse to the operator $\lambda I - G$. This is, recall, what we would like the resolvent to be, if the analogy with exponential functions is to hold good.

Remark 4: If we start from the semi-group K_t and obtain its generator G , the theorem tells us that it satisfies properties (1)–(3). If we start from an operator G and see that it satisfies (1)–(3), the theorem tells us that it generates some semi-group. It might seem, however, that it doesn't tell us how to construct that semi-group, since the Yosida approximation involves the resolvent R_λ , which is defined in terms of the semi-group, creating an air of circularity. In fact, when we start from G , we decree R_λ to be $(\lambda I - G)^{-1}$. The Yosida approximations then is defined in terms of G and λ alone.

Corollary 164 (Stochastic Approximation of Initial Value Problems)

Let $u(0) = f$, $u'(t) = Gu(t)$ be an initial value problem in L . Then a stochastic approximation to $u(t)$ can be founded by taking

$$\hat{u}(t) = \frac{1}{n} \sum_{i=1}^n f(X_i(t)) \quad (12.13)$$

where the X_i are independent copies of the Markov process corresponding to the semi-group K_t generated by G , with initial condition $X_i(0) = x$.

PROOF: Combine corollaries. \square

12.1 Exercises

Exercise 12.1 (Generators are Linear) Prove Lemma 152.

Exercise 12.2 (Semi-Groups Commute with Their Generators) Prove Lemma 155.

1. Prove Equation 12.3, restricted to $t' \downarrow t$ instead of $t' \rightarrow t$. Hint: Write T_t in terms of an integral over the corresponding transition kernel, and find a reason to exchange integration and limits.
2. Show that the limit as $t' \uparrow t$ also exists, and is equal to the limit from above. Hint: Re-write the quotient inside the limit so it only involves positive time-differences.
3. Prove Equation 12.4.

Exercise 12.3 (Generator of the Poisson Counting Process) Find the generator of the time-evolution operators of the Poisson counting process (Example 140).

Chapter 13

The Strong Markov Property and Martingale Problems

Section 13.1 introduces the strong Markov property — independence of the past and future conditional on the state at *random* (optional) times. It includes an example of a Markov process which is not *strongly* Markovian.

Section 13.2 describes “the martingale problem for Markov processes”, explains why it would be nice to solve the martingale problem, and how solutions are strong Markov processes.

13.1 The Strong Markov Property

A process is Markovian, with respect to a filtration $\{\mathcal{F}_t\}$, if for any *fixed* time t , the future of the process is independent of \mathcal{F}_t given X_t . This is not necessarily the case for a *random* time τ , because there could be subtle linkages between the random time and the evolution of the process. If these can be ruled out, we have a *strong* Markov process.

Definition 165 (Strongly Markovian at a Random Time) *Let X be a Markov process with respect to a filtration $\{\mathcal{F}_t\}$, with transition kernels $\mu_{t,s}$ and evolution operators $K_{t,s}$. Let τ be an $\{\mathcal{F}_t\}$ -optional time which is almost surely finite. Then X is strongly Markovian at τ when either of the two following (equivalent) conditions hold*

$$\mathbb{P}(X_{\tau+t} \in B | \mathcal{F}_\tau) = \mu_{\tau, \tau+t}(X_\tau, B) \quad (13.1)$$

$$\mathbf{E}[f(X_{\tau+t}) | \mathcal{F}_\tau] = (K_{\tau, \tau+t}f)(X_\tau) \quad (13.2)$$

for all $t \geq 0$, $B \in \mathcal{X}$ and bounded measurable functions f .

Definition 166 (Strong Markov Property) *If X is Markovian with respect to $\{\mathcal{F}_t\}$, and strongly Markovian at every $\{\mathcal{F}_t\}$ -optional time which is almost surely finite, then it is a strong Markov process with respect to $\{\mathcal{F}_t\}$.*

If the index set T is discrete, then the strong Markov property is implied by the ordinary Markov property (Definition 102). If time is continuous, this is not necessarily the case. It is generally true that, if X is Markov and τ takes on only *countably* many values, X is strongly Markov at τ (Exercise 13.1). In continuous time, however, the Markov property does not imply the strong Markov property.

Example 167 (A Markov Process Which Is Not Strongly Markovian) *(After Fristedt and Gray (1997, pp. 626–627).) We will construct an \mathbb{R}^2 -valued Markov process on $[0, \infty)$ which is not strongly Markovian. Begin by defining the following map from \mathbb{R} to \mathbb{R}^2 :*

$$f(w) = \begin{cases} (w, 0) & w \leq 0 \\ (\sin w, 1 - \cos w) & 0 < w < 2\pi \\ (w - 2\pi, 0) & w \geq 2\pi \end{cases} \quad (13.3)$$

When w is less than zero or above 2π , $f(w)$ moves along the x axis of the plane; in between, it moves along a circle of radius 1, centered at $(0, 1)$, which it enters and leaves at the origin. Notice that f is invertible everywhere except at the origin, which is ambiguous between $w = 0$ and $w = 2\pi$.

Let $X(t) = f(W(t) + \pi)$, where $W(t)$ is a standard Wiener process. At all t , $\mathbb{P}(W(t) + \pi = 0) = \mathbb{P}(W(t) + \pi = 2\pi) = 0$, so, with probability 1, $X(t)$ can be inverted to get $W(t)$. Since $W(t)$ is a Markov process, it follows that $\mathbb{P}(X(t+h) \in B | X(t) = x) = \mathbb{P}(X(t+h) \in B | \mathcal{F}_t^X)$ almost surely, i.e., X is Markov. Now consider $\tau = \inf_t X(t) = (0, 0)$, the hitting time of the origin. This is clearly an F^X -optional time, and equally clearly almost surely finite, because, with probability 1, $W(t)$ will leave the interval $(-\pi, \pi)$ within a finite time. But, equally clearly, the future behavior of X will be very different if it hits the origin because $W = \pi$ or because $W = -\pi$, which cannot be determined just from X . Hence, there is at least one optional time at which X is not strongly Markovian, so X is not a strong Markov process.

Since we often want to condition on the state of the process at random times, we would like to find conditions under which a process is strongly Markovian for all optional times.

13.2 Martingale Problems

One approach to getting strong Markov processes is through martingales, and more specifically through what is known as the martingale problem.

Notice the following consequence of Theorem 158:

$$K_t f(x) - f(x) = \int_0^t K_s G f(x) ds \quad (13.4)$$

for any $t \geq 0$ and $f \in \text{Dom}(G)$. The relationship between $K_t f$ and the conditional expectation of f suggests the following definition.

Definition 168 (Martingale Problem) *Let Ξ be a Polish space, \mathcal{D} a class of bounded, continuous, real-valued functions on Ξ , and G an operator from \mathcal{D} to bounded, measurable functions on Ξ . A Ξ -valued stochastic process on \mathbb{R}^+ is a solution to the martingale problem for G and \mathcal{D} if, for all $f \in \mathcal{D}$,*

$$f(X_t) - \int_0^t Gf(X_s) ds \quad (13.5)$$

is a martingale with respect to $\{\mathcal{F}_t^X\}$, the natural filtration of X .

Proposition 169 (Cadlag Nature of Functions in Martingale Problems) *Suppose X is a cadlag solution to the martingale problem for G, \mathcal{D} . Then for any $f \in \mathcal{D}$, the stochastic process given by Eq. 13.5 is also cadlag.*

PROOF: Follows from the assumption that f is continuous. \square

Lemma 170 (Alternate Formulation of Martingale Problem) *X is a solution to the martingale problem for G, \mathcal{D} if and only if, for all $t, s \geq 0$,*

$$\mathbf{E}[f(X_{t+s})|\mathcal{F}_t^X] - \mathbf{E}\left[\int_t^{t+s} Gf(X_u) du | \mathcal{F}_t^X\right] = f(X_t) \quad (13.6)$$

PROOF: Take the definition of a martingale and re-arrange the terms in Eq. 13.5. \square

Martingale problems are important because of the two following theorems (which can both be refined considerably).

Theorem 171 (Markov Processes Solve Martingale Problems) *Let X be a homogeneous Markov process with generator G and cadlag sample paths, and let \mathcal{D} be the continuous functions in $\text{Dom}(G)$. Then X solves the martingale problem for G, \mathcal{D} .*

PROOF: Exercise 13.2. \square

Theorem 172 (Solutions to the Martingale Problem are Strongly Markovian) *Suppose that for each $x \in \Xi$, there is a unique cadlag solution to the martingale problem for G, \mathcal{D} such that $X_0 = x$. Then the collection of these solutions is a homogeneous strong Markov family X , and the generator is equal to G on \mathcal{D} .*

PROOF: Exercise 13.3. \square

The main use of Theorem 171 is that it lets us prove convergence of some functions of Markov processes, by showing that they can be cast into the form of Eq. 13.5, and then applying the martingale convergence devices. The other use is in conjunction with Theorem 172. We will often want to show that a *sequence*

of Markov processes converges on a limit which is, itself, a Markov process. One approach is to show that the terms in the sequence solve martingale problems (via Theorem 171), argue that then the limiting process does too, and finally invoke Theorem 172 to argue that the limiting process must itself be strongly Markovian. This is often *much* easier than showing directly that the limiting process is Markovian, much less strongly Markovian. Theorem 172 itself is often a convenient way of showing that the strong Markov property holds.

13.3 Exercises

Exercise 13.1 (Strongly Markov at Discrete Times) *Let X be a homogeneous Markov process with respect to a filtration $\{\mathcal{F}_t\}$ and τ be an $\{\mathcal{F}_t\}$ -optional time. Prove that if $\mathbb{P}(\tau < \infty) = 1$, and τ takes on only countably many values, then X is strongly Markovian at τ . (Note: the requirement that X be homogeneous can be lifted, but requires some more technical machinery I want to avoid.)*

Exercise 13.2 (Markovian Solutions of the Martingale Problem) *Prove Theorem 171. Hints: Use Lemma 170, bounded convergence, and Theorem 158.*

Exercise 13.3 (Martingale Solutions are Strongly Markovian) *Prove Theorem 172. Hint: use the Optional Sampling Theorem (from 36-752, or from chapter 7 of Kallenberg).*

Chapter 14

Feller Processes

Section 14.1 makes explicit the idea that the transition kernels of a Markov process induce a kernel over sample paths, mostly to fix notation for later use.

Section 14.2 defines Feller processes, which link the cadlag and strong Markov properties.

14.1 Markov Families

We have been fairly cavalier about the idea of a Markov process having a particular initial state or initial distribution, basically relying on our familiarity with these ideas from elementary courses on stochastic processes. For future purposes, however, it is helpful to bring this notions formally within our general framework, and to fix some notation.

Definition 173 (Initial Distribution, Initial State) *Let Ξ be a Borel space with σ -field \mathcal{X} , T be a one-sided index set, and $\mu_{t,s}$ be a collection of Markovian transition kernels on Ξ . Then the Markov process with initial distribution ν , X_ν , is the Markov process whose finite-dimensional distributions are given by the action of $\mu_{t,s}$ on ν . That is, for all $0 \leq t_1 \leq t_2 \leq \dots \leq t_n$,*

$$X_\nu(0), X_\nu(t_1), X_\nu(t_2), \dots, X_\nu(t_n) \sim \nu \otimes \mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \dots \otimes \mu_{t_{n-1},t_n} \quad (14.1)$$

If $\nu = \delta(x - a)$, the delta distribution at a , then we write X_a and call it the Markov process with initial state a .

The existence of processes with given initial distributions and initial states is a trivial consequence of Theorem 106, our general existence result for Markov processes.

Lemma 174 (Kernel from Initial States to Paths) *For every initial state x , there is a probability distribution P_x on Ξ^T, \mathcal{X}^T . The function $P_x(A) : \Xi \times \mathcal{X}^T \rightarrow [0, 1]$ is a probability kernel.*

PROOF: The initial state fixes all the finite-dimensional distributions, so the existence of the probability distribution follows from Theorem 23. The fact that $P_x(A)$ is a kernel is a straightforward application of the definition of kernels (Definition 30). \square

Definition 175 (Markov Family) *The Markov family corresponding to a given set of transition kernels $\mu_{t,s}$ is the collection of all P_x .*

That is, rather than thinking of a different stochastic process for each initial state, we can simply think of different distributions over the path space Ξ^T . This suggests the following definition.

Definition 176 (Mixtures of Path Distributions (Mixed States)) *For a given initial distribution ν on Ξ , we define a distribution on the paths in a Markov family as, $\forall A \in \mathcal{X}^T$,*

$$P_\nu(A) \equiv \int_{\Xi} P_x(A) \nu(dx) \quad (14.2)$$

In physical contexts, we sometimes refer to distributions ν as *mixed states*, as opposed to the *pure states* x , because the path-space distributions induced by the former are mixtures of the distributions induced by the latter. You should check that the distribution over paths given by a Markov process with initial distribution ν , according to Definition 173, agrees with that given by Definition 176.

14.2 Feller Processes

Working in the early 1950s, Feller showed that, by imposing very reasonable conditions on the semi-group of evolution operators corresponding to a homogeneous Markov process, one could obtain very powerful results about the near-continuity of sample paths (namely, the existence of cadlag versions), about the strong Markov property, etc. Ever since, processes with such nice semi-groups have been known as *Feller processes*, or sometimes as *Feller-Dynkin processes*, in recognition of Dynkin's work in extending Feller's original approach. (This is also recognized in the name of Dynkin's Formula, Theorem 195 below.) Unfortunately, to first order there are as many definitions of a Feller semi-group as there are books on Markov processes. I am going to try to follow Kallenberg as closely as possible, because his version is pretty clearly motivated, and you've already got it.

Definition 177 (Feller Process) *A continuous-time homogeneous Markov family X is a Feller process when, for all $x \in \Xi$,*

$$\forall t, y \rightarrow x \Rightarrow X_y(t) \xrightarrow{d} X_x(t) \quad (14.3)$$

$$t \rightarrow 0 \Rightarrow X_x(t) \xrightarrow{P} x \quad (14.4)$$

Remark 1: The first property basically says that the dynamics are a smooth function of the initial state. Recall¹ that if we have an ordinary differential equation, $dx/dt = f(x)$, and the function f is reasonably well-behaved, the existence and uniqueness theorem tells us that there is a function $x(t, x_0)$ satisfying the equation, and such that $x(0, x_0) = x_0$. Moreover, $x(t, x_0)$ is a continuous function of x_0 for all t . The first Feller property is one counterpart of this for stochastic processes. This is a very natural assumption to make in physical or social modeling, that very similar initial conditions should lead to very similar developments.

Remark 2: The second property imposes a certain amount of smoothness on the trajectories themselves, and not just how they depend on the initial state. It's a pretty well-attested fact about the world that teleportation does not take place, and that its state changes in a reasonably continuous manner (“*natura non facit saltum*”, as they used to say). However, the second Feller property does *not* require that the sample paths actually be continuous. We will see below that they are, in general, merely cadlag. So a certain limited amount of teleportation, or *salti*, is allowed after all. We do not think this actually happens, but it is a convenience when using a discrete set of states to approximate a continuous variable.

In developing the theory of Feller processes, we will work mostly with the time-evolution operators, acting on observables, rather than the Markov operators, acting on distributions. This is traditional in this part of the theory, as it seems to let us get away with less technical machinery, in effect because the norm $\sup_x |f(x)|$ is stronger than the norm $\int |f(x)|dx$. Of course, because the two kinds of operators are adjoint, you *can* work out everything for the Markov operators, if you want to.

As usual, we warm up with some definitions. The first two apply to operators on any normed linear space L , which norm we generically write as $\|\cdot\|$. The second two apply specifically when L is a space of real-valued functions on some Ξ , such as L_p , p from 1 to ∞ inclusive.

Definition 178 (Contraction Operator) *An operator A is an L -contraction when $\|Af\| \leq \|f\|$.*

Definition 179 (Strongly Continuous Semigroup) *A semigroup of operators A_t is strongly continuous in the L sense on a set of functions D when, $\forall f \in D$*

$$\lim_{t \rightarrow 0} \|A_t f - f\| = 0 \quad (14.5)$$

Definition 180 (Positive Operator) *An operator A on a function space L is positive when $f \geq 0$ a.e. implies $Af \geq 0$ a.e.*

¹Or read Arnol'd (1973), if memory fails you.

Definition 181 (Conservative Operator) *An operator A is conservative when $A1_\Xi = 1_\Xi$.*

In these terms, our earlier Markov operators are linear, positive, conservative contractions, either on $L_1(\mu)$ (for densities) or $\mathcal{M}(\Xi)$ (for measures).

Lemma 182 (Continuous semi-groups produce continuous paths in function space) *If A_t is a strongly continuous semigroup of linear contractions on L , then, for each $f \in L$, $A_t f$ is a continuous function of t .*

PROOF: Continuity here means that $\lim_{t' \rightarrow t} \|A_{t'} f - A_t f\| = 0$ — we are using the norm $\|\cdot\|$ to define our metric in function space. Consider first the limit from above:

$$\|A_{t+h} f - A_t f\| = \|A_t(A_h f - f)\| \quad (14.6)$$

$$\leq \|O_h f - f\| \quad (14.7)$$

since the operators are contractions. Because they are strongly continuous, $\|A_h f - f\|$ can be made smaller than any $\epsilon > 0$ by taking h sufficiently small. Hence $\lim_{h \downarrow 0} A_{t+h} f$ exists and is $A_t f$. Similarly, for the limit from below,

$$\|A_{t-h} f - A_t f\| = \|A_t f - A_{t-h} f\| \quad (14.8)$$

$$= \|A_{t-h}(A_h f - f)\| \quad (14.9)$$

$$\leq \|A_h f - f\| \quad (14.10)$$

using the contraction property again. So $\lim_{h \downarrow 0} A_{t-h} f = A_t f$, also, and we can just say that $\lim_{t' \rightarrow t} A_{t'} f = A_t f$. \square

Remark: The result actually holds if we just assume strong continuity, without contraction, but the proof isn't so pretty; see Ethier and Kurtz (1986, ch. 1, corollary 1.2, p. 7).

There is one particular function space L we will find especially interesting.

Definition 183 (Continuous Functions Vanishing at Infinity) *Let Ξ be a locally compact and separable metric space. The class of functions C_0 will consist of functions $f : \Xi \rightarrow \mathbb{R}$ which are continuous and for which $\|x\| \rightarrow \infty$ implies $f(x) \rightarrow 0$. The norm on C_0 is $\sup_x |f(x)|$.*

Definition 184 (Feller Semigroup) *A semigroup of linear, positive, conservative contraction operators K_t is a Feller semigroup if, for every $f \in C_0$ and $x \in \Xi$, (Definition 183),*

$$K_t f \in C_0 \quad (14.11)$$

$$\lim_{t \rightarrow 0} K_t f(x) = f(x) \quad (14.12)$$

Remark: Some authors omit the requirement that K_t be conservative. Also, this is just the homogeneous case, and one can define inhomogeneous Feller semigroups. However, the homogeneous case will be plenty of work enough for us!

You can guess how Feller semi-groups relate to Feller processes.

Lemma 185 (The First Pair of Feller Properties) *Eq. 14.11 holds if and only if Eq. 14.3 does.*

PROOF: Exercise 14.2. \square

Lemma 186 (The Second Pair of Feller Properties) *Eq. 14.12 holds if and only if Eq. 14.4 does.*

PROOF: Exercise 14.3. \square

Theorem 187 (Feller Processes and Feller Semigroups) *A Markov process is a Feller process if and only if its evolution operators form a Feller semigroup.*

PROOF: Combine Lemmas 185 and 186. \square

Feller semigroups in continuous time have generators, as in Chapter 12. In fact, the generator is *especially* useful for Feller semigroups, as seen by this theorem.

Theorem 188 (Generator of a Feller Semigroup) *If K_t and H_t are Feller semigroups with generator G , then $K_t = H_t$.*

PROOF: Because Feller semigroups consist of contractions, the Hille-Yosida Theorem 163 applies, and, for every positive λ , the resolvent $R_\lambda = (\lambda I - G)^{-1}$. Hence, if K_t and H_t have the same generator, they have the same resolvent operators. But this means that, for every $f \in C_0$ and x , $K_t f(x)$ and $H_t f(x)$ have the same Laplace transforms. Since, by Eq. 14.12 $K_t f(x)$ and $H_t f(x)$ are both right-continuous, their Laplace transforms are unique, so $K_t f(x) = H_t f(x)$. \square

Theorem 189 (Feller Semigroups are Strongly Continuous) *Every Feller semigroup K_t with generator G is strongly continuous on $\text{Dom}(G)$.*

PROOF: From Corollary 159, we have, as seen in Chapter 13, for all $t \geq 0$,

$$K_t f - f = \int_0^t K_s G f ds \quad (14.13)$$

Clearly, the right-hand side goes to zero as $t \rightarrow 0$. \square

The two most important properties of Feller processes is that they are cadlag (or, rather, always have cadlag versions), and that they are strongly Markovian. First, let's look at the cadlag property. We need a result which I really should have put in Chapter 8.

Proposition 190 (Cadlag Modifications Implied by a Kind of Modulus of Continuity) *Let Ξ be a locally compact, separable metric space with metric ρ , and let X be a separable Ξ -valued stochastic process on T . For given $\epsilon, \delta > 0$, define $\alpha(\epsilon, \delta)$ to be*

$$\inf_{\Gamma \in \mathcal{F}_s^X: \mathbb{P}(\Gamma)=1} \sup_{s, t \in T: s \leq t \leq s+\delta} \mathbb{P}(\omega: \rho(X(s, \omega), X(t, \omega)) \geq \epsilon, \omega \in \Gamma | \mathcal{F}_s^X) \quad (14.14)$$

If, for all ϵ ,

$$\lim_{\delta \rightarrow 0} \alpha(\epsilon, \delta) = 0 \quad (14.15)$$

then X has a cadlag version.

PROOF: Combine Theorem 2 and Theorem 3 of Gikhman and Skorokhod (1965/1969, Chapter IV, Section 4). \square

Lemma 191 (Markov Processes Have Cadlag Versions When They Don't Move Too Fast (in Probability)) *Let X be a separable homogeneous Markov process. Define*

$$\alpha(\epsilon, \delta) = \sup_{t \in T: 0 \leq t \leq \delta; x \in \Xi} \mathbb{P}(\rho(X_x(t), x) \geq \epsilon) \quad (14.16)$$

If, for every $\epsilon > 0$,

$$\lim_{\delta \rightarrow 0} \alpha(\epsilon, \delta) = 0 \quad (14.17)$$

then X has a cadlag version.

PROOF: The α in this lemma is clearly the α in the preceding proposition (190), using the fact that X is Markovian with respect to its natural filtration (Theorem 112) and homogeneous. \square

Lemma 192 (Markov Processes Have Cadlag Versions If They Don't Move Too Fast (in Expectation)) *A separable homogeneous Markov process X has a cadlag version if*

$$\lim_{\delta \downarrow 0} \sup_{x \in \Xi, 0 \leq t \leq \delta} \mathbf{E}[\rho(X_x(t), x)] = 0 \quad (14.18)$$

PROOF: Start with the Markov inequality.

$$\forall x, t > 0, \epsilon > 0, \mathbb{P}(\rho(X_x(t), x) \geq \epsilon) \leq \frac{\mathbf{E}[\rho(X_x(t), x)]}{\epsilon} \quad (14.19)$$

$$\forall x, \delta > 0, \epsilon > 0, \sup_{0 \leq t \leq \delta} \mathbb{P}(\rho(X_x(t), x) \geq \epsilon) \leq \sup_{0 \leq t \leq \delta} \frac{\mathbf{E}[\rho(X_x(t), x)]}{\epsilon} \quad (14.20)$$

$$\forall \delta > 0, \epsilon > 0, \sup_{x, 0 \leq t \leq \delta} \mathbb{P}(\rho(X_x(\delta), x) \geq \epsilon) \leq \frac{1}{\epsilon} \sup_{x, 0 \leq t \leq \delta} \mathbf{E}[\rho(X_x(\delta), x)] \quad (14.21)$$

Taking the limit as $\delta \downarrow 0$, we have, for all $\epsilon > 0$,

$$\lim_{\delta \rightarrow 0} \alpha(\epsilon, \delta) \leq \frac{1}{\epsilon} \lim_{\delta \downarrow 0} \sup_{x, 0 \leq t \leq \delta} \mathbf{E}[\rho(X_x(\delta), x)] = 0 \quad (14.22)$$

So the preceding lemma (191) applies. \square

Theorem 193 (Feller Implies Cadlag) *Every Feller process X has a cadlag version.*

PROOF: First, by the usual arguments, we can get a separable version of X . Next, we want to show that the last lemma is satisfied. Notice that, because Ξ is compact, $\lim_x \rho(x_n, x) = 0$ if and only if $f_k(x_n) \rightarrow f_k(x)$, for all f_k in some countable dense subset of the continuous functions on the state space.² Since the Feller semigroup is strongly continuous on the domain of its generator (Theorem 189), and that domain is dense in C_0 by the Hille-Yosida Theorem (163), we can pick our f_k to be in this class. The strong continuity is with respect to the C_0 norm, so $\sup_x |K_t f(x) - K_s f(x)| = \sup_x |K_s(K_{t-s} f(x) - f(x))| \rightarrow 0$ as $t - s \rightarrow 0$, for every $f \in C_0$. But $\sup_x |K_t f(x) - K_s f(x)| = \sup_x \mathbf{E} [|f(X_x(t)) - f(X_x(s))|]$. So $\sup_x, 0 \leq t \leq \delta \mathbf{E} [|f(X_x(t)) - f(x)|] \rightarrow 0$ as $\delta \rightarrow 0$. Now Lemma 192 applies. \square

Remark: Kallenberg (Theorem 19.15, p. 379) gives a different proof, using the existence of cadlag paths for certain kinds of supermartingales, which he builds using the resolvent operator. This seems to be the favored approach among modern authors, but obscures, somewhat, the work which the Feller properties do in getting the conclusion.

Theorem 194 (Feller Processes are Strongly Markovian) *Any Feller process X is strongly Markovian with respect to \mathcal{F}^{X+} , the right-continuous version of its natural filtration.*

PROOF: The strong Markov property holds if and only if, for all bounded, continuous functions f , $t \geq 0$ and \mathcal{F}^{X+} -optional times τ ,

$$\mathbf{E} [f(X(\tau + t)) | \mathcal{F}_\tau^{X+}] = K_t f(X(\tau)) \quad (14.23)$$

We'll show this holds for arbitrary, fixed choices of f , t and τ . First, we discretize time, to exploit the fact that the Markov and strong Markov properties coincide for discrete parameter processes. For every $h > 0$, set

$$\tau_h \equiv \inf_u \{u \geq \tau : u = kh, k \in \mathbb{N}\} \quad (14.24)$$

Now τ_h is almost surely finite (because τ is), and $\tau_h \rightarrow \tau$ a.s. as $h \rightarrow 0$. We construct the discrete-parameter sequence $X_h(n) = X(nh)$, $n \in \mathbb{N}$. This is a Markov sequence with respect to the natural filtration, i.e., for every bounded continuous f and $m \in \mathbb{N}$,

$$\mathbf{E} [f(X_h(n + m)) | \mathcal{F}_n^X] = K_{mh} f(X_h(n)) \quad (14.25)$$

Since the Markov and strong Markov properties coincide for Markov sequences, we can now assert that

$$\mathbf{E} [f(X(\tau_h + mh)) | \mathcal{F}_{\tau_h}^X] = K_{mh} f(X(\tau_h)) \quad (14.26)$$

²Roughly speaking, if $f(x_n) \rightarrow f(x)$ for all continuous functions f , it should be obvious that there is no way to avoid having $x_n \rightarrow x$. Picking a countable dense subset of functions is still enough.

Since $\tau_h \geq \tau$, $\mathcal{F}_{\tau}^X \subseteq \mathcal{F}_{\tau_h}^X$. Now pick any set $B \in \mathcal{F}_{\tau}^{X+}$ and use smoothing:

$$\mathbf{E}[f(X(\tau_h + t))\mathbf{1}_B] = \mathbf{E}[K_t f(X(\tau_h))\mathbf{1}_B] \quad (14.27)$$

$$\mathbf{E}[f(X(\tau + t))\mathbf{1}_B] = \mathbf{E}[K_t f(X(\tau))\mathbf{1}_B] \quad (14.28)$$

where we let $h \downarrow 0$, and invoke the fact that $X(t)$ is right-continuous (Theorem 193) and $K_t f$ is continuous. Since this holds for arbitrary $B \in \mathcal{F}_{\tau}^{X+}$, and $K_t f(X(\tau))$ has to be \mathcal{F}_{τ}^{X+} -measurable, we have that

$$\mathbf{E}[f(X(\tau + t))|\mathcal{F}_{\tau}^{X+}] = K_t f(X(\tau)) \quad (14.29)$$

as required. \square

Here is a useful consequence of Feller property, related to the martingale-problem properties we saw last time.

Theorem 195 (Dynkin's Formula) *Let X be a Feller process with generator G . Let α and β be two almost-surely-finite \mathcal{F} -optional times, $\alpha \leq \beta$. Then, for every continuous $f \in \text{Dom}(G)$,*

$$\mathbf{E}[f(X(\beta)) - f(X(\alpha))] = \mathbf{E}\left[\int_{\alpha}^{\beta} Gf(X(t))dt\right] \quad (14.30)$$

PROOF: Exercise 14.4. \square

Remark: A large number of results very similar to Eq. 14.30 are *also* called “Dynkin’s formula”. For instance, Rogers and Williams (1994, ch. III, sec. 10, pp. 253–254) give that name to *three* different equations. Be careful about what people mean!

14.3 Exercises

Exercise 14.1 (Yet Another Interpretation of the Resolvents) *Consider again a homogeneous Markov process with transition kernel μ_t . Let τ be an exponentially-distributed random variable with rate λ , independent of X . Show that $\mathbf{E}[K_{\tau}f(x)] = \lambda R_{\lambda}f(x)$.*

Exercise 14.2 (The First Pair of Feller Properties) *Prove Lemma 185. Hint: you may use the fact that, for measures, $\nu_t \rightarrow \nu$ if and only if $\nu_t f \rightarrow \nu f$, for every bounded, continuous f .*

Exercise 14.3 (The Second Pair of Feller Properties) *Prove Lemma 186.*

Exercise 14.4 (Dynkin's Formula) *Prove Theorem 195.*

Exercise 14.5 (Lévy and Feller Processes) *Is every Lévy process a Feller process? If yes, prove it. If not, provide a counter-example, and try to find a sufficient condition for a Lévy process to be Feller.*

Chapter 15

Convergence of Feller Processes

This chapter looks at the convergence of sequences of Feller processes to a limiting process.

Section 15.1 lays some ground work concerning weak convergence of processes with cadlag sample paths.

Section 15.2 states and proves the central theorem about the convergence of sequences of Feller processes.

Section 15.3 examines a particularly important special case, the approximation of ordinary differential equations by pure-jump Markov processes.

15.1 Weak Convergence of Processes with Cadlag Paths (The Skorokhod Topology)

Recall that a sequence of random variables X_1, X_2, \dots converges in distribution on X , or weakly converges on X , $X_n \xrightarrow{d} X$, if and only if $\mathbf{E}[f(X_n)] \rightarrow \mathbf{E}[f(X)]$, for all bounded, continuous functions f . This is still true when X_n are random functions, i.e., stochastic processes, only now the relevant functions f are functionals of the sample paths.

Definition 196 (Convergence in Finite-Dimensional Distribution) *Random processes X_n on T converge in finite-dimensional distribution on X , $X_n \xrightarrow{fd} X$, when, $\forall J \in \text{Fin}(T)$, $X_n(J) \xrightarrow{d} X(J)$.*

Lemma 197 (Finite and Infinite Dimensional Distributional Convergence) *Convergence in finite-dimensional distribution is necessary but not sufficient for convergence in distribution.*

PROOF: Necessity is obvious: the coordinate projections π_t are continuous functionals of the sample path, so they must converge if the distributions converge. Insufficiency stems from the problem that, even if a sequence of X_n all have sample paths in some set U , the limiting process might not: recall our example (79) of the version of the Wiener process with unmeasurable suprema. \square

Definition 198 (The Space \mathbf{D}) By $\mathbf{D}(T, \Xi)$ we denote the space of all cadlag functions from T to Ξ . By default, \mathbf{D} will mean $\mathbf{D}(\mathbb{R}^+, \Xi)$.

\mathbf{D} admits of multiple topologies. For most purposes, the most convenient one is the *Skorokhod topology*, a.k.a. the J_1 topology or the *Skorokhod J_1 topology*, which makes $\mathbf{D}(\Xi)$ a complete separable metric space when Ξ is itself complete and separable. (See Appendix A2 of Kallenberg.) For our purposes, we need only the following notion and propositions.

Definition 199 (Modified Modulus of Continuity) The modified modulus of continuity of a function $x \in \mathbf{D}(T, \Xi)$ at time $t \in T$ and scale $h > 0$ is given by

$$w(x, t, h) \equiv \inf_{(I_k)} \max_k \sup_{r, s \in I_k} \rho(x(s), x(r)) \quad (15.1)$$

where the infimum is over partitions of $[0, t)$ into half-open intervals whose length is at least h (except possibly for the last one). Because x is cadlag, for fixed x and t , $w(x, t, h) \rightarrow 0$ as $h \rightarrow 0$.

Proposition 200 (Weak Convergence in $\mathbf{D}(\mathbb{R}^+, \Xi)$) Let Ξ be a complete, separable metric space. Then a sequence of random functions $X_1, X_2, \dots \in \mathbf{D}(\mathbb{R}^+, \Xi)$ converges in distribution to $X \in \mathbf{D}$ if and only if

- i The set $T_c = \{t \in T : X(t) = X(t^-)\}$ has a countable dense subset T_0 , and the finite-dimensional distributions of the X_n converge on those of X on T_0 .
- ii For every t ,

$$\lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbf{E}[w(X_n, t, h) \wedge 1] = 0 \quad (15.2)$$

PROOF: See Kallenberg, Theorem 16.10, pp. 313–314. \square

Proposition 201 (Sufficient Condition for Weak Convergence) The following three conditions are all equivalent, and all imply condition (ii) in Proposition 200.

- 1. For any sequence of a.s.-finite \mathcal{F}^{X_n} -optional times τ_n and positive constants $h_n \rightarrow 0$,

$$\rho(X_n(\tau_n), X_n(\tau_n + h_n)) \xrightarrow{P} 0 \quad (15.3)$$

2. For all $t > 0$, for all

$$\lim_{h \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{\sigma, \tau} \mathbf{E} [\rho(X_n(\sigma), X_n(\tau)) \wedge 1] = 0 \quad (15.4)$$

where σ and τ are \mathcal{F}^{X_n} -optional times $\sigma, \tau \leq t$, with $\sigma \leq \tau \leq \tau + h$.

3. For all $t > 0$,

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \sup_{\tau \leq t} \sup_{0 \leq h \leq \delta} \mathbf{E} [\rho(X_n(\tau), X_n(\tau + h)) \wedge 1] = 0 \quad (15.5)$$

where the supremum in τ runs over all \mathcal{F}^{X_n} -optional times $\leq t$.

PROOF: See Kallenberg, Theorem 16.11, pp. 314–315. \square

15.2 Convergence of Feller Processes

We need some technical notions about generators.

Definition 202 (Closed and Closable Generators, Closures) *A linear operator A on a Banach space \mathcal{B} is closed if its graph — $\{f, g \in \mathcal{B}^2 : f \in \text{Dom}(A), g = Af\}$ — is a closed set. An operator is closable if the closure of its graph is a function (and not just a relation). The closure of a closable operator is that function.*

Notice, by the way, that because A is linear, it is closable iff $f_n \rightarrow 0$ and $Af_n \rightarrow g$ implies $g = 0$.

Definition 203 (Core of an Operator) *Let A be a closed linear operator on a Banach space \mathcal{B} . A linear subspace $D \subseteq \text{Dom}(A)$ is a core of A if the closure of A restricted to D is, again A .*

The idea of a core is that we can get away with knowing how the operator works on a linear subspace, which is often much easier to deal with, rather than controlling how it acts on its whole domain.

Lemma 204 (Feller Generators Are Closed) *The generator of every Feller semigroup is closed.*

PROOF: We need to show that the graph of G contains all of its limit points, that is, if $f_n \in \text{Dom}(G)$ converges (in L_∞) on f , and $Gf_n \rightarrow g$, then $f \in \text{Dom}(G)$ and $Gf = g$. First we show that $f \in \text{Dom}(G)$.

$$\lim_{n \rightarrow \infty} (I - G)f_n = \lim_n f_n - \lim_n Gf_n \quad (15.6)$$

$$= f - g \quad (15.7)$$

But $(I - G)^{-1} = R_1$. Since this is a bounded linear operator, we can exchange applying the inverse and taking the limit, i.e.,

$$R_1 \lim_n (I - G)f_n = R_1(f - g) \quad (15.8)$$

$$\lim_n R_1(I - G)f_n = R_1(f - g) \quad (15.9)$$

$$\lim_n f_n = R_1(f - g) \quad (15.10)$$

$$f = R_1(f - g) \quad (15.11)$$

Since the range of the resolvents is contained in the domain of the generator, $f \in \text{Dom}(G)$. We can therefore say that $f - g = (I - G)f$, which implies that $Gf = g$. Hence, the graph of G contains all its limit points, and G is closed. \square

Theorem 205 (Convergence of Feller Processes) *Let X_n be a sequence of Feller processes with semigroups $K_{n,t}$ and generators G_n , and X be another Feller process with semigroup K_t and a generator G containing a core D . Then the following are equivalent.*

1. *If $f \in D$, there exists a sequence of $f_n \in \text{Dom}(G_n)$ such that $\|f_n - f\|_\infty \rightarrow 0$ and $\|A_n f_n - A f\|_\infty \rightarrow 0$.*
2. *For every $t > 0$, $K_{n,t}f \rightarrow K_t f$ for every $f \in C_0$*
3. *$\|K_{n,t}f - K_t f\|_\infty \rightarrow 0$ for each $f \in C_0$, uniformly in t for bounded positive t*
4. *If $X_n(0) \xrightarrow{d} X(0)$ in Ξ , then $X_n \xrightarrow{d} X$ in \mathbf{D} .*

PROOF: See Kallenberg, Theorem 19.25, p. 385. \square

Remark: The important versions of the property above are the second — convergence of the semigroups — and the fourth — converge in distribution of the processes. The other two are there to simplify the proof. The way the proof works is to first show that conditions (1)–(3) are all equivalent, i.e., that convergence of the operators in the semi-group implies the apparently-stronger conditions about uniform convergence as well as convergence on the core. Then one establishes the equivalence between the second condition and the fourth. To go from convergence in distribution to convergence of conditional expectations is fairly straightforward; to go the other way involves using both the first and third condition, and the first part of Proposition 201. This last step uses the Feller properties to bound, in expectation, the amount by which the process can move in a small amount of time.

Corollary 206 (Convergence of Discret-Time Markov Processes on Feller Processes) *Let X be a Feller process with semigroup K_t , generator G and core D as in Theorem 205. Let h_n be a sequence of positive real constants converging (not necessarily monotonically) to zero. Let Y_n be a sequence of discrete-time Markov processes with evolution operators H_n . Finally, let*

$X_n(t) \equiv Y_n(\lfloor t/h_n \rfloor)$, with corresponding semigroup $K_{n,t} = H_n^{\lfloor t/h_n \rfloor}$, and generator $A_n = (1/h_n)(H_n - I)$. Even though X_n is not in general a homogeneous Markov process, the conclusions of Theorem 205 remain valid.

PROOF: Kalleneberg, Theorem 19.28, pp. 387–388. \square

Remark 1: The basic idea is to show that the process X_n is close (in the Skorokhod-topology sense) to a Feller process \tilde{X}_n , whose generator is A_n . One then shows that \tilde{X}_n converges on X , using Theorem 205, and that $\tilde{X}_n \xrightarrow{d} X_n$.

Remark 2: Even though $K_{n,t}$ in the corollary above is not, strictly speaking, the time-evolution operator of X_n , because X_n is not a Markov process, it is a conditional expectation operator. Much more general theorems can be proved on when non-Markov processes converge on Markov processes, using the idea that $K_{n,t} \rightarrow K_t$. See Kurtz (1975).

15.3 Approximation of Ordinary Differential Equations by Markov Processes

The following result, due to Kurtz (1970, 1971), is essentially an application of Theorem 205.

First, recall that continuous-time, discrete-state Markov processes work essentially like a combination of a Poisson process (giving the time of transitions) with a Markov chain (giving the state moved to on transitions). This can be generalized to continuous-time, continuous-state processes, of what are called “pure jump” type.

Definition 207 (Pure Jump Markov Process) *A continuous-parameter Markov process is a pure jump process when its sample paths are piece-wise constant. For each state, there is an exponential distribution of times spent in that state, whose parameter is denoted $\lambda(x)$, and a transition probability kernel or exit distribution $\mu(x, B)$.*

Observe that pure-jump Markov processes always have cadlag sample paths. Also observe that the average amount of time the process spends in state x , once it jumps there, is $1/\lambda(x)$. So the time-average “velocity”, i.e., rate of change, starting from x ,

$$\lambda(x) \int_{\Xi} (y - x) \mu(x, dy)$$

Proposition 208 (Pure-Jump Markov Processes and ODEs) *Let X_n be a sequence of pure-jump Markov processes with state spaces Ξ_n , holding time parameters λ_n and transition probabilities μ_n . Suppose that, for all n Ξ_n is a Borel-measurable subset of \mathbb{R}^k for some k . Let Ξ be another measurable subset of \mathbb{R}^k , on which there exists a function $F(x)$ such that $|F(x) - F(y)| \leq M|x - y|$ for some constant M . Suppose all of the following conditions holds.*

1. The time-averaged rate of change is always finite:

$$\sup_n \sup_{x \in \Xi_n \cap \Xi} \lambda_n(x) \int_{\Xi_n} |y - x| \mu_n(x, dy) < \infty \quad (15.12)$$

2. There exists a positive sequence $\epsilon_n \rightarrow 0$ such that

$$\lim_{n \rightarrow \infty} \sup_{x \in \Xi_n \cap \Xi} \lambda_n(x) \int_{|y-x| > \epsilon} |y - x| \mu_n(x, dy) = 0 \quad (15.13)$$

3. The worst-case difference between $F(x)$ and the time-averaged rates of change goes to zero:

$$\lim_{n \rightarrow \infty} \sup_{x \in \Xi_n \cap \Xi} \left| F(x) - \lambda_n(x) \int (y - x) \mu_n(x, dy) \right| = 0 \quad (15.14)$$

Let $X(s, x_0)$ be the solution to the initial-value problem where the differential is given by F , i.e., for each $0 \leq s \leq t$,

$$\frac{\partial}{\partial s} X(s, x_0) = F(X(s, x_0)) \quad (15.15)$$

$$X(0, x_0) = x_0 \quad (15.16)$$

and suppose there exists an $\eta > 0$ such that, for all n ,

$$\Xi_n \cap \left\{ y \in \mathbb{R}^k : \inf_{0 \leq s \leq t} |y - X(s, x_0)| \leq \eta \right\} \subseteq \Xi \quad (15.17)$$

Then $\lim X_n(0) = x_0$ implies that, for every $\delta > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\sup_{0 \leq s \leq t} |X_n(s) - X(s, x_0)| > \delta \right) = 0 \quad (15.18)$$

The first conditions on the X_n basically make sure that they are Feller processes. The subsequent ones make sure that the mean time-averaged rate of change of the jump processes converges on the instantaneous rate of change of the differential equation, and that, if we're sufficiently close to the solution of the differential equation in \mathbb{R}^k , we're not in some weird way outside the relevant domains of definition. Even though Theorem 205 is about weak convergence, converging in distribution on a non-random object is the same as converging in probability, which is how we get uniform-in-time convergence in probability for a conclusion.

There are, broadly speaking, two kinds of uses for this result. One kind is practical, and has to do with justifying convenient approximations. If n is large, we can get away with using an ODE instead of the noisy stochastic scheme, or alternately we can use stochastic simulation to approximate the solutions of ugly ODEs. The other kind is theoretical, about showing that the large-population limit behaves deterministically, even when the individual behavior is stochastic and strongly dependent over time.

15.4 Exercises

Exercise 15.1 (Poisson Counting Process) *Show that the Poisson counting process is a pure jump Markov process.*

Exercise 15.2 (Exponential Holding Times in Pure-Jump Processes) *Prove that pure-jump Markov processes have exponentially-distributed holding times, i.e., that if X is a Markov process with piecewise-constant sample paths, and $\epsilon_x = \inf t > 0 X(t) \neq x$, that $\epsilon_x | X(0) = x$ is exponentially distributed.*

Exercise 15.3 (Solutions of ODEs are Feller Processes) *Let $F : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a sufficiently smooth vector field that the ordinary differential equation $dx/dt = F(x)$ has a unique solution for every initial condition x_0 . Prove that the set of solutions forms a Feller process.*

Chapter 16

Convergence of Random Walks

This lecture examines the convergence of random walks to the Wiener process. This is very important both physically and statistically, and illustrates the utility of the theory of Feller processes.

Section 16.1 finds the semi-group of the Wiener process, shows it satisfies the Feller properties, and finds its generator.

Section 16.2 turns random walks into cadlag processes, and gives a fairly easy proof that they converge on the Wiener process.

16.1 The Wiener Process is Feller

Recall that the Wiener process $W(t)$ is defined by starting at the origin, by independent increments over non-overlapping intervals, by the Gaussian distribution of increments, and by continuity of sample paths (Examples 38 and 79). The process is homogeneous, and the transition kernels are (Section 11.2)

$$\mu_t(w_1, B) = \int_B dw_2 \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2 - w_1)^2}{2t}} \quad (16.1)$$

$$\frac{d\mu_t(w_1, w_2)}{d\lambda} = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2 - w_1)^2}{2t}} \quad (16.2)$$

where the second line gives the density of the transition kernel with respect to Lebesgue measure.

Since the kernels are known, we can write down the corresponding evolution operators:

$$K_t f(w_1) = \int dw_2 f(w_2) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(w_2 - w_1)^2}{2t}} \quad (16.3)$$

We saw in Section 11.2 that the kernels have the semi-group property, so (Lemma 121) the evolution operators do too.

Let's check that $\{K_t\}, t \geq 0$ is a Feller semi-group. The first Feller property is easier to check in its probabilistic form, that, for all $t, y \rightarrow x$ implies $W_y(t) \xrightarrow{d} W_x(t)$. The distribution of $W_x(t)$ is just $\mathcal{N}(x, t)$, and it is indeed true that $y \rightarrow x$ implies $\mathcal{N}(y, t) \rightarrow \mathcal{N}(x, t)$. The second Feller property can be checked in its semi-group form: as $t \rightarrow 0, \mu_t(w_1, B)$ approaches $\delta(w - w_1)$, so $\lim_{t \rightarrow 0} K_t f(x) = f(x)$. Thus, the Wiener process is a Feller process. This implies that it has cadlag sample paths (Theorem 193), but we already knew that, since we know it's continuous. What we did not know was that the Wiener process is not just Markov but strong Markov, which follows from Theorem 194.

To find the generator of $\{K_t\}, t \geq 0$, it will help to re-write it in an equivalent form, as

$$K_t f(w) = \mathbf{E} \left[f(w + Z\sqrt{t}) \right] \quad (16.4)$$

where Z is an independent $\mathcal{N}(0, 1)$ random variable. (We saw that this was equivalent in Section 11.2.) Now let's pick an $f \in C_0$ which is also twice continuously differentiable, i.e., $f \in C_0 \cap C^2$. Look at $K_t f(w) - f(w)$, and apply Taylor's theorem, expanding around w :

$$K_t f(w) - f(w) = \mathbf{E} \left[f(w + Z\sqrt{t}) \right] - f(w) \quad (16.5)$$

$$= \mathbf{E} \left[f(w + Z\sqrt{t}) - f(w) \right] \quad (16.6)$$

$$= \mathbf{E} \left[Z\sqrt{t}f'(w) + \frac{1}{2}tZ^2f''(w) + R(Z\sqrt{t}) \right] \quad (16.7)$$

$$= \sqrt{t}f'(w)\mathbf{E}[Z] + t\frac{f''(w)}{2}\mathbf{E}[Z^2] + \mathbf{E} \left[R(Z\sqrt{t}) \right] \quad (16.8)$$

Recalling that $\mathbf{E}[Z] = 0, \mathbf{E}[Z^2] = 1$,

$$\lim_{t \downarrow 0} \frac{K_t f(w) - f(w)}{t} = \frac{1}{2}f''(w) + \lim_{t \downarrow 0} \frac{\mathbf{E} \left[R(Z\sqrt{t}) \right]}{t} \quad (16.9)$$

So, we need to investigate the behavior of the remainder term $R(Z\sqrt{t})$.

We know from Taylor's theorem that

$$R(Z\sqrt{t}) = \frac{tZ^2}{2} \int_0^1 du f''(w + uZ\sqrt{t}) - f''(w) \quad (16.10)$$

$$(16.11)$$

Since $f \in C_0 \cap C^2$, we know that $f'' \in C_0$. Therefore, f'' is uniformly continuous, and has a modulus of continuity,

$$m(f'', h) = \sup_{x, y: |x-y| \leq h} |f''(x) - f''(y)| \quad (16.12)$$

which goes to 0 as $h \downarrow 0$. Thus

$$\left| R(Z\sqrt{t}) \right| \leq \frac{tZ^2}{2} m(f'', Z\sqrt{t}) \quad (16.13)$$

$$\lim_{t \rightarrow 0} \frac{|R(Z\sqrt{t})|}{t} \leq \lim_{t \rightarrow 0} \frac{Z^2 m(f'', Z\sqrt{t})}{2} \quad (16.14)$$

$$= 0 \quad (16.15)$$

Plugging back in to Equation 16.9,

$$Gf(w) = \frac{1}{2}f''(w) + \lim_{t \downarrow 0} \frac{\mathbf{E}[R(Z\sqrt{t})]}{t} \quad (16.16)$$

$$= \frac{1}{2}f''(w) \quad (16.17)$$

That is, $G = \frac{1}{2} \frac{d^2}{dw^2}$, one half of the Laplacian. We have shown this only for $C_0 \cap C^2$, but this is clearly a linear subspace of C_0 , and, since C^2 is dense in C , it is dense in C_0 , i.e., this is a core for the generator. Hence the generator is really the extension of $\frac{1}{2} \frac{d^2}{dw^2}$ to the whole of C_0 , but this is too cumbersome to repeat all the time, so we just say it's the Laplacian.

16.2 Convergence of Random Walks

Let X_1, X_2, \dots be a sequence of IID variables with mean 0 and variance 1. The random walk process S_n is then just $\sum_{i=1}^n X_i$. It is a discrete-time Markov process, and consequently also a strong Markov process. Imagine each step of the walk takes some time h , and imagine this time interval becoming smaller and smaller. Then, between any two times t_1 and t_2 , the number of steps of the random walk will be about $\frac{t_2 - t_1}{h}$, which will go to infinity. The increment of the random walk from t_1 to t_2 will then be a sum of an increasingly large number of IID random variables, and by the central limit theorem will approach a Gaussian distribution. Moreover, if we look at the interval of time from t_2 to t_3 , we will see another Gaussian, but all of the random-walk steps going into it will be independent of those going into our first interval. So, we expect that the random walk will in some sense come to look like the Wiener process, no matter what the exact distribution of the X_1 . (We saw some of this in Section 11.3.) Let's consider this in more detail.

Definition 209 (Continuous-Time Random Walk (Cadlag)) Let X_1, X_2, \dots be an IID sequence of real-valued random variables with mean 0 and variance 1. Define $S(m)$ as $X_1 + X_2 + \dots + X_m$, and $X_0 = 0$. The corresponding continuous-time random walks (CTRWs) are the processes

$$Y_n(t) \equiv \frac{1}{n^{1/2}} \sum_{i=0}^{\lfloor nt \rfloor} X_i \quad (16.18)$$

$$= n^{-1/2} S(\lfloor nt \rfloor) \quad (16.19)$$

Remark 1: You should verify (Exercise 16.3) that continuous-time random walks are inhomogeneous Markov processes with cadlag sample paths.

Remark 2: We will later see CTRWs with continuous sample paths, obtained by linear interpolation, but these, with their piece-wise constant sample paths, will do for now.

As has been hinted at repeatedly, random walks converge in distribution on the Wiener process. There are, broadly speaking, two ways to show this. One is to use the Feller process machinery of Chapter 15, and apply Corollary 206. The other is to directly manipulate the criteria for convergence of cadlag processes. Both lead to the same conclusion.

16.2.1 Approach Through Feller Processes

The processes Y_n are not homogeneously Markovian, though the discrete-time processes $n^{-1/2}S(m)$ are. Nonetheless, we can find the equivalent of their evolution operators, and show that they converge on the evolution operators of the Wiener process. First, let's establish a nice property of the increments of Y_n .

Lemma 210 (Increments of Random Walks) *For a continuous-time random walk, for all n ,*

$$Y_n(t+h) - Y_n(t) = n^{-1/2}S'(\lfloor n(t+h) \rfloor - \lfloor nt \rfloor) \quad (16.20)$$

where $S'(m) \equiv \sum_{i=0}^m X'_i$ and the IID sequence X' is an independent copy of X .

PROOF: By explicit calculation. For any n , for any t and any $h > 0$,

$$Y_n(t+h) = \frac{1}{n^{1/2}} \sum_{i=0}^{\lfloor n(t+h) \rfloor} X_i \quad (16.21)$$

$$= Y_n(t) + n^{-1/2} \sum_{i=\lfloor nt \rfloor + 1}^{\lfloor n(t+h) \rfloor} X_i \quad (16.22)$$

$$= Y_n(t) + n^{-1/2} \sum_{i=0}^{\lfloor n(t+h) \rfloor - \lfloor nt \rfloor} X'_i \quad (16.23)$$

using, in the last line, the fact that the X_i are IID. Eq. 16.20 follows from the definition of $S(m)$. \square

Lemma 211 (Continuous-Time Random Walks are Pseudo-Feller) *Every continuous-time random walk has the Feller properties (but is not a homogeneous Markov process).*

PROOF: This is easiest using the process/probabilistic forms (Definition 177) of the Feller properties.

To see that the first property (Eq. 14.3) holds, we need the distribution of $Y_{n,y}(t)$, that is, the distribution of the state of Y_n at time t , when started from state y (rather than 0). By Lemma 210,

$$Y_{n,y}(t) \stackrel{d}{=} y + n^{-1/2}S'(\lfloor nt \rfloor) \quad (16.24)$$

Clearly, as $y \rightarrow x$, $y + n^{-1/2}S'(\lfloor nt \rfloor) \xrightarrow{d} x + n^{-1/2}S'(\lfloor nt \rfloor)$, so the first Feller property holds.

To see that the second Feller property holds, observe that, for each n , for all t , and for all ω , $Y_n(t+h, \omega) = Y_n(t, \omega)$ if $0 \geq h < 1/n$. This sure convergence implies almost-sure convergence, which implies convergence in probability. \square

Lemma 212 (Evolution Operators of Random Walks) *The “evolution” (i.e., conditional expectation) operators of the random walk Y_n , $K_{n,t}$, are given by*

$$K_{n,t}f(y) = \mathbf{E}[f(y + Y'_n(t))] \quad (16.25)$$

where Y'_n is an independent copy of Y_n .

PROOF: Substitute Lemma 210 into the definition of the evolution operator.

$$K_{n,h}f(y) \equiv \mathbf{E}[f(Y_n(t+h)) | Y_n(t) = y] \quad (16.26)$$

$$= \mathbf{E}[f(Y_n(t+h) + Y_n(t) - Y_n(t)) | Y_n(t) = y] \quad (16.27)$$

$$= \mathbf{E}\left[f\left(n^{-1/2}S'(\lfloor nt \rfloor) + Y_n(t) | Y_n(t) = y\right)\right] \quad (16.28)$$

$$= \mathbf{E}\left[f(y + n^{-1/2}S'(\lfloor nt \rfloor))\right] \quad (16.29)$$

In words, the transition operator simply takes the expectation over the increments of the random walk, just as with the Wiener process. Finally, substitution of $Y'_n(t)$ for $n^{-1/2}S'(\lfloor nt \rfloor)$ is licensed by Eq. 16.19. \square

Theorem 213 (Functional Central Limit Theorem (I)) $Y_n \xrightarrow{d} W$ in \mathbf{D} .

PROOF: Apply Theorem 206. Clause (4) of the theorem says that if any of the other three clauses are satisfied, and $Y_n(0) \xrightarrow{d} W(0)$ in \mathbb{R} , then $Y_n \xrightarrow{d} W$ in \mathbf{D} . Clause (2) is that $K_{n,t} \rightarrow K_t$ for all $t > 0$. That is, for any $t > 0$, and $f \in C_0$, $K_{n,t}f \rightarrow K_tf$ as $n \rightarrow \infty$. Pick any such t and f and consider $K_{n,t}f$. By Lemma 212,

$$K_{n,t}f(y) = \mathbf{E}\left[f(y + n^{-1/2}S'(\lfloor nt \rfloor))\right] \quad (16.30)$$

As $n \rightarrow \infty$, $n^{-1/2} \rightarrow t^{1/2} \lfloor nt \rfloor^{-1/2}$. Since the X_i (and so the X'_i) have variance 1, the central limit theorem applies, and $\lfloor nt \rfloor^{-1/2}S'(\lfloor nt \rfloor) \xrightarrow{d} \mathcal{N}(0, 1)$, say Z . Consequently $n^{-1/2}S'(\lfloor nt \rfloor) \xrightarrow{d} \sqrt{t}Z$. Since $f \in C_0$, it is continuous and bounded, hence, by the definition of convergence in distribution,

$$\mathbf{E}\left[f\left(y + n^{-1/2}S'(\lfloor nt \rfloor)\right)\right] \rightarrow \mathbf{E}\left[f\left(y + \sqrt{t}Z\right)\right] \quad (16.31)$$

But $\mathbf{E}[f(y + \sqrt{t}Z)] = K_t f(y)$, the time-evolution operator of the Wiener process applied to f at y . Since the evolution operators of the random walks converge on those of the Wiener process, and since their initial conditions match, by the theorem $Y_n \xrightarrow{d} W$ in \mathbf{D} . \square

16.2.2 Direct Approach

The alternate approach to the convergence of random walks works directly with the distributions, avoiding the Feller properties. It is not quite so slick, but provides a comparatively tractable example of how general results about convergence of stochastic processes go.

We want to find the limiting distribution of Y_n as $n \rightarrow \infty$. First of all, we should convince ourselves that a limit distribution exists. But this is not too hard. For any fixed t , $Y_n(t)$ approaches a Gaussian distribution by the central limit theorem. For any fixed finite collection of times $t_1 \leq t_2 \leq \dots \leq t_k$, $Y_n(t_1), Y_n(t_2), \dots, Y_n(t_k)$ approaches a limiting distribution if $Y_n(t_1), Y_n(t_2) - Y_n(t_1), \dots, Y_n(t_k) - Y_n(t_{k-1})$ does, but that again will be true by the (multivariate) central limit theorem. Since the limiting finite-dimensional distributions exist, some limiting distribution exists (via Theorem 23). It remains to convince ourselves that this limit is in \mathbf{D} , and to identify it.

Lemma 214 (Convergence of Random Walks in Finite-Dimensional Distribution) $Y_n \xrightarrow{fd} W$.

PROOF: For all n , $Y_n(0) = 0 = W(0)$. For any $t_2 > t_1$,

$$\mathcal{L}(Y_n(t_2) - Y_n(t_1)) = \mathcal{L}\left(\frac{1}{\sqrt{n}} \sum_{i=\lfloor nt_1 \rfloor}^{\lfloor nt_2 \rfloor} X_i\right) \quad (16.32)$$

$$\xrightarrow{d} \mathcal{N}(0, t_2 - t_1) \quad (16.33)$$

$$= \mathcal{L}(W(t_2) - W(t_1)) \quad (16.34)$$

Finally, for any three times $t_1 < t_2 < t_3$, $Y_n(t_3) - Y_n(t_2)$ and $Y_n(t_2) - Y_n(t_1)$ are independent for sufficiently large n , because they become sums of disjoint collections of independent random variables. The same applies to large groups of times. Thus, the limiting distribution of Y_n starts at the origin and has independent Gaussian increments. Since these properties determine the finite-dimensional distributions of the Wiener process, $Y_n \xrightarrow{fd} W$. \square

Theorem 215 (Functional Central Limit Theorem (II)) $Y_n \xrightarrow{d} W$.

PROOF: By Proposition 200, it is enough to show that $Y_n \xrightarrow{fd} W$, and that any of the properties in Proposition 201 hold. The lemma took care of the finite-dimensional convergence, so we can turn to the second part. A sufficient

condition is property (1) in the latter theorem, that $|Y_n(\tau_n + h_n) - Y_n(\tau_n)| \xrightarrow{P} 0$ for all finite optional times τ_n and any sequence of positive constants $h_n \rightarrow 0$.

$$|Y_n(\tau_n + h_n) - Y_n(\tau_n)| = n^{-1/2} |S(\lfloor n\tau_n + nh_n \rfloor) - S(\lfloor n\tau_n \rfloor)| \quad (16.35)$$

$$\stackrel{d}{=} n^{-1/2} |S(\lfloor nh_n \rfloor) - S(0)| \quad (16.36)$$

$$= n^{-1/2} |S(\lfloor nh_n \rfloor)| \quad (16.37)$$

$$= n^{-1/2} \left| \sum_{i=0}^{\lfloor nh_n \rfloor} X_i \right| \quad (16.38)$$

To see that this converges in probability to zero, we will appeal to Chebyshev's inequality: if Z_i have common mean 0 and variance σ^2 , then, for every positive ϵ ,

$$\mathbb{P} \left(\left| \sum_{i=1}^m Z_i \right| > \epsilon \right) \leq \frac{m\sigma^2}{\epsilon^2} \quad (16.39)$$

Here we have $Z_i = X_i/\sqrt{n}$, so $\sigma^2 = 1/n$, and $m = \lfloor nh_n \rfloor$. Thus

$$\mathbb{P} \left(n^{-1/2} |S(\lfloor nh_n \rfloor)| > \epsilon \right) \leq \frac{\lfloor nh_n \rfloor}{n\epsilon^2} \quad (16.40)$$

As $0 \leq \lfloor nh_n \rfloor / n \leq h_n$, and $h_n \rightarrow 0$, the bounding probability must go to zero for every fixed ϵ . Hence $n^{-1/2} |S(\lfloor nh_n \rfloor)| \xrightarrow{P} 0$. \square

16.2.3 Consequences of the Functional Central Limit Theorem

Corollary 216 (The Invariance Principle) *Let X_1, X_2, \dots be IID random variables with mean μ and variance σ^2 . Then*

$$Y_n(t) \equiv \frac{1}{\sqrt{n}} \sum_{i=0}^{\lfloor nt \rfloor} \frac{X_i - \mu}{\sigma} \xrightarrow{d} W(t) \quad (16.41)$$

PROOF: $(X_i - \mu)/\sigma$ has mean 0 and variance 1, so Theorem 215 applies. \square

This result is called “the invariance principle”, because it says that the limiting distribution of the sequences of sums depends only on the mean and variance of the individual terms, and is consequently *invariant* under changes which leave those alone. Both this result and the previous one are known as the “functional central limit theorem”, because convergence in distribution is the same as convergence of all bounded continuous *functionals* of the sample path. Another name is “Donsker’s Theorem”, which is sometimes associated however with the following corollary of Theorem 215.

Corollary 217 (Donsker’s Theorem) *Let $Y_n(t)$ and $W(t)$ be as before, but restrict the index set T to the unit interval $[0, 1]$. Let f be any function from $\mathbf{D}([0, 1])$ to \mathbb{R} which is measurable and a.s. continuous at W . Then $f(Y_n) \xrightarrow{d} f(W)$.*

PROOF: Exercise. \square

This version is especially important for statistical purposes, as we'll see a bit later.

16.3 Exercises

Exercise 16.1 (Example 167 Revisited) *Go through all the details of Example 167.*

1. Show that $\mathcal{F}_t^X \subseteq \mathcal{F}_t^W$ for all t , and that $\{\mathcal{F}_t^X\} \subset \{\mathcal{F}_t^W\}$.
2. Show that $\tau = \inf_t X(t) = (0, 0)$ is a $\{\mathcal{F}_t^X\}$ -optional time, and that it is finite with probability 1.
3. Show that X is Markov with respect to both its natural filtration and the natural filtration of the driving Wiener process.
4. Show that X is not strongly Markov at τ .
5. Which, if either, of the Feller properties does X have?

Exercise 16.2 (Generator of the d -dimensional Wiener Process) *Consider a d -dimensional Wiener process, i.e., an \mathbb{R}^d -valued process where each coordinate is an independent Wiener process. Find the generator.*

Exercise 16.3 (Continuous-time random walks are Markovian) *Show that every continuous-time random walk (as per Definition 209) is an inhomogeneous Markov process, with cadlag sample paths.*

Exercise 16.4 (Donsker's Theorem) *Prove Donsker's Theorem (Corollary 217).*

Exercise 16.5 (Diffusion equation) *The partial differential equation*

$$\frac{1}{2} \frac{\partial^2 f(x, t)}{\partial x^2} = \frac{\partial f(x, t)}{\partial t}$$

is called the diffusion equation. From our discussion of initial value problems in Chapter 12 (Corollary 159 and related material), it is clear that the function $f(x, t)$ solves the diffusion equation with initial condition $f(x, 0)$ if and only if $f(x, t) = K_t f(x, 0)$, where K_t is the evolution operator of the Wiener process.

1. Take $f(x, 0) = (2\pi 10^{-4})^{-1/2} e^{-\frac{x^2}{2 \cdot 10^{-4}}}$. $f(x, t)$ can be found analytically; do so.

2. Estimate $f(x, 10)$ over the interval $[-5, 5]$ stochastically. Use the fact that $K_t f(x) = \mathbf{E}[f(W(t)) | W(0) = x]$, and that random walks converge on the Wiener process. (Be careful that you scale your random walks the right way!) Give an indication of the error in this estimate.
3. Can you find an analytical form for $f(x, t)$ if $f(x, 0) = \mathbf{1}_{[-0.5, 0.5]}(x)$?
4. Find $f(x, 10)$, with the new initial conditions, by numerical integration on the domain $[-10, 10]$, and compare it to a stochastic estimate.

Exercise 16.6 (Functional CLT for Dependent Variables) Let X_i , $i = 1, 2, \dots$, be a weakly stationary but dependent sequence of real-valued random variables, with mean 0 and standard deviation 1. (Note that any weakly-stationary sequence with finite variance can be normalized into this form.) Let Y_n be the corresponding continuous-time random walk, i.e.,

$$Y_n(t) = \frac{1}{n^{1/2}} \sum_{i=0}^{\lfloor nt \rfloor} X_i$$

Suppose that, despite their interdependence, the X_i still obey the central limit theorem,

$$\frac{1}{n^{1/2}} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0, 1)$$

Are these conditions enough to prove a functional central limit theorem, that $Y_n \xrightarrow{d} W$? If so, prove it. If not, explain what the problem is, and suggest an additional sufficient condition on the X_i .

Part IV

Diffusions and Stochastic
Calculus

Chapter 17

Diffusions and the Wiener Process

Section 17.1 introduces the ideas which will occupy us for the next few lectures, the continuous Markov processes known as diffusions, and their description in terms of stochastic calculus.

Section 17.2 collects some useful properties of the most important diffusion, the Wiener process.

Section 17.3 shows, first heuristically and then more rigorously, that almost all sample paths of the Wiener process don't have derivatives.

17.1 Diffusions and Stochastic Calculus

So far, we have looked at Markov processes in general, and then paid particular attention to Feller processes, because the Feller properties are very natural continuity assumptions to make about stochastic models and have very important consequences, especially the strong Markov property and cadlag sample paths. The natural next step is to go to Markov processes with continuous sample paths. The most important case, overwhelmingly dominating the literature, is that of *diffusions*.

Definition 218 (Diffusion) *A stochastic process X adapted to a filtration \mathcal{F} is a diffusion when it is a strong Markov process with respect to \mathcal{F} , homogeneous in time, and has continuous sample paths.*

Remark: Having said that, I should confess that some authors don't insist that diffusions be homogeneous, and some even don't insist that they be *strong* Markov processes. But this is the general sense in which the term is used.

Diffusions matter to us for several reasons. First, they are very natural models of many important systems — the motion of physical particles (the

source of the term “diffusion”), fluid flows, noise in communication systems, financial time series, etc. Probabilistic and statistical studies of time-series data thus need to understand diffusions. Second, many discrete Markov models have large-scale limits which are diffusion processes: these are important in physics and chemistry, population genetics, queueing and network theory, certain aspects of learning theory¹, etc. These limits are often more tractable than more exact finite-size models. (We saw a hint of this in Section 15.3.) Third, many statistical-inferential problems can be described in terms of diffusions, most prominently ones which concern goodness of fit, the convergence of empirical distributions to true probabilities, and nonparametric estimation problems of many kinds.

The easiest way to get at diffusions is to through the theory of stochastic differential equations; the most important diffusions can be thought of as, roughly speaking, the result of adding a noise term to the right-hand side of a differential equation. A more exact statement is that, just as an autonomous ordinary differential equation

$$\frac{dx}{dt} = f(x), \quad x(t_0) = x_0 \quad (17.1)$$

has the solution

$$x(t) = x_0 + \int_{t_0}^t f(x(s))ds \quad (17.2)$$

a stochastic differential equation

$$\frac{dX}{dt} = f(X) + g(X)\frac{dY}{dt}, \quad X(t_0) = x_0 \text{ a.s.} \quad (17.3)$$

where X and Y are stochastic processes, is solved by

$$X(t) = x_0 + \int_{t_0}^t f(X(s))ds + \int_{t_0}^t g(X(s))dY \quad (17.4)$$

where $\int g(X, t)dY$ is a *stochastic integral*. It turns out that, properly constructed, this sort of integral, and so this sort of stochastic differential equation, makes sense even when dY/dt does not make sense as any sort of ordinary derivative, so that the more usual way of writing an SDE is

$$dX = f(X)dt + g(X)dY, \quad X(t_0) = x_0 \text{ a.s.} \quad (17.5)$$

even though this seems to invoke infinitesimals, which don't exist.²

¹Specifically, discrete-time reinforcement learning converges to the continuous-time replicator equation of evolutionary theory.

²Some people, like Ethier and Kurtz (1986), prefer to talk about stochastic *integral* equations, rather than stochastic *differential* equations, because things like 17.5 are really shortcuts for “find an X such that Eq. 17.4 holds”, and objects like dX don't really make much sense on their own. There's a certain logic to this, but custom is overwhelmingly against them.

The fully general theory of stochastic calculus considers integration with respect to a very broad range of stochastic processes, but the original case, which is still the most important, is integration with respect to the Wiener process, which corresponds to driving a system with white noise. In addition to its many applications in all the areas which use diffusions, the theory of integration against the Wiener process occupies a central place in modern probability theory; I simply would not be doing my job if this course did not cover it. We therefore begin our study of diffusions and stochastic calculus by reviewing some of the properties of the Wiener process — which is also the most important diffusion process.

17.2 Once More with the Wiener Process and Its Properties

To review, the standard Wiener process $W(t)$ is defined by

1. $W(0) = 0$,
2. centered Gaussian increments with linearly-growing variance, $\mathcal{L}(W(t_2) - W(t_1)) = \mathcal{N}(0, t_2 - t_1)$,
3. independent increments and
4. continuity of sample paths.

We have seen that it is a homogeneous Markov process (Section 11.2), and in fact (Section 16.1) a Feller process, and therefore a strong Markov process, whose generator (Equation 16.16) is $\frac{1}{2}\nabla^2$. By Definition 218, W is a diffusion.

This section proves a few more useful properties.

Lemma 219 (The Wiener Process Is a Martingale) *The Wiener process is a martingale with respect to its natural filtration.*

PROOF: This follows directly from the Gaussian increment property:

$$\mathbf{E}[W(t+h)|\mathcal{F}_t^X] = \mathbf{E}[W(t+h)|W(t)] \quad (17.6)$$

$$= \mathbf{E}[W(t+h) - W(t) + W(t)|W(t)] \quad (17.7)$$

$$= \mathbf{E}[W(t+h) - W(t)|W(t)] + W(t) \quad (17.8)$$

$$= 0 + W(t) = W(t) \quad (17.9)$$

where the first line uses the Markov property of W , and the last line the Gaussian increments property. \square

Definition 220 (Wiener Processes with Respect to Filtrations) *If $W(t, \omega)$ is adapted to a filtration $\{\mathcal{F}_t\}$ and is an $\{\mathcal{F}_t\}$ -filtration, it is an $\{\mathcal{F}_t\}$ Wiener process or $\{\mathcal{F}_t\}$ Brownian motion.*

It seems natural to speak of the Wiener process as a Gaussian process. This motivates the following definition.

Definition 221 (Gaussian Process) *A real-valued stochastic process is Gaussian when all its finite-dimensional distributions are multivariate Gaussian distributions.*

Lemma 222 (Wiener Process Is Gaussian) *The Wiener process is a Gaussian process.*

PROOF: Pick any k times $t_1 < t_2 < \dots < t_k$. Then the increments $W(t_1) - W(0)$, $W(t_2) - W(t_1)$, $W(t_3) - W(t_2)$, \dots , $W(t_k) - W(t_{k-1})$ are independent Gaussian random variables. If X and Y are independent Gaussians, then $X, X + Y$ is a multivariate Gaussian, so (recursively) $W(t_1) - W(0), W(t_2) - W(0), \dots, W(t_k) - W(0)$ has a multivariate Gaussian distribution. Since $W(0) = 0$, the Gaussian distribution property follows. Since t_1, \dots, t_k were arbitrary, as was k , all the finite-dimensional distributions are Gaussian. \square

Just as the distribution of a Gaussian random variable is determined by its mean and covariance, the distribution of a Gaussian process is determined by its mean over time, $\mathbf{E}[X(t)]$, and its covariance function, $\text{cov}(X(s), X(t))$. (You might find it instructive to prove this *without* looking at Lemma 13.1 in Kallenberg.) Clearly, $\mathbf{E}[W(t)] = 0$, and, taking $s \leq t$ without loss of generality,

$$\text{cov}(W(s), W(t)) = \mathbf{E}[W(s)W(t)] - \mathbf{E}[W(s)]\mathbf{E}[W(t)] \quad (17.10)$$

$$= \mathbf{E}[(W(t) - W(s) + W(s))W(s)] \quad (17.11)$$

$$= \mathbf{E}[(W(t) - W(s))W(s)] + \mathbf{E}[W(s)W(s)] \quad (17.12)$$

$$= \mathbf{E}[W(t) - W(s)]\mathbf{E}[W(s)] + s \quad (17.13)$$

$$= s \quad (17.14)$$

17.3 Wiener Measure; Most Continuous Curves Are Not Differentiable

We can regard the Wiener process as establishing a measure on the space $\mathbf{C}(\mathbb{R}^+)$ of continuous real-valued functions; this is one of the considerations which led Wiener to it (Wiener, 1958)³. This will be important when we want to do statistical inference for stochastic processes. All Bayesian methods, and most frequentist ones, will require us to have a likelihood for the model θ given data x , $f_\theta(x)$, but likelihoods are really Radon-Nikodym derivatives, $f_\theta(x) = \frac{d\nu_\theta}{d\mu}(x)$ with respect to some reference measure μ . When our sample space is \mathbb{R}^d , we generally use Lebesgue measure as our reference measure, since its support is

³The early chapters of this book form a wonderfully clear introduction to Wiener measure, starting from prescriptions on the measures of finite-dimensional cylinders and building from there, deriving the incremental properties we've started with as consequences.

the whole space, it treats all points uniformly, and it's reasonably normalizable. Wiener measure will turn out to play a similar role when our sample space is \mathbf{C} .

A mathematically important question, which will also turn out to matter to us very greatly when we try to set up stochastic differential equations, is whether, under this *Wiener measure*, most curves are differentiable. If, say, almost all curves were differentiable, then it would be easy to define dW/dt . Unfortunately, this is not the case; almost all curves are nowhere differentiable.

There is an easy heuristic argument to this conclusion. $W(t)$ is a Gaussian, whose variance is t . If we look at the ratio in a derivative

$$\frac{W(t+h) - W(t)}{(t+h) - t}$$

the numerator has variance h and the denominator is the constant h , so the ratio has variance $1/h$, which goes to infinity as $h \rightarrow 0$. In other words, as we look at the curve of $W(t)$ on smaller and smaller scales, it becomes more and more erratic, and the slope finally blows up into a completely unpredictable quantity. This is basically the shape of the more rigorous argument as well.

Theorem 223 (Almost All Continuous Curves Are Non-Differentiable)

With probability 1, $W(t)$ is nowhere-differentiable.

PROOF: Assume, by way of contradiction, that $W(t)$ is differentiable at t_0 . Then

$$\lim_{t \rightarrow t_0} \frac{W(t, \omega) - W(t_0, \omega)}{t - t_0} \quad (17.15)$$

must exist, for some set of ω of positive measure. Call its supposed value $W'(t_0, \omega)$. That is, for every $\epsilon > 0$, we must have some δ such that $t_0 - \delta \leq t \leq t_0 + \delta$ implies

$$\left| \frac{W(t, \omega) - W(t_0, \omega)}{t - t_0} - W'(t_0, \omega) \right| \leq \epsilon \quad (17.16)$$

Without loss of generality, take $t > t_0$. Then $W(t, \omega) - W(t_0, \omega)$ is independent of $W(t_0, \omega)$ and has a Gaussian distribution with mean zero and variance $t - t_0$. Therefore the differential ratio is $\mathcal{N}(0, \frac{1}{t-t_0})$. The quantity inside the absolute value sign in Eq. 17.16 is thus Gaussian with distribution $\mathcal{N}(-W'(t_0), \frac{1}{t-t_0})$. The probability that it exceeds any ϵ is therefore always positive, and in fact can be made arbitrarily large by taking t sufficiently close to t_0 . Hence, with probability 1, there is no point of differentiability. \square

Continuous curves which are nowhere differentiable are odd-looking beasts, but we've just established that such "pathological" cases are in fact typical, and non-pathological ones vanishingly rare in \mathbf{C} . What's worse, in the functional central limit theorem (215), we obtained W as the limit of piecewise constant, and so piecewise differentiable, random functions. We could even have linearly interpolated between the points of the random walk, and those random functions

would also have converged in distribution on W .⁴ The continuous, almost-everywhere-differentiable curves form a subset of \mathbf{C} , and now we have a sequence of measures which give them probability 1, converging on Wiener measure, which gives them probability 0. This sounds like trouble, especially if we want to use Wiener measure as a reference measure in likelihoods, because it sounds like lots of interesting measures, which *do* produce differentiable curves, will not be absolutely continuous...⁵

⁴Could we have used quadratic spline interpolation to get sample paths with continuous first derivatives, and still have had $\xrightarrow{d} W$?

⁵Doing Exercises 1.1 and 7.1, if you haven't already, should help you resolve this paradox.

Chapter 18

A First Look at Stochastic Integrals with the Wiener Process

Section 18.1 touches briefly on the martingale characterization of the Wiener process.

Section 18.2 gives a heuristic introduction to stochastic integrals, via Euler's method for approximating ordinary integrals.

18.1 Martingale Characterization of the Wiener Process

Because the Wiener process is a Lévy process (Example 139), it is self-similar in the sense of Definition 147. That is, for any $a > 0$, $W(at) \stackrel{d}{=} a^{1/2}W(t)$. In fact, if we define a new process W_a through $W_a(t, \omega) = a^{-1/2}W(at, \omega)$, then W_a is itself a Wiener process. Thus the whole process is self-similar. This is only one of several sorts of self-similarities in the Wiener process. Another is sometimes called *spatial homogeneity*: W_τ , defined through $W_\tau(t, \omega) = W(t + \tau, \omega) - W(\tau, \omega)$ is also a Wiener process. That is, if we “re-zero” to the state of the Wiener process $W(\tau)$ at an arbitrary time τ , the new process looks just like the old process. $-W(t)$, obviously, is also a Wiener process.

Related to these properties is the fact that $W^2(t) - t$ is a martingale with respect to $\{\mathcal{F}_t^W\}$. (This is easily shown with a little algebra.) What is more surprising is that this is enough to characterize the Wiener process.

Theorem 224 (Martingale Characterization of the Wiener Process) *If $M(t)$ is a continuous martingale, and $M^2(t) - t$ is also a martingale, then $M(t)$ is a Wiener process.*

There are some very clean proofs of this theorem¹ — but they require us to use stochastic calculus! Doob (1953, pp. 384ff) gives a proof which does not, however. The details of his proof are messy, but the basic idea is to get the central limit theorem to apply, using the martingale property of $M^2(t) - t$ to get the variance to grow linearly with time and to get independent increments, and then seeing that between any two times t_1 and t_2 , we can fit arbitrarily many little increments so we can use the CLT.

We will return to this result as an illustration of the stochastic calculus (Theorem 247).

18.2 A Heuristic Introduction to Stochastic Integrals

Euler's method is perhaps the most basic method for numerically approximating integrals. If we want to evaluate $I(x) \equiv \int_a^b x(t)dt$, then we pick n intervals of time, with boundaries $a = t_0 < t_1 < \dots < t_n = b$, and set

$$I_n(x) = \sum_{i=1}^n x(t_{i-1}) (t_i - t_{i-1})$$

Then $I_n(x) \rightarrow I(x)$, if x is well-behaved and the length of the largest interval $\rightarrow 0$. If we want to evaluate $\int_{t=a}^{t=b} x(t)dw$, where w is another function of t , the natural thing to do is to get the derivative of w , w' , replace the integrand by $x(t)w'(t)$, and perform the integral with respect to t . The approximating sums are then

$$\sum_{i=1}^n x(t_{i-1}) w'(t_{i-1}) (t_i - t_{i-1}) \quad (18.1)$$

Alternately, we could, if $w(t)$ is nice enough, approximate the integral by

$$\sum_{i=1}^n x(t_{i-1}) (w(t_i) - w(t_{i-1})) \quad (18.2)$$

even if w' doesn't exist.

(You may be more familiar with using Euler's method to solve ODEs, $dx/dt = f(x)$. Then one generally picks a Δt , and iterates

$$x(t + \Delta t) = x(t) + f(x)\Delta t \quad (18.3)$$

from the initial condition $x(t_0) = x_0$, and uses linear interpolation to get a continuous, almost-everywhere-differentiable curve. Remarkably enough, this converges on the actual solution as Δt shrinks (Arnol'd, 1973).)

Let's try to carry all this over to random functions of time $X(t)$ and $W(t)$. The integral $\int X(t)dt$ is generally not a problem — we just find a version of X

¹See especially Ethier and Kurtz (1986, Theorem 5.2.12, p. 290).

with measurable sample paths (Section 8.2). $\int X(t)dW$ is also comprehensible if dW/dt exists (almost surely). Unfortunately, we've seen that this is not the case for the Wiener process, which (as you can tell from the W) is what we'd really like to use here. So we can't approximate the integral with a sum like Eq. 18.1. But there's nothing preventing us from using one like Eq. 18.2, since that only demands increments of W . So what we would like to say is that

$$\int_{t=a}^{t=b} X(t)dW \equiv \lim_{n \rightarrow \infty} \sum_{i=1}^n X(t_{i-1})(W(t_i) - W(t_{i-1})) \quad (18.4)$$

This is a crude-but-workable approach to numerically evaluating stochastic integrals, and apparently how the first stochastic integrals were defined, back in the 1920s. Notice that it is going to make the integral a *random variable*, i.e., a measurable function of ω . Notice also that I haven't said anything yet which should lead you to believe that the limit on the right-hand side exists, in any sense, or that it is independent of the choice of partitions $a = t_0 < t_1 < \dots t_n b$. The next chapter will attempt to rectify this.

(When it comes to the SDE $dX = f(X)dt + g(X)dW$, the counterpart of Eq. 18.3 is

$$X(t + \Delta t) = X(t) + f(X(t))\Delta t + g(X(t))\Delta W \quad (18.5)$$

where $\Delta W = W(t + \Delta t) - W(t)$, and again we use linear interpolation in between the points, starting from $X(t_0) = x_0$.)

Chapter 19

Stochastic Integrals and Stochastic Differential Equations

Section 19.1 gives a rigorous construction for the Itô integral of a function with respect to a Wiener process.

Section 19.2 gives two easy examples of Itô integrals. The second one shows that there's something funny about change of variables, or if you like about the chain rule.

Section 19.3 explains how to do change of variables in a stochastic integral, also known as “Itô’s formula”.

Section 19.4 defines stochastic differential equations.

Section 19.5 sets up a more realistic model of Brownian motion, leading to an SDE called the Langevin equation, and solves it to get Ornstein-Uhlenbeck processes.

19.1 Integrals with Respect to the Wiener Process

The drill by now should be familiar: first we define integrals of step functions, then we approximate more general classes of functions by these elementary functions. We need some preliminary technicalities.

Definition 225 (Progressive Process) *A continuous-parameter stochastic process X adapted to a filtration $\{\mathcal{G}_t\}$ is progressively measurable or progressive when $X(s, \omega)$, $0 \leq s \leq t$, is always measurable with respect to $\mathcal{B}_t \times \mathcal{G}_t$, where \mathcal{B}_t is the Borel σ -field on $[0, t]$.*

If X has continuous sample paths, for instance, then it is progressive.

Definition 226 (Non-anticipating filtrations, processes) Let W be a standard Wiener process, $\{\mathcal{F}_t\}$ the right-continuous completion of the natural filtration of W , and \mathcal{G} any σ -field independent of $\{\mathcal{F}_t\}$. Then the non-anticipating filtrations are the ones of the form $\sigma(\mathcal{F}_t \cup \mathcal{G})$, $0 \leq t < \infty$. A stochastic process X is non-anticipating if it is adapted to some non-anticipating filtration.

The idea of the definition is that if X is non-anticipating, we allow it to depend on the history of W , and possibly some extra, independent random stuff, but none of that extra information is of any use in predicting the future development of W , since it's independent.

Definition 227 (Elementary process) A progressive, non-anticipating process X is elementary if there exist an increasing sequence of times t_i , starting at zero and tending to infinity, such that $X(t) = X(t_n)$ if $t \in [t_n, t_{n+1})$, i.e., if X is a step-function of time.

Remark: It is sometimes convenient to allow the break-points of the elementary process to be optional random times. We won't need this for our purposes, however.

Definition 228 (Mean square integrable) A random process X is mean-square-integrable from a to b if $\mathbf{E} \left[\int_a^b X^2(t) dt \right]$ is finite. The class of all such processes will be written $\mathcal{S}_2[a, b]$.

Notice that if X is bounded on $[a, b]$, in the sense that $|X(t)| \leq M$ with probability 1 for all $a \leq t \leq b$, then X is square-integrable from a to b .

Definition 229 (\mathcal{S}_2 norm) The norm of a process $X \in \mathcal{S}_2[a, b]$ is its root-mean-square time integral:

$$\|X\|_{\mathcal{S}_2} \equiv \left| \mathbf{E} \left[\int_a^b X^2(t) dt \right] \right|^{1/2} \quad (19.1)$$

Proposition 230 ($\|\cdot\|_{\mathcal{S}_2}$ is a norm) $\|\cdot\|_{\mathcal{S}_2}$ is a semi-norm on $\mathcal{S}_2[a, b]$; it is a full norm if processes such that $X(t) - Y(t) = 0$ a.s., for Lebesgue-almost-all t , are identified. Like any norm, it induces a metric on $\mathcal{S}_2[a, b]$, and by "a limit in \mathcal{S}_2 " we will mean a limit with respect to this metric. As a normed space, it is complete, i.e. every Cauchy-convergent sequence has a limit in the \mathcal{S}_2 sense.

PROOF: Recall that a semi-norm is a function from a vector space to the real numbers such that, for any vector X and any scalar a , $\|aX\| = |a|\|X\|$, and, for any two vectors X and Y , $\|X + Y\| \leq \|X\| + \|Y\|$. The root-mean-square time integral $\|X\|_{\mathcal{S}_2}$ clearly has both properties. To be a norm and not just a semi-norm, we need in addition that $\|X\| = 0$ if and only if $X = 0$. This is not true for random processes, because the process which is zero at irrational times $t \in [a, b]$ but 1 at rational times in the interval also has semi-norm 0. However,

by identifying two processes X and Y if $X(t) - Y(t) = 0$ a.s. for almost all t , we get a norm. This is exactly analogous to the way the L_2 norm for random variables is really only a norm on equivalence classes of variables which are equal almost always. The proof that the space with this norm is complete is exactly analogous to the proof that the L_2 space of random variables is complete, see e.g. Lemma 1.31 in Kallenberg (2002, p. 15). \square

Definition 231 (Itô integral of an elementary process) *If X is an elementary, progressive, non-anticipative process, square-integrable from a to b , then its Itô integral from a to b is*

$$\int_a^b X(t) dW \equiv \sum_{i \geq 0} X(t_i) (W(t_{i+1}) - W(t_i)) \quad (19.2)$$

where the t_i are as in Definition 227, truncated below by a and above by b .

Notice that this is basically a Riemann-Stieltjes integral. It's a random variable, but we don't have to worry about the existence of a limit. Now we set about approximating more general sorts of processes by elementary processes.

Lemma 232 (Approximation of Bounded, Continuous Processes by Elementary Processes) *Suppose X is progressive, non-anticipative, bounded on $[a, b]$, and has continuous sample paths. Then there exist bounded elementary processes X_n , Itô-integrable on $[a, b]$, such that X is the $\mathcal{S}_2[a, b]$ limit of X_n , i.e.,*

$$\lim_{n \rightarrow \infty} \|X - X_n\|_{\mathcal{S}_2} = 0 \quad (19.3)$$

$$\lim_{n \rightarrow \infty} \mathbf{E} \left[\int_a^b (X - X_n)^2 dt \right] = 0 \quad (19.4)$$

PROOF: Set

$$X_n(t) \equiv \sum_{i=0}^{\infty} X(i2^{-n}) \mathbf{1}_{[i/2^n, (i+1)/2^n)}(t) \quad (19.5)$$

This is clearly elementary, bounded and square-integrable on $[a, b]$. Moreover, for fixed ω , $\int_a^b (X(t, \omega) - X_n(t, \omega))^2 dt \rightarrow 0$, since $X(t, \omega)$ is continuous. So the expectation of the time-integral goes to zero by bounded convergence. \square

Remark: There is nothing special about using intervals of length 2^{-n} . Any division of $[a, b]$ into sub-intervals would do, provided the width of the largest sub-interval shrinks to zero.

Lemma 233 (Approximation by of Bounded Processes by Bounded, Continuous Processes) *Suppose X is progressive, non-anticipative, and bounded on $[a, b]$. Then there exist progressive, non-anticipative processes X_n which are bounded and continuous on $[a, b]$, and have X as their \mathcal{S}_2 limit,*

$$\lim_{n \rightarrow \infty} \|X - X_n\|_{\mathcal{S}_2} = 0 \quad (19.6)$$

PROOF: Let M be the bound on the absolute value of X . For each n , pick a probability density $f_n(t)$ on \mathbb{R} whose support is confined to the interval $(-1/n, 0)$. Set

$$X_n(t) \equiv \int_0^t f_n(s-t)X(s)ds \quad (19.7)$$

$X_n(t)$ is then a sort of moving average of X , over the interval $(t-1/n, t)$. Clearly, it's continuous, bounded, progressively measurable, and non-anticipative. Moreover, for each ω ,

$$\lim_{n \rightarrow \infty} \int_a^b (X_n(t, \omega) - X(t, \omega))^2 dt = 0 \quad (19.8)$$

because of the way we've set up f_n and X_n . By bounded convergence, this implies

$$\lim_{n \rightarrow \infty} \mathbf{E} \left[\int_a^b (X - X_n)^2 dt \right] = 0 \quad (19.9)$$

which is equivalent to Eq. 19.6. \square

Lemma 234 (Approximation of Square-Integrable Processes by Bounded Processes) *Suppose X is progressive, non-anticipative, and square-integrable on $[a, b]$. Then there exist a sequence of random processes X_n which are progressive, non-anticipative and bounded on $[a, b]$, which have X as their limit in \mathcal{S}_2 .*

PROOF: Set $X_n(t) = (-n \vee X(t)) \wedge n$. This has the desired properties, and the result follows from dominated (not bounded!) convergence. \square

Lemma 235 (Approximation of Square-Integrable Processes by Elementary Processes) *Suppose X is progressive, non-anticipative, and square-integrable on $[a, b]$. Then there exist a sequence of bounded elementary processes X_n with X as their limit in \mathcal{S}_2 .*

PROOF: Combine Lemmas 232, 233 and 234. \square

This lemma gets its force from the following result.

Lemma 236 (Itô Isometry for Elementary Processes) *Suppose X is as in Definition 231, and in addition bounded on $[a, b]$. Then*

$$\mathbf{E} \left[\left(\int_a^b X(t) dW \right)^2 \right] = \mathbf{E} \left[\int_a^b X^2(t) dt \right] = \|X\|_{\mathcal{S}_2}^2 \quad (19.10)$$

PROOF: Set $\Delta W_i = W(t_{i+1}) - W(t_i)$. Notice that ΔW_j is independent of $X(t_i)X(t_j)\Delta W_i$ if $i < j$, because of the non-anticipation properties of X . On the other hand, $\mathbf{E}[(\Delta W_i)^2] = t_{i+1} - t_i$, by the linear variance of the increments of W . So

$$\mathbf{E}[X(t_i)X(t_j)\Delta W_j\Delta W_i] = \mathbf{E}[X^2(t_i)](t_{i+1} - t_i)\delta_{ij} \quad (19.11)$$

Substituting Eq. 19.2 into the left-hand side of Eq. 19.10,

$$\mathbf{E} \left[\left(\int_a^b X(t) dW \right)^2 \right] = \mathbf{E} \left[\sum_{i,j} X(t_i) X(t_j) \Delta W_j \Delta W_i \right] \quad (19.12)$$

$$= \sum_{i,j} \mathbf{E} [X(t_i) X(t_j) \Delta W_j \Delta W_i] \quad (19.13)$$

$$= \sum_i \mathbf{E} [X^2(t_i)] (t_{i+1} - t_i) \quad (19.14)$$

$$= \mathbf{E} \left[\sum_i X^2(t_i) (t_{i+1} - t_i) \right] \quad (19.15)$$

$$= \mathbf{E} \left[\int_a^b X^2(t) dt \right] \quad (19.16)$$

where the last step uses the fact that X^2 must also be elementary. \square

Theorem 237 (Itô Integrals of Approximating Elementary Processes Converge) *Let X and X_n be as in Lemma 235. Then the sequence $I_n(X) \equiv$*

$$\int_a^b X_n(t) dW \quad (19.17)$$

has a limit in L_2 . Moreover, this limit is the same for any such approximating sequence X_n .

PROOF: For each X_n , $I_n(X(\omega))$ is an \mathcal{S}_2 function of ω , by the fact that X_n is square-integrable and Lemma 236. Now, the X_n are converging on X , in the sense that

$$\|X - X_n\|_{\mathcal{S}_2} \rightarrow 0 \quad (19.18)$$

Since (Proposition 230) the space $\mathcal{S}_2[a, b]$ is complete, every convergent sequence is also Cauchy, and so, for every $\epsilon > 0$, there exists an n such that

$$\|X_{n+k} - X_n\|_{\mathcal{S}_2} < \epsilon \quad (19.19)$$

for every positive k . Since X_n and X_{n+k} are both elementary processes, their difference is also elementary, and we can apply Lemma 236 to it. That is, for every $\epsilon > 0$, there is an n such that

$$\mathbf{E} \left[\left(\int_a^b (X_{n+k}(t) - X_n(t)) dW \right)^2 \right] < \epsilon \quad (19.20)$$

for all k . (Chose the ϵ in Eq. 19.19 to be the square root of the ϵ in Eq. 19.20.) But this is to say that $I_n(X)$ is a Cauchy sequence in L_2 , therefore it has a limit, which is also in L_2 (because L_2 is complete). If Y_n is another sequence

of approximations of X by elementary processes, parallel arguments show that the Itô integrals of Y_n are a Cauchy sequence, and that for every $\epsilon > 0$, there exist m and n such that $\|X_{n+k} - Y_{m+l}\|_{S_2} \leq \epsilon$, hence the integrals of Y_n must be converging on the same limit as the integrals of X_n . \square

Definition 238 (Itô integral) Let X be progressive, non-anticipative and square-integrable on $[a, b]$. Then its Itô integral is

$$\int_a^b X(t) dW \equiv \lim_n \int_a^b X_n(t) dW \quad (19.21)$$

taking the limit in L_2 , with X_n as in Lemma 235. We will say that X is Itô-integrable on $[a, b]$.

Corollary 239 (The Itô isometry) Eq. 19.10 holds for all Itô-integrable X .

PROOF: Obvious from the approximation by elementary processes and Lemma 236.

This would be a good time to do Exercises 19.1, 19.2 and 19.3.

19.2 Some Easy Stochastic Integrals, with a Moral

19.2.1 $\int dW$

Let's start with the easiest possible stochastic integral:

$$\int_a^b dW \quad (19.22)$$

i.e., the Itô integral of the function which is always 1, $\mathbf{1}_{\mathbb{R}^+}(t)$. If this is any kind of integral at all, it should be W — more exactly, because this is a definite integral, we want $\int_a^b dW = W(b) - W(a)$. Mercifully, this works. Pick any set of time-points t_i we like, and treat 1 as an elementary function with those times as its break-points. Then, using our definition of the Itô integral for elementary functions,

$$\int_a^b dW = \sum_{t_i} W(t_{i+1}) - W(t_i) \quad (19.23)$$

$$= W(b) - W(a) \quad (19.24)$$

as required. (This would be a good time to convince yourself that adding extra break-points to an elementary function doesn't change its integral (Exercise 19.5.)

19.2.2 $\int W dW$

Tradition dictates that the next example be $\int W dW$. First, we should convince ourselves that $W(t)$ is Itô-integrable: it's clearly measurable and non-anticipative, but is it square-integrable? Yes; by Fubini's theorem,

$$\mathbf{E} \left[\int_0^t W^2(s) ds \right] = \int_0^t \mathbf{E} [W^2(s)] ds \quad (19.25)$$

$$= \int_0^t s ds \quad (19.26)$$

which is clearly finite on finite intervals $[0, t]$. So, this integral should exist. Now, if the ordinary rules for change of variables held — equivalent, if the chain-rule worked the usual way — we could say that $W dW = \frac{1}{2} d(W^2)$, so $\int W dW = \frac{1}{2} \int dW^2$, and we'd expect $\int_0^t W dW = \frac{1}{2} W^2(t)$. But, alas, this can't be right. To see why, take the expectation: it'd be $\frac{1}{2}t$. But we know that it has to be zero, and it has to be a martingale in t , and this is neither. A bone-head would try to fix this by subtracting off the non-martingale part, i.e., a bone-head would guess that $\int_0^t W dW = \frac{1}{2} W^2(t) - t/2$. Annoyingly, in this case the bone-head is correct. The demonstration is fundamentally straightforward, if somewhat long-winded.

To begin, we need to approximate W by elementary functions. For each n , let $t_i = i \frac{t}{2^n}$, $0 \leq i \leq 2^n - 1$. Set $\phi_n(t) = \sum_{i=0}^{2^n-1} W(t_i) \mathbf{1}_{[t_i, t_{i+1})}$. Let's check that this converges to $W(t)$ as $n \rightarrow \infty$:

$$\mathbf{E} \left[\int_0^t (\phi_n(s) - W(s))^2 ds \right] = \mathbf{E} \left[\sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} (W(t_i) - W(s))^2 ds \right] \quad (19.27)$$

$$= \sum_{i=0}^{2^n-1} \mathbf{E} \left[\int_{t_i}^{t_{i+1}} (W(t_i) - W(s))^2 ds \right] \quad (19.28)$$

$$= \sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} \mathbf{E} \left[(W(t_i) - W(s))^2 \right] d\xi \quad (19.29)$$

$$= \sum_{i=0}^{2^n-1} \int_{t_i}^{t_{i+1}} (s - t_i) ds \quad (19.30)$$

$$= \sum_{i=0}^{2^n-1} \int_0^{2^{-n}} s ds \quad (19.31)$$

$$= \sum_{i=0}^{2^n-1} \left[\frac{t^2}{2} \right]_0^{2^{-n}} \quad (19.32)$$

$$= \sum_{i=0}^{2^n-1} 2^{-2n-1} \quad (19.33)$$

$$= 2^{-n-1} \quad (19.34)$$

which $\rightarrow 0$ as $n \rightarrow \infty$. Hence

$$\int_0^t W(s) dW = \lim_n \int_0^t \phi_n(s) dW \quad (19.35)$$

$$= \lim_n \sum_{i=0}^{2^n-1} W(t_i)(W(t_{i+1}) - W(t_i)) \quad (19.36)$$

$$= \lim_n \sum_{i=0}^{2^n-1} W(t_i) \Delta W(t_i) \quad (19.37)$$

where $\Delta W(t_i) \equiv W(t_{i+1}) - W(t_i)$, because I'm getting tired of writing both subscripts. Define $\Delta W^2(t_i)$ similarly. Since $W(0) = 0 = W^2(0)$, we have that

$$W(t) = \sum_i \Delta W(t_i) \quad (19.38)$$

$$W^2(t) = \sum_i \Delta W^2(t_i) \quad (19.39)$$

Now let's re-write ΔW^2 in such a way that we get a $W\Delta W$ term, which is what we want to evaluate our integral.

$$\Delta W^2(t_i) = W^2(t_{i+1}) - W^2(t_i) \quad (19.40)$$

$$= (\Delta W(t_i) + W(t_i))^2 - W^2(t_i) \quad (19.41)$$

$$= (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) + W^2(t_i) - W^2(t_i) \quad (19.42)$$

$$= (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) \quad (19.43)$$

This looks promising, because it's got $W\Delta W$ in it. Plugging in to Eq. 19.39,

$$W^2(t) = \sum_i (\Delta W(t_i))^2 + 2W(t_i)\Delta W(t_i) \quad (19.44)$$

$$\sum_i W(t_i)\Delta W(t_i) = \frac{1}{2}W^2(t) - \frac{1}{2}\sum_i (\Delta W(t_i))^2 \quad (19.45)$$

Now, it is possible to show (Exercise 19.4) that

$$\lim_n \sum_{i=0}^{2^n-1} (\Delta W(t_i))^2 = t \quad (19.46)$$

in L_2 , so we have that

$$\int_0^t W(s) dW = \lim_n \sum_{i=0}^{2^n-1} W(t_i) \Delta W(t_i) \quad (19.47)$$

$$= \frac{1}{2}W^2(t) - \lim_n \sum_{i=0}^{2^n-1} (\Delta W(t_i))^2 \quad (19.48)$$

$$= \frac{1}{2}W^2(t) - \frac{t}{2} \quad (19.49)$$

as required.

Clearly, something weird is going on here, and it would be good to get to the bottom of this. At the very least, we'd like to be able to use change of variables, so that we can find functions of stochastic integrals.

19.3 Itô's Formula

Integrating $\int W dW$ has taught us two things: first, we want to avoid evaluating Itô integrals directly from the definition; and, second, there's something funny about change of variables in Itô integrals. A central result of stochastic calculus, known as *Itô's formula*, gets us around both difficulties, by showing how to write functions of stochastic integrals as, themselves, stochastic integrals.

Definition 240 (Itô Process) *If A is a non-anticipating measurable process, B is Itô-integrable, and X_0 is an L_2 random variable independent of W , then $X(t) = X_0 + \int_0^t A(s)ds + \int_0^t B(s)dW$ is an Itô process. This is equivalently written $dX = Adt + BdW$.*

Lemma 241 (Itô processes are non-anticipating) *Every Itô process is non-anticipating.*

PROOF: Clearly, the non-anticipating processes are closed under linear operations, so it's enough to show that the three components of any Itô process are non-anticipating. But a process which is always equal to $X_0 \perp W(t)$ is clearly non-anticipating. Similarly, since $A(t)$ is non-anticipating, $\int A(s)ds$ is too: its natural filtration is smaller than that of A , so it cannot provide more information about $W(t)$, and A is, by assumption, non-anticipating. Finally, Itô integrals are always non-anticipating, so $\int B(s)dW$ is non-anticipating. \square

Theorem 242 (Itô's Formula in One Dimension) *Suppose X is an Itô process with $dX = Adt + BdW$. Let $f(t, x) : \mathbb{R}^+ \times \mathbb{R} \mapsto \mathbb{R}$ be a function with continuous partial time derivative $\frac{\partial f}{\partial t}$, and continuous second partial space derivative, $\frac{\partial^2 f}{\partial x^2}$. Then $F(t) = f(t, X(t))$ is an Itô process, and*

$$dF = \frac{\partial f}{\partial t}(t, X(t))dt + \frac{\partial f}{\partial x}(t, X(t))dX + \frac{1}{2}B^2(t)\frac{\partial^2 f}{\partial x^2}(t, X(t))dt \quad (19.50)$$

That is,

$$F(t) - F(0) = \int_0^t \left[\frac{\partial f}{\partial t}(s, X(s)) + A(s)\frac{\partial f}{\partial x}(s, X(s)) + \frac{1}{2}B^2(s)\frac{\partial^2 f}{\partial x^2}(s, X(s)) \right] dt + \int_0^t B(s)\frac{\partial f}{\partial x}(s, X(s))dW \quad (19.51)$$

PROOF: I will suppose first of all that f , and its partial derivatives appearing in Eq. 19.50, are all bounded. (You can show that the general case of C^2

functions can be uniformly approximated by functions with bounded derivatives.) I will further suppose that A and B are elementary processes, since in the last chapter we saw how to use them to approximate general Itô-integrable functions. (If you are worried about the interaction of all these approximations and simplifications, I commend your caution, and suggest you step through the proof in the general case.)

For each n , let $t_i = i \frac{t}{2^n}$, as in the last section. Define $\Delta t_i \equiv t_{i+1} - t_i$, $\Delta X(t_i) = X(t_{i+1}) - X(t_i)$, etc. Thus

$$F(t) = f(t, X(t)) = f(0, X(0)) + \sum_{i=0}^{2^n-1} \Delta f(t_i, X(t_i)) \quad (19.52)$$

Now we'll approximate the increments of F by a Taylor expansion:

$$\begin{aligned} F(t) &= f(0, X(0)) + \sum_{i=0}^{2^n-1} \frac{\partial f}{\partial t} \Delta t_i \\ &\quad + \sum_{i=0}^{2^n-1} \frac{\partial f}{\partial x} \Delta X(t_i) \\ &\quad + \frac{1}{2} \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t^2} (\Delta t_i)^2 \\ &\quad + \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} \Delta t_i \Delta X(t_i) \\ &\quad + \frac{1}{2} \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} (\Delta X(t_i))^2 \\ &\quad + \sum_{i=0}^{2^n-1} R_i \end{aligned} \quad (19.53)$$

Because the derivatives are bounded, all the remainder terms R_i are of third order,

$$R_i = O\left((\Delta t_i)^3 + \Delta X(t_i)(\Delta t_i)^2 + (\Delta X(t_i))^2 \Delta t_i + (\Delta X(t_i))^3\right) \quad (19.54)$$

We will come back to showing that the remainders are harmless, but for now let's concentrate on the leading-order components of the Taylor expansion.

First, as $n \rightarrow \infty$,

$$\sum_{i=0}^{2^n-1} \frac{\partial f}{\partial t} \Delta t_i \rightarrow \int_0^t \frac{\partial f}{\partial t} ds \quad (19.55)$$

$$\sum_{i=0}^{2^n-1} \frac{\partial f}{\partial x} \Delta X(t_i) \rightarrow \int_0^t \frac{\partial f}{\partial x} dX \quad (19.56)$$

$$\equiv \int_0^t \frac{\partial f}{\partial x} A(s) dt + \int_0^t \frac{\partial f}{\partial x} B(s) dW \quad (19.57)$$

[You can use the definition in the last line to build up a theory of stochastic integrals with respect to arbitrary Itô processes, not just Wiener processes.]

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t^2} (\Delta t_i)^2 \rightarrow 0 \int_0^t \frac{\partial^2 f}{\partial t^2} ds = 0 \quad (19.58)$$

Next, since A and B are (by assumption) elementary,

$$\begin{aligned} \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} (\Delta X(t_i))^2 &= \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A^2(t_i) (\Delta t_i)^2 \\ &+ 2 \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i) B(t_i) \Delta t_i \Delta W(t_i) \\ &+ \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} B^2(t_i) (\Delta W(t_i))^2 \end{aligned} \quad (19.59)$$

The first term on the right-hand side, in $(\Delta t)^2$, goes to zero as n increases. Since A is square-integrable and $\frac{\partial^2 f}{\partial x^2}$ is bounded, $\sum \frac{\partial^2 f}{\partial x^2} A^2(t_i) \Delta t_i$ converges to a finite value as $\Delta t \rightarrow 0$, so multiplying by another factor Δt , as $n \rightarrow \infty$, gives zero. (This is the same argument as the one for Eq. 19.58.) Similarly for the second term, in $\Delta t \Delta X$:

$$\lim_n \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i) B(t_i) \Delta t_i \Delta W(t_i) = \lim_n \frac{t}{2^n} \int_0^t \frac{\partial^2 f}{\partial x^2} A(s) B(s) dW \quad (19.60)$$

because A and B are elementary and the partial derivative is bounded. Now apply the Itô isometry:

$$\mathbf{E} \left[\left(\frac{t}{2^n} \int_0^t \frac{\partial^2 f}{\partial x^2} A(s) B(s) dW \right)^2 \right] = \frac{t^2}{2^{2n}} \mathbf{E} \left[\int_0^t \left(\frac{\partial^2 f}{\partial x^2} \right)^2 A^2(s) B^2(s) ds \right]$$

The time-integral on the right-hand side is finite, since A and B are square-integrable and the partial derivative is bounded, and so, as n grows, both sides go to zero. But this means that, in L_2 ,

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} A(t_i) B(t_i) \Delta t_i \Delta W(t_i) \rightarrow 0 \quad (19.61)$$

The third term, in $(\Delta X)^2$, does *not* vanish, but rather converges in L_2 to a time integral:

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial x^2} B^2(t_i) (\Delta W(t_i))^2 \rightarrow \int_0^t \frac{\partial^2 f}{\partial x^2} B^2(s) ds \quad (19.62)$$

You will prove this in Exercise 19.4.

The mixed partial derivative term has no counterpart in Itô's formula, so it needs to go away.

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} \Delta t_i \Delta X(t_i) = \sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} \left[A(t_i)(\Delta t_i)^2 + B(t_i)\Delta t_i \Delta W(t_i) \right] \quad (19.63)$$

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} A(t_i)(\Delta t_i)^2 \rightarrow 0 \quad (19.64)$$

$$\sum_{i=0}^{2^n-1} \frac{\partial^2 f}{\partial t \partial x} B(t_i)\Delta t_i \Delta W(t_i) \rightarrow 0 \quad (19.65)$$

where the argument for Eq. 19.65 is the same as that for Eq. 19.58, while that for Eq. 19.65 follows the pattern of Eq. 19.61.

Let us, as promised, dispose of the remainder terms, given by Eq. 19.54, re-stated here for convenience:

$$R_i = O\left((\Delta t)^3 + \Delta X(\Delta t)^2 + (\Delta X)^2 \Delta t + (\Delta X)^3\right) \quad (19.66)$$

Taking $\Delta X = A\Delta t + B\Delta W$, expanding the powers of ΔX , and using the fact that everything is inside a big O to let us group together terms with the same powers of Δt and ΔW , we get

$$R_i = O\left((\Delta t)^3 + \Delta W(\Delta t)^2 + (\Delta W)^2 \Delta t + (\Delta W)^3\right) \quad (19.67)$$

From our previous uses of Exercise 19.4, it's clear that in the limit $(\Delta W)^2$ terms will behave like Δt terms, so

$$R_i = O\left((\Delta t)^3 + \Delta W(\Delta t)^2 + (\Delta t)^2 + \Delta W \Delta t\right) \quad (19.68)$$

Now, by our previous arguments, the sum of terms which are $O((\Delta t)^2) \rightarrow 0$, so the first three terms all go to zero; similarly we have seen that a sum of terms which are $O(\Delta W \Delta t) \rightarrow 0$. We may conclude that the sum of the remainder terms goes to 0, in L_2 , as n increases.

Putting everything together, we have that

$$F(t) - F(0) = \int_0^t \left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} A + \frac{1}{2} B^2 \frac{\partial^2 f}{\partial x^2} \right] dt + \int_0^t \frac{\partial f}{\partial x} B dW \quad (19.69)$$

exactly as required. This completes the proof, under the stated restrictions on f , A and B ; approximation arguments extend the result to the general case. \square

Remark 1. Our manipulations in the course of the proof are often summarized in the following multiplication rules for differentials: $dt dt = 0$, $dW dt = 0$, $dt dW = 0$, and, most important of all,

$$dW dW = dt$$

This last is of course related to the linear growth of the variance of the increments of the Wiener process. If we used a different driving noise term, it would be replaced by the corresponding rule for the growth of that noise's variance.

Remark 2. Re-arranging Itô's formula a little yields

$$dF = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dX + \frac{1}{2} B^2 \frac{\partial^2 f}{\partial x^2} dt \quad (19.70)$$

The first two terms are what we expect from the ordinary rules of calculus; it's the third term which is new and strange. Notice that it disappears if $\frac{\partial^2 f}{\partial x^2} = 0$. When we come to stochastic differential equations, this will correspond to state-independent noise.

Remark 3. One implication of Itô's formula is that *Itô processes are closed under the application of C^2 mappings.*

Example 243 (Section 19.2.2 summarized) *The integral $\int W dW$ is now trivial. Let $X(t) = W(t)$ (by setting $A = 0$, $B = 1$ in the definition of an Itô process), and $f(t, x) = x^2/2$. Applying Itô's formula,*

$$dF = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dW + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} dt \quad (19.71)$$

$$\frac{1}{2} dW^2 = W dW + \frac{1}{2} dt \quad (19.72)$$

$$\frac{1}{2} \int dW^2 = \int W dW + \frac{1}{2} \int dt \quad (19.73)$$

$$\int_0^t W(s) dW = \frac{1}{2} W^2(t) - \frac{t}{2} \quad (19.74)$$

All of this extends naturally to higher dimensions.

Definition 244 (Multidimensional Itô Process) *Let A be an n -dimensional vector of non-anticipating processes, B an $n \times m$ matrix of Itô-integrable processes, and W an m -dimensional Wiener process. Then*

$$X(t) = X(0) + \int_0^t A(s) ds + \int_0^t B(s) dW \quad (19.75)$$

$$dX = A(t) dt + B(t) dW \quad (19.76)$$

is an n -dimensional Itô process.

Theorem 245 (Itô's Formula in Multiple Dimensions) *Let $X(t)$ be an n -dimensional Itô process, and let $f(t, x) : \mathbb{R}^+ \times \mathbb{R}^n \mapsto \mathbb{R}^m$ have a continuous partial time derivative and continuous second partial space derivatives. Then $F(t) = f(t, X(t))$ is an m -dimensional Itô process, whose k^{th} component F_k is given by*

$$dF_k = \frac{\partial g_k}{\partial t} dt + \frac{\partial g_k}{\partial x_i} dX_i + \frac{1}{2} \frac{\partial^2 g_k}{\partial X_i \partial X_j} dX_i dX_j \quad (19.77)$$

summing over repeated indices, with the understanding that $dW_i dW_j = \delta_{ij} dt$, $dW_i dt = dt dW_i = dt dt = 0$.

PROOF: Entirely parallel to the one-dimensional case, only with even more algebra. \square

It is also possible to define Wiener processes and stochastic integrals on arbitrary curved manifolds, but this would take us way, way too far afield.

19.3.1 Stratonovich Integrals

It is possible to make the extra term in Eq. 19.70 go away, and have stochastic differentials which work just like the ordinary ones. This corresponds to making stochastic integrals limits of sums of the form

$$\sum_i X\left(\frac{t_{i+1} + t_i}{2}\right) \Delta W(t_i)$$

rather than the Itô sums we are using,

$$\sum_i X(t_i) \Delta W(t_i)$$

That is, we could evade the Itô formula if we evaluated our test function in the middle of intervals, rather than at their beginning. This leads to what are called *Stratonovich integrals*. However, while Stratonovich integrals give simpler change-of-variable formulas, they have many other inconveniences: they are not martingales, for instance, and the nice connections between the form of an SDE and its generator, which we will see and use in the next chapter, go away. Fortunately, every Stratonovich SDE can be converted into an Itô SDE, and vice versa, by adding or subtracting the appropriate noise term.

19.3.2 Martingale Representation

One property of the Itô integral is that it is always a continuous square-integrable martingale. Remarkably enough, the converse is also true. In the interest of time, I omit the proof of the following theorem; there is one using only tools we've seen so far in Øksendal (1995, ch. 4), but it builds up from some auxiliary results.

Theorem 246 (Representation of Martingales as Stochastic Integrals (Martingale Representation Theorem)) *Let $M(t)$ be a continuous martingale, with $\mathbf{E}[M^2(t)] < \infty$ for all $t \geq 0$. Then there exists a unique process $M'(t)$, Itô-integrable for all finite positive t , such that*

$$M(t) = \mathbf{E}[M(0)] + \int_0^t M'(t) dW \text{ a.s.} \quad (19.78)$$

A direct consequence of this and Itô's formula is the promised martingale characterization of the Wiener process.

Theorem 247 (Martingale Characterization of the Wiener Process)
If M is continuous martingale, and $M^2 - t$ is also a martingale, then M is a Wiener process.

PROOF: Since M is a continuous martingale, by Theorem 246 there is a unique process M' such that $dM = M'dw$. That is, M is an Itô process with $A = 0$, $B = M'$. Now define $f(t, x) = x^2 - t$. Clearly, $M^2(t) - t = f(t, M) \equiv Y(t)$. Since f is smooth, we can apply Itô's formula (Theorem 242):

$$dY = \frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dM + \frac{1}{2} B^2 \frac{\partial^2 f}{\partial x^2} dt \quad (19.79)$$

$$= -dt + 2MM'dW + (M')^2 dt \quad (19.80)$$

Since Y is itself a martingale, $dY = Y'dW$, and this is the unique representation as an Itô process, hence the dt terms must cancel. Therefore

$$0 = -1 + (M'(t))^2 \quad (19.81)$$

$$\pm 1 = M'(t) \quad (19.82)$$

Since $-W$ is also a Wiener process, it follows that $M \stackrel{d}{=} W$ (plus a possible additive). \square

19.4 Stochastic Differential Equations

Definition 248 (Stochastic Differential Equation, Solutions) *Let $a(x) : \mathbb{R}^n \mapsto \mathbb{R}^n$ and $b(x) : \mathbb{R}^n \mapsto \mathbb{R}^{nm}$ be measurable functions (vector and matrix valued, respectively), W an m -dimensional Wiener process, and X_0 an L_2 random variable in \mathbb{R}^n , independent of W . Then an \mathbb{R}^n -valued stochastic process X on \mathbb{R}^+ is a solution to the autonomous stochastic differential equation*

$$dX = a(X)dt + b(X)dW, \quad X(0) = X_0 \quad (19.83)$$

when, with probability 1, it is equal to the corresponding Itô process,

$$X(t) = X_0 + \int_0^t a(X(s))ds + \int_0^t b(X(s))dW \text{ a.s.} \quad (19.84)$$

The a term is called the drift, and the b term the diffusion.

Remark 1: A given process X can fail to be a solution either because it happens not to agree with Eq. 19.84, or, perhaps more seriously, because the integrals on the right-hand side don't even exist. This can, in particular, happen if $b(X(t))$ is anticipating. For a *fixed* choice of Wiener process, there are circumstances where otherwise reasonable SDEs have no solution, for basically

this reason — the Wiener process is constructed in such a way that the class of Itô processes is impoverished. This leads to the idea of a *weak solution* to Eq. 19.83, which is a *pair* X, W such that W is a Wiener process, with respect to the appropriate filtration, and X then is given by Eq. 19.84. I will avoid weak solutions in what follows.

Remark 2: In a non-autonomous SDE, the coefficients would be explicit functions of time, $a(t, X)dt + b(t, X)dW$. The usual trick for dealing with non-autonomous n -dimensional ODEs is turn them into autonomous $n + 1$ -dimensional ODEs, making $x_{n+1} = t$ by decreeing that $x_{n+1}(t_0) = t_0$, $x'_{n+1} = 1$ (Arnol'd, 1973). This works for SDEs, too: add time as an extra variable with constant drift 1 and constant diffusion 0. Without loss of generality, therefore, I'll only consider autonomous SDEs.

Let's now prove the existence of unique solutions to SDEs. First, recall how we do this for ordinary differential equations. There are several approaches, most of which carry over to SDEs, but one of the most elegant is the “method of successive approximations”, or “Picard's method” (Arnol'd, 1973, SS30–31)). To construct a solution to $dx/dt = f(x)$, $x(0) = x_0$, this approach uses functions $x_n(t)$, with $x_{n+1}(t) = x_0 + \int_0^t f(x_n(s))ds$, starting with $x_0(t) = x_0$. That is, there is an operator P such that $x_{n+1} = Px_n$, and x solves the ODE iff it is a fixed point of the operator. Step 1 is to show that the sequence x_n is Cauchy on finite intervals $[0, T]$. Step 2 uses the fact that the space of continuous functions is complete, with the topology of uniform convergence of compact sets — which, for \mathbb{R}^+ , is the same as uniform convergence on finite intervals. So, x_n has a limit. Step 3 is to show that the limit point must be a fixed point of P , that is, a solution. Uniqueness is proved by showing that there cannot be more than one fixed point.

Before plunging in to the proof, we need some lemmas: an algebraic triviality, a maximal inequality for martingales, a consequent maximal inequality for Itô processes, and an inequality from real analysis about integral equations.

Lemma 249 (A Quadratic Inequality) *For any real numbers a and b , $(a + b)^2 \leq 2a^2 + 2b^2$.*

PROOF: No matter what a and b are, a^2 , b^2 , and $(a - b)^2$ are non-negative, so

$$(a - b)^2 \geq 0 \quad (19.85)$$

$$a^2 + b^2 - 2ab \geq 0 \quad (19.86)$$

$$a^2 + b^2 \geq 2ab \quad (19.87)$$

$$2a^2 + 2b^2 \geq a^2 + 2ab + b^2 = (a + b)^2 \quad (19.88)$$

□

Definition 250 (Maximum Process) *Given a stochastic process $X(t)$, we define its maximum process $X^*(t)$ as $\sup_{0 \leq s \leq t} |X(s)|$.*

Remark: Example 79 was of course designed with malice aforethought.

Definition 251 (The Space $\mathcal{QM}(T)$) Let $\mathcal{QM}(T)$, $T > 0$, be the space of all non-anticipating processes, square-integrable on $[0, T]$, with norm $\|X\|_{\mathcal{QM}(T)} \equiv \|X^*(T)\|_2$.

Technically, this is only a norm on equivalence classes of processes, where the equivalence relation is “is a version of”. You may make that amendment mentally as you read what follows.

Lemma 252 (Completeness of $\mathcal{QM}(T)$) $\mathcal{QM}(T)$ is a complete normed space for each T .

PROOF: Identical to the usual proof that L_p spaces are complete, see, e.g., Lemma 1.31 of Kallenberg (2002, p. 15). \square

Proposition 253 (Doob’s Martingale Inequalities) If $M(t)$ is a continuous martingale, then, for all $p \geq 1$, $t \geq 0$ and $\epsilon > 0$,

$$\mathbb{P}(M^*(t) \geq \epsilon) \leq \frac{\mathbf{E}[|M(t)|^p]}{\epsilon^p} \quad (19.89)$$

$$\|M^*(t)\|_p \leq q \|M(t)\|_p \quad (19.90)$$

where $q^{-1} + p^{-1} = 1$. In particular, for $p = q = 2$,

$$\mathbf{E}[(M^*(t))^2] \leq 4\mathbf{E}[M^2(t)]$$

PROOF: See Propositions 7.15 and 7.16 in Kallenberg (pp. 128 and 129). \square

These can be thought of as versions of the Markov inequality, only for martingales. They accordingly get used *all the time*.

Lemma 254 (A Maximal Inequality for Itô Processes) Let $X(t)$ be an Itô process, $X(t) = X_0 + \int_0^t A(s)ds + \int_0^t B(s)dW$. Then there exists a constant C , depending only on T , such that, for all $t \in [0, T]$,

$$\|X\|_{\mathcal{QM}(t)}^2 \leq C \left(\mathbf{E}[X_0^2] + \mathbf{E} \left[\int_0^t A^2(s) + B^2(s)ds \right] \right) \quad (19.91)$$

PROOF: Clearly,

$$X^*(t) \leq |X_0| + \int_0^t |A(s)|ds + \sup_{0 \leq s \leq t} \left| \int_0^s B(s)dW \right| \quad (19.92)$$

$$(X^*(t))^2 \leq 2X_0^2 + 2 \left(\int_0^t |A(s)|ds \right)^2 + 2 \left(\sup_{0 \leq s \leq t} \left| \int_0^s B(s')dW \right| \right)^2 \quad (19.93)$$

by Lemma 249. By Jensen’s inequality¹,

$$\left(\int_0^t |A(s)|ds \right)^2 \leq t \int_0^t A^2(s)ds \quad (19.94)$$

¹Remember that Lebesgue measure isn’t a probability measure on $[0, t]$, but $\frac{1}{t}ds$ is a probability measure, so we can apply Jensen’s inequality to that. This is where the t on the right-hand side will come from.

Writing $I(t)$ for $\int_0^t B(s)dW$, and noticing that it is a martingale, we have, from Doob's inequality (Proposition 253), $\mathbf{E} \left[(I^*(t))^2 \right] \leq 4\mathbf{E} [I^2(t)]$. But, from Itô's isometry (Corollary 239), $\mathbf{E} [I^2(t)] = \mathbf{E} \left[\int_0^t B^2(s)ds \right]$. Putting all the parts together, then,

$$\mathbf{E} \left[(X^*(t))^2 \right] \leq 2\mathbf{E} [X_0^2] + 2\mathbf{E} \left[t \int_0^t A^2(s)ds + \int_0^t B^2(s)ds \right] \quad (19.95)$$

and the conclusion follows, since $t \leq T$. \square

Remark: The lemma also holds for multidimensional Itô processes, and for powers greater than two (though then the Doob inequality needs to be replaced by a different one: see Rogers and Williams (2000, Ch. V, Lemma 11.5, p. 129)).

Definition 255 (Picard operator) *Given an SDE $dX = a(X)dt + b(X)dW$ with initial condition X_0 , the corresponding integral operator $P_{X_0,a,b}$ is defined for all Itô processes Y as*

$$P_{X_0,a,b}Y(t) = X_0 + \int_0^t a(Y(s))ds + \int_0^t b(Y(s))dW \quad (19.96)$$

Lemma 256 (Solutions are fixed points of the Picard operator) *Y is a solution of $dX = a(X)dt + b(X)dW$, $X(0) = X_0$, if and only if $P_{X_0,a,b}Y = Y$ a.s.*

PROOF: Obvious from the definitions. \square

Lemma 257 (A maximal inequality for Picard iterates) *If a and b are uniformly Lipschitz continuous, with constants K_a and K_B , then, for some positive D depending only on T , K_a and K_b ,*

$$\|P_{X_0,a,b}X - P_{X_0,a,b}Y\|_{\mathcal{QM}(t)}^2 \leq D \int_0^t \|X - Y\|_{\mathcal{QM}(s)}^2 ds \quad (19.97)$$

PROOF: Since the SDE is understood to be fixed, abbreviate $P_{X_0,a,b}$ by P . Let X and Y be any two Itô processes. We want to find the $\mathcal{QM}(t)$ norm of

$PX - PY$.

$$|PX(t) - PY(t)| \quad (19.98)$$

$$= \left| \int_0^t a(X(s)) - a(Y(s)) dt + \int_0^t b(X(s)) - b(Y(s)) dW \right|$$

$$\leq \int_0^t |a(X(s)) - a(Y(s))| ds + \int_0^t |b(X(s)) - b(Y(s))| dW \quad (19.99)$$

$$\leq \int_0^t K_a |X(s) - Y(s)| ds + \int_0^t K_b |X(s) - Y(s)| dW \quad (19.100)$$

$$\|PX - PY\|_{\mathcal{QM}(t)}^2 \quad (19.101)$$

$$\leq C(K_a^2 + K_b^2) \mathbf{E} \left[\int_0^t |X(s) - Y(s)|^2 ds \right]$$

$$\leq C(K_a^2 + K_b^2) t \int_0^t \|X - Y\|_{\mathcal{QM}(s)}^2 ds \quad (19.102)$$

which, as $t \leq T$, completes the proof. \square

Lemma 258 (Gronwall's Inequality) *If f is continuous function on $[0, T]$ such that $f(t) \leq c_1 + c_2 \int_0^t f(s) ds$, then $f(t) \leq c_1 e^{c_2 t}$.*

PROOF: See Kallenberg, Lemma 21.4, p. 415. \square

Theorem 259 (Existence and Uniqueness of Solutions to SDEs in One Dimension) *Let X_0 , a , b and W be as in Definition 248, and let a and b be uniformly Lipschitz continuous. Then there exists a square-integrable, non-anticipating $X(t)$ which solves $dX = a(X)dt + b(X)dW$ with initial condition X_0 , and this solution is unique (almost surely).*

PROOF: I'll first prove existence, along with square-integrability, and then uniqueness. That X is non-anticipating follows from the fact that it is an Itô process (Lemma 241). For concision, abbreviate $P_{X_0, a, b}$ by P .

As with ODEs, iteratively construct approximate solutions. Fix a $T > 0$, and, for $t \in [0, T]$, set

$$X_0(t) = X_0 \quad (19.103)$$

$$X_{n+1}(t) = PX_n(t) \quad (19.104)$$

The first step is showing that X_n is Cauchy in $\mathcal{QM}(T)$. Define $\phi_n(t) \equiv \|X_{n+1} - X_n\|_{\mathcal{QM}(t)}^2$. Notice that $\phi_n(t) = \|PX_n - PX_{n-1}\|_{\mathcal{QM}(t)}^2$, and that, for each n , $\phi_n(t)$ is non-decreasing in t (because of the supremum embedded in

its definition). So, using Lemma 257,

$$\phi_n(t) \leq D \int_0^t \|X_n - X_{n-1}\|_{\mathcal{QM}(s)}^2 ds \quad (19.105)$$

$$\leq D \int_0^t \phi_{n-1}(s) ds \quad (19.106)$$

$$\leq D \int_0^t \phi_{n-1}(t) ds \quad (19.107)$$

$$= Dt\phi_{n-1}(0) \quad (19.108)$$

$$\leq \frac{D^n t^n}{n!} \phi_0(t) \quad (19.109)$$

$$\leq \frac{D^n t^n}{n!} \phi_0(T) \quad (19.110)$$

Since, for any constant c , $c^n/n! \rightarrow 0$, to get the successive approximations to be Cauchy, we just need to show that $\phi_0(T)$ is finite, using Lemma 254.

$$\phi_0(T) = \|P_{X_0, a, b} X_0 - X_0\|_{\mathcal{QM}(T)}^2 \quad (19.111)$$

$$= \left\| \int_0^t a(X_0) ds + \int_0^t b(X_0) dW \right\|_{\mathcal{QM}(T)}^2 \quad (19.112)$$

$$\leq C\mathbf{E} \left[\int_0^T a^2(X_0) + b^2(X_0) ds \right] \quad (19.113)$$

$$\leq C\mathbf{TE} [a^2(X_0) + b^2(X_0)] \quad (19.114)$$

Because a and b are Lipschitz, this will be finite if X_0 has a finite second moment, which, by assumption, it does. So X_n is a Cauchy sequence in $\mathcal{QM}(T)$, which is a complete space, so X_n has a limit in $\mathcal{QM}(T)$, call it X :

$$\lim_{n \rightarrow \infty} \|X - X_n\|_{\mathcal{QM}(T)} = 0 \quad (19.115)$$

The next step is to show that X is a fixed point of the operator P . This is true because PX is also a limit of the sequence X_n .

$$\|PX - X_{n+1}\|_{\mathcal{QM}(T)}^2 = \|PX - PX_n\|_{\mathcal{QM}(T)}^2 \quad (19.116)$$

$$\leq DT \|X - X_n\|_{\mathcal{QM}(T)}^2 \quad (19.117)$$

which $\rightarrow 0$ as $n \rightarrow \infty$ (by Eq. 19.115). So PX is the limit of X_{n+1} , which means it is the limit of X_n , and, since X is also a limit of X_n and limits are unique, $PX = X$. Thus, by Lemma 256, X is a solution.

To prove uniqueness, suppose that there were another solution, Y . By Lemma 256, $PY = Y$ as well. So, with Lemma 257,

$$\|X - Y\|_{\mathcal{QM}(t)}^2 = \|PX - PY\|_{\mathcal{QM}(t)}^2 \quad (19.118)$$

$$\leq D \int_0^t \|X - Y\|_{\mathcal{QM}(s)}^2 ds \quad (19.119)$$

So, from Gronwall's inequality (Lemma 258), we have that $\|X - Y\|_{\mathcal{QM}(t)} \leq 0$ for all t , implying that $X(t) = Y(t)$ a.s. \square

Remark: For an alternative approach, based on Euler's method (rather than Picard's), see Fristedt and Gray (1997, §33.4). It has a certain appeal, but it also involves some uglier calculations. For a side-by-side comparison of the two methods, see Lasota and Mackey (1994).

Theorem 260 (Existence and Uniqueness for Multidimensional SDEs)

Theorem 259 also holds for multi-dimensional stochastic differential equations, provided a and b are uniformly Lipschitz in the appropriate Euclidean norms.

PROOF: Entirely parallel to the one-dimensional case, only with more algebra. \square

The conditions on the coefficients can be reduced to something like “locally Lipschitz up to a stopping time”, but it does not seem profitable to pursue this here. See Rogers and Williams (2000, Ch. V, Sec. 12).

19.5 Brownian Motion, the Langevin Equation, and Ornstein-Uhlenbeck Processes

The Wiener process is not a realistic model of Brownian motion, because it implies that Brownian particles do not have well-defined velocities, which is absurd. Setting up a more realistic model will eliminate this absurdity, and illustrate how SDEs can be used as models.² I will first need to summarize classical mechanics in one paragraph.

Classical mechanics starts with Newton's laws of motion. The zeroth law, implicit in everything, is that the laws of nature are differential equations in position variables with respect to time. The first law says that they are not first-order differential equations. The second law says that they are second-order differential equations. The usual trick for higher-order differential equations is to introduce supplementary variables, so that we have a higher-dimensional system of first-order differential equations. The supplementary variable here is momentum. Thus, for particle i , with mass m_i ,

$$\frac{d\vec{x}_i}{dt} = \frac{\vec{p}_i}{m_i} \quad (19.120)$$

$$\frac{d\vec{p}_i}{dt} = \frac{F(\mathbf{x}, \mathbf{p}, t)}{m_i} \quad (19.121)$$

constitute the laws of motion. All the physical content comes from specifying the force function $F(\mathbf{x}, \mathbf{p}, t)$. We will consider only autonomous systems, so we do not need to deal with forces which are explicit functions of time. Newton's

²See Selmeczi *et al.* (2006) for an account of the physical theory of Brownian motion, including some of its history and some fascinating biophysical applications. Wax (1954) collects classic papers on this and related subjects, still very much worth reading.

third law says that total momentum is conserved, when all bodies are taken into account.

Consider a large particle of (without loss of generality) mass 1, such as a pollen grain, sitting in a still fluid at thermal equilibrium. What forces act on it? One is drag. At a molecular level, this is due to the particle colliding with the molecules (mass m) of the fluid, whose average momentum is zero. This typically results in momentum being transferred from the pollen to the fluid molecules, and the amount of momentum lost by the pollen is proportional to what it had, i.e., one term in $d\vec{p}/dt$ is $-\gamma\vec{p}$. In addition, however, there will be fluctuations, which will be due to the fact that the fluid molecules are not all at rest. In fact, because the fluid is at equilibrium, the momenta of the molecules will follow a Maxwell-Boltzmann distribution,

$$f(\vec{p}_{\text{molec}}) = (2\pi mk_B T)^{-3/2} e^{-\frac{1}{2} \frac{p_{\text{molec}}^2}{mk_B T}}$$

where which is a zero-mean Gaussian with variance $mk_B T$. Tracing this through, we expect that, over short time intervals in which the pollen grain nonetheless collides with a large number of molecules, there will be a random impulse (i.e., random change in momentum) which is Gaussian, but uncorrelated over shorter sub-intervals (by the functional CLT). That is, we would like to write

$$d\vec{p} = -\gamma\vec{p}dt + DI dW \quad (19.122)$$

where D is the *diffusion constant*, I is the 3×3 identity matrix, and W of course is the standard three-dimensional Wiener process. This is known as the *Langevin equation* in the physics literature, as this model was introduced by Langevin in 1907 as a correction to Einstein's 1905 model of Brownian motion. (Of course, Langevin didn't use Wiener processes and Itô integrals, which came much later, but the spirit was the same.) If you like time-series models, you might recognize this as a continuous-time version of a mean-reverting AR(1) model, which explains why it also shows up as an interest rate model in financial theory.

We can consider each component of the Langevin equation separately, because they decouple, and solve them easily with Itô's formula:

$$d(e^{\gamma t} p) = D e^{\gamma t} dW \quad (19.123)$$

$$e^{\gamma t} p(t) = p_0 + D \int_0^t e^{\gamma s} dW \quad (19.124)$$

$$p(t) = p_0 e^{-\gamma t} + D \int_0^t e^{-\gamma(t-s)} dW \quad (19.125)$$

We will see in the next chapter a general method of proving that solutions of equations like 19.122 are Markov processes; for now, you can either take that on faith, or try to prove it yourself.

Assuming p_0 is itself Gaussian, with mean 0 and variance σ^2 , then (using Exercise 19.6), \vec{p} always has mean zero, and the covariance is

$$\text{cov}(\vec{p}(t), \vec{p}(s)) = \sigma^2 e^{-\gamma(s+t)} + \frac{D^2}{2\gamma} (e^{-\gamma|s-t|} - e^{-\gamma(s+t)}) \quad (19.126)$$

If $\sigma^2 = D^2/2\gamma$, then the covariance is a function of $|s - t|$ alone, and the process is weakly stationary. Such a solution of Eq. 19.122 is known as a *stationary Ornstein-Uhlenbeck process*. (Ornstein and Uhlenbeck provided the Wiener processes and Itô integrals.)

Weak stationarity, and the fact that the Ornstein-Uhlenbeck process is Markovian, allow us to say that the distribution $\mathcal{N}(0, D^2/2\gamma)$ is invariant. Now, if the Brownian particle began in equilibrium, we expect its energy to have a Maxwell-Boltzmann distribution, which means that its momentum has a Gaussian distribution, and the variance is (as with the fluid molecules) $k_B T$. Thus, $k_B T = D^2/2\gamma$, or $D^2 = 2\gamma k_B T$. This is an example of what the physics literature calls a *fluctuation-dissipation relation*, since one side of the equation involves the magnitude of fluctuations (the diffusion coefficient D) and the other the response to fluctuations (the frictional damping coefficient γ). Such relationships turn out to hold quite generally at or near equilibrium, and are often summarized by the saying that “systems respond to forcing just like fluctuations”. (Cf. 19.125.)

Oh, and that story I told you before about Brownian particles following Wiener processes? It’s something of a lie told to children, or at least to probability theorists, but see Exercise 19.9.

For more on the physical picture of Brownian motion, fluctuation-dissipation relations, and connections to more general thermodynamic processes in and out of equilibrium, see Keizer (1987).³

19.6 Exercises

Exercise 19.1 (Basic Properties of the Itô Integral) *Prove the following, first for elementary Itô-integrable processes, and then in general.*

1.

$$\int_a^c X(t) dW = \int_a^b X(t) dW + \int_b^c X(t) dW \quad (19.127)$$

almost surely.

2. *If c is any real constant, then, almost surely,*

$$\int_a^b (cX(t) + Y(t)) dW = c \int_a^b X dW + \int_a^b Y(t) dW \quad (19.128)$$

Exercise 19.2 (Martingale Properties of the Itô Integral) *Suppose X is Itô-integrable on $[a, b]$. Show that*

$$I_x(t) \equiv \int_a^t X(s) dW \quad (19.129)$$

$a \leq t \leq b$, is a martingale. What is $E[I_x(t)]$?

³Be warned that he perversely writes the probability of event A conditional on event B as $\mathbb{P}(B|A)$, not $\mathbb{P}(A|B)$.

Exercise 19.3 (Continuity of the Itô Integral) Show that $I_x(t)$ has (a modification with) continuous sample paths.

Exercise 19.4 (“The square of dW ”) Use the notation of Section 19.2 here.

1. Show that $\sum_i (\Delta W(t_i))^2$ converges on t (in L_2) as n grows. Hint: Show that the terms in the sum are IID, and that their variance shrinks sufficiently fast as n grows. (You will need the fourth moment of a Gaussian distribution.)
2. If $X(t)$ is measurable and non-anticipating, show that

$$\lim_n \sum_{i=0}^{2^n-1} X(t_i) (\Delta W(t_i))^2 = \int_0^t X(s) ds \quad (19.130)$$

in L_2 .

Exercise 19.5 (Itô integrals of elementary processes do not depend on the break-points) Let X and Y be two elementary processes which are versions of each other. Show that $\int_a^b X dW = \int_a^b Y dW$ a.s.

Exercise 19.6 (Itô integrals are Gaussian processes) For any fixed, non-random cadlag function f on \mathbb{R}^+ , let $I_f(t) = \int_0^t f(s) dW$.

1. Show that $\mathbf{E}[I_f(t)] = 0$ for all t .
2. Show $\text{cov}(I_f(t), I_f(s)) = \int_0^{t \wedge s} f^2(u) du$.
3. Show that $I_f(t)$ is a Gaussian process.

Exercise 19.7 (A Solvable SDE) Consider

$$dX = \frac{1}{2} X dt + \sqrt{1+X^2} dW \quad (19.131)$$

1. Show that there is a unique solution for every initial value $X(0) = x_0$.
2. It happens (you do not have to show this) that, for fixed x_0 , the the solution has the form $X(t) = \phi(W(t))$, where ϕ is a C^2 function. Use Itô's formula to find the first two derivatives of ϕ , and then solve the resulting second-order ODE to get ϕ .
3. Verify that, with the ϕ you found in the previous part, $\phi(W(t))$ solves Eq. 19.131 with initial condition $X(0) = x_0$.

Exercise 19.8 (Building Martingales from SDEs) Let X be an Itô process given by $dX = Adt + BdW$, and f any C^2 function. Use Itô's formula to prove that

$$f(X(t)) - f(X(0)) - \int_0^t \left[A \frac{\partial f}{\partial x} + \frac{1}{2} B^2 \frac{\partial^2 f}{\partial x^2} \right] dt$$

is a martingale.

Exercise 19.9 (Brownian Motion and the Ornstein-Uhlenbeck Process) Consider a Brownian particle whose momentum follows a stationary Ornstein-Uhlenbeck process, in one spatial dimension (for simplicity). Assume that its initial position $x(0)$ is fixed at the origin, and then $x(t) = \int_0^t p(t)dt$. Show that as $D \rightarrow \infty$ and $D/\gamma \rightarrow 1$, the distribution of $x(t)$ converges to a standard Wiener process. Explain why this limit is a physically reasonable one.

Exercise 19.10 (Again with the martingale characterization of the Wiener process) Try to prove Theorem 247, starting from the integral representation of $M^2 - t$ and using Itô's lemma to get the integral representation of M .

Chapter 20

More on Stochastic Differential Equations

Section 20.1 shows that the solutions of SDEs are diffusions, and how to find their generators. Our previous work on Feller processes and martingale problems pays off here. Some other basic properties of solutions are sketched, too.

Section 20.2 explains the “forward” and “backward” equations associated with a diffusion (or other Feller process). We get our first taste of finding invariant distributions by looking for stationary solutions of the forward equation.

For the rest of this lecture, whenever I say “an SDE”, I mean “an SDE satisfying the requirements of the existence and uniqueness theorem”, either Theorem 259 (in one dimension) or Theorem 260 (in multiple dimensions). And when I say “a solution”, I mean “a strong solution”. If you are really curious about what has to be changed to accommodate weak solutions, see Rogers and Williams (2000, ch. V, sec. 16–18).

20.1 Solutions of SDEs are Diffusions

Solutions of SDEs are diffusions: i.e., continuous, homogeneous strong Markov processes.

Theorem 261 (Solutions of SDEs are Non-Anticipating and Continuous) *The solution of an SDE is non-anticipating, and has a version with continuous sample paths. If $X(0) = x$ is fixed, then $X(t)$ is \mathcal{F}_t^W -adapted.*

PROOF: Every solution is an Itô process, so it is non-anticipating by Lemma 241. The adaptation for non-random initial conditions follows similarly. (Informally: there’s nothing else for it to depend on.) In the proof of the existence

of solutions, each of the successive approximations is continuous, and we bound the maximum deviation over time, so the solution must be continuous too. \square

Theorem 262 (Solutions of SDEs are Strongly Markov) *Let X_x be the process solving a one-dimensional SDE with non-random initial condition $X(0) = x$. Then X_x forms a homogeneous strong Markov family.*

PROOF: By Exercise 19.8, for every C^2 function f ,

$$f(X(t)) - f(X(0)) - \int_0^t \left[a(X(s)) \frac{\partial f}{\partial x}(X(s)) + \frac{1}{2} b^2(X(s)) \frac{\partial^2 f}{\partial x^2}(X(s)) \right] ds \quad (20.1)$$

is a martingale. Hence, for every x_0 , there is a unique, continuous solution to the martingale problem with operator $G = a(x) \frac{\partial}{\partial x} + \frac{1}{2} b^2(x) \frac{\partial^2}{\partial x^2}$ and function class $\mathcal{D} = C^2$. Since the process is continuous, it is also cadlag. Therefore, by Theorem 172, X is a homogeneous strong Markov family, whose generator equals G on C^2 . \square

Similarly, for a multi-dimensional SDE, where a is a vector and b is a matrix, the generator extends¹ $a_i(x) \partial_i + \frac{1}{2} (bb^T)_{ij}(x) \partial_{ij}^2$. Notice that the coefficients are *outside* the differential operators.

Remark: To see what it is like to try to prove this without using our more general approach, read pp. 103–114 in Øksendal (1995).

Theorem 263 (SDEs and Feller Diffusions) *The processes which solve SDEs are all Feller diffusions.*

PROOF: Theorem 262 shows that solutions are homogeneous strong Markov processes, the previous theorem shows that they are continuous (or can be made so), and so by Definition 218, solutions are diffusions. For them to be Feller, we need (i) for every $t > 0$, $X_y(t) \xrightarrow{d} X_x(t)$ as $y \rightarrow x$, and (ii) $X_x(t) \xrightarrow{P} x$ as $t \rightarrow 0$. But, since solutions are a.s. continuous, $X_x(t) \rightarrow x$ with probability 1, automatically implying convergence in probability, so (ii) is automatic.

¹Here, and elsewhere, I am going to freely use the Einstein conventions for vector calculus: repeated indices in a term indicate that you should sum over those indices, ∂_i abbreviates $\frac{\partial}{\partial x_i}$, ∂_{ij}^2 means $\frac{\partial^2}{\partial x_i \partial x_j}$, etc. Also, $\partial_t \equiv \frac{\partial}{\partial t}$.

To get (i), prove convergence in mean square (i.e. in L_2), which implies convergence in distribution.

$$\mathbf{E} \left[|X_x(t) - X_y(t)|^2 \right] \quad (20.2)$$

$$= \mathbf{E} \left[\left| x - y + \int_0^t a(X_x(s)) - a(X_y(s)) ds + \int_0^t b(X_x(s)) - b(X_y(s)) dW \right|^2 \right]$$

$$\leq |x - y|^2 + \mathbf{E} \left[\left| \int_0^t a(X_x(s)) - a(X_y(s)) ds \right|^2 \right] \quad (20.3)$$

$$+ \mathbf{E} \left[\left| \int_0^t b(X_x(s)) - b(X_y(s)) dW \right|^2 \right]$$

$$= |x - y|^2 + \mathbf{E} \left[\left| \int_0^t a(X_x(s)) - a(X_y(s)) ds \right|^2 \right] \quad (20.4)$$

$$+ \int_0^t \mathbf{E} \left[|b(X_x(s)) - b(X_y(s))|^2 \right] ds$$

$$\leq |x - y|^2 + K \int_0^t \mathbf{E} \left[|X_x(s) - X_y(s)|^2 \right] ds \quad (20.5)$$

for some $K \geq 0$, using the Lipschitz properties of a and b . So, by Gronwall's Inequality (Lemma 258),

$$\mathbf{E} \left[|X_x(t) - X_y(t)|^2 \right] \leq |x - y|^2 e^{Kt} \quad (20.6)$$

This clearly goes to zero as $y \rightarrow x$, so $X_y(t) \rightarrow X_x(t)$ in L_2 , which implies convergence in distribution. \square

Corollary 264 (Convergence of Initial Conditions and of Processes)

For a given SDE, convergence in distribution of the initial condition implies convergence in distribution of the trajectories: if $Y \xrightarrow{d} X_0$, then $X_Y \xrightarrow{d} X_{X_0}$.

PROOF: For every initial condition, the generator of the semi-group is the same (Theorem 262, proof). Since the process is Feller for every initial condition (Theorem 263), and a Feller semi-group is determined by its generator (Theorem 188), the process has the same evolution operator for every initial condition. Hence, condition (ii) of Theorem 205 holds. This implies condition (iv) of that theorem, which is the stated convergence. \square

20.2 Forward and Backward Equations

You will often see probabilists, and applied stochastic people, write about “forward” and “backward” equations for Markov processes, sometimes with the eponym “Kolmogorov” attached. We have already seen a version of the

“backward” equation for Markov processes, with semi-group K_t and generator G , in Theorem 158:

$$\partial_t K_t f(x) = G K_t f(x) \quad (20.7)$$

Let’s unpack this a little, which will help see where the “backwards” comes from. First, remember that the operators K_t are really just conditional expectation:

$$\partial_t \mathbf{E}[f(X_t)|X_0 = x] = G \mathbf{E}[f(X_t)|X_0 = x] \quad (20.8)$$

Next, turn the expectations into integrals with respect to the transition probability kernels:

$$\partial_t \int \mu_t(x, dy) f(y) = G \int \mu_t(x, dy) f(y) \quad (20.9)$$

Finally, assume that there is some reference measure $\lambda \gg \mu_t(x, \cdot)$, for all $t \in T$ and $x \in \Xi$. Denote the corresponding transition densities by $\kappa_t(x, y)$.

$$\partial_t \int d\lambda \kappa_t(x, y) f(y) = G \int d\lambda \kappa_t(x, y) f(y) \quad (20.10)$$

$$\int d\lambda f(y) \partial_t \kappa_t(x, y) = \int d\lambda f(y) G \kappa_t(x, y) \quad (20.11)$$

$$\int d\lambda f(y) [\partial_t \kappa_t(x, y) - G \kappa_t(x, y)] = 0 \quad (20.12)$$

Since this holds for arbitrary nice test functions f ,

$$\partial_t \kappa_t(x, y) = G \kappa_t(x, y) \quad (20.13)$$

The operator G alters the way a function depends on x , the *initial* state. That is, this equation is about how the transition density κ depends on the starting point, “backwards” in time. Generally, we’re in a position to know $\kappa_0(x, y) = \delta(x - y)$; what we want, rather, is $\kappa_t(x, y)$ for some positive value of t . To get this, we need the “forward” equation.

We obtain this from Lemma 155, which asserts that $G K_t = K_t G$.

$$\partial_t \int d\lambda \kappa_t(x, y) f(y) = K_t G f(x) \quad (20.14)$$

$$= \int d\lambda \kappa_t(x, y) G f(y) \quad (20.15)$$

Notice that here, G is altering the dependence on the y coordinate, i.e. the state at time t , not the initial state at time 0. Writing the adjoint² operator as G^\dagger ,

$$\partial_t \int d\lambda \kappa_t(x, y) f(y) = \int d\lambda G^\dagger \kappa_t(x, y) f(y) \quad (20.16)$$

$$\partial_t \kappa_t(x, y) = G^\dagger \kappa_t(x, y) \quad (20.17)$$

²Recall that, in a vector space with an inner product, such as L_2 , the adjoint of an operator A is another operator, defined through $\langle f, Ag \rangle = \langle A^\dagger f, g \rangle$. Further recall that L_2 is an inner-product space, where $\langle f, g \rangle = \mathbf{E}[f(X)g(X)]$.

N.B., G^\dagger is acting on the y -dependence of the transition density, i.e., it says how the probability density is going to change *going forward from t* .

In the physics literature, this is called the Fokker-Planck equation, because Fokker and Planck discovered it, at least in the special case of Langevin-type equations, in 1913, about 20 years before Kolmogorov's work on Markov processes. Notice that, writing ν_t for the distribution of X_t , $\nu_t = \nu_0 \mu_t$. Assuming ν_t has density ρ_t w.r.t. λ , one can get, by integrating the forward equation over space,

$$\partial_t \rho_t(x) = G^\dagger \rho_t(x) \quad (20.18)$$

and this, too, is sometimes called the “Fokker-Planck equation”.

We saw, in the last section, that a diffusion process solving an equation with drift terms $a_i(x)$ and diffusion terms $b_{ij}(x)$ has the generator

$$Gf(x) = a_i(x)\partial_i f(x) + \frac{1}{2}(bb^T)_{ij}(x)\partial_{ij}^2 f(x) \quad (20.19)$$

You can show — it's an exercise in vector calculus, integration by parts, etc. — that the adjoint to G is the differential operator

$$G^\dagger f(x) = -\partial_i a_i(x)f(x) + \frac{1}{2}\partial_{ij}^2 (bb^T)_{ij}(x)f(x) \quad (20.20)$$

Notice that the space-dependence of the SDE's coefficients now appears *inside* the derivatives. Of course, if a and b are independent of x , then they simply pull outside the derivatives, giving us, in that special case,

$$G^\dagger f(x) = -a_i \partial_i f(x) + \frac{1}{2}(bb^T)_{ij} \partial_{ij}^2 f(x) \quad (20.21)$$

Let's interpret this physically, imagining a large population of independent tracer particles wandering around the state space Ξ , following independent copies of the diffusion process. The second derivative term is easy: diffusion tends to smooth out the probability density, taking probability mass away from maxima (where $f'' < 0$) and adding it to minima. (Remember that bb^T is positive semi-definite.) If a_i is positive, then particles tend to move in the positive direction along the i^{th} axis. If $\partial_i \rho$ is also positive, this means that, on average, the point x sends more particles up along the axis than wander down, against the gradient, so the density at x will tend to decline.

Example 265 (Wiener process, heat equation) Notice that (for diffusions produced by SDEs) $G^\dagger = G$ when $a = 0$ and b is constant over the state space. This is the case with Wiener processes, where $G = G^\dagger = \frac{1}{2}\nabla^2$. Thus, the heat equation holds both for the evolution of observable functions of the Wiener process, and for the evolution of the Wiener process's density. You should convince yourself that there is no non-negative integrable ρ such that $G\rho(x) = 0$.

Example 266 (Ornstein-Uhlenbeck process) For the one-dimensional Ornstein-Uhlenbeck process, the generator may be read off from the Langevin equation,

$$Gf(p) = -\gamma p \partial_p f(p) + \frac{1}{2} D^2 \partial_{pp}^2 f(p)$$

and the Fokker-Planck equation becomes

$$\partial_t \rho(p) = \gamma \partial_p (p \rho(p)) + D^2 \frac{1}{2} \partial_{pp}^2 \rho(p)$$

It's easily checked that $\rho(p) = \mathcal{N}(0, D^2/2\gamma)$ gives $\partial_t \rho = 0$. That is, the long-run invariant distribution can be found as a stationary solution of the Fokker-Planck equation. See also Exercise 20.1.

20.3 Exercises

Exercise 20.1 (Langevin equation with a conservative force) A conservative force is one derived from an external potential, i.e., there is a function $\phi(x)$ giving energy, and $F(x) = -d\phi/dx$. The equations of motion for a body subject to a conservative force, drag, and noise read

$$dx = \frac{p}{m} dt \quad (20.22)$$

$$dp = -\gamma p dt + F(x) dt + \sigma dW \quad (20.23)$$

1. Find the corresponding forward (Fokker-Planck) equation.
2. Find a stationary density for this equation, at least up to normalization constants. Hint: use separation of variables, i.e., $\rho(x, p) = u(x)v(p)$. You should be able to find the normalizing constant for the momentum density $v(p)$, but not for the position density $u(x)$. (Its general form should however be familiar from theoretical statistics: what is it?)
3. Show that your stationary solution reduces to that of the Ornstein-Uhlenbeck process, if $F(x) = 0$.

Chapter 21

Large Deviations for Small-Noise Stochastic Differential Equations

This lecture is at once the end of our main consideration of diffusions and stochastic calculus, and a first taste of large deviations theory. Here we study the divergence between the trajectories produced by an ordinary differential equation, and the trajectories of the same system perturbed by a small amount of uncorrelated Gaussian noise (“white noise”; see Sections 22.1 and 22.2.1).

Section 21.1 establishes that, in the small noise limit, the SDE’s trajectories converge in probability on the ODE’s trajectory. This uses Feller-process convergence.

Section 21.2 upper bounds the rate at which the probability of large deviations goes to zero as the noise vanishes. The methods are elementary, but illustrate deeper themes to which we will recur once we have the tools of ergodic and information theory.

In this chapter, we will use the results we have already obtained about SDEs to give a rough estimate of a basic problem, frequently arising in practice¹ namely taking a system governed by an ordinary differential equation and seeing how much effect injecting a small amount of white noise has. More exactly, we will put an upper bound on the probability that the perturbed trajectory goes very far from the unperturbed trajectory, and see the rate at which this

¹For applications in statistical physics and chemistry, see Keizer (1987). For applications in signal processing and systems theory, see Kushner (1984). For applications in nonparametric regression and estimation, and also radio engineering (!) see Ibragimov and Has’minskii (1979/1981). The last book is especially recommended for those who care about the connections between stochastic process theory and statistical inference, but unfortunately expounding the results, or even just the problems, would require a too-long detour through asymptotic statistical theory.

probability goes to zero as the amplitude ϵ of the noise shrinks; this will be $O(e^{-C\epsilon^2})$. This will be our first illustration of a large deviations calculation. It will be crude, but it will also introduce some themes to which we will return (inshallah!) at greater length towards the end of the course. Then we will see that the major improvement of the more refined tools is to give a lower bound to match the upper bound we will calculate now, and see that we at least got the logarithmic rate right.

I should say before going any further that this example is shamelessly ripped off from Freidlin and Wentzell (1998, ch. 3, sec. 1, pp. 70–71), which is *the* book on the subject of large deviations for continuous-time processes.

21.1 Convergence in Probability of SDEs to ODEs

To begin with, consider an unperturbed ordinary differential equation:

$$\frac{d}{dt}x(t) = a(x(t)) \quad (21.1)$$

$$x(0) = x_0 \in \mathbb{R}^d \quad (21.2)$$

Assume that a is uniformly Lipschitz-continuous (as in the existence and uniqueness theorem for ODEs, and more to the point for SDEs). Then, for the given, non-random initial condition, there exists a unique continuous function x which solves the ODE.

Now, for $\epsilon > 0$, consider the SDE

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon dW \quad (21.3)$$

where W is a standard d -dimensional Wiener process, with non-random initial condition $X_\epsilon(0) = x_0$. Theorem 260 clearly applies, and consequently so does Theorem 263, meaning X_ϵ is a Feller diffusion with generator $G_\epsilon f(x) = a_i(x)\partial_i f(x) + \frac{\epsilon^2}{2}\nabla^2 f(x)$.

Write X_0 for the deterministic solution of the ODE.

Our first assertion is that $X_\epsilon \xrightarrow{d} X_0$ as $\epsilon \rightarrow 0$. Notice that X_0 is a Feller process², whose generator is $G_0 = a_i(x)\partial_i$. We can apply Theorem 205 on convergence of Feller processes. Take the class of functions with bounded second derivatives. This is clearly a core for G_0 , and for every G_ϵ . For every function f in this class,

$$\|G_\epsilon f - G_0 f\|_\infty = \left\| a_i \partial_i f(x) + \frac{\epsilon^2}{2} \nabla^2 f(x) - a_i \partial_i f(x) \right\|_\infty \quad (21.4)$$

$$= \frac{\epsilon^2}{2} \|\nabla^2 f(x)\|_\infty \quad (21.5)$$

²You can amuse yourself by showing this. Remember that $X_y(t) \xrightarrow{d} X_x(t)$ is equivalent to $\mathbf{E}[f(X_t)|X_0 = y] \rightarrow \mathbf{E}[f(X_t)|X_0 = x]$ for all bounded *continuous* f , and the solution of an ODE depends continuously on its initial condition.

which goes to zero as $\epsilon \rightarrow 0$. But this is condition (i) of the convergence theorem, which is equivalent to condition (iv), that convergence in distribution of the initial condition implies convergence in distribution of the whole trajectory. Since the initial condition is the same non-random point x_0 for all ϵ , we have $X_\epsilon \xrightarrow{d} X_0$ as $\epsilon \rightarrow 0$. In fact, since X_0 is non-random, we have that $X_\epsilon \xrightarrow{P} X_0$. That last assertion really needs some consideration of metrics on the space of continuous random functions to make sense (see Appendix A2 of Kallenberg), but once that's done, the upshot is

Theorem 267 (Small-Noise SDEs Converge in Probability on No-Noise ODEs) *Let $\Delta_\epsilon(t) = |X_\epsilon(t) - X_0(t)|$. For every $T > 0$, $\delta > 0$,*

$$\lim_{\epsilon \rightarrow 0} \mathbb{P} \left(\sup_{0 \leq t \leq T} \Delta_\epsilon(t) > \delta \right) = 0 \quad (21.6)$$

Or, using the maximum-process notation, for every $T > 0$,

$$\Delta(T)^* \xrightarrow{P} 0 \quad (21.7)$$

PROOF: See above. \square

This is a version of the weak law of large numbers, and nice enough in its own way. One crucial limitation, however, is that it tells us nothing about the *rate* of convergence. That is, it leaves us clueless about how big the noise can be, while still leaving us in the small-noise limit. If the rate of convergence were, say, $O(\epsilon^{1/100})$, then this would not be very useful. (In fact, if the convergence were that slow, we should be really suspicious of numerical solutions of the *unperturbed* ODE.)

21.2 Rate of Convergence; Probability of Large Deviations

Large deviations theory is essentially a study of rates of convergence in probabilistic limit theorems. Here, we will estimate the rate of convergence: our methods will be crude, but it will turn out that even more refined estimates won't change the rate, at least not by more than log factors.

Let's go back to the difference between the perturbed and unperturbed trajectories, going through our now-familiar procedure.

$$X_\epsilon(t) - X_0(t) = \int_0^t [a(X_\epsilon(s)) - a(X_0(s))] ds + \epsilon W(t) \quad (21.8)$$

$$\Delta_\epsilon(t) \leq \int_0^t |a(X_\epsilon(s)) - a(X_0(s))| ds + \epsilon |W(t)| \quad (21.9)$$

$$\leq K_a \int_0^t \Delta_\epsilon(s) ds + \epsilon |W(t)| \quad (21.10)$$

$$\Delta_\epsilon^*(T) \leq \epsilon W^*(T) + K_a \int_0^T \Delta_\epsilon^*(s) ds \quad (21.11)$$

Applying Gronwall's Inequality (Lemma 258),

$$\Delta_\epsilon^*(T) \leq \epsilon W^*(T) e^{K_a T} \quad (21.12)$$

The only random component on the RHS is the supremum of the Wiener process, so we're in business, at least once we take on two standard results, one about the Wiener process itself, the other just about multivariate Gaussians.

Lemma 268 (A Maximal Inequality for the Wiener Process) *For a standard Wiener process, $\mathbb{P}(W^*(t) > a) = 2\mathbb{P}(|W(t)| > a)$.*

PROOF: Proposition 13.13 (pp. 256–257) in Kallenberg. \square

Lemma 269 (A Tail Bound for Maxwell-Boltzmann Distributions) *If Z is a d -dimensional standard Gaussian (i.e., mean 0 and covariance matrix I), then*

$$\mathbb{P}(|Z| > z) \leq \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} \quad (21.13)$$

for sufficiently large z .

PROOF: Each component of Z , $Z_i \sim \mathcal{N}(0, 1)$. So $|Z| = \sqrt{\sum_{i=1}^d Z_i^2}$ has the density function (see, e.g., Cramér (1945, sec. 18.1, p. 236))

$$f(z) = \frac{2}{2^{d/2}\sigma^d\Gamma(d/2)} z^{d-1} e^{-\frac{z^2}{2\sigma^2}}$$

This is the d -dimensional Maxwell-Boltzmann distribution, sometimes called the χ -distribution, because $|Z|^2$ is χ^2 -distributed with d degrees of freedom. Notice that $\mathbb{P}(|Z| \geq z) = \mathbb{P}(|Z|^2 \geq z^2)$, so we will be able to solve this problem in terms of the χ^2 distribution. Specifically, $\mathbb{P}(|Z|^2 \geq z^2) = \Gamma(d/2, z^2/2)/\Gamma(d/2)$, where $\Gamma(r, a)$ is the upper incomplete gamma function. For said function, for every r , $\Gamma(r, a) \leq a^{r-1}e^{-a}$ for sufficiently large a (Abramowitz and Stegun, 1964, Eq. 6.5.32, p. 263). Hence (for sufficiently large z)

$$\mathbb{P}(|Z| \geq z) = \mathbb{P}(|Z|^2 \geq z^2) \quad (21.14)$$

$$= \frac{\Gamma(d/2, z^2/2)}{\Gamma(d/2)} \quad (21.15)$$

$$\leq \frac{(z^2)^{d/2-1} 2^{1-d/2} e^{-z^2/2}}{\Gamma(d/2)} \quad (21.16)$$

$$= \frac{2z^{d-2}e^{-z^2/2}}{2^{d/2}\Gamma(d/2)} \quad (21.17)$$

\square

Theorem 270 (Upper Bound on the Rate of Convergence of Small-Noise SDEs) *In the limit as $\epsilon \rightarrow 0$, for every $\delta > 0$, $T > 0$,*

$$\log \mathbb{P}(\Delta_\epsilon^*(T) > \delta) \leq O(\epsilon^{-2}) \quad (21.18)$$

PROOF: Start by directly estimating the probability of the deviation, using preceding lemmas.

$$\mathbb{P}(\Delta_\epsilon^*(T) > \delta) \leq \mathbb{P}\left(|W|^*(T) > \frac{\delta e^{-K_a T}}{\epsilon}\right) \quad (21.19)$$

$$= 2\mathbb{P}\left(|W(T)| > \frac{\delta e^{-K_a T}}{\epsilon}\right) \quad (21.20)$$

$$\leq \frac{4}{2^{d/2}\Gamma(d/2)} \left(\frac{\delta^2 e^{-2K_a T}}{\epsilon^2}\right)^{d/2-1} e^{-\frac{\delta^2 e^{-2K_a T}}{2\epsilon^2}} \quad (21.21)$$

if ϵ is sufficiently small, so that ϵ^{-1} is sufficiently large to apply Lemma 269. Now take the log and multiply through by ϵ^2 :

$$\epsilon^2 \log \mathbb{P}(\Delta_\epsilon^*(T) > \delta) \quad (21.22)$$

$$\leq \epsilon^2 \log \frac{4}{2^{d/2}\Gamma(d/2)} + \epsilon^2 \left(\frac{d}{2} - 1\right) [\log \delta^2 e^{-2K_a T} - 2 \log \epsilon] - \delta^2 e^{-2K_a T}$$

$$\lim_{\epsilon \downarrow 0} \epsilon^2 \log \mathbb{P}(\Delta_\epsilon^*(T) > \delta) \leq -\delta^2 e^{-2K_a T} \quad (21.23)$$

since $\epsilon^2 \log \epsilon \rightarrow 0$, and the conclusion follows. \square

Notice several points.

1. Here ϵ gauges the size of the noise, and we take a small noise limit. In many forms of large deviations theory, we are concerned with large-sample ($N \rightarrow \infty$) or long-time ($T \rightarrow \infty$) limits. In every case, we will identify some asymptotic parameter, and obtain limits on the asymptotic probabilities. There are deviations inequalities which hold *non*-asymptotically, but they have a different flavor, and require different machinery. (Some people are made uncomfortable by an ϵ^2 rate, and prefer to write the SDE $dX = a(X)dt + \sqrt{\epsilon}dW$ so as to avoid it. I don't get this.)
2. The magnitude of the deviation δ does not change as the noise becomes small. This is basically what makes this a *large* deviations result. There is also a theory of *moderate* deviations.
3. We only have an upper bound. This is enough to let us know that the probability of large deviations becomes exponentially small. But we might be wrong about the rate — it could be even faster than we've estimated. In this case, however, it'll turn out that we've got at least the order of magnitude correct.

4. We also don't have a *lower* bound on the probability, which is something that would be *very* useful in doing reliability analyses. It will turn out that, under many circumstances, one can obtain a lower bound on the probability of large deviations, which has the *same* asymptotic dependence on ϵ as the upper bound.
5. Suppose we're right about the rate (which, it will turn out, we are), and it holds both from above and below. It would be nice to be able to say something like

$$\mathbb{P}(\Delta_\epsilon^*(T) > \delta) \rightarrow C_1(\delta, T)e^{-C_2(\delta, T)\epsilon^{-2}} \quad (21.24)$$

rather than

$$\epsilon^2 \log \mathbb{P}(\Delta_\epsilon^*(T) > \delta) \rightarrow -C_2(\delta, T) \quad (21.25)$$

The difficulty with making an assertion like 21.24 is that the large deviation *probability* actually converges on *any* function which goes to asymptotically to zero! So, to extract the actual rate of dependence, we need to get a result like 21.25.

More generally, one consequence of Theorem 270 is that SDE trajectories which are far from the trajectory of the ODE have exponentially small probabilities. The vast majority of the probability will be concentrated around the unperturbed trajectory. Reasonable sample-path functionals can therefore be well-approximated by averaging their value over some small (δ) neighborhood of the unperturbed trajectory. This should sound very similar to Laplace's method for the evaluate of asymptotic integrals in Euclidean space, and in fact one of the key parts of large deviations theory is an extension of Laplace's method to infinite-dimensional function spaces.

In addition to this mathematical content, there is also a close connection to the principle of least action in physics. In classical mechanics, the system follows the trajectory of least action, the "action" of a trajectory being the integral of the kinetic minus potential energy along that path. In quantum mechanics, this is no longer an axiom but a *consequence* of the dynamics: the action-minimizing trajectory is the most probable one, and large deviations from it have exponentially small probability. Similarly, the theory of large deviations can be used to establish quite general *stochastic* principles of least action for Markovian systems.³

³For a fuller discussion, see Eyink (1996), Freidlin and Wentzell (1998, ch. 3).

Chapter 22

Spectral Analysis and L_2 Ergodicity

Section 22.1 makes sense of the idea of white noise. This forms the bridge from the ideas of Wiener integrals, in the previous lectures, and spectral and ergodic theory, which we will pursue here.

Section 22.2 introduces the spectral representation of weakly stationary processes, and the central Wiener-Khinchin theorem connecting autocovariance to the power spectrum. Subsection 22.2.1 explains why white noise is “white”.

Section 22.3 gives our first classical ergodic result, the “mean square” (L_2) ergodic theorem for weakly stationary processes. Subsection 22.3.1 gives an easy proof of a sufficient condition, just using the autocovariance. Subsection 22.3.2 gives a necessary and sufficient condition, using the spectral representation.

Any reasonable real-valued function $x(t)$ of time, $t \in \mathbb{R}$, has a Fourier transform, that is, we can write

$$\tilde{x}(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt e^{i\nu t} x(t)$$

which can usually be inverted to recover the original function,

$$x(t) = \int_{-\infty}^{\infty} d\nu e^{-i\nu t} \tilde{x}(\nu)$$

This one example of an “analysis”, in the original sense of resolving into parts, of a function into a collection of orthogonal basis functions. (You can find the details in any book on Fourier analysis, as well as the varying conventions on where the 2π goes, which side gets the $e^{-i\nu t}$, the constraints on \tilde{x} which arise from the fact that x is real, etc.)

There are various reasons to prefer the trigonometric basis functions $e^{i\nu t}$ over other possible choices. One is that they are invariant under translation in time, which just changes phases¹. This suggests that the Fourier basis will be particularly useful when dealing with time-invariant systems. For stochastic processes, however, time-invariance is stationarity. This suggests that there should be some useful way of doing Fourier analysis on stationary random functions. In fact, it turns out that stationary and even weakly-stationary processes *can* be productively Fourier-transformed. This is potentially a huge topic, especially when it's expanded to include representing random functions in terms of (countable) series of orthogonal functions. The spectral theory of random functions connects Fourier analysis, disintegration of measures, Hilbert spaces and ergodicity. This lecture will do no more than scratch the surface, covering, in succession, white noise, the basics of the spectral representation of weakly-stationary random functions and the fundamental Wiener-Khinchin theorem linking covariance functions to power spectra, why white noise is called “white”, and the mean-square ergodic theorem.

Good sources, if you want to go further, are the books of Bartlett (1955, ch. 6) (from whom I've stolen shamelessly), the historically important and inspiring Wiener (1949, 1961), and of course Doob (1953). Loève (1955, ch. X) is highly edifying, particular his discussion of Karhunen-Loève transforms, and the associated construction of the Wiener process as a Fourier series with random phases.

22.1 White Noise

Scientists and engineers are often uncomfortable with the SDEs in the way probabilists write them, because they want to divide through by dt and have the result mean something. The trouble, of course, is that dW/dt does not, in any ordinary sense, exist. They, however, are often happier ignoring this inconvenient fact, and talking about “white noise” as what dW/dt ought to be. This is not totally crazy. Rather, one can define $\xi \equiv dW/dt$ as a *generalized* derivative, one whose value at any given time is a random real linear functional, rather than a random real number. Consequently, it only really makes sense in integral expressions (like the solutions of SDEs!), but it can, in many ways, be formally manipulated like an ordinary function.

One way to begin to make sense of this is to start with a standard Wiener process $W(t)$, and a C^1 non-random function $u(t)$, and to use integration by

¹If $t \mapsto t + \tau$, then $\tilde{x}(\nu) \mapsto e^{i\nu\tau} \tilde{x}(\nu)$.

parts:

$$\frac{d}{dt}(uW) = u \frac{dW}{dt} + \frac{du}{dt}W \quad (22.1)$$

$$= u(t)\xi(t) + \dot{u}(t)W(t) \quad (22.2)$$

$$\int_0^t \frac{d}{dt}(uW)ds = \int_0^t \dot{u}(s)W(s) + u(s)\xi(s)ds \quad (22.3)$$

$$u(t)W(t) - u(0)W(0) = \int_0^t \dot{u}(s)W(s)ds + \int_0^t u(s)\xi(s)ds \quad (22.4)$$

$$\int_0^t u(s)\xi(s)ds \equiv u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds \quad (22.5)$$

We can take the last line to *define* ξ , and time-integrals within which it appears. Notice that the terms on the RHS are well-defined *without* the Itô calculus: one is just a product of two measurable random variables, and the other is the time-integral of a continuous random function. With this definition, we can establish some properties of ξ .

Proposition 271 (Linearity of White Noise Integrals) $\xi(t)$ is a linear functional:

$$\int_0^t (a_1 u_1(s) + a_2 u_2(s))\xi(s)ds = a_1 \int_0^t u_1(s)\xi(s)ds + a_2 \int_0^t u_2(s)\xi(s)ds \quad (22.6)$$

PROOF:

$$\int_0^t (a_1 u_1(s) + a_2 u_2(s))\xi(s)ds \quad (22.7)$$

$$= (a_1 u_1(t) + a_2 u_2(t))W(t) - \int_0^t (a_1 \dot{u}_1(s) + a_2 \dot{u}_2(s))W(s)ds$$

$$= a_1 \int_0^t u_1(s)\xi(s)ds + a_2 \int_0^t u_2(s)\xi(s)ds \quad (22.8)$$

□

Proposition 272 (White Noise Has Mean Zero) For all t , $\mathbf{E}[\xi(t)] = 0$.

PROOF:

$$\int_0^t u(s)\mathbf{E}[\xi(s)]ds = \mathbf{E}\left[\int_0^t u(s)\xi(s)ds\right] \quad (22.9)$$

$$= \mathbf{E}\left[u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds\right] \quad (22.10)$$

$$= \mathbf{E}[u(t)W(t)] - \int_0^t \dot{u}(s)\mathbf{E}[W(s)]ds \quad (22.11)$$

$$= 0 - 0 = 0 \quad (22.12)$$

Proposition 273 (White Noise and Itô Integrals) For all $u \in C^1$, $\int_0^t u(s)\xi(s)ds = \int_0^t u(s)dW$.

PROOF: Apply Itô's formula to the function $f(t, W) = u(t)W(t)$:

$$d(uW) = W(t)\dot{u}(t)dt + u(t)dW \quad (22.13)$$

$$u(t)W(t) = \int_0^t \dot{u}(s)W(s)ds + \int_0^t u(s)dW \quad (22.14)$$

$$\int_0^t u(s)dW = u(t)W(t) - \int_0^t \dot{u}(s)W(s)ds \quad (22.15)$$

$$= \int_0^t u(s)\xi(s)ds \quad (22.16)$$

□

This could be used to extend the definition of white-noise integrals to any Itô-integrable process.

Proposition 274 (White Noise is Uncorrelated) ξ has delta-function covariance: $\text{cov}(\xi(t_1), \xi(t_2)) = \delta(t_1 - t_2)$.

PROOF: Since $\mathbf{E}[\xi(t)] = 0$, we just need to show that $\mathbf{E}[\xi(t_1)\xi(t_2)] = \delta(t_1 - t_2)$. Remember (Eq. 17.14 on p. 127) that $\mathbf{E}[W(t_1)W(t_2)] = t_1 \wedge t_2$.

$$\int_0^t \int_0^t u(t_1)u(t_2)\mathbf{E}[\xi(t_1)\xi(t_2)]dt_1dt_2 \quad (22.17)$$

$$= \mathbf{E} \left[\int_0^t u(t_1)\xi(t_1)dt_1 \int_0^t u(t_2)\xi(t_2)dt_2 \right] \quad (22.18)$$

$$= \mathbf{E} \left[\left(\int_0^t u(t_1)\xi(t_1)dt_1 \right)^2 \right] \quad (22.19)$$

$$= \int_0^t \mathbf{E}[u^2(t_1)]dt_1 = \int_0^t u^2(t_1)dt_1 \quad (22.20)$$

using the preceding proposition, the Itô isometry, and the fact that u is non-random. But

$$\int_0^t \int_0^t u(t_1)u(t_2)\delta(t_1 - t_2)dt_1dt_2 = \int_0^t u^2(t_1)dt_1 \quad (22.21)$$

so $\delta(t_1 - t_2) = \mathbf{E}[\xi(t_1)\xi(t_2)] = \text{cov}(\xi(t_1), \xi(t_2))$. □

Proposition 275 (White Noise is Gaussian and Stationary) ξ is a strongly stationary Gaussian process.

PROOF: To show that it is Gaussian, use Exercise 19.6. The mean is constant for all times, and the covariance depends only on $|t_1 - t_2|$, so it satisfies Definition 50 and is weakly stationary. But a weakly stationary Gaussian process is also strongly stationary. □

22.2 Spectral Representation of Weakly Stationary Processes

This section will only handle spectral representations of real- and complex-valued one-parameter processes in continuous time. Generalizations to vector-valued and multi-parameter processes are straightforward; handling discrete time is actually in some ways more irritating, because of limitations on allowable frequencies of Fourier components (to the range from $-\pi$ to π).

Definition 276 (Autocovariance Function) *Suppose that, for all $t \in T$, X is real and $\mathbf{E}[X^2(t)]$ is finite. Then $\Gamma(t_1, t_2) \equiv \mathbf{E}[X(t_1)X(t_2)] - \mathbf{E}[X(t_1)]\mathbf{E}[X(t_2)]$ is the autocovariance function of the process. If the process is weakly stationary, so that $\Gamma(t, t + \tau) = \Gamma(0, \tau)$ for all t, τ , write $\Gamma(\tau)$. If $X(t) \in \mathbb{C}$, then $\Gamma(t_1, t_2) \equiv \mathbf{E}[X^\dagger(t_1)X(t_2)] - \mathbf{E}[X^\dagger(t_1)]\mathbf{E}[X(t_2)]$, where \dagger is complex conjugation.*

Lemma 277 (Autocovariance and Time Reversal) *If X is real and weakly stationary, then $\Gamma(\tau) = \Gamma(-\tau)$; if X is complex and weakly stationary, then $\Gamma(\tau) = \Gamma^\dagger(-\tau)$.*

PROOF: Direct substitution into the definitions. \square

Remarks on terminology. It is common, when only dealing with one stochastic process, to drop the qualifying “auto” and just speak of the covariance function; I probably will myself. It is also common (especially in the time series literature) to switch to the (auto)correlation function, i.e., to normalize by the standard deviations. Finally, be warned that the statistical physics literature (e.g. Forster, 1975) uses “correlation function” to mean $\mathbf{E}[X(t_1)X(t_2)]$, i.e., the *uncentered* mixed second moment. This is a matter of tradition, not (despite appearances) ignorance.

Definition 278 (Second-Order Process) *A complex-valued process X is second order when $\mathbf{E}[|X|^2(t)] < \infty$ for all t .*

Definition 279 (Spectral Representation, Power Spectrum) *A real-valued process X on T has a complex-valued spectral process \tilde{X} , if it has a spectral representation:*

$$X(t) \equiv \int_{-\infty}^{\infty} e^{-i\nu t} d\tilde{X}(\nu) \quad (22.22)$$

The power spectrum $V(\nu) \equiv \mathbf{E}\left[\left|\tilde{X}(\nu)\right|^2\right]$.

Remark 1. The name “power spectrum” arises because this is proportional to the amount of power (energy per unit time) carried by oscillations of frequency $\leq \nu$, at least in a linear system.

Remark 2. Notice that the only part of the right-hand side of Equation 22.22 which depends on t is the integrand, $e^{-i\nu t}$, which just changes the phase of each Fourier component deterministically. Roughly speaking, for a fixed ω the amplitudes of the different Fourier components in $X(t, \omega)$ are fixed, and shifting forward in time just involves changing their phases. (Making this simple is why we have to allow \tilde{X} to have complex values.)

The spectral representation is another stochastic integral, like the Itô integral we saw in Section 19.1. There, the measure of the time interval $[t_1, t_2]$ was given by the increment of the Wiener process, $W(t_2) - W(t_1)$. Here, for any frequency interval $[\nu_1, \nu_2]$, the increment $\tilde{X}(\nu_2) - \tilde{X}(\nu_1)$ defines a random set function (admittedly, a complex-valued one). Since those intervals are a generating class for the Borel σ -field, the usual arguments extend this set function uniquely to a random complex measure on \mathbb{R}, \mathcal{B} . When we write something like $\int G(\nu) d\tilde{X}(\nu)$, we mean an integral with respect to this measure.

Rather than dealing directly with this measure, we can, as with the Itô integral, use approximation by elementary processes. That is, we should interpret

$$\int_{-\infty}^{\infty} G(t, \nu) d\tilde{X}(\nu)$$

as the L_2 limit of sums

$$\sum_{\nu_i=-\infty}^{\infty} G(t, \nu_i) (\tilde{X}(\nu_{i+1}) - \tilde{X}(\nu_i))$$

as $\sup \nu_{i+1} - \nu_i$ goes to zero. This limit can be shown to exist in pretty much exactly the same way we showed the corresponding limit for Itô integrals to exist.²

Lemma 280 (Regularity of the Spectral Process) *When it exists, $\tilde{X}(\nu)$ has right and left limits at every point ν , and limits as $\nu \rightarrow \pm\infty$.*

PROOF: See Loève (1955, §34.4). You can prove this yourself, however, using the material on characteristic functions in 36-752. \square

Definition 281 (Jump of the Spectral Process) *The jump of the spectral process at ν is the difference between the right- and left- hand limits at ν , $\Delta\tilde{X}(\nu) \equiv \tilde{X}(\nu+0) - \tilde{X}(\nu-0)$.*

Remark 1: As usual, $\tilde{X}(\nu+0) \equiv \lim_{h \downarrow 0} \tilde{X}(\nu+h)$, and $\tilde{X}(\nu-0) \equiv \lim_{h \downarrow 0} \tilde{X}(\nu-h)$.

Remark 2: Some people call the set of points at which the jump is non-zero the “spectrum”. This usage comes from functional analysis, but seems needlessly confusing in the present context.

Proposition 282 (Spectral Representations of Weakly Stationary Processes) *Every weakly-stationary process has a spectral representation.*

²There is a really excellent discussion of such stochastic integrals, and L_2 stochastic calculus more generally, in Loève (1955, §34).

PROOF: See Loève (1955, §34.4), or Bartlett (1955, §6.2). \square

The following property will be very important for us, since when the spectral process has it, many nice consequences follow.

Definition 283 (Orthogonal Increments) *A one-parameter random function (real or complex) has orthogonal increments if, for $t_1 \leq t_2 \leq t_3 \leq t_4 \in T$, the covariance of the increment from t_1 to t_2 and the increment from t_3 to t_4 is always zero:*

$$\mathbf{E} \left[\left(\tilde{X}(\nu_4) - \tilde{X}(\nu_3) \right) \left(\tilde{X}(\nu_2) - \tilde{X}(\nu_1) \right)^\dagger \right] = 0 \quad (22.23)$$

Lemma 284 (Orthogonal Spectral Increments and Weak Stationarity)

The spectral process of a second-order process has orthogonal increments if and only if the process is weakly stationary.

SKETCH PROOF: Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$, so $\mathbf{E}[\tilde{X}(\nu)] = 0$. “If”: Pick any arbitrary t . We can write, using the fact that $X(t) = X^\dagger(t)$ for real-valued processes,

$$\Gamma(\tau) = \Gamma(t, t + \tau) \quad (22.24)$$

$$= \mathbf{E} [X^\dagger(t)X(t + \tau)] \quad (22.25)$$

$$= \mathbf{E} \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i\nu_1 t} e^{-i\nu_2(t+\tau)} d\tilde{X}_{\nu_1}^\dagger d\tilde{X}_{\nu_2} \right] \quad (22.26)$$

$$= \mathbf{E} \left[\lim_{\Delta\nu \rightarrow 0} \sum_{\nu_1} \sum_{\nu_2} e^{it(\nu_1 - \nu_2)} e^{-i\nu_2 \tau} \Delta\tilde{X}^\dagger(\nu_1) \Delta\tilde{X}(\nu_2) \right] \quad (22.27)$$

$$= \lim_{\Delta\nu \rightarrow 0} \sum_{\nu_1} \sum_{\nu_2} e^{it(\nu_1 - \nu_2)} e^{-i\nu_2 \tau} \mathbf{E} [\Delta\tilde{X}^\dagger(\nu_1) \Delta\tilde{X}(\nu_2)] \quad (22.28)$$

where $\Delta\tilde{X}(\nu) = \tilde{X}(\nu + \Delta\nu) - \tilde{X}(\nu)$. Since t was arbitrary, every term on the right must be independent of t . When $\nu_1 = \nu_2$, $e^{it(\nu_1 - \nu_2)} = 1$, so $\mathbf{E} [\Delta\tilde{X}^\dagger(\nu) \Delta\tilde{X}(\nu)]$ is unconstrained. If $\nu_1 \neq \nu_2$, however, we must have $\mathbf{E} [\Delta\tilde{X}^\dagger(\nu_1) \Delta\tilde{X}(\nu_2)] = 0$, which is to say (Definition 283) we must have orthogonal increments.

“Only if”: if the increments are orthogonal, then clearly the steps of the argument can be reversed to conclude that $\Gamma(t_1, t_2)$ depends only on $t_2 - t_1$. \square

Definition 285 (Spectral Function, Spectral Density) *The spectral function of a weakly stationary process is the function $S(\nu)$ appearing in the spectral representation of its autocovariance:*

$$\Gamma(\tau) = \int_{-\infty}^{\infty} e^{-i\nu\tau} dS_\nu \quad (22.29)$$

Remark. Some people prefer to talk about the spectral function as the Fourier transform of the autocorrelation function, rather than of the autocovariance. This has the advantage that the spectral function turns out to be a normalized cumulative distribution function (see Theorem 286 immediately below), but is otherwise inconsequential.

Theorem 286 (Weakly Stationary Processes Have Spectral Functions)

The spectral function exists for every weakly stationary process, if $\Gamma(\tau)$ is continuous. Moreover, $S(\nu) \geq 0$, S is non-decreasing, $S(-\infty) = 0$, $S(\infty) = \Gamma(0)$, and $\lim_{h \downarrow 0} S(\nu + h)$ and $\lim_{h \downarrow 0} S(\nu - h)$ exist for every ν .

PROOF: Usually, by an otherwise-obscure result in Fourier analysis called Bochner's theorem. A more direct proof is due to Loève. Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$.

Start by defining

$$H_T(\nu) \equiv \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} e^{i\nu t} X(t) dt \quad (22.30)$$

and define $f_T(\nu)$ through H :

$$2\pi f_T(\nu) \equiv \mathbf{E} \left[H_T(\nu) H_T^\dagger(\nu) \right] \quad (22.31)$$

$$= \mathbf{E} \left[\frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} e^{i\nu t_1} X(t_1) e^{-i\nu t_2} X^\dagger(t_2) dt_1 dt_2 \right] \quad (22.32)$$

$$= \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} e^{i\nu(t_1-t_2)} \mathbf{E}[X(t_1)X(t_2)] dt_1 dt_2 \quad (22.33)$$

$$= \frac{1}{T} \int_{-T/2}^{T/2} \int_{-T/2}^{T/2} e^{i\nu(t_1-t_2)} \Gamma(t_1 - t_2) dt_1 dt_2 \quad (22.34)$$

$$= \int_{-T}^T \left(1 - \frac{|\tau|}{T}\right) \Gamma(\tau) e^{i\nu\tau} d\tau \quad (22.35)$$

Recall that $\Gamma(\tau)$ defines a non-negative quadratic form, meaning that

$$\sum_{s,t} a_s^\dagger a_t \Gamma(t - s) \geq 0$$

for any sets of times and any complex numbers a_t . This will in particular work if the complex numbers lie on the unit circle and can be written $e^{i\nu t}$. This means that integrals

$$\int \int e^{i\nu(t_1-t_2)} \Gamma(t_1 - t_2) dt_1 dt_2 \geq 0 \quad (22.36)$$

so $f_T(\nu) \geq 0$.

Define $\phi_T(\tau)$ as the integrand in Eq. 22.35, so that

$$f_T(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau) e^{i\nu\tau} d\tau \quad (22.37)$$

which is recognizable as a proper Fourier transform. Now pick some $N > 0$ and massage the equation so it starts to look like an inverse transform.

$$f_T(\nu)e^{-i\nu t} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau)e^{i\nu\tau}e^{-i\nu t}d\tau \quad (22.38)$$

$$\left(1 - \frac{|\nu|}{N}\right) f_T(\nu)e^{-i\nu t} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau)e^{i\nu\tau}e^{-i\nu t} \left(1 - \frac{|\nu|}{N}\right) d\tau \quad (22.39)$$

Integrating over frequencies,

$$\int_{-N}^N \left(1 - \frac{|\nu|}{N}\right) f_T(\nu)e^{-i\nu t} d\nu \quad (22.40)$$

$$\begin{aligned} &= \int_{-N}^N \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau)e^{i\nu\tau}e^{-i\nu t} \left(1 - \frac{|\nu|}{N}\right) d\tau d\nu \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_T(\tau) \left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2} \right)^2 N d\tau \end{aligned} \quad (22.41)$$

For fixed N , it is easy to verify that

$$\int_{-\infty}^{\infty} N \left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2} \right)^2 dt = 1$$

and that

$$\lim_{t \rightarrow \tau} N \left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2} \right)^2 = N$$

On the other hand, if $\tau \neq t$,

$$\lim_{N \rightarrow \infty} N \left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2} \right)^2 = 0$$

uniformly over any bounded interval in t . (You might find it instructive to try plotting this function; you will need to be careful near the origin!) In other words, this is a representation of the Dirac delta function, so that

$$\lim_{N \rightarrow \infty} \int_{-\infty}^{\infty} \phi_T(\tau) \left(\frac{\sin N(\tau-t)/2}{N(\tau-t)/2} \right)^2 N d\tau = \phi_T(\tau)$$

and in fact the convergence is uniform.

Turning to the other side of Equation 22.41, $\left(1 - \frac{|\nu|}{N}\right) f_T(\nu) \geq 0$, so

$$\int_{-N}^N \left(1 - \frac{|\nu|}{N}\right) f_T(\nu)e^{-i\nu t} d\nu$$

is like a characteristic function of a distribution, up to, perhaps, an over-all normalizing factor, which will be (given the right-hand side) $\phi_T(0) = \Gamma(0) > 0$. Since $\Gamma(\tau)$ is continuous, $\phi_T(\tau)$ is too, and so, as $N \rightarrow \infty$, the right-hand

side converges uniformly on $\phi_T(t)$, but a uniform limit of characteristic functions is still a characteristic function. Thus $\phi_T(t)$, too, can be obtained from a characteristic function. Finally, since $\Gamma(t)$ is the uniform limit of $\phi_T(t)$ on every bounded interval, $\Gamma(t)$ has a characteristic-function representation of the stated form. This allows us to further conclude that $S(\nu)$ is real-valued, non-decreasing, $S(-\infty) = 0$ and $S(\infty) = \Gamma(0)$, and has both right and left limits everywhere. \square

There is a converse, with a cute constructive proof.

Theorem 287 (Existence of Weakly Stationary Processes with Given Spectral Functions) *Let $S(\nu)$ be any function with the properties described at the end of Theorem 286. Then there is a weakly stationary process whose autocovariance is of the form given in Eq. 22.29.*

PROOF: Define $\sigma^2 = \Gamma(0)$, $F(\nu) = S(\nu)/\sigma^2$. Now $F(\nu)$ is a properly normalized cumulative distribution function. Let N be a random variable distributed according to F , and $\Phi \sim U(0, 2\pi)$ be independent of N . Set $X(t) \equiv \sigma e^{i(\Phi - Nt)}$. Then $\mathbf{E}[X(t)] = \sigma \mathbf{E}[e^{i\Phi}] \mathbf{E}[e^{-iNt}] = 0$. Moreover,

$$\mathbf{E}[X^\dagger(t_1)X(t_2)] = \sigma^2 \mathbf{E}\left[e^{-i(\Phi - Nt_1)} e^{i(\Phi - Nt_2)}\right] \quad (22.42)$$

$$= \sigma^2 \mathbf{E}\left[e^{-iN(t_1 - t_2)}\right] \quad (22.43)$$

$$= \sigma^2 \int_{-\infty}^{\infty} e^{-i\nu(t_1 - t_2)} dF_\nu \quad (22.44)$$

$$= \Gamma(t_1 - t_2) \quad (22.45)$$

\square

Definition 288 (Jump of the Spectral Function) *The jump of the spectral function at ν , $\Delta S(\nu)$, is $S(\nu + 0) - S(\nu - 0)$.*

Lemma 289 (Spectral Function Has Non-Negative Jumps) $\Delta S(\nu) \geq 0$.

PROOF: Obvious from the fact that $S(\nu)$ is non-decreasing. \square

Theorem 290 (Wiener-Khinchin Theorem) *If X is a weakly stationary process, then its power spectrum is equal to its spectral function.*

$$V(\nu) \equiv \mathbf{E}\left[\left|\tilde{X}(\nu)\right|^2\right] = S(\nu) \quad (22.46)$$

PROOF: Assume, without loss of generality, that $\mathbf{E}[X(t)] = 0$. Substitute the spectral representation of X into the autocovariance, using Fubini's theorem

to turn a product of integrals into a double integral.

$$\Gamma(\tau) = \mathbf{E}[X(t)X(t+\tau)] \quad (22.47)$$

$$= \mathbf{E}[X^\dagger(t)X(t+\tau)] \quad (22.48)$$

$$= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(t+\tau)\nu_1} e^{it\nu_2} d\tilde{X}_{\nu_1} d\tilde{X}_{\nu_2}\right] \quad (22.49)$$

$$= \mathbf{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-it(\nu_1-\nu_2)} e^{-i\tau\nu_2} d\tilde{X}_{\nu_1} d\tilde{X}_{\nu_2}\right] \quad (22.50)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-it(\nu_1-\nu_2)} e^{-i\tau\nu_2} \mathbf{E}[d\tilde{X}_{\nu_1} d\tilde{X}_{\nu_2}] \quad (22.51)$$

using the fact that integration and expectation commute to (formally) bring the expectation inside the integral. Since \tilde{X} has orthogonal increments, $\mathbf{E}[d\tilde{X}_{\nu_1}^\dagger d\tilde{X}_{\nu_2}] = 0$ unless $\nu_1 = \nu_2$. This turns the double integral into a single integral, and kills the $e^{-it(\nu_1-\nu_2)}$ factor, which had to go away because t was arbitrary.

$$\Gamma(\tau) = \int_{-\infty}^{\infty} e^{-i\tau\nu} \mathbf{E}[d(\tilde{X}_\nu^\dagger \tilde{X}_\nu)] \quad (22.52)$$

$$= \int_{-\infty}^{\infty} e^{-i\tau\nu} dV_\nu \quad (22.53)$$

using the definition of the power spectrum. Since $\Gamma(\tau) = \int_{-\infty}^{\infty} e^{-i\tau\nu} dV_\nu$, it follows that S_ν and V_ν differ by a constant, namely the value of $V(-\infty)$, which can be chosen to be zero without affecting the spectral representation of X . \square

22.2.1 How the White Noise Lost Its Color

Why is white noise, as defined in Section 22.1, called “white”? The answer is easy, given the Wiener-Khinchin relation in Theorem 290.

Recall from Proposition 274 that the autocovariance function of white noise is $\delta(t_1 - t_2)$. Recall from general analysis that one representation of the delta function is the following Fourier integral:

$$\delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\nu e^{i\nu t}$$

(This can be “derived” from inserting the definition of the Fourier transform into the inverse Fourier transform, among other, more respectable routes.) Appealing then to the theorem, $S(\nu) = \frac{1}{2\pi}$ for all ν . That is, there is equal power at all frequencies, just as white light is composed of light of all colors (frequencies), mixed with equal intensity.

Relying on this analogy, there is an elaborate taxonomy red, pink, black, brown, and other variously-colored noises, depending on the shape of their power spectra. The value of this terminology has honestly never been very clear to me, but the curious reader is referred to the (very fun) book of Schroeder (1991) and references therein.

22.3 The Mean-Square Ergodic Theorem

Ergodic theorems relate functionals calculated along individual sample paths (say, the time average, $T^{-1} \int_0^T dt X(t)$, or the maximum attained value) to functionals calculated over the whole distribution (say, the expectation, $\mathbf{E}[X(t)]$, or the expected maximum). The basic idea is that the two should be close, and they should get closer the longer the trajectory we use, because in some sense any one sample path, carried far enough, is representative of the whole distribution. Since there are many different kinds of functionals, and many different modes of stochastic convergence, there are many different kinds of ergodic theorem. The classical ergodic theorems say that time averages converge on expectations³, either in L_p or a.s. (both implying convergence in distribution or in probability). The traditional centerpiece of ergodic theorem is Birkhoff's "individual" ergodic theorem, asserting a.s. convergence. We will see its proof, but it will need a lot of preparatory work, and it requires strict stationarity. By contrast, the L_2 , or "mean square", ergodic theorem, attributed to von Neumann⁴ is already in our grasp, and holds for weakly stationary processes.

We will actually prove it twice, once with a fairly transparent sufficient condition, and then again with a more complicated necessary-and-sufficient condition. The more complicated proof will wait until next lecture.

22.3.1 Mean-Square Ergodicity Based on the Autocovariance

First, the easy version, which gives an estimate of the rate of convergence. (What I say here is ripped off from the illuminating discussion in (Frisch, 1995, sec. 4.4, especially pp. 49–50).)

Definition 291 (Time Averages) *When X is a one-sided, continuous-parameter random process, we say that its time average between times T_1 and T_2 is $\bar{X}(T_1, T_2) \equiv (T_2 - T_1)^{-1} \int_{T_1}^{T_2} dt X(t)$. When we only mention one time argument, by default the time average is from 0 to T , $\bar{X}(T) \equiv \bar{X}(0, T)$.*

(Only considering time averages starting from zero involves no loss of generality for weakly stationary processes: why?)

Definition 292 (Integral Time Scale) *The integral time scale of a weakly-stationary random process is*

$$\tau_{\text{int}} \equiv \frac{\int_0^\infty d\tau |\Gamma(\tau)|}{\Gamma(0)} \quad (22.54)$$

³Proverbially: "time averages converge on space averages", the space in question being the state space Ξ ; or "converge on phase averages", since physicists call certain kinds of state space "phase space".

⁴See von Plato (1994, ch. 3) for a fascinating history of the development of ergodic theory through the 1930s, and its place in the history of mathematical probability.

Notice that τ_{int} does, indeed, have units of time.

As a particular example, suppose that $\Gamma(\tau) = \Gamma(0)e^{-\tau/A}$, where the constant A is known as the *autocorrelation time*. Then simple calculus shows that $\tau_{\text{int}} = A$.

Theorem 293 (Mean-Square Ergodic Theorem (Finite Autocovariance Time)) *Let $X(t)$ be a weakly stationary process with $\mathbf{E}[X(t)] = 0$. If $\tau_{\text{int}} < \infty$, then $\bar{X}(T) \xrightarrow{L_2} 0$ as $T \rightarrow \infty$.*

PROOF: Use Fubini's theorem to to the square of the integral into a double integral, and then bring the expectation inside it:

$$\mathbf{E} \left[\left(\frac{1}{T} \int_0^T dt X(t) \right)^2 \right] = \mathbf{E} \left[\frac{1}{T^2} \int_0^T \int_0^T dt_1 dt_2 X(t_1) X(t_2) \right] \quad (22.55)$$

$$= \frac{1}{T^2} \int_0^T \int_0^T dt_1 dt_2 \mathbf{E}[X(t_1) X(t_2)] \quad (22.56)$$

$$= \frac{1}{T^2} \int_0^T \int_0^T dt_1 dt_2 \Gamma(t_1 - t_2) \quad (22.57)$$

$$= \frac{2}{T^2} \int_0^T dt_1 \int_0^{t_1} d\tau \Gamma(\tau) \quad (22.58)$$

$$\leq \frac{2}{T^2} \int_0^T dt_1 \int_0^\infty d\tau |\Gamma(\tau)| \quad (22.59)$$

$$= \frac{2}{T} \int_0^\infty d\tau |\Gamma(\tau)| \quad (22.60)$$

Since the integral in the final inequality is $\Gamma(0)\tau_{\text{int}}$, which is finite, everything must go to zero as $T \rightarrow \infty$. \square

Remark. From the proof, we can see that the rate of convergence of the mean-square of $\|\bar{X}(T)\|_2^2$ is (at least) $O(1/T)$. This would give a root-mean-square (rms) convergence rate of $O(1/\sqrt{T})$, which is what the naive statistician who ignored inter-temporal dependence would expect from the central limit theorem. (This ergodic theorem says *nothing* about the form of the distribution of $\bar{X}(T)$ for large T . We will see that, under some circumstances, it *is* Gaussian, but that needs stronger assumptions [forms of “mixing”] than we have imposed.) The naive statistician would expect that the mean-square time average would go like $\Gamma(0)/T$, since $\Gamma(0) = \mathbf{E}[X^2(t)] = \mathbf{Var}[X(t)]$. The proportionality constant is instead $\int_0^\infty d\tau |\Gamma(\tau)|$. This is equal to the naive guess for white noise, and for other collections of IID variables, but not in the general case. This leads to the following

Corollary 294 (Convergence Rate in the Mean-Square Ergodic Theorem) *Under the conditions of Theorem 293,*

$$\mathbf{Var}[\bar{X}(T)] \leq 2\mathbf{Var}[X(0)] \frac{\tau_{\text{int}}}{T} \quad (22.61)$$

PROOF: Since $X(t)$ is centered, $\mathbf{E} [\bar{X}(T)] = 0$, and $\|\bar{X}(T)\|_2^2 = \mathbf{Var} [\bar{X}(T)]$. Everything else follows from re-arranging the bound in the proof of Theorem 293, Definition 292, and the fact that $\Gamma(0) = \mathbf{Var} [X(0)]$. \square

As a consequence of the corollary, if $T \gg \tau_{\text{int}}$, then the variance of the time average is negligible compared to the variance at any one time.

22.3.2 Mean-Square Ergodicity Based on the Spectrum

Let's warm up with some lemmas of a technical nature. The first relates the jumps of the spectral process $\tilde{X}(\nu)$ to the jumps of the spectral function $S(\nu)$.

Lemma 295 (Mean-Square Jumps in the Spectral Process) *For a weakly stationary process, $\mathbf{E} \left[\left| \Delta \tilde{X}(\nu) \right|^2 \right] = \Delta S(\nu)$.*

PROOF: This follows directly from the Wiener-Khinchin relation (Theorem 290). \square

Lemma 296 (The Jump in the Spectral Function) *The jump of the spectral function at ν is given by*

$$\Delta S(\nu) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \Gamma(\tau) e^{i\nu\tau} d\tau \quad (22.62)$$

PROOF: This is a basic inversion result for characteristic functions. It should become plausible by thinking of this as getting the Fourier transform of Γ as T grows. \square

Lemma 297 (Existence of L_2 Limits for Time Averages) *If X is weakly stationary, then for any real f , $\bar{e^{ift}} X(T)$ converges in L_2 to $\Delta \tilde{X}(f)$.*

PROOF: Start by looking at the squared modulus of the time average for finite time.

$$\left| \frac{1}{T} \int_0^T e^{ift} X(t) dt \right|^2 \quad (22.63)$$

$$\begin{aligned} &= \frac{1}{T^2} \int_0^T \int_0^T e^{-if(t_1-t_2)} X^\dagger(t_1) X(t_2) dt_1 dt_2 \\ &= \frac{1}{T^2} \int_0^T \int_0^T e^{-if(t_1-t_2)} \int_{-\infty}^{\infty} e^{i\nu_1 t_1} d\tilde{X}_{\nu_1} \int_{-\infty}^{\infty} e^{-i\nu_2 t_2} d\tilde{X}_{\nu_2} \quad (22.64) \end{aligned}$$

$$= \frac{1}{T^2} \int_0^T \int_{-\infty}^{\infty} dt_1 d\tilde{X}_{\nu_1} e^{it_1(f-\nu_1)} \int_0^T \int_{-\infty}^{\infty} dt_2 d\tilde{X}_{\nu_2} e^{-it_2(f-\nu_2)} \quad (22.65)$$

As $T \rightarrow \infty$, these integrals pick out $\Delta \tilde{X}(f)$ and $\Delta \tilde{X}^\dagger(f)$. So, $\overline{e^{ift} X}(T) \xrightarrow{L_2} \Delta \tilde{X}(f)$. \square

Notice that the limit provided by the lemma is a random quantity. What's really desired, in most applications, is convergence to a *deterministic* limit, which here would mean convergence (in L_2) to zero.

Lemma 298 (The Mean-Square Ergodic Theorem) *If X is weakly stationary, and $\mathbf{E}[X(t)] = 0$, then $\bar{X}(t)$ converges in L_2 to 0 iff*

$$\lim T^{-1} \int_0^T d\tau \Gamma(\tau) = 0 \quad (22.66)$$

PROOF: Taking $f = 0$ in Lemma 297, $\bar{X}(T) \xrightarrow{L_2} \Delta\tilde{X}(0)$, the jump in the spectral function at zero. Let's show that the (i) expectation of this jump is zero, and that (ii) its variance is given by the integral expression on the LHS of Eq. 22.66. For (i), because $\bar{X}(T) \xrightarrow{L_2} Y$, we know that $\mathbf{E}[\bar{X}(T)] \rightarrow \mathbf{E}[Y]$. But $\mathbf{E}[\bar{X}(T)] = \overline{\mathbf{E}[X]}(T) = 0$. So $\mathbf{E}[\Delta\tilde{X}(0)] = 0$. For (ii), Lemma 295, plus the fact that $\mathbf{E}[\Delta\tilde{X}(0)] = 0$, shows that the variance is equal to the jump in the spectrum at 0. But, by Lemma 296 with $\nu = 0$, that jump is exactly the LHS of Eq. 22.66. \square

Remark 1: Notice that if the integral time is finite, then the integral condition on the autocovariance is automatically satisfied, but not vice versa, so the hypotheses here are strictly weaker than in Theorem 293.

Remark 2: One interpretation of the theorem is that the time-average is converging on the zero-frequency component of the spectral process. Intuitively, all the other components are being killed off by the time-averaging, because the time-integral of a sinusoidal function is bounded, but the denominator in a time average is unbounded. The only part left is the zero-frequency component, whose time integral can also grow linearly with time. If there is a jump at 0, then this has finite variance; if not, not.

Remark 3: Lemma 297 establishes the L_2 convergence of time-averages of the form

$$\frac{1}{T} \int_0^T e^{ift} X(t) dt$$

for any real f . Specifically, from Lemma 295, the mean-square of this variable is converging on the jump in the spectrum at f . Multiplying $X(t)$ by e^{ift} makes the old frequency f component the new frequency zero component, so it is the surviving term. While the ergodic theorem itself only needs the $f = 0$ case, this result is useful in connection with estimating spectra from time series (Doob, 1953, ch. X, §7).

22.4 Exercises

Exercise 22.1 (Mean-Square Ergodicity in Discrete Time) *It is often convenient to have a mean-square ergodic theorem for discrete-time sequences rather than continuous-time processes. If the dt in the definition of \bar{X} is re-interpreted as counting measure on \mathbb{N} , rather than Lebesgue measure on \mathbb{R}^+ , does the proof of Theorem 293 remain valid? (If yes, say why; if no, explain where the argument fails.)*

Exercise 22.2 (Mean-Square Ergodicity with Non-Zero Mean) *State and prove a version of Theorem 293 which does not assume that $\mathbf{E}[X(t)] = 0$.*

Exercise 22.3 (Functions of Weakly Stationary Processes) *Suppose X is a weakly stationary process, and f is a measurable function such that $\|f(X_0)\|_2 < \infty$. Is $f(X)$ a weakly stationary process? (If yes, prove it; if not, give a counter-example.)*

Exercise 22.4 (Ergodicity of the Ornstein-Uhlenbeck Process?) *Suppose the Ornstein-Uhlenbeck process has its invariant distribution as its initial distribution, and is therefore weakly stationary. Does Theorem 293 apply?*

Part V

Ergodic Theory

Chapter 23

Ergodic Properties and Ergodic Limits

Section 23.1 gives a general orientation to ergodic theory, which we will study in discrete time.

Section 23.2 introduces dynamical systems and their invariants, the setting in which we will prove our ergodic theorems.

Section 23.3 considers time averages, defines what we mean for a function to have an ergodic property (its time average converges), and derives some consequences.

Section 23.4 defines asymptotic mean stationarity, and shows that, with AMS dynamics, the limiting time average is equivalent to conditioning on the invariant sets.

23.1 General Remarks

To begin our study of ergodic theory, let us consider a famous¹ line from Gnedenko and Kolmogorov (1954, p. 1):

In fact, all epistemological value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, nonrandom regularity.

Now, this is how Gnedenko and Kolmogorov introduced their classic study of the limit laws for *independent* random variables, but most of the random phenomena we encounter around us are not independent. Ergodic theory is a study of how large-scale *dependent* random phenomena nonetheless create nonrandom regularity. The classical limit laws for IID variables X_1, X_2, \dots assert that,

¹Among mathematical scientists, anyway.

under the right conditions, sample averages converge on expectations,

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mathbf{E}[X_i]$$

where the sense of convergence can be “almost sure” (strong law of large numbers), “ L_p ” (p^{th} mean), “in probability” (weak law), etc., depending on the hypotheses we put on the X_i . One meaning of this convergence is that sufficiently large random samples are representative of the entire population — that the sample mean makes a good estimate of $\mathbf{E}[X]$.

The ergodic theorems, likewise, assert that for *dependent* sequences X_1, X_2, \dots , time averages converge on expectations

$$\frac{1}{t} \sum_{i=1}^t X_i \rightarrow \mathbf{E}[X_\infty | \mathcal{I}]$$

where X_∞ is some limiting random variable, or in the most useful cases a *non*-random variable, and \mathcal{I} is a σ -field representing some sort of asymptotic information. Once again, the mode of convergence will depend on the kind of hypotheses we make about the random sequence X . Once again, the interpretation is that a *single* sample path is representative of an entire distribution over sample paths, *if* it goes on long enough. The IID laws of large numbers are, in fact, special cases of the corresponding ergodic theorems.

Section 22.3 proved a mean-square (L_2) ergodic theorem for weakly stationary continuous-parameter processes.² The next few chapters, by contrast, will develop ergodic theorems for non-stationary discrete-parameter processes.³ This is a little unusual, compared to most probability books, so let me say a word or two about why. (1) Results we get will include stationary processes as special cases, but stationarity fails for many applications where ergodicity (in a suitable sense) holds. So this is more general and more broadly applicable. (2) Our results will all have continuous-time analogs, but the algebra is a lot cleaner in discrete time. (3) Some of the most important applications (for people like you!) are to statistical inference and learning with dependent samples, and to Markov chain Monte Carlo, and both of those are naturally discrete-parameter processes. We will, however, stick to continuous state spaces.

23.2 Dynamical Systems and Their Invariants

It is a very remarkable fact — but one with deep historical roots (von Plato, 1994, ch. 3) — that the way to get regular limits for stochastic processes is to first turn them into irregular deterministic dynamical systems, and then let averaging smooth away the irregularity. This section will begin by laying out dynamical systems, and their invariant sets and functions, which will be the foundation for what follows.

²Can you identify X_∞ and \mathcal{I} for this case?

³In doing so, I’m ripping off Gray (1988), especially chapters 6 and 7.

Definition 299 (Dynamical System) A dynamical system consists of a measurable space Ξ , a σ -field \mathcal{X} on Ξ , a probability measure μ defined on \mathcal{X} , and a mapping $T : \Xi \mapsto \Xi$ which is \mathcal{X}/\mathcal{X} -measurable.

Remark: Measure-preserving transformations (Definition 53) are special cases of dynamical systems. Since (Theorem 52) every strongly stationary process can be represented by a measure-preserving transformation, namely the shift (Definition 48), the theory of ergodicity for dynamical systems which we'll develop is easily seen to include the usual ergodic theory of strictly-stationary processes as a special case. Thus, at the cost of going to the infinite-dimensional space of sample paths, we can always make it the case that the time-evolution is completely deterministic, and the only stochastic component to the process is its initial condition.

Lemma 300 (Dynamical Systems are Markov Processes) Let Ξ, \mathcal{X}, μ, T be a dynamical system. Let $\mathcal{L}(X_1) = \mu$, and define $X_t = T^{t-1}X_1$. Then the X_t form a Markov process, with evolution operator K defined through $Kf(x) = f(Tx)$.

PROOF: For every $x \in \Xi$ and $B \in \mathcal{X}$, define $\kappa(x, B) \equiv \mathbf{1}_B(Tx)$. For fixed x , this is clearly a probability measure (specifically, the δ measure at Tx). For fixed B , this is a measurable function of x , because T is a measurable mapping. Hence, $\kappa(x, B)$ is a probability kernel. So, by Theorem 106, the X_t form a Markov process. By definition, $\mathbf{E}[f(X_1)|X_0 = x] = Kf(x)$. But the expectation is in this case just $f(Tx)$. \square

Notice that, as a consequence, there is a corresponding operator, call it U , which takes signed measures (defined over \mathcal{X}) to signed measures, and specifically takes probability measures to probability measures.

Definition 301 (Observable) A function $f : \Xi \mapsto \mathbb{R}$ which is \mathbb{B}/\mathcal{X} measurable is an observable of the dynamical system Ξ, \mathcal{X}, μ, T .

Pretty much all of what follows would work if the observables took values in any real or complex vector space, but that situation can be built up from this one.

Definition 302 (Invariant Functions, Sets and Measures) A function is invariant, under the action of a dynamical system, if $f(Tx) = f(x)$ for all $x \in \Xi$, or equivalently if $Kf = f$ everywhere. An event $B \in \mathcal{X}$ is invariant if its indicator function is an invariant function. A measure ν is invariant if it is preserved by T , i.e. if $\nu(C) = \nu(T^{-1}C)$ for all $C \in \mathcal{X}$, equivalently if $U\nu = \nu$.

Lemma 303 (Invariant Sets are a σ -Algebra) The class \mathcal{I} of all measurable invariant sets in Ξ forms a σ -algebra.

PROOF: Clearly, Ξ is invariant. The other properties of a σ -algebra follow because set-theoretic operations (union, complementation, etc.) commute with taking inverse images. \square

Lemma 304 (Invariant Sets and Observables) *An observable is invariant if and only if it is \mathcal{I} -measurable. Consequently, \mathcal{I} is the σ -field generated by the invariant observables.*

PROOF: “If”: Pick any Borel set B . Since $f = f \circ T$, $f^{-1}(B) = (f \circ T)^{-1}(B) = T^{-1}f^{-1}(B)$. Hence $f^{-1}(B) \in \mathcal{I}$. Since the inverse image of every Borel set is in \mathcal{I} , f is \mathcal{I} -measurable. “Only if”: Again, pick any Borel set B . By assumption, $f^{-1}(B) \in \mathcal{I}$, so $f^{-1}(B) = T^{-1}f^{-1}(B) = (f \circ T)^{-1}(B)$, so the inverse image of under Tf of any Borel set is an invariant set, implying that $f \circ T$ is \mathcal{I} -measurable. Since, for every B , $f^{-1}(B) = (f \circ T)^{-1}(B)$, we must have $f \circ T = f$. The consequence follows. \square

Definition 305 (Infinitely Often, i.o.) *For any set $C \in \mathcal{X}$, the set C infinitely often, $C_{i.o.}$, consists of all those points in Ξ whose trajectories visit C infinitely often, $C_{i.o.} \equiv \limsup_t T^{-t}C$.*

Lemma 306 (“Infinitely often” implies invariance) *For every $C \in \mathcal{X}$, $C_{i.o.}$ is invariant.*

PROOF: Exercise 23.1. \square

Definition 307 (Invariance Almost Everywhere) *A measurable function is invariant μ -a.e., or almost invariant, when*

$$\mu \{x \in \Xi | \forall n, f(x) = K^n f(x)\} = 1 \quad (23.1)$$

A measurable set is invariant μ -a.e., when its indicator function is almost invariant.

Remark 1: Some of the older literature annoyingly calls these objects *totally invariant*.

Remark 2: Invariance implies invariance μ -almost everywhere, for any μ .

Lemma 308 (Almost-invariant sets form a σ -algebra) *The almost-invariant sets form a σ -field, \mathcal{I}' , and an observable is almost invariant if and only if it is measurable with respect to this field.*

PROOF: Entirely parallel to that for the strict invariants. \square

Let’s close this section with a simple lemma, which will however be useful in approximation-by-simple-function arguments in building up expectations.

Lemma 309 (Invariance for simple functions) *A simple function, $f(x) = \sum_{k=1}^m a_k \mathbf{1}_{C_k}(x)$, is invariant if and only if all the sets $C_k \in \mathcal{I}$. Similarly, a simple function is almost invariant iff all the defining sets are almost invariant.*

PROOF: Exercise 23.2. \square

23.3 Time Averages and Ergodic Properties

For convenience, let's re-iterate the definition of a time average. The notation differs here a little from that given earlier, in a way which helps with discrete time.

Definition 310 (Time Averages) *The time-average of an observable f is the real-valued function*

$$A_t f(x) \equiv \frac{1}{t} \sum_{i=0}^{t-1} f(T^i x) \quad (23.2)$$

where A_t is the operator taking functions to their time-averages.

Lemma 311 (Time averages are observables) *For every t , the time-average of an observable is an observable.*

PROOF: The class of measurable functions is closed under finite iterations of arithmetic operations. \square

Definition 312 (Ergodic Property) *An observable f has the ergodic property when $A_t f(x)$ converges as $t \rightarrow \infty$ for μ -almost-all x . An observable has the mean ergodic property when $A_t f(x)$ converges in $L_1(\mu)$, and similarly for the other L_p ergodic properties. If for some class of functions \mathcal{D} , every $f \in \mathcal{D}$ has an ergodic property, then the class \mathcal{D} has that ergodic property.*

Remark. Notice that what is required for f to have the ergodic property is that almost every initial point has *some* limit for its time average,

$$\mu \left\{ x \in \Xi \mid \exists r \in \mathbb{R} : \lim_{t \rightarrow \infty} A_t f(x) = r \right\} = 1 \quad (23.3)$$

This does *not* mean that there is some *common* limit for almost every initial point,

$$\exists r \in \mathbb{R} : \mu \left\{ x \in \Xi \mid \lim_{t \rightarrow \infty} A_t f(x) = r \right\} = 1 \quad (23.4)$$

Similarly, a class of functions has the ergodic property if all of their time averages converge; they do not have to converge uniformly.

Definition 313 (Ergodic Limit) *If an observable f has the ergodic property, define $Af(x)$ to be the limit of $A_t f(x)$ where that exists, and 0 elsewhere. The domain of A consists of all and only the functions with ergodic properties.*

Observe that

$$Af(x) = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^t K^n f(x) \quad (23.5)$$

That is, A is the limit of an arithmetic mean of conditional expectations. This suggests that it should itself have many of the properties of conditional expectations. In fact, under a reasonable condition, we will see that $Af = \mathbf{E}[f|\mathcal{I}]$, expectation conditional on the σ -algebra of invariant sets. We'll check first that A has the properties we'd want from a conditional expectation.

Lemma 314 (Linearity of Ergodic Limits) *A is a linear operator, and its domain is a linear space.*

PROOF: If c is any real number, then $A_t c f(x) = c A_t f(x)$, and so clearly, if the limit exists, $A c f(x) = c A f(x)$. Similarly, $A_t(f + g)(x) = A_t f(x) + A_t g(x)$, so if f and g both have ergodic properties, then so does $f + g$, and $A(f + g)(x) = A f(x) + A g(x)$. \square

Lemma 315 (Non-negativity of ergodic limits) *If $f \in \text{Dom} A$, and, for all $n \geq 0$, $f T^n \geq 0$ a.e., then $A f(x) \geq 0$ a.e.*

PROOF: The event $A f(x) < 0$ is a sub-event of $\bigcup_n \{f(T^n(x)) < 0\}$. Since the union of a countable collection of measure zero events has measure zero, $A f(x) \geq 0$ almost everywhere. \square

Notice that our hypothesis is that $f T^n \geq 0$ a.e. for all n , not just that $f \geq 0$. The reason we need the stronger assumption is that the transformation T might map every point to the bad set of f ; the lemma guards against that. Of course, if $f(x) \geq 0$ for all, and not just almost all, x , then the bad set is non-existent, and $A f \geq 0$ follows automatically.

Lemma 316 (Constant functions have the ergodic property) *The constant function 1 has the ergodic property. Consequently, so does every other constant function.*

PROOF: For every n , $1(T^n x) = 1$. Hence $A_t 1(x) = 1$ for all t , and so $A 1(x) = 1$. Extension to other constants follows by linearity (Lemma 314). \square

Remember that for any time-evolution operator K , $K 1 = 1$.

Lemma 317 (Ergodic limits are invariantifying) *If $f \in \text{Dom}(A)$, then, for all n , $f \circ T^n$ is too, and $A f(x) = A f \circ T^n(x)$. Or, $A K^n f(x) = A f(x)$.*

PROOF: Start with $n = 1$, and show that the discrepancy goes to zero.

$$A K f(x) - A f(x) = \lim_t \frac{1}{t} \sum_{i=0}^t (K^{i+1} f(x) - K^i f(x)) \quad (23.6)$$

$$= \lim_t \frac{1}{t} (K^t f(x) - f(x)) \quad (23.7)$$

Since $A f(x)$ exists a.e., we know that the series $t^{-1} \sum_{i=0}^{t-1} K^i f(x)$ converges a.e., implying that $(t+1)^{-1} K^t f(x) \rightarrow 0$ a.e.. But $t^{-1} = \frac{t+1}{t} (t+1)^{-1}$, and for large t , $t+1/t < 2$. Hence $(t+1)^{-1} K^t f(x) \leq t^{-1} K^t f(x) \leq 2(t+1)^{-1} K^t f(x)$, implying that $t^{-1} K^t f(x)$ itself goes to zero (a.e.). Similarly, $t^{-1} f(x)$ must go to zero. Thus, overall, we have $A K f(x) = A f(x)$ a.e., and $K f(x) \in \text{Dom}(A)$. \square

Lemma 318 (Ergodic limits are invariant functions) *If $f \in \text{Dom}(A)$, then $A f$ is an invariant, and \mathcal{I} -measurable.*

PROOF: Af exists, so (previous lemma) AKf exists and is equal to Af (almost everywhere). But $AKf(x) = Af(Tx)$, by definition, hence Af is invariant, i.e., $KAf = AKf = Af$. Measurability follows from Lemma 304. \square

Lemma 319 (Ergodic limits and invariant indicator functions) *If $f \in \text{Dom}(A)$, and B is any set in \mathcal{I} , then $A(\mathbf{1}_B(x)f(x)) = \mathbf{1}_B(x)Af(x)$.*

PROOF: For every n , $\mathbf{1}_B(T^n x)f(T^n x) = \mathbf{1}_B(x)f(T^n x)$, since $x \in B$ iff $T^n x \in B$. So, for all finite t , $A_t(\mathbf{1}_B(x)f(x)) = \mathbf{1}_B(x)A_t f(x)$, and the lemma follows by taking the limit. \square

Lemma 320 (Ergodic properties of sets and observables) *All indicator functions of measurable sets have ergodic properties if and only if all bounded observables have ergodic properties.*

PROOF: A standard approximation-by-simple-functions argument, as in the construction of Lebesgue integrals. \square

Lemma 321 (Expectations of ergodic limits) *Let f be bounded and have the ergodic property. Then Af is μ -integrable, and $\mathbf{E}[Af(X)] = \mathbf{E}[f(X)]$, where $\mathcal{L}(X) = \mu$.*

PROOF: Since f is bounded, it is integrable. Hence $A_t f$ is bounded, too, for all t , and $A_t f(X)$ is an integrable random variable. A sequence of bounded, integrable random variables is uniformly integrable. Uniform integrability, plus the convergence $A_t f(x) \rightarrow Af(x)$ for μ -almost-all x , gives us that $\mathbf{E}[Af(X)]$ exists and is equal to $\lim \mathbf{E}[A_t f(X)]$ via Fatou's lemma. (See e.g., Theorem 117 in the notes to 36-752.)

Now use the invariance of Af , i.e., the fact that $Af(X) = Af(TX)$ μ -a.s.

$$0 = \mathbf{E}[Af(TX)] - \mathbf{E}[Af(X)] \quad (23.8)$$

$$= \lim \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[K^n f(TX)] - \lim \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[K^n f(X)] \quad (23.9)$$

$$= \lim \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[K^n f(TX)] - \mathbf{E}[K^n f(X)] \quad (23.10)$$

$$= \lim \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[K^{n+1} f(X)] - \mathbf{E}[K^n f(X)] \quad (23.11)$$

$$= \lim \frac{1}{t} (\mathbf{E}[K^t f(X)] - \mathbf{E}[f(X)]) \quad (23.12)$$

Hence

$$\mathbf{E}[Af] = \lim \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[K^n f(X)] = \mathbf{E}[f(X)] \quad (23.13)$$

as was to be shown. \square

Lemma 322 (Convergence of Ergodic Limits) *If f is as in Lemma 321, then $A_t f \rightarrow f$ in $L_1(\mu)$.*

PROOF: From Lemma 321, $\lim \mathbf{E}[A_t f(X)] = \mathbf{E}[f(X)]$. Since the variables $A_t f(X)$ are uniformly integrable (as we saw in the proof of that lemma), it follows (Proposition 4.12 in Kallenberg, p. 68) that they also converge in $L_1(\mu)$. \square

Lemma 323 (Cesàro Mean of Expectations) *Let f be as in Lemmas 321 and 322, and $B \in \mathcal{X}$ be an arbitrary measurable set. Then*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mathbf{E}[\mathbf{1}_B(X) K^n f(X)] = \mathbf{E}[\mathbf{1}_B(X) f(X)] \quad (23.14)$$

where $\mathcal{L}(X) = \mu$.

PROOF: Let's write out the expectations explicitly as integrals.

$$\left| \int_B f(x) d\mu - \frac{1}{t} \sum_{n=0}^{t-1} \int_B K^n f(x) d\mu \right| \quad (23.15)$$

$$= \left| \int_B f(x) - \frac{1}{t} \sum_{n=0}^{t-1} K^n f(x) d\mu \right|$$

$$= \left| \int_B f(x) - A_t f(x) d\mu \right| \quad (23.16)$$

$$\leq \int_B |f(x) - A_t f(x)| d\mu \quad (23.17)$$

$$\leq \int |f(x) - A_t f(x)| d\mu \quad (23.18)$$

$$= \|f - A_t f\|_{L_1(\mu)} \quad (23.19)$$

But (previous lemma) these functions converge in $L_1(\mu)$, so the limit of the norm of their difference is zero. \square

Boundedness is not essential.

Corollary 324 (Replacing Boundedness with Uniform Integrability) *Lemmas 321, 322 and 323 hold for any integrable observable $f \in \text{Dom}(A)$, bounded or not, provided that $A_t f$ is a uniformly integrable sequence.*

PROOF: Examining the proofs shows that the boundedness of f was important only to establish the uniform integrability of $A_t f$. \square

23.4 Asymptotic Mean Stationarity

Next, we come to an important concept which will prove to be necessary and sufficient for the most important ergodic properties to hold.

Definition 325 (Asymptotically Mean Stationary) *A dynamical system is asymptotically mean stationary when, for every $C \in \mathcal{X}$, the limit*

$$m(C) \equiv \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(T^{-n}C) \quad (23.20)$$

exists, and the set function m is its stationary mean.

Remark 1: It might've been more logical to call this "asymptotically measure stationary", or something, but I didn't make up the names...

Remark 2: Symbolically, we can write

$$m = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} U^n \mu$$

where U is the operator taking measures to measures. This leads us to the next proposition.

Lemma 326 (Stationary Implies Asymptotically Mean Stationary) *If a dynamical system is stationary, i.e., T preserves the measure μ , then it is asymptotically mean stationary, with stationary mean μ .*

PROOF: If T preserves μ , then for every measurable set, $\mu(C) = \mu(T^{-1}C)$. Hence every term in the sum in Eq. 23.20 is $\mu(C)$, and consequently the limit exists and is equal to $\mu(C)$. \square

Proposition 327 (Vitali-Hahn Theorem) *If m_t are a sequence of probability measures on a common σ -algebra \mathcal{X} , and $m(C)$ is a set function such that $\lim_t m_t(C) = m(C)$ for all $C \in \mathcal{X}$, then m is a probability measure on \mathcal{X} .*

PROOF: This is a standard result from measure theory. \square

Theorem 328 (Stationary Means are Invariant Measures) *If a dynamical system is asymptotically mean stationary, then its stationary mean is an invariant probability measure.*

PROOF: For every t , let $m_t(C) = \frac{1}{t} \sum_{n=0}^{t-1} \mu(T^{-n}(C))$. Then m_t is a linear combination of probability measures, hence a probability measure itself. Since, for every $C \in \mathcal{X}$, $\lim m_t(C) = m(C)$, by Definition 325, Proposition 327 says that $m(C)$ is also a probability measure. It remains to check invariance.

$$m(C) - m(T^{-1}C) \quad (23.21)$$

$$\begin{aligned} &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(T^{-n}(C)) - \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(T^{-n}(T^{-1}C)) \\ &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(T^{-n-1}C) - \mu(T^{-n}C) \end{aligned} \quad (23.22)$$

$$= \lim_{t \rightarrow \infty} \frac{1}{t} (\mu(T^{-t}C) - \mu(C)) \quad (23.23)$$

Since the probability measure of any set is at most 1, the difference between two probabilities is at most 1, and so $m(C) = m(T^{-1}C)$, for all $C \in \mathcal{X}$. But this means that m is invariant under T (Definition 53). \square

Remark: Returning to the symbolic manipulations, if μ is AMS with stationary mean m , then $Um = m$ (because m is invariant), and so we can write $\mu = m + (\mu - m)$, knowing that $\mu - m$ goes to zero under averaging. Speaking loosely (this can be made precise, at the cost of a fairly long excursion) m is an eigenvector of U (with eigenvalue 1), and $\mu - m$ lies in an orthogonal direction, along which U is contracting, so that, under averaging, it goes away, leaving only m , which is like the projection of the original measure μ on to the invariant manifold of U .

The relationship between an AMS measure μ and its stationary mean m is particularly simple on invariant sets: they are equal there. A slightly more general theorem is actually just as easy to prove, however, so we'll do that.

Lemma 329 (Expectations of Almost-Invariant Functions) *If μ is AMS with limit m , and f is an observable which is invariant μ -a.e., then $\mathbf{E}_\mu[f] = \mathbf{E}_m[f]$.*

PROOF: Let C be any almost invariant set. Then, for any t , C and $T^{-t}C$ differ by, at most, a set of μ -measure 0, so that $\mu(C) = \mu(T^{-t}C)$. The definition of the stationary mean (Equation 23.20) then gives $\mu(C) = m(C)$, or $\mathbf{E}_\mu[\mathbf{1}_C] = \mathbf{E}_m[\mathbf{1}_C]$, i.e., the result holds for indicator functions. By Lemma 309, this then extends to simple functions. The usual arguments then take us to all functions which are measurable with respect to \mathcal{I}' , the σ -field of almost-invariant sets, but this (Lemma 308) is the class of all almost-invariant functions. \square

Lemma 330 (Limit of Cesàro Means of Expectations) *If μ is AMS with stationary mean m , and f is a bounded observable,*

$$\lim_{t \rightarrow \infty} \mathbf{E}_\mu[A_t f] = \mathbf{E}_m[f] \quad (23.24)$$

PROOF: By Eq. 23.20, this must hold when f is an indicator function. By the linearity of A_t and of expectations, it thus holds for simple functions, and so for general measurable functions, using boundedness to exchange limits and expectations where necessary. \square

Lemma 331 (Expectation of the Ergodic Limit is the AMS Expectation) *If f is a bounded observable in $\text{Dom}(A)$, and μ is AMS with stationary mean m , then $\mathbf{E}_\mu[Af] = \mathbf{E}_m[f]$.*

PROOF: From Lemma 323, $\mathbf{E}_\mu[Af] = \lim_{t \rightarrow \infty} \mathbf{E}_\mu[A_t f]$. From Lemma 330, the latter is $\mathbf{E}_m[f]$. \square

Remark: Since Af is invariant, we've got $\mathbf{E}_\mu[Af] = \mathbf{E}_m[Af]$, from Lemma 329, but that's not the same.

Corollary 332 (Replacing Boundedness with Uniform Integrability) *Lemmas 330 and 331 continue to hold if f is not bounded, but $A_t f$ is uniformly integrable (μ).*

PROOF: As in Corollary 324. \square

Theorem 333 (Ergodic Limits of Bounded Observables are Conditional Expectations) *If μ is AMS, with stationary mean m , and the dynamics have ergodic properties for all the indicator functions, then, for any bounded observable f ,*

$$Af = \mathbf{E}_m[f|\mathcal{I}] \quad (23.25)$$

with probability 1 under both μ and m .

PROOF: Begin by showing this holds for the indicator functions of measurable sets, i.e., when $f = \mathbf{1}_C$ for arbitrary measurable C . By assumption, this function has the ergodic property.

By Lemma 318, $A\mathbf{1}_C$ is an invariant function. Pick any set $B \in \mathcal{I}$, so that $\mathbf{1}_B$ is also invariant. By Lemma 319, $A(\mathbf{1}_B\mathbf{1}_C) = \mathbf{1}_BA\mathbf{1}_C$, which is invariant (as a product of invariant functions). So Lemma 329 gives

$$\mathbf{E}_\mu[\mathbf{1}_BA\mathbf{1}_C] = \mathbf{E}_m[\mathbf{1}_BA\mathbf{1}_C] \quad (23.26)$$

while Lemma 331 says

$$\mathbf{E}_\mu[A(\mathbf{1}_B\mathbf{1}_C)] = \mathbf{E}_m[\mathbf{1}_B\mathbf{1}_C] \quad (23.27)$$

Since the left-hand sides are equal, the right-hand sides must be equal as well, so

$$m(B \cap C) = \mathbf{E}_m[\mathbf{1}_B\mathbf{1}_C] \quad (23.28)$$

$$= \mathbf{E}_m[\mathbf{1}_BA\mathbf{1}_C] \quad (23.29)$$

Since this holds for all invariant sets $B \in \mathcal{I}$, we conclude that $A\mathbf{1}_C$ must be a version of the conditional probability $m(C|\mathcal{I})$.

From Lemma 320, every bounded observable has the ergodic property. The proof above then extends naturally to arbitrary bounded f . \square

Corollary 334 (Ergodic Limits of Integrable Observables are Conditional Expectations) *Equation 23.25 continues to hold if $A_t f$ are uniformly μ -integrable, or f is m -integrable.*

PROOF: Exercise 23.3. \square

23.5 Exercises

Exercise 23.1 (“Infinitely often” implies invariant) *Prove Lemma 306.*

Exercise 23.2 (Invariant simple functions) *Prove Lemma 309.*

Exercise 23.3 (Ergodic limits of integrable observables) *Prove Corollary 334.*

Chapter 24

The Almost-Sure Ergodic Theorem

This chapter proves Birkhoff's ergodic theorem, on the almost-sure convergence of time averages to expectations, under the assumption that the dynamics are asymptotically mean stationary.

This is not the usual proof of the ergodic theorem, as you will find in e.g. Kallenberg. Rather, it uses the AMS machinery developed in the last lecture, following Gray (1988, sec. 7.2), in turn following Katznelson and Weiss (1982). The central idea is that of “blocking”: break the infinite sequence up into non-overlapping blocks, show that each block is well-behaved, and conclude that the whole sequence is too. This is a very common technique in modern ergodic theory, especially among information theorists. In pure probability theory, the usual proof of the ergodic theorem uses a result called the “maximal ergodic lemma”, which is clever but somewhat obscure, and doesn't seem to generalize well to non-stationary processes: see Kallenberg, ch. 10.

We saw, at the end of the last chapter, that if time-averages converge in the long run, they converge on conditional expectations. Our work here is showing that they (almost always) converge. We'll do this by showing that their lim infs and lim sups are (almost always) equal. This calls for some preliminary results about the upper and lower limits of time-averages.

Definition 335 (Lower and Upper Limiting Time Averages) *For any observable f , define the lower and upper limits of its time averages as, respectively,*

$$\underline{A}f(x) \equiv \liminf_{t \rightarrow \infty} A_t f(x) \quad (24.1)$$

$$\overline{A}f(x) \equiv \limsup_{t \rightarrow \infty} A_t f(x) \quad (24.2)$$

Define L_f as the set of x where the limits coincide:

$$L_f \equiv \{x \mid \underline{A}f(x) = \overline{A}f(x)\} \quad (24.3)$$

Lemma 336 (Lower and Upper Limiting Time Averages are Invariant) *For each observable f , $\underline{A}f$ and $\overline{A}f$ are invariant functions.*

PROOF: Use our favorite trick, and write $A_tf(Tx) = \frac{t+1}{t}A_{t+1}f(x) - f(x)/t$. Clearly, the limsup and liminf of this expression will equal the limsup and liminf of $A_{t+1}f(x)$, which is the same as that of $A_tf(x)$. \square

Lemma 337 (“Limits Coincide” is an Invariant Event) *For each f , the set L_f is invariant.*

PROOF: Since $\underline{A}f$ and $\overline{A}f$ are both invariant, they are both measurable with respect to \mathcal{I} (Lemma 304), so the set of x such that $\underline{A}f(x) = \overline{A}f(x)$ is in \mathcal{I} , therefore it is invariant (Definition 303). \square

Lemma 338 (Ergodic properties under AMS measures) *An observable f has the ergodic property with respect to an AMS measure μ if and only if it has it with respect to the stationary limit m .*

PROOF: By Lemma 337, L_f is an invariant set. But then, by Lemma 329, $m(L_f) = \mu(L_f)$. (Take $f = \mathbf{1}_{L_f}$ in the lemma.) f has the ergodic property with respect to μ iff $\mu(L_f) = 1$, so f has the ergodic property with respect to μ iff it has it with respect to m . \square

Theorem 339 (Birkhoff’s Almost-Sure Ergodic Theorem) *If a dynamical system is AMS with stationary mean m , then all bounded observables have the ergodic property, and with probability 1 (under both μ and m),*

$$Af = \mathbf{E}_m[f|\mathcal{I}] \quad (24.4)$$

for all $f \in L_1(m)$.

PROOF: From Theorem 333 and its corollaries, it is enough to prove that all $L_1(m)$ observables have ergodic properties to get Eq. 24.4. From Lemma 338, it is enough to show that the observables have ergodic properties in the stationary system Ξ, \mathcal{X}, m, T . (Accordingly, all expectations in the rest of this proof will be with respect to m .) Since any observable can be decomposed into its positive and negative parts, $f = f^+ - f^-$, assume, without loss of generality, that f is positive. Since $\overline{A}f \geq \underline{A}f$ everywhere, it suffices to show that $\mathbf{E}[\overline{A}f - \underline{A}f] \leq 0$. This in turn will follow from $\mathbf{E}[\overline{A}f] \leq \mathbf{E}[f] \leq \mathbf{E}[\underline{A}f]$. (Since f is bounded, the integrals exist.)

We’ll prove that $\mathbf{E}[\overline{A}f] \leq \mathbf{E}[f]$, by showing that the time average comes close to its limsup, but *from above* (in the mean). Proving that $\mathbf{E}[\underline{A}f] \geq \mathbf{E}[f]$ will be entirely parallel.

Since f is bounded, we may assume that $\overline{f} \leq M$ everywhere.

For every $\epsilon > 0$, for every x there exists a *finite* t such that

$$A_tf(x) \geq \overline{f}(x) - \epsilon \quad (24.5)$$

This is because \bar{f} is the limit of the *least* upper bounds. (You can see where this is going already — the time-average has to be close to its lim sup, but close *from above*.)

Define $t(x, \epsilon)$ to be the smallest t such that $\bar{f}(x) \leq \epsilon + A_t f(x)$. Then, since \bar{f} is invariant, we can add from time 0 to time $t(x, \epsilon) - 1$ and get:

$$\sum_{n=0}^{t(x, \epsilon)-1} K^n f(x) + \epsilon t(x, \epsilon) \geq \sum_{n=0}^{t(x, \epsilon)-1} K^n \bar{f}(x) \quad (24.6)$$

Define $B_N \equiv \{x | t(x, \epsilon) \geq N\}$, the set of “bad” x , where the sample average fails to reach a reasonable (ϵ) distance of the lim sup before time N . Because $t(x, \epsilon)$ is finite, $m(B_N)$ goes to zero as $N \rightarrow \infty$. Chose a N such that $m(B_N) \leq \epsilon/M$, and, for the corresponding bad set, drop the subscript. (We’ll see why this level is important presently.)

We’ll find it convenient to not deal directly with f , but with a related function which is better-behaved on the bad set B . Set $\tilde{f}(x) = M$ when $x \in B$, and $= f(x)$ elsewhere. Similarly, define $\tilde{t}(x, \epsilon)$ to be 1 if $x \in B$, and $t(x, \epsilon)$ elsewhere. Notice that $\tilde{t}(x, \epsilon) \leq N$ for all x . Something like Eq. 24.6 still holds for the nice-ified function \tilde{f} , specifically,

$$\sum_{n=0}^{\tilde{t}(x, \epsilon)-1} K^n \bar{f}(x) \leq \sum_{n=0}^{\tilde{t}(x, \epsilon)-1} K^n \tilde{f}(x) + \epsilon \tilde{t}(x, \epsilon) \quad (24.7)$$

If $x \in B$, this reduces to $\bar{f}(x) \leq M + \epsilon$, which is certainly true because $\bar{f}(x) \leq M$. If $x \notin B$, it will follow from Eq. 24.6, provided that $T^n x \notin B$, for all $n \leq \tilde{t}(x, \epsilon) - 1$. To see that this, in turn, must be true, suppose that $T^n x \in B$ for some such n . Because (we’re assuming) $n < t(x, \epsilon)$, it must be the case that

$$A_n f(x) < \bar{f}(x) - \epsilon \quad (24.8)$$

Otherwise, $t(x, \epsilon)$ would not be the *first* time at which Eq. 24.5 held true. Similarly, because $T^n x \in B$, while $x \notin B$, $t(T^n x, \epsilon) > N \geq t(x, \epsilon)$, and so

$$A_{t(x, \epsilon)-n} f(T^n x) < \bar{f}(x) - \epsilon \quad (24.9)$$

Combining the last two displayed equations,

$$A_{t(x, \epsilon)} f(x) < \bar{f}(x) - \epsilon \quad (24.10)$$

contradicting the definition of $t(x, \epsilon)$. Consequently, there can be no $n < t(x, \epsilon)$ such that $T^n x \in B$.

We are now ready to consider the time average $A_L f$ over a stretch of time of some considerable length L . We’ll break the time indices over which we’re averaging into blocks, each block ending when $T^t x$ hits B again. We need to make sure that L is sufficiently large, and it will turn out that $L \geq N/(\epsilon/M)$ suffices, so that $NM/L \leq \epsilon$. The end-points of the blocks are defined recursively,

starting with $b_0 = 0$, $b_{k+1} = b_k + \tilde{t}(T^{b_k}x, \epsilon)$. (Of course the b_k are implicitly dependent on x and ϵ and N , but suppress that for now, since these are constant through the argument.) The number of completed blocks, C , is the largest k such that $L - 1 \geq b_k$. Notice that $L - b_C \leq N$, because $\tilde{t}(x, \epsilon) \leq N$, so if $L - b_C > N$, we could squeeze in another block after b_C , contradicting its definition.

Now let's examine the sum of the lim sup over the trajectory of length L .

$$\sum_{n=0}^{L-1} K^n \bar{f}(x) = \sum_{k=1}^C \sum_{n=b_{k-1}}^{b_k} K^n \bar{f}(x) + \sum_{n=b_C}^{L-1} K^n \bar{f}(x) \quad (24.11)$$

For each term in the inner sum, we may assert that

$$\sum_{n=0}^{\tilde{t}(T^{b_k}x, \epsilon)-1} K^n \bar{f}(T^{b_k}x) \leq \sum_{n=0}^{\tilde{t}(T^{b_k}x, \epsilon)-1} K^n \tilde{f}(T^{b_k}x) + \epsilon \tilde{t}(T^{b_k}x, \epsilon) \quad (24.12)$$

on the strength of Equation 24.7, so, returning to the over-all sum,

$$\sum_{n=0}^{L-1} K^n \bar{f}(x) \leq \sum_{k=1}^C \sum_{n=b_{k-1}}^{b_k-1} K^n \tilde{f}(x) + \epsilon(b_k - b_{k-1}) + \sum_{n=b_C}^{L-1} K^n \bar{f}(x) \quad (24.13)$$

$$= \epsilon b_C + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) + \sum_{n=b_C}^{L-1} K^n \bar{f}(x) \quad (24.14)$$

$$\leq \epsilon b_C + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) + \sum_{n=b_C}^{L-1} M \quad (24.15)$$

$$\leq \epsilon b_C + M(L - 1 - b_C) + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) \quad (24.16)$$

$$\leq \epsilon b_C + M(N - 1) + \sum_{n=0}^{b_C-1} K^n \tilde{f}(x) \quad (24.17)$$

$$\leq \epsilon L + M(N - 1) + \sum_{n=0}^{L-1} K^n \tilde{f}(x) \quad (24.18)$$

where the last step, going from b_C to L , uses the fact that both ϵ and \tilde{f} are non-negative. Taking expectations of both sides,

$$\mathbf{E} \left[\sum_{n=0}^{L-1} K^n \bar{f}(X) \right] \leq \mathbf{E} \left[\epsilon L + M(N - 1) + \sum_{n=0}^{L-1} K^n \tilde{f}(X) \right] \quad (24.19)$$

$$\sum_{n=0}^{L-1} \mathbf{E} [K^n \bar{f}(X)] \leq \epsilon L + M(N - 1) + \sum_{n=0}^{L-1} \mathbf{E} [K^n \tilde{f}(X)] \quad (24.20)$$

$$L \mathbf{E} [\bar{f}(x)] \leq \epsilon L + M(N - 1) + L \mathbf{E} [\tilde{f}(X)] \quad (24.21)$$

using the fact that \bar{f} is invariant on the left-hand side, and that m is stationary on the other. Now divide both sides by L .

$$\mathbf{E}[\bar{f}(x)] \leq \epsilon + \frac{M(N-1)}{L} + \mathbf{E}[\tilde{f}(X)] \quad (24.22)$$

$$\leq 2\epsilon + \mathbf{E}[\tilde{f}(X)] \quad (24.23)$$

since $MN/L \leq \epsilon$. Now let's bound $\mathbf{E}[\tilde{f}(X)]$ in terms of $\mathbf{E}[f]$:

$$\mathbf{E}[\tilde{f}] = \int \tilde{f}(x) dm \quad (24.24)$$

$$= \int_{B^c} \tilde{f}(x) dm + \int_B \tilde{f}(x) dm \quad (24.25)$$

$$= \int_{B^c} f(x) dm + \int_B M dm \quad (24.26)$$

$$\leq \mathbf{E}[f] + \int_B M dm \quad (24.27)$$

$$= \mathbf{E}[f] + Mm(B) \quad (24.28)$$

$$\leq \mathbf{E}[f] + M \frac{\epsilon}{M} \quad (24.29)$$

$$= \mathbf{E}[f] + \epsilon \quad (24.30)$$

using the definition of \tilde{f} in Eq. 24.26, the non-negativity of f in Eq. 24.27, and the bound on $m(B)$ in Eq. 24.29. Substituting into Eq. 24.23,

$$\mathbf{E}[\bar{f}] \leq \mathbf{E}[f] + 3\epsilon \quad (24.31)$$

Since ϵ can be made arbitrarily small, we conclude that

$$\mathbf{E}[\bar{f}] \leq \mathbf{E}[f] \quad (24.32)$$

as was to be shown.

The proof of $\mathbf{E}[\bar{f}] \geq \mathbf{E}[f]$ proceeds in parallel, only the nice-ified function \tilde{f} is set equal to 0 on the bad set.

Since $\mathbf{E}[\bar{f}] \geq \mathbf{E}[f] \geq \mathbf{E}[\tilde{f}]$, we have that $\mathbf{E}[\bar{f} - \tilde{f}] \geq 0$. Since however it is always true that $\bar{f} - \tilde{f} \geq 0$, we may conclude that $\bar{f} - \tilde{f} = 0$ m -almost everywhere. Thus $m(L_f) = 1$, i.e., the time average converges m -almost everywhere. Since this is an invariant event, it has the same measure under μ and its stationary limit m , and so the time average converges μ -almost-everywhere as well. By Corollary 334, $Af = \mathbf{E}_m[f|I]$, as promised. \square

Corollary 340 (Birkhoff's Ergodic Theorem for Integrable Observables) *Under the assumptions of Theorem 339, all $L_1(m)$ functions have ergodic properties, and Eq. 24.4 holds a.e. m and μ .*

PROOF: We need merely show that the ergodic properties hold, and then the equation follows. To do so, define $\bar{f}_M(x) \equiv \bar{f}(x) \wedge M$, an upper-limited version of the lim sup. Reasoning entirely parallel to the proof of Theorem 339 leads to the conclusion that $\mathbf{E}[\bar{f}_M] \leq \mathbf{E}[f]$. Then let $M \rightarrow \infty$, and apply the monotone convergence theorem to conclude that $\mathbf{E}[\bar{f}] \leq \mathbf{E}[f]$; the rest of the proof goes through as before. \square

Chapter 25

Ergodicity and Metric Transitivity

Section 25.1 explains the ideas of ergodicity (roughly, there is only one invariant set of positive measure) and metric transitivity (roughly, the system has a positive probability of going from anywhere to anywhere), and why they are (almost) the same.

Section 25.2 gives some examples of ergodic systems.

Section 25.3 deduces some consequences of ergodicity, most importantly that time averages have deterministic limits (§25.3.1), and an asymptotic approach to independence between events at widely separated times (§25.3.2), admittedly in a very weak sense.

25.1 Metric Transitivity

Definition 341 (Ergodic Systems, Processes, Measures and Transformations) *A dynamical system Ξ, \mathcal{X}, μ, T is ergodic, or an ergodic system or an ergodic process when $\mu(C) = 0$ or $\mu(C) = 1$ for every T -invariant set C . μ is called a T -ergodic measure, and T is called a μ -ergodic transformation, or just an ergodic measure and ergodic transformation, respectively.*

Remark: Most authorities require a μ -ergodic transformation to also be measure-preserving for μ . But (Corollary 54) measure-preserving transformations are necessarily stationary, and we want to minimize our stationarity assumptions. So what most books call “ergodic”, we have to qualify as “stationary and ergodic”. (Conversely, when other people talk about processes being “stationary and ergodic”, they mean “stationary with only one ergodic component”; but of that, more later.

Definition 342 (Metric Transitivity) *A dynamical system is metrically transitive, metrically indecomposable, or irreducible when, for any two sets $A, B \in \mathcal{X}$, if $\mu(A), \mu(B) > 0$, there exists an n such that $\mu(T^{-n}A \cap B) > 0$.*

Remark: In dynamical systems theory, metric transitivity is contrasted with *topological* transitivity: T is topologically transitive on a domain D if for any two open sets $U, V \subseteq D$, the images of U and V remain in D , and there is an n such that $T^n U \cap V \neq \emptyset$. (See, e.g., Devaney (1992).) The “metric” in “metric transitivity” refers not to a distance function, but to the fact that a measure is involved. Under certain conditions, metric transitivity in fact implies topological transitivity: e.g., if D is a subset of a Euclidean space and μ has a positive density with respect to Lebesgue measure. The converse is not generally true, however: there are systems which are transitive topologically but not metrically.

A dynamical system is *chaotic* if it is topologically transitive, and it contains dense periodic orbits (Banks *et al.*, 1992). The two facts together imply that a trajectory can start out arbitrarily close to a periodic orbit, and so remain near it for some time, only to eventually find itself arbitrarily close to a *different* periodic orbit. This is the source of the fabled “sensitive dependence on initial conditions”, which paradoxically manifests itself in the fact that all typical trajectories look pretty much the same, at least in the long run. Since metric transitivity generally implies topological transitivity, there is a close connection between ergodicity and chaos; in fact, most of the well-studied chaotic systems are also ergodic (Eckmann and Ruelle, 1985), including the logistic map. However, it is possible to be ergodic without being chaotic: the one-dimensional rotations with irrational shifts are, because their periodic orbits do not exist, and *a fortiori* are not dense.

Lemma 343 (Metric transitivity implies ergodicity) *If a dynamical system is metrically transitive, then it is ergodic.*

PROOF: By contradiction. Suppose there was an invariant set A whose μ -measure was neither 0 nor 1; then A^c is also invariant, and has strictly positive measure. By metric transitivity, for some n , $\mu(T^{-n}A \cap A^c) > 0$. But $T^{-n}A = A$, and $\mu(A \cap A^c) = 0$. So metrically transitive systems are ergodic. \square

There is a partial converse.

Lemma 344 (Stationary Ergodic Systems are Metrically Transitive) *If a dynamical system is ergodic and stationary, then it is metrically transitive.*

PROOF: Take any $\mu(A), \mu(B) > 0$. Let $A_{\text{ever}} \equiv \bigcup_{n=0}^{\infty} T^{-n}A$ — the union of A with all its pre-images. This set contains its pre-images, $T^{-1}A_{\text{ever}} \subseteq A_{\text{ever}}$, since if $x \in T^{-n}A$, $T^{-1}x \in T^{-(n+1)}A$. The sequence of pre-images is thus non-increasing, and so tends to a limiting set, $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} T^{-k}A = A_{\text{i.o.}}$, the set of points which not only visit A eventually, but visit A infinitely often. This is an invariant set (Lemma 306), so by ergodicity it has either measure 0 or measure 1. By the Poincaré recurrence theorem (Corollaries 67 and 68), since $\mu(A) > 0$, $\mu(A_{\text{i.o.}}) = 1$. Hence, for any B , $\mu(A_{\text{i.o.}} \cap B) = \mu(B)$. But this means that, for some n , $\mu(T^{-n}A \cap B) > 0$, and the process is metrically transitive. \square

25.2 Examples of Ergodicity

Example 345 (IID Sequences, Strong Law of Large Numbers) *Every IID sequence is ergodic. This is because the Kolmogorov 0-1 law states that every tail event has either probability 0 or 1, and (Exercise 25.3) every invariant event is a tail event. The strong law of large numbers is thus a two-line corollary of the Birkhoff ergodic theorem.*

Example 346 (Markov Chains) *In the elementary theory of Markov chains, an ergodic chain is one which is irreducible, aperiodic and positive recurrent. To see that such a chain corresponds to an ergodic process in the present sense, look at the shift operator on the sequence space. For consistency of notation, let S_1, S_2, \dots be the values of the Markov chain in Σ , and X be the semi-infinite sequence in sequence space Ξ , with shift operator T , and distribution μ over sequences. μ is the product of an initial distribution $\nu \sim S_1$ and the Markov-family kernel. Now, “irreducible” means that one goes from every state to every other state with positive probability at some lag, i.e., for every $s_1, s_2 \in \Sigma$, there is an n such that $\mathbb{P}(S_n = s_2 | S_1 = s_1) > 0$. But, writing $[s]$ for the cylinder set in Ξ with base s , this means that, for every $[s_1], [s_2]$, $\mu(T^{-n}[s_2] \cap [s_1]) > 0$, provided $\mu([s_1]) > 0$. The Markov property of the S chain, along with positive recurrence, can be used to extend this to all finite-dimensional cylinder sets (Exercise 25.4), and so, by a generating-class argument, to all measurable sets.*

Example 347 (Deterministic Ergodicity: The Logistic Map) *We have seen that the logistic map, $Tx = 4x(1-x)$, has an invariant density (with respect to Lebesgue measure). It has an infinite collection of invariant sets, but the only invariant interval is the whole state space $[0, 1]$ — any smaller interval is not invariant. From this, it is easy to show that all the invariant sets either have measure 0 or measure 1 — they differ from \emptyset or from $[0, 1]$ by only a countable collection of points. Hence, the invariant measure is ergodic. Notice, too, that the Lebesgue measure on $[0, 1]$ is ergodic, but not invariant.*

Example 348 (Invertible Ergodicity: Rotations) *Let $\Xi = [0, 1)$, $Tx = x + \phi \bmod 1$, and let μ be the Lebesgue measure on Ξ . (This corresponds to a rotation, where the angle advances by $2\pi\phi$ radians per unit time.) Clearly, T preserve μ . If ϕ is rational, then, for any x , the sequence of iterates will visit only finitely many points, and the process is not ergodic, because one can construct invariant sets whose measure is neither 0 nor 1. (You may construct such a set by taking any one of the periodic orbits, and surrounding its points by intervals of sufficiently small, yet positive, width.) If, on the other hand, ϕ is irrational, then $T^n x$ never repeats, and it is easy to show that the process is ergodic, because it is metrically transitive. Nonetheless, T is invertible.*

This example (suitably generalized to multiple coordinates) is very important in physics, because many mechanical systems can be represented in terms of

“action-angle” variables, the speed of rotation of the angular variables being set by the actions, which are conserved, energy-like quantities. See Mackey (1992); Arnol’d and Avez (1968) for the ergodicity of rotations and its limitations, and Arnol’d (1978) for action-angle variables. Astonishingly, the result for the one-dimensional case was proved by Nicholas Oresme in the 14th century (von Plato, 1994).

Example 349 (Ergodicity when the Distribution Does Not Converge)

Ergodicity does not ensure a uni-directional evolution of the distribution. (Some people (Mackey, 1992) believe this has great bearing on the foundations of thermodynamics.) For a particularly extreme example, which also illustrates why elementary Markov chain theory insists on aperiodicity, consider the period-two deterministic chain, where state A goes to state B with probability 1, and vice versa. Every sample path spends just much time in state A as in state B , so every time average will converge on $\mathbf{E}_m[f]$, where m puts equal probability on both states. It doesn’t matter what initial distribution we use, because they are all ergodic (the only invariant sets are the whole space and the empty set, and every distribution gives them probability 1 and 0, respectively). The uniform distribution is the unique stationary distribution, but other distributions do not approach it, since $U^{2n}\nu = \nu$ for every integer n . So, $A_tf \rightarrow \mathbf{E}_m[f]$ a.s., but $\mathcal{L}(X_n) \not\rightarrow m$. We will see later that aperiodicity of Markov chains connects to “mixing” properties, which do guarantee stronger forms of distributional convergence.

25.3 Consequences of Ergodicity

The most basic consequence of ergodicity is that all invariant functions are constant almost everywhere; this in fact characterizes ergodicity. This in turn implies that time-averages converge to deterministic, rather than random, limits. Another important consequence is that events widely separated in time become nearly independent, in a somewhat funny-looking sense.

Theorem 350 (Ergodicity and the Triviality of Invariant Functions)

A T transformation is μ -ergodic if and only if all T -invariant observables are constant μ -almost-everywhere.

PROOF: “Only if”: Because invariant observables are \mathcal{I} -measurable (Lemma 304), the pre-image under an invariant observable f of any Borel set B is an invariant set. Since every invariant set has μ -probability 0 or 1, the probability that $f(x) \in B$ is either 0 or 1, hence f is constant with probability 1. “If”: The indicator function of an invariant set is an invariant function. If all invariant functions are constant μ -a.s., then for any $A \in \mathcal{I}$, either $\mathbf{1}_A(x) = 0$ or $\mathbf{1}_A(x) = 1$ for μ -almost all x , which is the same as saying that either $\mu(A) = 0$ or $\mu(A) = 1$, as required. \square

25.3.1 Deterministic Limits for Time Averages

Theorem 351 (The Ergodic Theorem for Ergodic Processes) *Suppose μ is AMS, with stationary mean m , and T -ergodic. Then, almost surely,*

$$\lim_{t \rightarrow \infty} A_t f(x) = \mathbf{E}_m[f] \quad (25.1)$$

for μ - and m - almost all x , for any $L_1(m)$ observable f .

PROOF: Because every invariant set has μ -probability 0 or 1, it likewise has m -probability 0 or 1 (Lemma 329). Hence, $\mathbf{E}_m[f]$ is a version of $\mathbf{E}_m[f|\mathcal{I}]$. Since $A_t f$ is also a version of $\mathbf{E}_m[f|\mathcal{I}]$ (Corollary 340), they are equal almost surely. \square

An important consequence is the following. Suppose S_t is a strictly stationary random sequence. Let $\Phi_t(S) = f(S_{t+\tau_1}, S_{t+\tau_2}, \dots, S_{t+\tau_n})$ for some fixed collection of shifts τ_n . Then Φ_t is another strictly stationary random sequence. Every strictly stationary random sequence can be represented by a measure-preserving transformation (Theorem 52), where X is the sequence S_1, S_2, \dots , the mapping T is just the shift, and the measure μ is the infinite-dimensional measure of the original stochastic process. Thus $\Phi_t = \phi(X_t)$, for some measurable function ϕ . If the measure is ergodic, and $\mathbf{E}[\Phi]$ is finite, then the time-average of Φ converges almost surely to its expectation. In particular, let $\Phi_t = S_t S_{t+\tau}$. Then, assuming the mixed moments are finite, $t^{-1} \sum_{t=1}^{\infty} S_t S_{t+\tau} \rightarrow \mathbf{E}[S_t S_{t+\tau}]$ almost surely, and so the sample covariance converges on the true covariance. More generally, for a stationary ergodic process, if the n -point correlation functions exist, the sample correlation functions converge a.s. on the true correlation functions.

25.3.2 Ergodicity and the approach to independence

Lemma 352 (Ergodicity Implies Approach to Independence) *If μ is T -ergodic, and μ is AMS with stationary mean m , then*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(B \cap T^{-n}C) = \mu(B)m(C) \quad (25.2)$$

for any measurable events B, C .

PROOF: Exercise 25.1. \square

Theorem 353 (Approach to Independence Implies Ergodicity) *Suppose \mathcal{X} is generated by a field \mathcal{F} . Then an AMS measure μ , with stationary mean m , is ergodic if and only if, for all $F \in \mathcal{F}$,*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^{t-1} \mu(F \cap T^{-n}F) = \mu(F)m(F) \quad (25.3)$$

i.e., iff Eq. 25.2 holds, taking $B = C = F \in \mathcal{F}$.

PROOF: “Only if”: Lemma 352. “If”: Exercise 25.2. \square

25.4 Exercises

Exercise 25.1 (Ergodicity implies an approach to independence) *Prove Lemma 352.*

Exercise 25.2 (Approach to independence implies ergodicity) *Prove the “if” part of Theorem 353.*

Exercise 25.3 (Invariant events and tail events) *Prove that every invariant event is a tail event. Does the converse hold?*

Exercise 25.4 (Ergodicity of ergodic Markov chains) *Complete the argument in Example 346, proving that ergodic Markov chains are ergodic processes (in the sense of Definition 341).*

Chapter 26

Decomposition of Stationary Processes into Ergodic Components

This chapter is concerned with the decomposition of asymptotically-mean-stationary processes into ergodic components.

Section 26.1 introduces some preliminary ideas about convex mixtures of invariant distributions, and in particular characterizes ergodic distributions as extremal points.

Section 26.2 shows how to write the stationary distribution as a mixture of distributions, each of which is stationary and ergodic, and each of which is supported on a distinct part of the state space. This is connected to ideas in nonlinear dynamics, each ergodic component being a different basin of attraction.

Section 26.3 lays out some connections to statistical inference: ergodic components can be seen as minimal sufficient statistics, and lead to powerful tests.

26.1 Preliminaries to Ergodic Decompositions

It is always the case, with a dynamical system, that if x lies within some invariant set A , then all its future iterates stay within A as well. In general, therefore, one might expect to be able to make some predictions about the future trajectory by knowing which invariant sets the initial condition lies within. An ergodic process is one where this is actually not possible. Because all invariant sets have probability 0 or 1, they are all independent of each other, and indeed of every other set. Therefore, knowing which invariant sets x falls into is *completely uninformative* about its future behavior. In the more general non-ergodic case, a limited amount of prediction is however possible on this basis, the limitations

being set by the way the state space breaks up into invariant sets of points with the same long-run average behavior — the ergodic components. Put slightly differently, the long-run behavior of an AMS system can be represented as a mixture of stationary, ergodic distributions, and the ergodic components are, in a sense, a minimal parametrically sufficient statistic for this distribution. (They are not in generally *predictively* sufficient.)

The idea of an ergodic decomposition goes back to von Neumann, but was considerably refined subsequently, especially by the Soviet school, who seem to have introduced most of the talk of predictions, and all of the talk of ergodic components as minimal sufficient statistics. Our treatment will follow Gray (1988, ch. 7), and Dynkin (1978).

Proposition 354 *Any convex combination of invariant probability measures is an invariant probability measure.*

PROOF: Let μ_1 and μ_2 be two invariant probability measures. It is elementary that for every $0 \leq a \leq 1$, $\nu \equiv a\mu_1 + (1-a)\mu_2$ is a probability measure. Now consider the measure under ν of the pre-image of an arbitrary measurable set $B \in \mathcal{X}$:

$$\nu(T^{-1}B) = a\mu_1(T^{-1}B) + (1-a)\mu_2(T^{-1}B) \quad (26.1)$$

$$= a\mu_1(B) + (1-a)\mu_2(B) \quad (26.2)$$

$$= \nu(B) \quad (26.3)$$

so ν is also invariant. \square

Proposition 355 *If μ_1 and μ_2 are invariant ergodic measures, then either $\mu_1 = \mu_2$, or they are singular, meaning that there is a set B on which $\mu_1(B) = 0$, $\mu_2(B) = 1$.*

PROOF: Suppose $\mu_1 \neq \mu_2$. Then there is at least one set C where $\mu_1(C) \neq \mu_2(C)$. Because both μ_i are stationary and ergodic, $A_t \mathbf{1}_C(x)$ converges to $\mu_i(C)$ for μ_i -almost-all x . So the set

$$\left\{ x \mid \lim_t A_t \mathbf{1}_C(x) = \mu_2(C) \right\}$$

has a μ_2 measure of 1, and a μ_1 measure of 0 (since, by hypothesis, $\mu_1(C) \neq \mu_2(C)$). \square

Proposition 356 *Ergodic invariant measures are extremal points of the convex set of invariant measures, i.e., they cannot be written as combinations of other invariant measures.*

PROOF: By contradiction. That is, suppose μ is ergodic and invariant, and that there were invariant measures ν and λ , and an $a \in (0, 1)$, such that $\mu = a\nu + (1-a)\lambda$. Let C be any invariant set; then $\mu(C) = 0$ or $\mu(C) = 1$. Suppose $\mu(C) = 0$. Then, because a is strictly positive, it must be the case that

$\nu(C) = \lambda(C) = 0$. If $\mu(C) = 1$, then C^c is also invariant and has μ -measure 0, so $\nu(C^c) = \lambda(C^c) = 0$, i.e., $\nu(C) = \lambda(C) = 1$. So ν and λ would both have to be ergodic, with the same support as μ . But then (Proposition 355 preceding) $\lambda = \nu = \mu$. \square

Remark: The converse is left as an exercise (26.1).

26.2 Construction of the Ergodic Decomposition

In the last section, we saw that the stationary distributions of a given dynamical system form a convex set, with the ergodic distributions as the extremal points. A standard result in convex analysis is that any point in a convex set can be represented as a convex combination of the extremal points. Thus, any stationary distribution can be represented as a mixture of stationary and ergodic distributions. We would like to be able to determine the weights used in the mixture, and even more to give them some meaningful stochastic interpretation.

Let's begin by thinking about the effective distribution we get from taking time-averages starting from a given point. For every measurable set B , and every finite t , $A_t \mathbf{1}_B(x)$ is a well-defined measurable function. As B ranges over the σ -field \mathcal{X} , holding x and t fixed, we get a set function, and one which, moreover, meets the requirements for being a probability measure. Suppose we go further and pass to the limit.

Definition 357 (Long-Run Distribution) *The long-run distribution starting from the point x is the set function $\lambda(x)$, defined through $\lambda(x, B) = \lim_t A_t \mathbf{1}_B(x)$, when the limit exists for all $B \in \mathcal{X}$. If $\lambda(x)$ exists, x is an ergodic point. The set of all ergodic points is E .*

Notice that whether or not $\lambda(x)$ exists depends *only* on x (and T and \mathcal{X}); the initial distribution has nothing to do with it. Let's look at some properties of the long-run distributions. (The name "ergodic point" is justified by one of them, Proposition 359.)

Proposition 358 *If $x \in E$, then $\lambda(x)$ is a probability distribution.*

PROOF: For every t , the set function given by $A_t \mathbf{1}_B(x)$ is clearly a probability measure. Since $\lambda(x)$ is defined by passage to the limit, the Vitali-Hahn Theorem (Proposition 327) says $\lambda(x)$ must be as well. \square

Proposition 359 *If $x \in E$, then $\lambda(x)$ is an ergodic distribution.*

PROOF: For every invariant set I , $\mathbf{1}_I(T^n x) = \mathbf{1}_I(x)$ for all n . Hence $A \mathbf{1}_I(x)$ exists and is either 0 or 1. This means $\lambda(x)$ assigns every invariant set either probability 0 or probability 1, so by Definition 341 it is ergodic. \square

Proposition 360 *If $x \in E$, then $\lambda(x)$ is an invariant function of x , i.e., $\lambda(x) = \lambda(Tx)$.*

PROOF: By Lemma 317, $A\mathbf{1}_B(x) = A\mathbf{1}_B(Tx)$, when the appropriate limit exists. Since, by assumption, it does in this case, for every measurable set $\lambda(x, B) = \lambda(Tx, B)$, and the set functions are thus equal. \square

Proposition 361 *If $x \in E$, then $\lambda(x)$ is a stationary distribution.*

PROOF: For all B and x , $\mathbf{1}_{T^{-1}B}(x) = \mathbf{1}_B(Tx)$. So $\lambda(x, T^{-1}B) = \lambda(Tx, B)$. Since, by Proposition 360, $\lambda(Tx, B) = \lambda(x, B)$, it finally follows that $\lambda(x, B) = \lambda(x, T^{-1}B)$, which proves that $\lambda(x)$ is an invariant distribution. \square

Proposition 362 *If $x \in E$ and $f \in L_1(\lambda(x))$, then $\lim_t A_t f(x)$ exists, and is equal to $\mathbf{E}_{\lambda(x)}[f]$.*

PROOF: This is true, by the definition of $\lambda(x)$, for the indicator functions of all measurable sets. Thus, by linearity of A_t and of expectation, it is true for all simple functions. Standard arguments then let us pass to all the functions integrable with respect to the long-run distribution. \square

At this point, you should be tempted to argue as follows. If μ is an AMS distribution with stationary mean m , then $Af(x) = \mathbf{E}_m[f|T]$ for almost all x . So, it's reasonable to hope that m is a combination of the $\lambda(x)$, and yet further that

$$Af(x) = \mathbf{E}_{\lambda(x)}[f]$$

for μ -almost-all x . This is basically true, but will take some extra assumptions to get it to work.

Definition 363 (Ergodic Component) *Two ergodic points $x, y \in E$ belong to the same ergodic component when $\lambda(x) = \lambda(y)$. We will write the ergodic components as C_i , and the function mapping x to its ergodic component as $\phi(x)$. $\phi(x)$ is not defined if x is not an ergodic point. By a slight abuse of notation, we will write $\lambda(C_i, B)$ for the common long-run distribution of all points in C_i .*

Obviously, the ergodic components partition the set of ergodic points. (The partition is not necessarily countable, and in some important cases, such as that of Hamiltonian dynamical systems in statistical mechanics, it must be uncountable (Khinchin, 1949).) Intuitively, they form the coarsest partition which is still fully informative about the long-run distribution. It's also pretty clear that the partition is left alone with the dynamics.

Proposition 364 *For all ergodic points x , $\phi(x) = \phi(Tx)$.*

PROOF: By Lemma 360, $\lambda(x) = \lambda(Tx)$, and the result follows. \square

Notice that I have been careful not to say that the ergodic components are invariant sets, because we've been using that to mean sets which are both left along by the dynamics *and* are measurable, i.e. members of the σ -field \mathcal{X} , and we have not established that any ergodic component is measurable, which in turn is because we have not established that $\lambda(x)$ is a measurable function.

Let's look a little more closely at the difficulty. If B is a measurable set, then $A_t \mathbf{1}_B(x)$ is a measurable function. If the limit exists, then $A \mathbf{1}_B(x)$ is also a measurable function, and consequently the set $\{y : A \mathbf{1}_B(y) = A \mathbf{1}_B(x)\}$ is a measurable set. Then

$$\phi(x) = \bigcap_{B \in \mathcal{X}} \{y : A \mathbf{1}_B(x) = A \mathbf{1}_B(y)\} \quad (26.4)$$

gives the ergodic component to which x belongs. The difficulty is that the intersection is over *all* measurable sets B , and there are generally an uncountable number of them (even if Ξ is countable!), so we have no guarantee that the intersection of uncountably many measurable sets is measurable. Consequently, we can't say that any of the ergodic components is measurable.

The way out, as so often in mathematics, is to cheat; or, more politely, to make an assumption which is strong enough to force open an exit, but not so strong that we can't support it or verify it¹. What we will assume is that there is a *countable* collection of sets \mathcal{G} such that $\lambda(x) = \lambda(y)$ if and only if $\lambda(x, G) = \lambda(y, G)$ for every $G \in \mathcal{G}$. Then the intersection in Eq. 26.4 need only run over the countable class \mathcal{G} , rather than all of \mathcal{X} , which will be enough to reassure us that $\phi(x)$ is a measurable set.

Definition 365 (Countable Extension Space) *A measurable space Ω, \mathcal{F} is a countable extension space when there is a countable field \mathcal{G} of sets in Ω such that $\mathcal{F} = \sigma(\mathcal{G})$, i.e., \mathcal{G} is the generating field of the σ -field, and any normalized, non-negative, finitely-additive set function on \mathcal{G} has a unique extension to a probability measure on \mathcal{F} .*

The reason the countable extension property is important is that it lets us get away with just checking properties of measures on a countable class (the generating field \mathcal{G}). Here are a few important facts about countable extension spaces; proofs, along with a much more detailed treatment of the general theory, are given by Gray (1988, chs. 2 and 3), who however calls them “standard” spaces.

Proposition 366 *Every countable space is a countable extension space.*

Proposition 367 *Every Borel space is a countable extension space.*

Remember that finite-dimensional Euclidean spaces are Borel spaces.

Proposition 368 *A countable product of countable extension spaces is a countable extension space.*

¹For instance, we *could* just assume that uncountable intersections of measurable sets are measurable, but you will find it instructive to try to work out the consequences of this assumption, and to examine whether it holds for the Borel σ -field \mathcal{B} — say on the unit interval, to keep things easy.

The last proposition is important for us: if Σ is a countable extension space, it means that $\Xi \equiv \Sigma^{\mathbb{N}}$ is too. So if we have a discrete- or Euclidean- valued random sequence, we can switch to the sequence space, and still appeal to generating-class arguments based on countable fields. Without further ado, then, let's assume that Ξ , the state space of our dynamical system, is a countable extension space, with countable generating field \mathcal{G} .

Lemma 369 $x \in E$ iff $\lim_t A_t \mathbf{1}_G(x)$ converges for every $G \in \mathcal{G}$.

PROOF: “If”: A direct consequence of Definition 365, since the set function $A\mathbf{1}_G(x)$ extends to a unique measure. “Only if”: a direct consequence of Definition 357, since every member of the generating field is a measurable set. \square

Lemma 370 The set of ergodic points is measurable: $E \in \mathcal{X}$.

PROOF: For each $G \in \mathcal{G}$, the set of x where $A_t \mathbf{1}_G(x)$ converges is measurable, because G is a measurable set. The set where those relative frequencies converge for all $G \in \mathcal{G}$ is the intersection of countably many measurable sets, hence itself measurable. This set is, exactly, the set of ergodic points (Lemma 369). \square

Lemma 371 All the ergodic components are measurable sets, and $\phi(x)$ is a measurable function. Thus, all $C_i \in \mathcal{I}$.

PROOF: For each G , the set $\{y : \lambda(y, G) = \lambda(x, G)\}$ is measurable. So their intersection over all $G \in \mathcal{G}$ is also measurable. But, by the countable extension property, this intersection is precisely the set $\{y : \lambda(y) = \lambda(x)\}$. So the ergodic components are measurable sets, and, since $\phi^{-1}(C_i) = C_i$, ϕ is measurable. Since we have already seen that $T^{-1}C_i = C_i$, and now that $C_i \in \mathcal{X}$, we may say that $C_i \in \mathcal{I}$. \square

Remark: Because C_i is a (measurable) invariant set, $\lambda(x, C_i) = 1$ for every $x \in C_i$. However, it does not follow that there might not be a smaller set, also with long-run measure 1, i.e., there might be a $B \subset C_i$ such that $\lambda(x, B) = 1$. For an extreme example, consider the uniform contraction on \mathbb{R} , with $Tx = ax$ for some $0 \leq a \leq 1$. Every trajectory converges on the origin. The only ergodic invariant measure is the Dirac delta function. Every point belongs to a single ergodic component.

More generally, if a little roughly², the ergodic components correspond to the dynamical systems idea of *basins of attraction*, while the support of the long-run distributions corresponds to the actual *attractors*. Basins of attraction typically contain points which are not actually parts of the attractor.

Theorem 372 (Ergodic Decomposition of AMS Processes) Suppose Ξ, \mathcal{X} is a countable extension space. If μ is an asymptotically mean stationary measure on Ξ , with stationary mean m , then $\mu(E) = m(E) = 1$, and, for any $f \in L_1(m)$, and μ - and m - almost all x ,

$$Af(x) = \mathbf{E}_{\lambda(x)}[f] = \mathbf{E}_m[f|\mathcal{I}] \quad (26.5)$$

²I don't want to get into subtleties arising from the dynamicists tendency to define things topologically, rather than measure-theoretically.

so that

$$m(B) = \int \lambda(x, B) d\mu(x) \quad (26.6)$$

PROOF: For every set $G \in \mathcal{G}$, $A_t \mathbf{1}_G(x)$ converges for μ - and m - almost all x (Theorem 339). Since there are only countably many G , the set on which they all converge also has probability 1; this set is E . Since (Proposition 362) $Af(x) = \mathbf{E}_{\lambda(x)}[f]$, and (Theorem 339 again) $Af(x) = \mathbf{E}_m[f|\mathcal{I}]$ a.s., we have that $\mathbf{E}_{\lambda(x)}[f] = \mathbf{E}_m[f|\mathcal{I}]$ a.s.

Now let $f = \mathbf{1}_B$. As we know (Lemma 331), $\mathbf{E}_\mu[A\mathbf{1}_B(X)] = \mathbf{E}_m[\mathbf{1}_B(X)] = m(B)$. But, for each x , $A\mathbf{1}_B(x) = \lambda(x, B)$, so $m(B) = \mathbf{E}_\mu[\lambda(X, B)]$. \square

In words, we have decomposed the stationary mean m into the long-run distributions of the ergodic components, with weights given by the fraction of the initial measure μ falling into each component. Because of Propositions 354 and 356, we may be sure that by mixing stationary ergodic measures, we obtain an ergodic measure, and that our decomposition is unique.

26.3 Statistical Aspects

26.3.1 Ergodic Components as Minimal Sufficient Statistics

The connection between sufficient statistics and ergodic decompositions is a very pretty one. First, recall the idea of parametric statistical sufficiency.³

Definition 373 (Sufficiency, Necessity) *Let \mathcal{P} be a class of probability measures on a common measurable space Ω, \mathcal{F} , indexed by a parameter θ . A σ -field $\mathcal{S} \subseteq \mathcal{F}$ is parametrically sufficient for θ , or just sufficient, when $\mathbb{P}_\theta(A|\mathcal{S}) = \mathbb{P}_{\theta'}(A|\mathcal{S})$ for all θ, θ' . That is, all the distributions in \mathcal{P} have the same distribution, conditional on \mathcal{S} . A random variable such that $\mathcal{S} = \sigma(S)$ is called a sufficient statistic. A σ -field is necessary (for the parameter θ) if it is a sub- σ -field of every sufficient σ -field; a necessary statistic is defined similarly. A σ -field which is both necessary and sufficient is minimal sufficient.*

Remark: The idea of sufficiency originates with Fisher; that of necessity, so far as I can work out, with Dynkin. This definition (after Dynkin (1978)) is based on what ordinary theoretical statistics texts call the “Neyman factorization criterion” for sufficiency. We will see all these concepts again when we do information theory.

³There is also a related idea of predictive statistical sufficiency, which we unfortunately will not be able to get to. Also, note that most textbooks on theoretical statistics state things in terms of random variables and measurable functions thereof, rather than σ -fields, but this is the more general case (Blackwell and Girshick, 1954).

Lemma 374 \mathcal{S} is sufficient for θ if and only if there exists an \mathcal{F} -measurable function $\lambda(\omega, A)$ such that

$$\mathbb{P}_\theta(A|\mathcal{S}) = \lambda(\omega, A) \quad (26.7)$$

almost surely, for all θ .

PROOF: Nearly obvious. “Only if”: since the conditional probability exists, there must be some such function (it’s a version of the conditional probability), and since all the conditional probabilities are versions of one another, the function cannot depend on θ . “If”: In this case, we have a single function which is a version of all the conditional probabilities, so it must be true that $\mathbb{P}_\theta(A|\mathcal{S}) = \mathbb{P}_{\theta'}(A|\mathcal{S})$. \square

Theorem 375 If a process on a countable extension space is asymptotically mean stationary, then ϕ is a minimal sufficient statistic for its long-run distribution.

PROOF: The set of distributions \mathcal{P} is now the set of all long-run distributions generated by the dynamics, and θ is an index which tracks them all unambiguously. We need to show both sufficiency and necessity. *Sufficiency*: The σ -field generated by ϕ is the one generated by the ergodic components, $\sigma(\{C_i\})$. (Because the C_i are mutually exclusive, this is a particularly simple σ -field.) Clearly, $\mathbb{P}_\theta(A|\sigma(\{C_i\})) = \lambda(\phi(x), A)$ for all x and θ , so (Lemma 374), ϕ is a sufficient statistic. *Necessity*: Follows from the fact that a given ergodic component contains *all* the points with a given long-run distribution. Coarser σ -fields will not, therefore, preserve conditional probabilities. \square

This theorem may not seem particularly exciting, because there isn’t, necessarily, anything whose distribution matches the long-run distribution. However, it has deeper meaning under two circumstances when $\lambda(x)$ really is the asymptotic distribution of random variables.

1. If Ξ is really a sequence space, so that $X = S_1, S_2, S_3, \dots$, then $\lambda(x)$ really *is* the asymptotic marginal distribution of the S_t , conditional on the starting point.
2. Even if Ξ is not a sequence space, if stronger conditions than ergodicity known as “mixing”, “asymptotic stability”, etc., hold, there are reasonable senses in which $\mathcal{L}(X_t)$ does converge, and converges on the long-run distribution.⁴

In both these cases, knowing the ergodic component thus turns out to be necessary and sufficient for knowing the asymptotic distribution of the observables. (Cf. Corollary 378 below.)

⁴Lemma 352 already gave us a *kind* of distributional convergence, but it is of a very weak sort, known as “convergence in Cesàro mean”, which was specially invented to handle sequences which are not convergent in normal senses! We will see that there is a direct correspondence between levels of distributional convergence and levels of decay of correlations.

26.3.2 Testing Ergodic Hypotheses

Finally, let's close with an application to hypothesis testing, inspired by Badino (2004).

Theorem 376 *Let Ξ, \mathcal{X} be a measurable space, and let μ_0 and μ_1 be two infinite-dimensional distributions of one-sided, discrete-parameter strictly-stationary Σ -valued stochastic processes, i.e., μ_0 and μ_1 are distributions on $\Xi^{\mathbb{N}}, \mathcal{X}^{\mathbb{N}}$, and they are invariant under the shift operator. If they are also ergodic under the shift, then there exists a sequence of sets $R_t \in \mathcal{X}^t$ such that $\mu_0(R_t) \rightarrow 0$ while $\mu_1(R_t) \rightarrow 1$.*

PROOF: By Proposition 355, there exists a set $R \in \mathcal{X}^{\mathbb{N}}$ such that $\mu_0(R) = 0$, $\mu_1(R) = 1$. So we just need to approximate R by sets which are defined on the first t observations in such a way that $\mu_i(R_t) \rightarrow \mu_i(R)$. If $R_t \downarrow R$, then monotone convergence will give us the necessary convergence of probabilities. Here is a construction with cylinder sets⁵ that gives us the necessary sequence of approximations. Let

$$R_t \equiv R \cup \prod_{n=t+1}^{\infty} \Xi_n \quad (26.8)$$

Clearly, R_t forms a non-increasing sequence, so it converges to a limit, which equally clearly must be R . Hence $\mu_i(R_t) \rightarrow \mu_i(R) = i$. \square

Remark: “ R ” is for “rejection”. Notice that the regions R_t will in general depend on the actual sequence $X_1, X_2, \dots, X_t \equiv X_1^t$, and not necessarily be permutation-invariant. When we come to the asymptotic equipartition theorem in information theory, we will see a more explicit way of constructing such tests.

Corollary 377 *Let H_0 be “ X_i are IID with distribution p_0 ” and H_1 be “ X_i are IID with distribution p_1 ”. Then, as $t \rightarrow \infty$, there exists a sequence of tests of H_0 against H_1 whose size goes to 0 while their power goes to 1.*

PROOF: Let μ_0 be the product measure induced by p_0 , and μ_1 the product measure induced p_1 , and apply the previous theorem. \square

Corollary 378 *If X is a strictly stationary (one-sided) random sequence whose shift representation has countably-many ergodic components, then there exists a sequence of functions ϕ_t , each \mathcal{X}_t -measurable, such that $\phi_t(X_1^t)$ converges on the ergodic component with probability 1.*

PROOF: From Theorem 52, we can write $X_1^t = \pi_{1:t}U$, for a sequence-valued random variable U , using the projection operators of Chapter 2. For each ergodic component, by Theorem 376, there exists a sequence of sets $R_{t,i}$ such that $\mathbb{P}(X_1^t \in R_{t,i}) \rightarrow 1$ if $U \in C_i$, and goes to zero otherwise. Let $\phi(X_1^t)$ be the

⁵Introduced in Chapters 2 and 3. It's possible to give an alternative construction using the Hilbert space of all square-integrable random variables, and then projecting onto the subspace of those which are \mathcal{X}^t measurable.

set of all C_i for which $X_1^t \in R_{t,i}$. By Theorem 372, U is in some component with probability 1, and, since there are only countably many ergodic components, with probability 1 X_1^t will eventually leave all but one of the $R_{t,i}$. The remaining one is the ergodic component. \square

26.4 Exercises

Exercise 26.1 *Prove the converse to Proposition 356: every external point of the convex set of invariant measures is an ergodic measure.*

Chapter 27

Mixing

A stochastic process is mixing if its values at widely-separated times are asymptotically independent.

Section 27.1 defines mixing, and shows that it implies ergodicity.

Section 27.2 gives some examples of mixing processes, both deterministic and non-deterministic.

Section 27.3 looks at the weak convergence of distributions produced by mixing, and the resulting decay of correlations.

Section 27.4 defines *strong* mixing, and the “mixing coefficient” which measures it. It then states, but does not prove, a central limit theorem for strongly mixing sequences. (The proof would demand first working through the central limit theorem for martingales.)

For stochastic processes, “mixing” means “asymptotically independent”: that is, the statistical dependence between $X(t_1)$ and $X(t_2)$ goes to zero as $|t_1 - t_2|$ increases. To make this precise, we need to specify how we measure the dependence between $X(t_1)$ and $X(t_2)$. The most common and natural choice (first used by Rosenblatt, 1956) is the total variation distance between their joint distribution and the product of their marginal distributions, but there are other ways of measuring such “decay of correlations”¹. Under all reasonable choices, IID processes are, naturally enough, special cases of mixing processes. This suggests that many of the properties of IID processes, such as laws of large numbers and central limit theorems, should continue to hold for mixing processes, at least if the approach to independence is sufficiently rapid. This in turn means that many statistical methods originally developed for the IID case will continue to work when the data-generating process is mixing; this is true both of parametric methods, such as linear regression, ARMA models being mixing (Doukhan, 1995, sec. 2.4.1), and of nonparametric methods like kernel prediction (Bosq, 1998). Considerations of time will prevent us from going into

¹The term is common, but slightly misleading: lack of correlation, in the ordinary covariance-normalized-by-standard-deviations sense, implies independence only in special cases, like Gaussian processes. Nonetheless, see Theorem 391.

the purely statistical aspects of mixing processes, but the central limit theorem at the end of this chapter will give some idea of the flavor of results in this area: much like IID results, only with the true sample size replaced by an effective sample size, with a smaller discount the faster the rate of decay of correlations.

27.1 Definition and Measurement of Mixing

Definition 379 (Mixing) A dynamical system Ξ, \mathcal{X}, μ, T is mixing when, for any $A, B \in \mathcal{X}$,

$$\lim_{t \rightarrow \infty} |\mu(A \cap T^{-t}B) - \mu(A)\mu(T^{-t}B)| = 0 \quad (27.1)$$

Lemma 380 If μ is T -invariant, mixing is equivalent to

$$\lim_{t \rightarrow \infty} \mu(A \cap T^{-t}B) = \mu(A)\mu(B) \quad (27.2)$$

PROOF: By stationarity, $\mu(T^{-t}B) = \mu(B)$, so $\mu(A)\mu(T^{-t}B) = \mu(A)\mu(B)$. The result follows. \square

Theorem 381 Mixing implies ergodicity.

PROOF: Let A be any invariant set. By mixing, $\lim_t \mu(T^{-t}A \cap A) = \mu(T^{-t}A)\mu(A)$. But $T^{-t}A = A$ for every t , so we have $\lim_t \mu(A) = \mu^2(A)$, or $\mu(A) = \mu^2(A)$. This can only be true if $\mu(A) = 0$ or $\mu(A) = 1$, i.e., only if μ is T -ergodic. \square

Everything we have established about ergodic processes, then, applies to mixing processes.

Definition 382 A dynamical system is asymptotically stationary, with stationary limit m , when $\lim_t \mu(T^{-t}A) = m(A)$ for all $A \in \mathcal{X}$.

Lemma 383 An asymptotically stationary system is mixing iff

$$\lim_{t \rightarrow \infty} \mu(A \cap T^{-t}B) = \mu(A)m(B) \quad (27.3)$$

for all $A, B \in \mathcal{X}$.

PROOF: Directly from the fact that in this case $m(B) = \lim_t \mu(T^{-t}B)$. \square

Theorem 384 Suppose \mathcal{G} is a π -system, and μ is an asymptotically stationary measure. If

$$\lim_t |\mu(A \cap T^{-t}B) - \mu(A)\mu(T^{-t}B)| = 0 \quad (27.4)$$

for all $A, B \in \mathcal{G}$, then it holds for all pairs of sets in $\sigma(\mathcal{G})$. If $\sigma(\mathcal{G}) = \mathcal{X}$, then the process is mixing.

PROOF(after Durrett, 1991, Lemma 6.4.3): Via the π - λ theorem, of course. Let Λ_A be the class of all B such that the equation holds, for a given $A \in \mathcal{G}$. We need to show that Λ_A really is a λ -system.

$\Xi \in \Lambda_A$ is obvious. $T^{-t}\Xi = \Xi$ so $\mu(A \cap \Xi) = \mu(A) = \mu(A)\mu(\Xi)$.

Closure under complements. Let B_1 and B_2 be two sets in Λ_A , and assume $B_1 \subset B_2$. Because set-theoretic operations commute with taking inverse images, $T^{-t}(B_2 \setminus B_1) = T^{-t}B_2 \setminus T^{-t}B_1$. Thus

$$0 \leq |\mu(A \cap T^{-t}(B_2 \setminus B_1)) - \mu(A)\mu(T^{-t}(B_2 \setminus B_1))| \quad (27.5)$$

$$\begin{aligned} &= |\mu(A \cap T^{-t}B_2) - \mu(A \cap T^{-t}B_1) - \mu(A)\mu(T^{-t}B_2) + \mu(A)\mu(T^{-t}B_1)| \\ &\leq |\mu(A \cap T^{-t}B_2) - \mu(A)\mu(T^{-t}B_2)| \\ &\quad + |\mu(A \cap T^{-t}B_1) - \mu(A)\mu(T^{-t}B_1)| \end{aligned} \quad (27.6)$$

Taking limits of both sides, we get that $\lim_t |\mu(A \cap T^{-t}(B_2 \setminus B_1)) - \mu(A)\mu(T^{-t}(B_2 \setminus B_1))| = 0$, so that $B_2 \setminus B_1 \in \Lambda_A$.

Closure under monotone limits: Let B_n be any monotone increasing sequence in Λ_A , with limit B . Thus, $\mu(B_n) \uparrow \mu(B)$, and at the same time $m(B_n) \uparrow m(B)$, where m is the stationary limit of μ . Using Lemma 383, it is enough to show that

$$\lim_t \mu(A \cap T^{-t}B) = \mu(A)m(B) \quad (27.7)$$

Since $B_n \subset B$, we can always use the following trick:

$$\mu(A \cap T^{-t}B) = \mu(A \cap T^{-t}B_n) + \mu(A \cap T^{-t}(B \setminus B_n)) \quad (27.8)$$

$$\lim_t \mu(A \cap T^{-t}B) = \mu(A)m(B_n) + \lim_t \mu(A \cap T^{-t}(B \setminus B_n)) \quad (27.9)$$

For any $\epsilon > 0$, $\mu(A)m(B_n)$ can be made to come within ϵ of $\mu(A)m(B)$ by taking n sufficiently large. Let us now turn our attention to the second term.

$$0 \leq \lim_t \mu(A \cap T^{-t}(B \setminus B_n)) = \lim_t \mu(T^{-t}(B \setminus B_n)) \quad (27.10)$$

$$= \lim_t \mu(T^{-t}B \setminus T^{-t}B_n) \quad (27.11)$$

$$= \lim_t \mu(T^{-t}B) - \lim_t \mu(T^{-t}B_n) \quad (27.12)$$

$$= m(B) - m(B_n) \quad (27.13)$$

which again can be made less than any positive ϵ by taking n large. So, for sufficiently large n , $\lim_t \mu(A \cap T^{-t}B)$ is always within 2ϵ of $\mu(A)m(B)$. Since ϵ can be made arbitrarily small, we conclude that $\lim_t \mu(A \cap T^{-t}B) = \mu(A)m(B)$. Hence, $B \in \Lambda_A$.

We conclude, from the π - λ theorem, that Eq. 27.4 holds for all $A \in \mathcal{G}$ and all $B \in \sigma(\mathcal{G})$. The same argument can be turned around for A , to show that Eq. 27.4 holds for all pairs $A, B \in \sigma(\mathcal{G})$. If \mathcal{G} generates the whole σ -field \mathcal{X} , then clearly Definition 379 is satisfied and the process is mixing. \square

27.2 Examples of Mixing Processes

Example 385 (IID Sequences) *IID sequences are mixing from Theorem 384, applied to finite-dimensional cylinder sets.*

Example 386 (Ergodic Markov Chains) *Another application of Theorem 384 shows that ergodic Markov chains are mixing.*

Example 387 (Irrational Rotations of the Circle are Not Mixing) *Irrational rotations of the circle, $Tx = x + \phi \bmod 1$, ϕ irrational, are ergodic (Example 348), and stationary under the Lebesgue measure. They are not, however, mixing. Recall that $T^t x$ is dense in the unit interval, for arbitrary initial x . Because it is dense, there is a sequence t_n such that $t_n \phi \bmod 1$ goes to $1/2$. Now let $A = [0, 1/4]$. Because T maps intervals to intervals (of equal length), it follows that $T^{-t_n} A$ becomes an interval disjoint from A , i.e., $\mu(A \cap T^{-t_n} A) = 0$. But mixing would imply that $\mu(A \cap T^{-t_n} A) \rightarrow 1/16 > 0$, so the process is not mixing.*

Example 388 (Deterministic, Reversible Mixing: The Cat Map) *Here $\Xi = [0, 1)^2$, \mathcal{X} are the appropriate Borel sets, μ is Lebesgue measure on the square, and $Tx = (x_1 + x_2, x_1 + 2x_2) \bmod 1$. This is known as the cat map. It is a deterministic, invertible transformation, but it can be shown that it is actually mixing. (For a proof, which uses Theorem 390, the Fibonacci numbers and a clever trick with Fourier transforms, see Lasota and Mackey (1994, example 4.4.3, pp. 77–78).) The origins of the name lie with a figure in Arnol'd and Avez (1968), illustrating the mixing action of the map by successively distorting an image of a cat.*

27.3 Convergence of Distributions Under Mixing

To show how distributions converge (weakly) under mixing, we need to recall some properties of Markov operators. Remember that, for a Markov process, the time-evolution operator for observables, K , was defined through $Kf(x) = \mathbf{E}[f(X_1)|X_0 = x]$. Remember also that it induces an adjoint operator for the evolution of distributions, taking signed measures to signed measures, through the intermediary of the transition kernel. We can view the measure-updating operator U as a linear operator on $L_1(\mu)$, which takes non-negative μ -integrable functions to non-negative μ -integrable functions, and probability densities to probability densities. Since dynamical systems are Markov processes, all of this remains valid; we have K defined through $Kf(x) = f(Tx)$, and U through the adjoint relationship, $\mathbf{E}_\mu[f(X)Kg(X)] = \mathbf{E}[Uf(X)g(X)]\mu$, where $g \in L_\infty$ and $f \in L_1(\mu)$. These relations continue to remain valid for powers of the operators.

Lemma 389 *In any Markov process, $U^n d$ converges weakly to 1, for all initial probability densities d , if and only if $U^n f$ converges weakly to $\mathbf{E}_\mu[f]$, for all initial L_1 functions f , i.e. $\mathbf{E}_\mu[U^n f(X)g(X)] \rightarrow \mathbf{E}_\mu[f(X)]\mathbf{E}_\mu[g(X)]$ for all bounded, measurable g .*

PROOF: “If”: If d is a probability density with respect to μ , then $\mathbf{E}_\mu[d] = 1$. “Only if”: Re-write an arbitrary $f \in L_1(\mu)$ as the difference of its positive and negative parts, $f = f^+ - f^-$. A positive f is a re-scaling of some density, $f = cd$ for constant $c = \mathbf{E}_\mu[f]$ and a density d . Through the linearity of U and its powers,

$$\lim U^t f = \lim U^t f^+ - \lim U^t f^- \quad (27.14)$$

$$= \mathbf{E}_\mu[f^+] \lim U^t d^+ - \mathbf{E}_\mu[f^-] \lim U^t d^- \quad (27.15)$$

$$= \mathbf{E}_\mu[f^+] - \mathbf{E}_\mu[f^-] \quad (27.16)$$

$$= \mathbf{E}_\mu[f^+ - f^-] = \mathbf{E}_\mu[f] \quad (27.17)$$

using the linearity of expectations at the last step. \square

Theorem 390 *A T -invariant probability measure μ is T -mixing if and only if any initial probability measure $\nu \ll \mu$ converges weakly to μ under the action of T , i.e., iff, for all bounded, measurable f ,*

$$\mathbf{E}_{U^t \nu}[f(X)] \rightarrow \mathbf{E}_\mu[f(X)] \quad (27.18)$$

PROOF: Exercise. The way to go is to use the previous lemma, of course. With that tool, one can prove that the convergence holds for indicator functions, and then for simple functions, and finally, through the usual arguments, for all L_1 densities.

Theorem 391 (Decay of Correlations) *A stationary system is mixing if and only if*

$$\lim_{t \rightarrow \infty} \text{cov}(f(X_0), g(X_t)) = 0 \quad (27.19)$$

for all bounded observables f, g .

PROOF: Exercise, from the fact that convergence in distribution implies convergence of expectations of all bounded measurable functions. \square

It is natural to ask what happens if $U^t \nu \rightarrow \mu$ not weakly but strongly. This is known as *asymptotic stability* or (especially in the nonlinear dynamics literature) *exactness*. Remarkably enough, it is equivalent to the requirement that $\mu(T^t A) \rightarrow 1$ whenever $\mu(A) > 0$. (Notice that for once the expression involves *images* rather than *pre-images*.) There is a kind of hierarchy here, where different levels of convergence of distribution (Cesàro, weak, strong) match different sorts of ergodicity (metric transitivity, mixing, exactness). For more details, see Lasota and Mackey (1994).

27.4 A Central Limit Theorem for Mixing Sequences

Notice that I say “a central limit theorem”, rather than “the central limit theorem”. In the IID case, the necessary and sufficient condition for the CLT is well-known (you saw it in 36-752) and reasonably comprehensible. In the mixing case, a necessary and sufficient condition is known², but not commonly used, because quite opaque and hard to check. Rather, the common practice is to rely upon a large set of distinct sufficient conditions. Some of these, it must be said, are pretty ugly, but they are more susceptible of verification.

Recall the notation that X_t^- consists of the entire past of the process, including X_t , and X_t^+ its entire future.

Definition 392 (Mixing Coefficients) For a stochastic process X_t , define the strong-, Rosenblatt- or α - mixing coefficients as

$$\alpha(t_1, t_2) = \sup \{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \sigma(X_{t_1}^-), B \in \sigma(X_{t_2}^+) \} \quad (27.20)$$

If the system is conditionally stationary, then $\alpha(t_1, t_2) = \alpha(t_2, t_1) = \alpha(|t_1 - t_2|) \equiv \alpha(\tau)$. If $\alpha(\tau) \rightarrow 0$, then the process is strong-mixing or α -mixing. If $\alpha(\tau) = O(e^{-b\tau})$ for some $b > 0$, the process is exponentially mixing, b is the mixing rate, and $1/b$ is the mixing time. If $\alpha(\tau) = O(\tau^{-k})$ for some $k > 0$, then the process is polynomially mixing.

Notice that $\alpha(t_1, t_2)$ is just the total variation distance between the joint distribution, $\mathcal{L}(X_{t_1}^-, X_{t_2}^+)$, and the product of the marginal distributions, $\mathcal{L}(X_{t_1}^-) \times \mathcal{L}(X_{t_2}^+)$. Thus, it is a natural measure of the degree to which the future of the system depends on its past. However, there are at least four other mixing coefficients (β , ϕ , ψ and ρ) regularly used in the literature. Since any of these others going to zero implies that α goes to zero, we will stick with α -mixing, as in Rosenblatt (1956).

Also notice that if X_t is a Markov process (e.g., a dynamical system) then the Markov property tells us that we only need to let the supremum run over measurable sets in $\sigma(X_{t_1})$ and $\sigma(X_{t_2})$.

Lemma 393 If a dynamical system is α -mixing, then it is mixing.

PROOF: α is the supremum of the quantity appearing in the definition of mixing.

□

Notation: For the remainder of this section,

$$S_n \equiv \sum_{k=1}^n X_k \quad (27.21)$$

$$\sigma_n^2 \equiv \mathbf{Var}[S_n] \quad (27.22)$$

$$Y_n(t) \equiv \frac{S_{[nt]}}{\sigma_n} \quad (27.23)$$

²Doukhan (1995, p. 47) cites Jakubowski and Szewczak (1990) as the source, but I have not verified the reference.

where n is any positive integer, and $t \in [0, 1]$.

Definition 394 X_t obeys the central limit theorem when

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \quad (27.24)$$

for some positive σ .

Definition 395 X_t obeys the functional central limit theorem or the invariance principle when

$$Y_n \xrightarrow{d} W \quad (27.25)$$

where W is a standard Wiener process on $[0, 1]$, and the convergence is in the Skorokhod topology of Sec. 15.1.

Theorem 396 (Central Limit Theorem for α -Mixing Sequences) Let X_t be a stationary sequence with $\mathbf{E}[X_t] = 0$. Suppose X is α -mixing, and that for some $\delta > 0$

$$\mathbf{E}[|X_t|^{2+\delta}] \leq \infty \quad (27.26)$$

$$\sum_{n=0}^{\infty} \alpha^{\frac{\delta}{2+\delta}}(n) \leq \infty \quad (27.27)$$

Then

$$\lim_{n \rightarrow \infty} \frac{\sigma_n^2}{n} = \mathbf{E}[|X_1|^2] + 2 \sum_{k=1}^{\infty} \mathbf{E}[X_1 X_k] \equiv \sigma^2 \quad (27.28)$$

If $\sigma^2 > 0$, moreover, X_t obeys both the central limit theorem with variance σ^2 , and the functional central limit theorem.

PROOF: Complicated, and based on a rather technical central limit theorem for martingale difference arrays. See Doukhan (1995, sec. 1.5), or, for a simplified presentation, Durrett (1991, sec. 7.7). \square

For the rate of convergence of $\mathcal{L}(S_n/\sqrt{n})$ to a Gaussian distribution, in the total variation metric, see Doukhan (1995, sec. 1.5.2), summarizing several works. Polynomially-mixing sequences converge polynomially in n , and exponentially-mixing sequences converge exponentially.

There are a number of results on central limit theorems and functional central limit theorems for deterministic dynamical systems. A particularly strong one was recently proved by Tyran-Kamińska (2005), in a friendly paper which should be accessible to anyone who's followed along this far, but it's too long for us to do more than note its existence.

Part VI

Information Theory

Chapter 28

Shannon Entropy and Kullback-Leibler Divergence

Section 28.1 introduces Shannon entropy and its most basic properties, including the way it measures how close a random variable is to being uniformly distributed.

Section 28.2 describes relative entropy, or Kullback-Leibler divergence, which measures the discrepancy between two probability distributions, and from which Shannon entropy can be constructed. Section 28.2.1 describes some statistical aspects of relative entropy, especially its relationship to expected log-likelihood and to Fisher information.

Section 28.3 introduces the idea of the mutual information shared by two random variables, and shows how to use it as a measure of serial dependence, like a nonlinear version of autocovariance (Section 28.3.1).

Information theory studies stochastic processes as sources of information, or as models of communication channels. It appeared in essentially its modern form with Shannon (1948), and rapidly proved to be an extremely useful mathematical tool, not only for the study of “communication and control in the animal and the machine” (Wiener, 1961), but more technically as a vital part of probability theory, with deep connections to statistical inference (Kullback, 1968), to ergodic theory, and to large deviations theory. In an introduction that’s so limited it’s almost a crime, we will do little more than build enough theory to see how it can fit in with the theory of inference, and then get what we need to progress to large deviations. If you want to learn more (and you should!), the deservedly-standard modern textbook is Cover and Thomas (1991), and a good treatment, at something more like our level of mathematical rigor, is Gray

(1990).¹

28.1 Shannon Entropy

The most basic concept of information theory is that of the *entropy* of a random variable, or its distribution, often called Shannon entropy to distinguish it from the many other sorts. This is a measure of the uncertainty or variability associated with the random variable. Let's start with the discrete case, where the variable takes on only a finite or countable number of values, and everything is easier.

Definition 397 (Shannon Entropy (Discrete Case)) *The Shannon entropy, or just entropy, of a discrete random variable X is*

$$H[X] \equiv - \sum_x \mathbb{P}(X = x) \log \mathbb{P}(X = x) = -\mathbf{E}[\log \mathbb{P}(X)] \quad (28.1)$$

when the sum exists. Entropy has units of bits when the logarithm has base 2, and nats when it has base e .

The joint entropy of two random variables, $H[X, Y]$, is the entropy of their joint distribution.

The conditional entropy of X given Y , $H[X|Y]$ is

$$H[X|Y] \equiv \sum_y \mathbb{P}(Y = y) \sum_x \mathbb{P}(X = x|Y = y) \log \mathbb{P}(X = x|Y = y) \quad (28.2)$$

$$= -\mathbf{E}[\log \mathbb{P}(X|Y)] \quad (28.3)$$

$$= H[X, Y] - H[Y] \quad (28.4)$$

Here are some important properties of the Shannon entropy, presented without proofs (which are not hard).

1. $H[X] \geq 0$
2. $H[X] = 0$ iff $\exists x_0 : X = x_0$ a.s.
3. If X can take on $n < \infty$ different values (with positive probability), then $H[X] \leq \log n$. $H[X] = \log n$ iff X is uniformly distributed.
4. $H[X] + H[Y] \geq H[X, Y]$, with equality iff X and Y are independent. (This comes from the logarithm in the definition.)

¹Remarkably, almost all of the post-1948 development has been either amplifying or refining themes first sounded by Shannon. For example, one of the fundamental results, which we will see in the next chapter, is the “Shannon-McMillan-Breiman theorem”, or “asymptotic equipartition property”, which says roughly that the log-likelihood per unit time of a random sequence converges to a constant, characteristic of the data-generating process. Shannon's original version was convergence in probability for ergodic Markov chains; the modern form is almost sure convergence for any stationary and ergodic process. Pessimistically, this says something about the decadence of modern mathematical science; optimistically, something about the value of getting it right the first time.

5. $H[X, Y] \geq H[X]$.
6. $H[X|Y] \geq 0$, with equality iff X is a.s. constant given Y , for almost all Y .
7. $H[X|Y] \leq H[X]$, with equality iff X is independent of Y . (“Conditioning reduces entropy”.)
8. $H[f(X)] \leq H[X]$, for any measurable function f , with equality iff f is invertible.

The first three properties can be summarized by saying that $H[X]$ is maximized by a uniform distribution, and minimized, to zero, by a degenerate one which is a.s. constant. We can then think of $H[X]$ as the variability of X , something like the log of the effective number of values it can take on. We can also think of it as how uncertain we are about X 's value.² $H[X, Y]$ is then how much variability or uncertainty is associated with the pair variable X, Y , and $H[Y|X]$ is how much uncertainty remains about Y once X is known, averaging over Y . Similarly interpretations follow for the other properties. The fact that $H[f(X)] = H[X]$ if f is invertible is nice, because then f just relabels the possible values, meshing nicely with this interpretation.

A simple consequence of the above results is particularly important for later use.

Lemma 398 (Chain Rule for Shannon Entropy) *Let X_1, X_2, \dots, X_n be discrete-valued random variables on a common probability space. Then*

$$H[X_1, X_2, \dots, X_n] = H[X_1] + \sum_{i=2}^n H[X_i | X_1, \dots, X_{i-1}] \quad (28.5)$$

PROOF: From the definitions, it is easily seen that $H[X_2|X_1] = H[X_2, X_1] - H[X_1]$. This establishes the chain rule for $n = 2$. A simple argument by induction does the rest. \square

For non-discrete random variables, it is necessary to introduce a reference measure, and many of the nice properties go away.

Definition 399 (Shannon Entropy (General Case)) *The Shannon entropy of a random variable X with distribution μ , with respect to a reference measure ρ , is*

$$H_\rho[X] \equiv -\mathbf{E}_\mu \left[\log \frac{d\mu}{d\rho} \right] \quad (28.6)$$

²This line of reasoning is sometimes supplemented by saying that we are more “surprised” to find that $X = x$ the less probable that event is, supposing that surprise should go as the log of one over that probability, and defining entropy as expected surprise. The choice of the logarithm, rather than any other increasing function, is of course retroactive, though one might cobble together some kind of psychophysical justification, since the perceived intensity of a sensation often grows logarithmically with the physical magnitude of the stimulus. More dubious, to my mind, is the idea that there is any surprise *at all* when a fair coin coming up heads.

when $\mu \ll \rho$. Joint and conditional entropies are defined similarly. We will also write $H_\rho[\mu]$, with the same meaning. This is sometimes called differential entropy when ρ is Lebesgue measure on Euclidean space, especially \mathbb{R} , and then is written $h(X)$ or $h[X]$.

It remains true, in the general case, that $H_\rho[X|Y] = H_\rho[X, Y] - H_\rho[Y]$, provided all of the entropies are finite. The chain rule remains valid, conditioning still reduces entropy, and the joint entropy is still \leq the sum of the marginal entropies, with equality iff the variables are independent. However, depending on the reference measure, $H_\rho[X]$ can be negative; e.g., if ρ is Lebesgue measure and $\mathcal{L}(X) = \delta(x)$, then $H_\rho[X] = -\infty$.

28.2 Relative Entropy or Kullback-Leibler Divergence

Some of the difficulties associated with Shannon entropy, in the general case, can be evaded by using relative entropy.

Definition 400 (Relative Entropy, Kullback-Leibler Divergence) *Given two probability distributions, $\nu \ll \mu$, the relative entropy of ν with respect to μ , or the Kullback-Leibler divergence of ν from μ , is*

$$D(\mu\|\nu) = -\mathbf{E}_\mu \left[\log \frac{d\nu}{d\mu} \right] \quad (28.7)$$

If ν is not absolutely continuous with respect to μ , then $D(\mu\|\nu) = \infty$.

Lemma 401 *$D(\mu\|\nu) \geq 0$, with equality iff $\nu = \mu$ almost everywhere (μ).*

PROOF: From Jensen's inequality, $\mathbf{E}_\mu \left[\log \frac{d\nu}{d\mu} \right] \leq \log \mathbf{E}_\mu \left[\frac{d\nu}{d\mu} \right] = \log 1 = 0$. The second part follows from the conditions for equality in Jensen's inequality. \square

Lemma 402 (Divergence and Total Variation) *For any two distributions, $D(\mu\|\nu) \geq \frac{1}{2\ln 2} \|\mu - \nu\|_1^2$.*

PROOF: Algebra. See, e.g., Cover and Thomas (1991, Lemma 12.6.1, pp. 300–301). \square

Definition 403 *The conditional relative entropy, $D(\mu(Y|X)\|\nu(Y|X))$ is*

$$D(\mu(Y|X)\|\nu(Y|X)) \equiv -\mathbf{E}_\mu \left[\log \frac{d\nu(Y|X)}{d\mu(Y|X)} \right] \quad (28.8)$$

Lemma 404 (Chain Rule for Relative Entropy) *$D(\mu(X, Y)\|\nu(X, Y)) = D(\mu(X)\|\nu(X)) + D(\mu(Y|X)\|\nu(Y|X))$*

PROOF: Algebra. \square

Shannon entropy can be constructed from the relative entropy.

Lemma 405 *The Shannon entropy of a discrete-valued random variable X , with distribution μ , is*

$$H[X] = \log n - D(\mu \| \nu) \quad (28.9)$$

where n is the number of values X can take on (with positive probability), and ν is the uniform distribution over those values.

PROOF: Algebra. \square

A similar result holds for the entropy of a variable which takes values in a finite subset, of volume V , of a Euclidean space, i.e., $H_\lambda[X] = \log V - D(\mu \| \nu)$, where λ is Lebesgue measure and ν is the uniform probability measure on the range of X .

28.2.1 Statistical Aspects of Relative Entropy

From Lemma 402, “convergence in relative entropy”, $D(\mu \| \nu_n) \rightarrow 0$ as $n \rightarrow \infty$, implies convergence in the total variation (L_1) metric. Because of Lemma 401, we can say that KL divergence has some of the properties of a metric on the space of probability distribution: it’s non-negative, with equality only when the two distributions are equal (a.e.). Unfortunately, however, it is not symmetric, and it does not obey the triangle inequality. (This is why it’s the KL *divergence* rather than the KL *distance*.) Nonetheless, it’s *enough* like a metric that it can be used to construct a kind of geometry on the space of probability distributions, and so of statistical models, which can be extremely useful. While we will not be able to go very far into this information geometry³, it will be important to indicate a few of the connections between information-theoretic notions, and the more usual ones of statistical theory.

Definition 406 (Cross-entropy) *The cross-entropy of ν and μ , $Q(\mu \| \nu)$, is*

$$Q_\rho(\mu \| \nu) \equiv -\mathbf{E}_\mu \left[\log \frac{d\nu}{d\rho} \right] \quad (28.10)$$

where ν is absolutely continuous with respect to the reference measure ρ . If the domain is discrete, we will take the reference measure to be uniform and drop the subscript, unless otherwise noted.

Lemma 407 *Suppose ν and μ are the distributions of two probability models, and $\nu \ll \mu$. Then the cross-entropy is the expected negative log-likelihood of the model corresponding to ν , when the actual distribution is μ . The actual or empirical negative log-likelihood of the model corresponding to ν is $Q_\rho(\nu \| \eta)$, where η is the empirical distribution.*

³See Kass and Vos (1997) or Amari and Nagaoka (1993/2000). For applications to statistical inference for stochastic processes, see Taniguchi and Kakizawa (2000). For an easier general introduction, Kulhavy (1996) is hard to beat.

PROOF: Obvious from the definitions. \square

Lemma 408 *If $\nu \ll \mu \ll \rho$, then $Q_\rho(\mu\|\nu) = H_\rho[\mu] + D(\mu\|\nu)$.*

PROOF: By the chain rule for densities,

$$\frac{d\nu}{d\rho} = \frac{d\mu}{d\rho} \frac{d\nu}{d\mu} \quad (28.11)$$

$$\log \frac{d\nu}{d\rho} = \log \frac{d\mu}{d\rho} + \log \frac{d\nu}{d\mu} \quad (28.12)$$

$$\mathbf{E}_\mu \left[\log \frac{d\nu}{d\rho} \right] = \mathbf{E}_\mu \left[\log \frac{d\mu}{d\rho} \right] + \mathbf{E}_\mu \left[\log \frac{d\nu}{d\mu} \right] \quad (28.13)$$

The result follows by applying the definitions. \square

Corollary 409 (Gibbs's Inequality) *$Q_\rho(\mu\|\nu) \geq H_\rho[\mu]$, with equality iff $\nu = \mu$ a.e.*

PROOF: Insert the result of Lemma 401 into the preceding proposition. \square

The statistical interpretation of the proposition is this: The log-likelihood of a model, leading to distribution ν , can be broken into two parts. One is the divergence of ν from μ ; the other just the entropy of μ , i.e., it is the same for all models. If we are considering the expected log-likelihood, then μ is the actual data-generating distribution. If we are considering the empirical log-likelihood, then μ is the empirical distribution. In either case, to maximize the likelihood is to minimize the relative entropy, or divergence. What we would like to do, as statisticians, is minimize the divergence from the data-generating distribution, since that will let us predict future values. What we *can* do is minimize divergence from the empirical distribution. The consistency of maximum likelihood methods comes down, then, to finding conditions under which a shrinking divergence from the empirical distribution guarantees a shrinking divergence from the true distribution.⁴

Definition 410 *Let $\theta \in \mathbb{R}^k$, $k < \infty$, be the parameter indexing a set \mathcal{M} of statistical models, where for every θ , $\nu_\theta \ll \rho$, with densities p_θ . Then the Fisher information matrix is*

$$I_{ij}(\theta) \equiv \mathbf{E}_{\nu_\theta} \left[\left(\frac{\partial \log p_\theta}{\partial \theta_i} \right) \left(\frac{\partial \log p_\theta}{\partial \theta_j} \right) \right] \quad (28.14)$$

Corollary 411 *The Fisher information matrix is equal to the Hessian (second partial derivative) matrix of the relative entropy:*

$$I_{ij}(\theta_0) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} D(\nu_{\theta_0} \|\nu_\theta) \quad (28.15)$$

⁴If we did have a triangle inequality, then we could say $D(\mu\|\nu) \leq D(\mu\|\eta) + D(\eta\|\nu)$, and it would be enough to make sure that both the terms on the RHS went to zero, say by some combination of maximizing the likelihood in-sample, so $D(\eta\|\nu)$ is small, and ergodicity, so that $D(\mu\|\eta)$ is small. While, as noted, there is no triangle inequality, under some conditions this idea is roughly right; there are nice diagrams in Kulhavy (1996).

PROOF: It is a classical result (see, e.g., Lehmann and Casella (1998, sec. 2.6.1)) that $I_{ij}(\theta) = -\mathbf{E}_{\nu_\theta} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta \right]$. The present result follows from this, Lemma 407, Lemma 408, and the fact that $H_\rho[\nu_{\theta_0}]$ is independent of θ . \square

28.3 Mutual Information

Definition 412 (Mutual Information) *The mutual information between two random variables, X and Y , is the divergence of the product of their marginal distributions from their actual joint distribution:*

$$I[X; Y] \equiv D(\mathcal{L}(X, Y) \parallel \mathcal{L}(X) \times \mathcal{L}(Y)) \quad (28.16)$$

Similarly, the mutual information among n random variables X_1, X_2, \dots, X_n is

$$I[X_1; X_2; \dots; X_n] \equiv D(\mathcal{L}(X_1, X_2, \dots, X_n) \parallel \prod_{i=1}^n \mathcal{L}(X_i)) \quad (28.17)$$

the divergence of the product distribution from the joint distribution.

Proposition 413 $I[X; Y] \geq 0$, with equality iff X and Y are independent.

PROOF: Directly from Lemma 401. \square

Proposition 414 *If all the entropies involved are finite,*

$$I[X; Y] = H[X] + H[Y] - H[X, Y] \quad (28.18)$$

$$= H[X] - H[X|Y] \quad (28.19)$$

$$= H[Y] - H[Y|X] \quad (28.20)$$

so $I[X; Y] \leq H[X] \wedge H[Y]$.

PROOF: Calculation. \square

This leads to the interpretation of the mutual information as the reduction in uncertainty or effective variability of X when Y is known, averaging over their joint distribution. Notice that in the discrete case, we can say $H[X] = I[X; X]$, which is why $H[X]$ is sometimes known as the *self-information*.

28.3.1 Mutual Information Function

Just as with the autocovariance function, we can define a mutual information function for one-parameter processes, to serve as a measure of serial dependence.

Definition 415 (Mutual Information Function) *The mutual information function of a one-parameter stochastic process X is*

$$\iota(t_1, t_2) \equiv I[X_{t_1}; X_{t_2}] \quad (28.21)$$

which is symmetric in its arguments. If the process is stationary, it is a function of $|t_1 - t_2|$ alone.

Notice that, unlike the autocovariance function, ι includes *nonlinear* dependencies between X_{t_1} and X_{t_2} . Also notice that $\iota(\tau) = 0$ means that the two variables are strictly independent, not just uncorrelated.

Theorem 416 *A stationary process is mixing if $\iota(\tau) \rightarrow 0$.*

PROOF: Because then the total variation distance between the joint distribution, $\mathcal{L}(X_{t_1}X_{t_2})$, and the product of the marginal distributions, $\mathcal{L}(X_{t_1})\mathcal{L}(X_{t_2})$, is being forced down towards zero, which implies mixing (Definition 379). \square

Chapter 29

Entropy Rates and Asymptotic Equipartition

Section 29.1 introduces the entropy rate — the asymptotic entropy per time-step of a stochastic process — and shows that it is well-defined; and similarly for information, divergence, etc. rates.

Section 29.2 proves the Shannon-MacMillan-Breiman theorem, a.k.a. the asymptotic equipartition property, a.k.a. the entropy ergodic theorem: asymptotically, almost all sample paths of a stationary ergodic process have the same log-probability per time-step, namely the entropy rate. This leads to the idea of “typical” sequences, in Section 29.2.1.

Section 29.3 discusses some aspects of asymptotic likelihood, using the asymptotic equipartition property, and allied results for the divergence rate.

29.1 Information-Theoretic Rates

Definition 417 (Entropy Rate) *The entropy rate of a random sequence X is*

$$h(X) \equiv \lim_n H_\rho[X_1^n]n \quad (29.1)$$

when the limit exists.

Definition 418 (Limiting Conditional Entropy) *The limiting conditional entropy of a random sequence X is*

$$h'(X) \equiv \lim_n H_\rho[X_n|X_1^{n-1}] \quad (29.2)$$

when the limit exists.

Lemma 419 *For a stationary sequence, $H_\rho[X_n|X_1^{n-1}]$ is non-increasing in n . Moreover, its limit exists if X takes values in a discrete space.*

PROOF: Because “conditioning reduces entropy”, $H_\rho[X_{n+1}|X_1^n] \leq H[X_{n+1}|X_2^n]$. By stationarity, $H_\rho[X_{n+1}|X_2^n] = H_\rho[X_n|X_1^{n-1}]$. If X takes discrete values, then conditional entropy is non-negative, and a non-increasing sequence of non-negative real numbers always has a limit. \square

Remark: Discrete values are a *sufficient* condition for the existence of the limit, not a necessary one.

We now need a natural-looking, but slightly technical, result from real analysis.

Theorem 420 (Cesàro) *For any sequence of real numbers $a_n \rightarrow a$, the sequence $b_n = n^{-1} \sum_{i=1}^n a_i$ also converges to a .*

PROOF: For every $\epsilon > 0$, there is an $N(\epsilon)$ such that $|a_n - a| < \epsilon$ whenever $n > N(\epsilon)$. Now take b_n and break it up into two parts, one summing the terms below $N(\epsilon)$, and the other the terms above.

$$\lim_n |b_n - a| = \lim_n \left| n^{-1} \sum_{i=1}^n a_i - a \right| \quad (29.3)$$

$$\leq \lim_n n^{-1} \sum_{i=1}^n |a_i - a| \quad (29.4)$$

$$\leq \lim_n n^{-1} \left(\sum_{i=1}^{N(\epsilon)} |a_i - a| + (n - N(\epsilon))\epsilon \right) \quad (29.5)$$

$$\leq \lim_n n^{-1} \left(\sum_{i=1}^{N(\epsilon)} |a_i - a| + n\epsilon \right) \quad (29.6)$$

$$= \epsilon + \lim_n n^{-1} \sum_{i=1}^{N(\epsilon)} |a_i - a| \quad (29.7)$$

$$= \epsilon \quad (29.8)$$

Since ϵ was arbitrary, $\lim b_n = a$. \square

Theorem 421 (Entropy Rate) *For a stationary sequence, if the limiting conditional entropy exists, then it is equal to the entropy rate, $h(X) = h'(X)$.*

PROOF: Start with the chain rule to break the joint entropy into a sum of conditional entropies, use Lemma 419 to identify their limit as $h^{\text{prime}}(X)$, and

then use Cesàro's theorem:

$$h(X) = \lim_n \frac{1}{n} H_\rho[X_1^n] \quad (29.9)$$

$$= \lim_n \frac{1}{n} \sum_{i=1}^n H_\rho[X_i | X_1^{i-1}] \quad (29.10)$$

$$= h'(X) \quad (29.11)$$

as required. \square

Because $h(X) = h'(X)$ for stationary processes (when both limits exist), it is not uncommon to find what I've called the limiting conditional entropy referred to as the entropy rate.

Lemma 422 *For a stationary sequence $h(X) \leq H[X_1]$, with equality iff the sequence is IID.*

PROOF: Conditioning reduces entropy, unless the variables are independent, so $H[X_n | X_1^{n-1}] < H[X_n]$, unless $X_n \perp\!\!\!\perp X_1^{n-1}$. For this to be true of all n , which is what's needed for $h(X) = H[X_1]$, all the values of the sequence must be independent of each other; since the sequence is stationary, this would imply that it's IID. \square

Example 423 (Markov Sequences) *If X is a stationary Markov sequence, then $h(X) = H_\rho[X_2 | X_1]$, because, by the chain rule, $H_\rho[X_1^n] = H_\rho[X_1] + \sum_{t=2}^n H_\rho[X_t | X_1^{t-1}]$. By the Markov property, however, $H_\rho[X_t | X_1^{t-1}] = H_\rho[X_t | X_{t-1}]$, which by stationarity is $H_\rho[X_2 | X_1]$. Thus, $H_\rho[X_1^n] = H_\rho[X_1] + (n-1)H_\rho[X_2 | X_1]$. Dividing by n and taking the limit, we get $H_\rho[X_1^n] = H_\rho[X_2 | X_1]$.*

Example 424 (Higher-Order Markov Sequences) *If X is a k^{th} order Markov sequence, then the same reasoning as before shows that $h(X) = H_\rho[X_{k+1} | X_1^k]$ when X is stationary.*

Definition 425 (Divergence Rate) *The divergence rate or relative entropy rate of the infinite-dimensional distribution Q from the infinite-dimensional distribution P , $d(P||Q)$, is*

$$d(P||Q) = \lim_n \mathbf{E}_P \left[\log \left(\frac{dP}{dQ} \Big|_{\sigma(X_{-n}^0)} \right) \right] \quad (29.12)$$

if all the finite-dimensional distributions of Q dominate all the finite-dimensional distributions of P . If P and Q have densities, respectively p and q , with respect to a common reference measure, then

$$d(P||Q) = \lim_n \mathbf{E}_P \left[\log \frac{p(X_0 | X_{-n}^{-1})}{q(X_0 | X_{-n}^{-1})} \right] \quad (29.13)$$

29.2 The Shannon-McMillan-Breiman Theorem or Asymptotic Equipartition Property

This is a central result in information theory, acting as a kind of ergodic theorem for the entropy. That is, we want to say that, for almost all ω ,

$$-\frac{1}{n} \log \mathbb{P}(X_1^n(\omega)) \rightarrow \lim_n \frac{1}{n} \mathbf{E}[-\log \mathbb{P}(X_1^n)] = h(X)$$

At first, it looks like we should be able to make a nice time-averaging argument. We can always factor the joint probability,

$$\frac{1}{n} \log \mathbb{P}(X_1^n) = \frac{1}{n} \sum_{t=1}^n \log \mathbb{P}(X_t | X_1^{t-1})$$

with the understanding that $\mathbb{P}(X_1 | X_1^0) = \mathbb{P}(X_1)$. This looks rather like the sort of Cesàro average that we became familiar with in ergodic theory. The problem is, there we were averaging $f(T^t\omega)$ for a *fixed* function f . This is not the case here, because we are conditioning on long and longer stretches of the past. There's no problem if the sequence is Markovian, because then the remote past is irrelevant, by the Markov property, and we can just condition on a fixed-length stretch of the past, so we're averaging a fixed function shifted in time. (This is why Shannon's original argument was for Markov chains.) The result nonetheless more broadly, but requires more subtlety than might otherwise be thought. Breiman's original proof of the general case was fairly involved¹, requiring both martingale theory, and a sort of dominated convergence theorem for ergodic time averages. (You can find a simplified version of his argument in Kallenberg, at the end of chapter 11.) We will go over the "sandwiching" argument of Algoet and Cover (1988), which is, to my mind, more transparent.

The idea of the sandwich argument is to show that, for large n , $-\frac{1}{n} \log \mathbb{P}(X_1^n)$ must lie between an upper bound, h_k , obtained by approximating the sequence by a Markov process of order k , and a lower bound, which will be shown to be h . Once we establish that $h_k \downarrow h$, we will be done.

Definition 426 (Markov Approximation) *For each k , define the order k Markov approximation to X by*

$$\mu_k(X_1^n) = \mathbb{P}(X_1^k) \prod_{t=k+1}^n \mathbb{P}(X_t | X_{t-k}^{t-1}) \quad (29.14)$$

μ_k is the distribution of a stationary Markov process of order k , where the distribution of X_1^{k+1} matches that of the original process.

¹Notoriously, the proof in his original paper was actually invalid, forcing him to publish a correction.

Lemma 427 *For each k , the entropy rate of the order k Markov approximation is equal to $H[X_{k+1}|X_1^k]$.*

PROOF: Under the approximation (but not under the original distribution of X), $H[X_1^n] = H[X_1^k] + (n-k)H[X_{k+1}|X_1^k]$, by the Markov property and stationarity (as in Examples 423 and 424). Dividing by n and taking the limit as $n \rightarrow \infty$ gives the result. \square

Lemma 428 *If X is a stationary two-sided sequence, then $Y_t = f(X_{-\infty}^t)$ defines a stationary sequence, for any measurable f . If X is also ergodic, then Y is ergodic too.*

PROOF: Because X is stationary, it can be represented as a measure-preserving shift on sequence space. Because it is measure-preserving, $\theta X_{-\infty}^t \stackrel{d}{=} X_{-\infty}^t$, so $Y(t) \stackrel{d}{=} Y(t+1)$, and similarly for all finite-length blocks of Y . Thus, all of the finite-dimensional distributions of Y are shift-invariant, and these determine the infinite-dimensional distribution, so Y itself must be stationary.

To see that Y must be ergodic if X is ergodic, recall that a random sequence is ergodic iff its corresponding shift dynamical system is ergodic. A dynamical system is ergodic iff all invariant functions are a.e. constant (Theorem 350). Because the Y sequence is obtained by applying a measurable function to the X sequence, a shift-invariant function of the Y sequence is a shift-invariant function of the X sequence. Since the latter are all constant a.e., the former are too, and Y is ergodic. \square

Lemma 429 *If X is stationary and ergodic, then, for every k ,*

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mu_k(X_1^n(\omega)) = h_k \right) = 1 \quad (29.15)$$

i.e., $-\frac{1}{n} \log \mu_k(X_1^n(\omega))$ converges a.s. to h_k .

PROOF: Start by factoring the approximating Markov measure in the way suggested by its definition:

$$-\frac{1}{n} \log \mu_k(X_1^n) = -\frac{1}{n} \log \mathbb{P}(X_1^k) - \frac{1}{n} \sum_{t=k+1}^n \log \mathbb{P}(X_t | X_{t-k}^{t-1}) \quad (29.16)$$

As n grows, $\frac{1}{n} \log \mathbb{P}(X_1^k) \rightarrow 0$, for every fixed k . On the other hand, $-\log \mathbb{P}(X_t | X_{t-k}^{t-1})$ is a measurable function of the past of the process, and since X is stationary and ergodic, it, too, is stationary and ergodic (Lemma 428). So

$$-\frac{1}{n} \log \mu_k(X_1^n) \rightarrow -\frac{1}{n} \sum_{t=k+1}^n \log \mathbb{P}(X_t | X_{t-k}^{t-1}) \quad (29.17)$$

$$\stackrel{a.s.}{\rightarrow} \mathbf{E}[-\log \mathbb{P}(X_{k+1} | X_1^k)] \quad (29.18)$$

$$= h_k \quad (29.19)$$

by Theorem 351. \square

Definition 430 *The infinite-order approximation to the entropy rate of a discrete-valued stationary process X is*

$$h_\infty(X) \equiv \mathbf{E}[-\log \mathbb{P}(X_0|X_{-\infty}^{-1})] \quad (29.20)$$

Lemma 431 *If X is stationary and ergodic, then*

$$\lim_n -\frac{1}{n} \log \mathbb{P}(X_1^n|X_{-\infty}^0) = h_\infty \quad (29.21)$$

almost surely.

PROOF: Via Theorem 351 again, as in Lemma 429. \square

Lemma 432 *For a stationary, ergodic, finite-valued random sequence, $h_k(X) \downarrow h_\infty(X)$.*

PROOF: By the martingale convergence theorem, for every $x_0 \in \Xi$,

$$\mathbb{P}(X_0 = x_0|X_n^{-1}) \xrightarrow{a.s.} \mathbb{P}(X_0 = x_0|X_\infty^{-1}) \quad (29.22)$$

Since Ξ is finite, the probability of any point in Ξ is between 0 and 1 inclusive, and $p \log p$ is bounded and continuous. So we can apply bounded convergence to get that

$$h_k = \mathbf{E} \left[- \sum_{x_0} \mathbb{P}(X_0 = x_0|X_{-k}^{-1}) \log \mathbb{P}(X_0 = x_0|X_{-k}^{-1}) \right] \quad (29.23)$$

$$\rightarrow \mathbf{E} \left[- \sum_{x_0} \mathbb{P}(X_0 = x_0|X_{-\infty}^{-1}) \log \mathbb{P}(X_0 = x_0|X_{-\infty}^{-1}) \right] \quad (29.24)$$

$$= h_\infty \quad (29.25)$$

Lemma 433 *$h_\infty(X)$ is the entropy rate of X , i.e. $h_\infty(X) = h(X)$.*

PROOF: Clear from Theorem 421 and the definition of conditional entropy. \square

We are almost ready for the proof, but need one technical lemma first.

Lemma 434 *If $R_n \geq 0$, $\mathbf{E}[R_n] \leq 1$ for all n , then*

$$\limsup_n \frac{1}{n} \log R_n \leq 0 \quad (29.26)$$

almost surely.

PROOF: Pick any $\epsilon > 0$.

$$\mathbb{P} \left(\frac{1}{n} \log R_n \geq \epsilon \right) = \mathbb{P}(R_n \geq e^{n\epsilon}) \quad (29.27)$$

$$\leq \frac{\mathbf{E}[R_n]}{e^{n\epsilon}} \quad (29.28)$$

$$\leq e^{-n\epsilon} \quad (29.29)$$

by Markov's inequality. Since $\sum_n e^{-n\epsilon} \leq \infty$, by the Borel-Cantelli lemma, $\limsup_n \frac{1}{n} \log R_n \leq \epsilon$. Since ϵ was arbitrary, this concludes the proof. \square

Theorem 435 (Asymptotic Equipartition Property) *For a stationary, ergodic, finite-valued random sequence X ,*

$$-\frac{1}{n} \log \mathbb{P}(X_1^n) \rightarrow h(X) \text{ a.s.} \quad (29.30)$$

PROOF: For every k , $\mu_k(X_1^n)/\mathbb{P}(X_1^n) \geq 0$, and $\mathbf{E}[\mu_k(X_1^n)/\mathbb{P}(X_1^n)] \leq 1$. Hence, by Lemma 434,

$$\limsup_n \frac{1}{n} \log \frac{\mu_k(X_1^n)}{\mathbb{P}(X_1^n)} \leq 0 \quad (29.31)$$

a.s. Manipulating the logarithm,

$$\limsup_n \frac{1}{n} \log \mu_k(X_1^n) \leq -\limsup_n \frac{1}{n} \log \mathbb{P}(X_1^n) \quad (29.32)$$

From Lemma 429, $\limsup_n \frac{1}{n} \log \mu_k(X_1^n) = \lim_n \frac{1}{n} \log \mu_k(X_1^n) = -h_k(X)$, a.s. Hence, for each k ,

$$h_k(X) \geq \limsup_n -\frac{1}{n} \log \mathbb{P}(X_1^n) \quad (29.33)$$

almost surely.

A similar manipulation of $\mathbb{P}(X_1^n)/\mathbb{P}(X_1^n | X_{-\infty}^0)$ gives

$$h_\infty(X) \leq \liminf_n -\frac{1}{n} \log \mathbb{P}(X_1^n) \quad (29.34)$$

a.s.

As $h_k \downarrow h_\infty$, it follows that the liminf and the limsup of the normalized log likelihood must be equal almost surely, and so equal to h_∞ , which is to say to $h(X)$. \square

Why is this called the AEP? Because, to within an $o(n)$ term, all sequences of length n have the same log-likelihood (to within factors of $o(n)$, if they have positive probability at all. In this sense, the likelihood is “equally partitioned” over those sequences.

29.2.1 Typical Sequences

Let’s turn the result of the AEP around. For large n , the probability of a given sequence is either approximately 2^{-nh} or approximately zero². To get the total probability to sum up to one, there need to be about 2^{nh} sequences with positive probability. If the size of the alphabet is s , then the fraction of sequences which are actually exhibited is $2^{n(h-\log s)}$, an increasingly small fraction (as $h \leq \log s$). Roughly speaking, these are the *typical* sequences, any one of which, via ergodicity, can act as a representative of the complete process.

²Of course that assumes using base-2 logarithms in the definition of entropy.

29.3 Asymptotic Likelihood

29.3.1 Asymptotic Equipartition for Divergence

Using methods analogous to those we employed on the AEP for entropy, it is possible to prove the following.

Theorem 436 *Let P be an asymptotically mean-stationary distribution, with stationary mean \bar{P} , with ergodic component function ϕ . Let M be a homogeneous finite-order Markov process, whose finite-dimensional distributions dominate those of P and \bar{P} ; denote the densities with respect to M by p and \bar{p} , respectively. If $\lim_n n^{-1} \log \bar{p}(X_1^n)$ is an invariant function \bar{P} -a.e., then*

$$-\frac{1}{n} \log p(X_1^n(\omega)) \xrightarrow{a.s.} d(\bar{P}_{\phi(\omega)} \| M) \quad (29.35)$$

where $\bar{P}_{\phi(\omega)}$ is the stationary, ergodic distribution of the ergodic component.

PROOF: See Algoet and Cover (1988, theorem 4), Gray (1990, corollary 8.4.1).

Remark. The usual AEP is in fact a consequence of this result, with the appropriate reference measure. (Which?)

29.3.2 Likelihood Results

It is left as an exercise for you to obtain the following result, from the AEP for relative entropy, Lemma 408 and the chain rules.

Theorem 437 *Let P be a stationary and ergodic data-generating process, whose entropy rate, with respect to some reference measure ρ , is h . Further let M be a finite-order Markov process which dominates P , whose density, with respect to the reference measure, is m . Then*

$$-\frac{1}{n} \log m(X_1^n) \rightarrow h + d(P \| M) \quad (29.36)$$

P -almost surely.

29.4 Exercises

Exercise 29.1 *Markov approximations are maximum-entropy approximations. (You may assume that the process X takes values in a finite set.)*

a Prove that μ_k , as defined in Definition 426, gets the distribution of sequences of length $k + 1$ correct, i.e., for any set $A \in \mathcal{X}^{k+1}$, $\nu(A) = \mathbb{P}(X_1^{k+1} \in A)$.

b Prove that $\mu_{k'}$, for any $k' > k$, also gets the distribution of length $k + 1$ sequences right.

- c* In a slight abuse of notation, let $H[\nu(X_1^n)]$ stand for the entropy of a sequence of length n when distributed according to ν . Show that $H[\mu_k(X_1^n)] \geq H[\mu_{k'}(X_1^n)]$ if $k' > k$. (Note that the $n \leq k$ case is easy!)
- d* Is it true that that if ν is any other measure which gets the distribution of sequences of length $k + 1$ right, then $H[\mu_k(X_1^n)] \geq H[\nu(X_1^n)]$? If yes, prove it; if not, find a counter-example.

Exercise 29.2 Prove Theorem 437.

Part VII

Large Deviations

Chapter 30

General Theory of Large Deviations

A family of random variables follows the *large deviations principle* if the probability of the variables falling into “bad” sets, representing large deviations from expectations, declines exponentially in some appropriate limit. Section 30.1 makes this precise, using some associated technical machinery, and explores a few consequences. The central one is Varadhan’s Lemma, for the asymptotic evaluation of exponential integrals in infinite-dimensional spaces.

Having found one family of random variables which satisfy the large deviations principle, many other, related families do too. Section 30.2 lays out some ways in which this can happen.

As the great forensic statistician C. Chan once remarked, “Improbable events permit themselves the luxury of occurring” (reported in Biggers, 1928). Large deviations theory, as I have said, studies these little luxuries.

30.1 Large Deviation Principles: Main Definitions and Generalities

Some technicalities:

Definition 438 (Level Sets) *For any real-valued function $f : \Xi \mapsto \mathbb{R}$, the level sets are the inverse images of intervals from $-\infty$ to c inclusive, i.e., all sets of the form $\{x \in \Xi : f(x) \leq c\}$.*

Definition 439 (Lower Semi-Continuity) *A real-valued function $f : \Xi \mapsto \mathbb{R}$ is lower semi-continuous if $x_n \rightarrow x$ implies $\liminf f(x_n) \geq f(x)$.*

Lemma 440 *A function is lower semi-continuous iff either of the following equivalent properties hold.*

- i For all $x \in \Xi$, the infimum of f over increasingly small open balls centered at x approaches $f(x)$:

$$\lim_{\delta \rightarrow 0} \inf_{y: d(y,x) < \delta} f(y) = f(x) \quad (30.1)$$

- ii f has closed level sets.

PROOF: A character-building exercise in real analysis, left to the reader. \square

Lemma 441 *A lower semi-continuous function attains its minimum on every non-empty compact set, i.e., if C is compact and $\neq \emptyset$, there is an $x \in C$ such that $f(x) = \inf_{y \in C} f(y)$.*

PROOF: Another character-building exercise in real analysis. \square

Definition 442 (Logarithmic Equivalence) *Two sequences of positive real numbers a_n and b_n are logarithmically equivalent, $a_n \simeq b_n$, when*

$$\lim_{n \rightarrow \infty} \frac{1}{n} (\log a_n - \log b_n) = 0 \quad (30.2)$$

Similarly, for continuous parameterizations by $\epsilon > 0$, $a_\epsilon \simeq b_\epsilon$ when

$$\lim_{\epsilon \rightarrow 0} \epsilon (\log a_\epsilon - \log b_\epsilon) = 0 \quad (30.3)$$

Lemma 443 (“Fastest rate wins”) *For any two sequences of positive numbers, $(a_n + b_n) \simeq a_n \vee b_n$.*

PROOF: A character-building exercise in elementary analysis. \square

Definition 444 (Large Deviation Principle) *A parameterized family of random variables, X_ϵ , $\epsilon > 0$, taking values in a metric space Ξ with Borel σ -field \mathcal{X} , obeys a large deviation principle with rate $1/\epsilon$, or just obeys an LDP, when, for any set $B \in \mathcal{X}$,*

$$-\inf_{x \in \text{int} B} J(x) \leq \liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) \leq \limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) \leq -\inf_{x \in \text{cl} B} J(x) \quad (30.4)$$

for some non-negative function $J : \Xi \mapsto [0, \infty]$, its raw rate function. If J is lower semi-continuous, it is just a rate function. If J is lower semi-continuous and has compact level sets, it is a good rate function.¹ By a slight abuse of notation, we will write $J(B) = \inf_{x \in B} J(x)$.

¹Sometimes what Kallenberg and I are calling a “good rate function” is just “a rate function”, and our “rate function” gets demoted to “weak rate function”.

Remark: The most common choices of ϵ are $1/n$, in sample-size or discrete sequence problems, or ϵ^2 , in small-noise problems (as in Chapter 21).

Lemma 445 (Uniqueness of Rate Functions) *If X_ϵ obeys the LDP with raw rate function J , then it obeys the LDP with a unique rate function J' .*

PROOF: First, show that a raw rate function can always be replaced by a lower semi-continuous function, i.e. a non-raw (cooked?) rate function. Then, show that non-raw rate functions are unique.

For any raw rate function J , define $J'(x) = \liminf_{y \rightarrow x} J(y)$. This is clearly lower semi-continuous, and $J'(x) \leq J(x)$. However, for any open set B , $\inf_{x \in B} J'(x) = \inf_{x \in B} J(x)$, so J and J' are equivalent for purposes of the LDP.

Now assume that J is a lower semi-continuous rate function, and suppose that $K \neq J$ was too; without loss of generality, assume that $J(x) > K(x)$ at some point x . We can use semi-continuity to find an open neighborhood B of x such that $J(\text{cl}B) > K(x)$. But, substituting into Eq. 30.4, we obtain a contradiction:

$$-K(x) \leq -K(B) \quad (30.5)$$

$$\leq \liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) \quad (30.6)$$

$$\leq -J(\text{cl}B) \quad (30.7)$$

$$\leq -K(x) \quad (30.8)$$

Hence there can be no such rate function K , and J is the unique rate function. \square

Lemma 446 *If X_ϵ obeys an LDP with rate function J , then $J(x) = 0$ for some x .*

PROOF: Because $\mathbb{P}(X_\epsilon \in \Xi) = 1$, we must have $J(\Xi) = 0$, and rate functions attain their infima. \square

Definition 447 *A Borel set B is J -continuous, for some rate function J , when $J(\text{int}B) = J(\text{cl}B)$.*

Lemma 448 *If X_ϵ satisfies the LDP with rate function J , then for every J -continuous set B ,*

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in B) = -J(B) \quad (30.9)$$

PROOF: By J -continuity, the right and left hand extremes of Eq. 30.4 are equal, so the limsup and the liminf sandwiched between them are equal; consequently the limit exists. \square

Remark: The obvious implication is that, for small ϵ , $\mathbb{P}(X_\epsilon \in B) \approx ce^{-J(B)/\epsilon}$, which explains why we say that the LDP has rate $1/\epsilon$. (Actually, c need not be constant, but it must be at least $o(\epsilon)$, i.e., it must go to zero faster than ϵ itself does.)

There are several equivalent ways of defining the large deviation principle. The following is especially important, because it often simplifies proofs.

Lemma 449 X_ϵ obeys the LDP with rate $1/\epsilon$ and rate function $J(x)$ if and only if

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in C) \leq -J(C) \quad (30.10)$$

$$\liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in O) \geq -J(O) \quad (30.11)$$

for every closed Borel set C and every open Borel set $O \subset \Xi$.

PROOF: “If”: The closure of any set is closed, and the interior of any set is open, so Eqs. 30.10 and 30.11 imply

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in \text{cl}B) \leq -J(\text{cl}B) \quad (30.12)$$

$$\liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in \text{int}B) \geq -J(\text{int}B) \quad (30.13)$$

but $\mathbb{P}(X_\epsilon \in B) \leq \mathbb{P}(X_\epsilon \in \text{cl}B)$ and $\mathbb{P}(X_\epsilon \in B) \geq \mathbb{P}(X_\epsilon \in \text{int}B)$, so the LDP holds. “Only if”: every closed set is equal to its own closure, and every open set is equal to its own interior, so the upper bound in Eq. 30.4 implies Eq. 30.10, and the lower bound Eq. 30.11. \square

A deeply important consequence of the LDP is the following, which can be thought of as a version of Laplace’s method for infinite-dimensional spaces.

Theorem 450 (Varadhan’s Lemma) If X_ϵ are random variables in a metric space Ξ , obeying an LDP with rate $1/\epsilon$ and rate function J , and $f : \Xi \mapsto \mathbb{R}$ is continuous and bounded from above, then

$$\Lambda_f \equiv \lim_{\epsilon \rightarrow 0} \epsilon \log \mathbf{E} \left[e^{f(X_\epsilon)/\epsilon} \right] = \sup_{x \in \Xi} f(x) - J(x) \quad (30.14)$$

PROOF: We’ll find the limsup and the liminf, and show that they are both $\sup f(x) - J(x)$.

First the limsup. Pick an arbitrary positive integer n . Because f is continuous and bounded above, there exist finitely closed sets, call them B_1, \dots, B_m , such that $f \leq -n$ on the complement of $\bigcup_i B_i$, and within each B_i , f varies by at most $1/n$. Now

$$\limsup \epsilon \log \mathbf{E} \left[e^{f(X_\epsilon)/\epsilon} \right] \quad (30.15)$$

$$\leq (-n) \vee \max_{i \leq m} \limsup \epsilon \log \mathbf{E} \left[e^{f(X_\epsilon)/\epsilon} \mathbf{1}_{B_i}(X_\epsilon) \right] \quad (30.16)$$

$$\leq (-n) \vee \max_{i \leq m} \sup_{x \in B_i} f(x) - \inf_{x \in B_i} J(x) \quad (30.17)$$

$$\leq (-n) \vee \max_{i \leq m} \sup_{x \in B_i} f(x) - J(x) + 1/n \quad (30.18)$$

$$\leq (-n) \vee \sup_{x \in \Xi} f(x) - J(x) + 1/n$$

Letting $n \rightarrow \infty$, we get $\limsup \epsilon \log \mathbf{E} [e^{f(X_\epsilon)/\epsilon}] = \sup f(x) - J(x)$.

To get the liminf, pick any $x \in X$ and an arbitrary ball of radius δ around it, $B_{\delta,x}$. We have

$$\liminf \epsilon \log \mathbf{E} [e^{f(X_\epsilon)/\epsilon}] \geq \liminf \epsilon \log \mathbf{E} [e^{f(X_\epsilon)/\epsilon} \mathbf{1}_{B_{\delta,x}}(X_\epsilon)] \quad (30.19)$$

$$\geq \inf_{y \in B_{\delta,x}} f(y) - \inf_{y \in B_{\delta,x}} J(y) \quad (30.20)$$

$$\geq \inf_{y \in B_{\delta,x}} f(y) - J(x) \quad (30.21)$$

Since δ was arbitrary, we can let it go to zero, so (by continuity of f) $\inf_{y \in B_{\delta,x}} f(y) \rightarrow f(x)$, or

$$\liminf \epsilon \log \mathbf{E} [e^{f(X_\epsilon)/\epsilon}] \geq f(x) - J(x) \quad (30.22)$$

Since this holds for arbitrary x , we can replace the right-hand side by a supremum over all x . Hence $\sup f(x) - J(x)$ is both the liminf and the limsup. \square

Remark: The implication of Varadhan's lemma is that, for small ϵ , $\mathbf{E} [e^{f(X_\epsilon)/\epsilon}] \approx c(\epsilon)e^{\epsilon^{-1}(\sup_{x \in \Xi} f(x) - J(x))}$, where $c(\epsilon) = o(\epsilon)$. So, we can replace the exponential integral with its value at the extremal points, at least to within a multiplicative factor and to first order in the exponent.

An important, if heuristic, consequence of the LDP is that “Highly improbable events tend to happen in the least improbable way”. Let us consider two events $B \subset A$, and suppose that $\mathbb{P}(X_\epsilon \in A) > 0$ for all ϵ . Then $\mathbb{P}(X_\epsilon \in B | X_\epsilon \in A) = \mathbb{P}(X_\epsilon \in B) / \mathbb{P}(X_\epsilon \in A)$. Roughly speaking, then, this conditional probability will vanish exponentially, with rate $J(A) - J(B)$. That is, even if we are looking at an exponentially-unlikely large deviation, the vast majority of the probability is concentrated around the *least unlikely* part of the event. More formal statements of this idea are sometimes known as “conditional limit theorems” or “the Gibbs conditioning principle”.

30.2 Breeding Large Deviations

Often, the easiest way to prove that one family of random variables obeys a large deviations principle is to prove that another, related family does.

Theorem 451 (Contraction Principle) *If X_ϵ , taking values in a metric space Ξ , obeys an LDP, with rate ϵ and rate function J , and $f : \Xi \rightarrow \Upsilon$ is a continuous function from that metric space to another, then $Y_\epsilon = f(X_\epsilon)$ also obeys an LDP, with rate ϵ and raw rate function $K(y) = J(f^{-1}(y))$. If J is a good rate function, then so is K .*

PROOF: Since f is continuous, f^{-1} takes open sets to open sets, and closed sets to closed sets. Pick any closed $C \subset \Upsilon$. Then

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(f(X_\epsilon) \in C) \quad (30.23)$$

$$= \limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in f^{-1}(C))$$

$$\leq -J(f^{-1}(C)) \quad (30.24)$$

$$= -\inf_{x \in f^{-1}(C)} J(x) \quad (30.25)$$

$$= -\inf_{y \in C} \inf_{x \in f^{-1}(y)} J(x) \quad (30.26)$$

$$= -\inf_{y \in C} K(y) \quad (30.27)$$

as required. The argument for open sets in Υ is entirely parallel, establishing that K , as defined, is a raw rate function. By Lemma 445, K can be modified to be lower semi-continuous without affecting the LDP, i.e., we can make a rate function from it. If J is a good rate function, then it has compact level sets. But continuous functions take compact sets to compact sets, so $K = J \circ f^{-1}$ will also have compact level sets, i.e., it will also be a good rate function. \square

There are a bunch of really common applications of the contraction principle, relating the large deviations at one level of description to those at coarser levels. To make the most frequent set of implications precise, let's recall a couple of definitions.

Definition 452 (Empirical Mean) If X_1, \dots, X_n are random variables in a common vector space Ξ , their empirical mean is $\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i$.

We have already encountered this as the sample average or, in ergodic theory, the finite time average. (Notice that nothing is said about the X_i being IID, or even having a common expectation.)

Definition 453 (Empirical Distribution) Let X_1, \dots, X_n be random variables in a common measurable space Ξ (not necessarily a vector or metric space). The empirical distribution is $\hat{P}_n \equiv \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability measure that puts all its probability on the point x , i.e., $\delta_x(B) = \mathbf{1}_B(x)$. \hat{P}_n is a random variable taking values in $\mathcal{P}(\Xi)$, the space of all probability measures on Ξ . (Cf. Example 10 in chapter 1 and Example 43 in chapter 4.) $\mathcal{P}(\Xi)$ is a metric space under any of several distances, and a complete separable metric space (i.e., Polish) under, for instance, the total variation metric.

Definition 454 (Finite-Dimensional Empirical Distributions) For each k , the k -dimensional empirical distribution is

$$\hat{P}_n^k \equiv \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, X_{i+1}, \dots, X_{i+k})} \quad (30.28)$$

where the addition of indices for the delta function is to be done modulo n , i.e., $\hat{P}_3^2 = \frac{1}{3} (\delta_{(X_1, X_2)} + \delta_{(X_2, X_3)} + \delta_{(X_3, X_1)})$. \hat{P}_n^k takes values in $\mathcal{P}(\Xi^k)$.

Definition 455 (Empirical Process Distribution) *With a finite sequence of random variables X_1, \dots, X_n , the empirical process is the periodic, infinite random sequence \tilde{X}_n as the repetition of the sample without limit, i.e., $\tilde{X}_n(i) = X_{i \bmod n}$. If T is the shift operator on the sequence space, then the empirical process distribution is*

$$\hat{P}_n^\infty \equiv \frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i \tilde{X}_n} \quad (30.29)$$

\hat{P}_n^∞ takes values in the space of infinite-dimensional distributions for one-sided sequences, $\mathcal{P}(\Xi^{\mathbb{N}})$. (In fact, it is always a stationary distribution, because by construction it is invariant under the shift T .)

Be careful not to confuse this “empirical process” with the quite distinct “empirical process” of Example 43.

Corollary 456 *The following chain of implications hold:*

- i If the empirical process distribution obeys an LDP, so do all the finite-dimensional distributions.*
- ii If the n -dimensional distribution obeys an LDP, all $m < n$ dimensional distributions do.*
- iii If any finite-dimensional distribution obeys an LDP, the empirical distribution does.*
- iv If the empirical distribution obeys an LDP, the empirical mean does.*

PROOF: In each case, we obtain the lower-level statistic from the higher-level one by applying a continuous function, hence the contraction principle applies. For the distributions, the continuous functions are the projection operators of Chapter 2. \square

Corollary 457 (“Tilted” LDP) *In set-up of Theorem 450, let $\mu_\epsilon = \mathcal{L}(X_\epsilon)$. Define the probability measures $\mu_{f,\epsilon}$ via*

$$\mu_{f,\epsilon}(B) \equiv \frac{\mathbf{E}[e^{f(X_\epsilon)/\epsilon} \mathbf{1}_B(X_\epsilon)]}{\mathbf{E}[e^{f(X_\epsilon)/\epsilon}]} \quad (30.30)$$

Then $Y_\epsilon \sim \mu_{f,\epsilon}$ obeys an LDP with rate $1/\epsilon$ and rate function

$$J_F(x) = -(f(x) - J(x)) + \sup_{y \in \Xi} f(y) - J(y) \quad (30.31)$$

PROOF: Define a set function $F_\epsilon(B) = \mathbf{E} [e^{f(X_\epsilon)/\epsilon} \mathbf{1}_B(X_\epsilon)]$; then $\mu_{f,\epsilon}(B) = F_\epsilon(B)/F_\epsilon(\Xi)$. From Varadhan's Lemma, we know that $F_\epsilon(\Xi)$ has asymptotic logarithm $\sup_{y \in \Xi} f(y) - J(y)$, so it is just necessary to show that

$$\limsup_{\epsilon} \epsilon \log F_\epsilon(B) \leq \sup_{x \in \text{cl} B} f(x) - J(x) \quad (30.32)$$

$$\liminf_{\epsilon} \epsilon \log F_\epsilon(B) \geq \sup_{x \in \text{int} B} f(x) - J(x) \quad (30.33)$$

which can be done by imitating the proof of Varadhan's Lemma itself. \square

Remark: "Tilting" here refers to some geometrical analogy which, in all honesty, has never made any sense to me.

Because the LDP is about exponential decay of probabilities, it is not surprising that several ways of obtaining it require a sort of exponential bound on the dispersion of the probability measure.

Definition 458 (Exponentially Tight) *The parameterized family of random variables X_ϵ , $\epsilon > 0$, is exponentially tight if, for every finite real M , there exists a compact set $C \subset \Xi$ such that*

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \notin C) \leq -M \quad (30.34)$$

The first use of exponential tightness is a converse to the contraction principle: a high-level LDP is implied by the combination of a low-level LDP and high-level exponential tightness.

Theorem 459 (Inverse Contraction Principle) *If X_ϵ are exponentially tight, f is continuous and injective, and $Y_\epsilon = f(X_\epsilon)$ obeys an LDP with rate function K , then X_ϵ obeys an LDP with a good rate function $J(x) = K(f(x))$.*

PROOF: See Kallenberg, Theorem 27.11 (ii). Notice, by the way, that the proof of the upper bound on probabilities (i.e. that $\limsup_{\epsilon} \epsilon \log \mathbb{P}(X_\epsilon \in B) \leq -J(B)$ for closed $B \subseteq \Xi$) does not depend on exponential tightness, just the continuity of f . Exponential tightness is only needed for the lower bound. \square

Theorem 460 (Bryc's Theorem) *If X_ϵ are exponentially tight, and, for all bounded continuous f , the limit*

$$\Lambda_f \equiv \lim_{\epsilon \rightarrow 0} \epsilon \log \mathbf{E} [e^{f(X_\epsilon/\epsilon)}] \quad (30.35)$$

exists, then X_ϵ obeys the LDP with good rate function

$$J(x) \equiv \sup_f f(x) - \Lambda_f \quad (30.36)$$

where the supremum extends over all bounded, continuous functions.

PROOF: See Kallenberg, Theorem 27.10, part (ii). \square

Remark: This is a converse to Varadhan's Lemma.

Theorem 461 (Projective Limit Theorem (Dawson-Gärtner)) *Let Ξ_1, Ξ_2, \dots be a countable sequence of metric spaces, and let X_ϵ be a random sequence from this space. If, for every n , $X_\epsilon^n = \pi_n X_\epsilon$ obeys the LDP with good rate function J_n , then X_ϵ obeys the LDP with good rate function*

$$J(x) \equiv \sup_n J_n(\pi_n x) \quad (30.37)$$

PROOF: See Kallenberg, Theorem 27.12. \square

Definition 462 (Exponentially Equivalent Random Variables) *Two families of random variables, X_ϵ and Y_ϵ , taking values in a common metric space, are exponentially equivalent when, for all positive δ ,*

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(d(X_\epsilon, Y_\epsilon) > \delta) = -\infty \quad (30.38)$$

Lemma 463 *If X_ϵ and Y_ϵ are exponentially equivalent, one of them obeys the LDP with a good rate function J iff the other does as well.*

PROOF: It is enough to prove that the LDP for X_ϵ implies the LDP for Y_ϵ , with the same rate function. (Draw a truth-table if you don't believe me!) As usual, first we'll get the upper bound, and then the lower.

Pick any closed set C , and let C_δ be its closed δ neighborhood, i.e., $C_\delta = \{x : \exists y \in C, d(x, y) \leq \delta\}$. Now

$$\mathbb{P}(Y_\epsilon \in C_\delta) \leq \mathbb{P}(X_\epsilon \in C_\delta) + \mathbb{P}(d(X_\epsilon, Y_\epsilon) > \delta) \quad (30.39)$$

Using Eq. 30.38 from Definition 462, the LDP for X_ϵ , and Lemma 443

$$\limsup \epsilon \log \mathbb{P}(Y_\epsilon \in C) \quad (30.40)$$

$$\leq \limsup \epsilon \log \mathbb{P}(X_\epsilon \in C_\delta) + \epsilon \log \mathbb{P}(d(X_\epsilon, Y_\epsilon) > \delta) \quad (30.41)$$

$$\leq -J(C_\delta) \vee -\infty \quad (30.42)$$

$$= -J(C_\delta) \quad (30.43)$$

Since J is a good rate function, we have $J(C_\delta) \uparrow J(C)$ as $\delta \downarrow 0$; since δ was arbitrary to start with,

$$\limsup \epsilon \log \mathbb{P}(Y_\epsilon \in C) \leq -J(C) \quad (30.44)$$

As usual, to obtain the lower bound on open sets, pick any open set O and any point $x \in O$. Because O is open, there is a $\delta > 0$ such that, for some open

neighborhood U of x , not only is $U \subset O$, but $U_\delta \subset O$. In which case, we can say that

$$\mathbb{P}(X_\epsilon \in U) \leq \mathbb{P}(Y_\epsilon \in O) + \mathbb{P}(d(X_\epsilon, Y_\epsilon) > h) \quad (30.45)$$

Proceeding as for the upper bound,

$$-J(x) \leq -J(U) \quad (30.46)$$

$$\leq \liminf \epsilon \log \mathbb{P}(X_\epsilon \in U) \quad (30.47)$$

$$\leq \liminf \epsilon \log \mathbb{P}(Y_\epsilon \in O) \vee \limsup \epsilon \log \mathbb{P}(d(X_\epsilon, Y_\epsilon) > \delta) \quad (30.48)$$

$$= \liminf \epsilon \log \mathbb{P}(Y_\epsilon \in O) \quad (30.49)$$

(Notice that the initial arbitrary choice of δ has dropped out.) Taking the supremum over all x gives $-J(O) \leq \liminf \epsilon \log \mathbb{P}(Y_\epsilon \in O)$, as required. \square

Chapter 31

Large Deviations for IID Sequences: The Return of Relative Entropy

Section 31.1 introduces the exponential version of the Markov inequality, which will be our major calculating device, and shows how it naturally leads to both the cumulant generating function and the Legendre transform, which we should suspect (correctly) of being the large deviations rate function. We also see the reappearance of relative entropy, as the Legendre transform of the cumulant generating *functional* of distributions.

Section 31.2 proves the large deviations principle for the empirical mean of IID sequences in finite-dimensional Euclidean spaces (Cramér's Theorem).

Section 31.3 proves the large deviations principle for the empirical distribution of IID sequences in Polish spaces (Sanov's Theorem), using Cramér's Theorem for a well-chosen collection of bounded continuous functions on the Polish space, and the tools of Section 30.2. Here the rate function is the relative entropy.

Section 31.4 proves that even the infinite-dimensional empirical process distribution of an IID sequence in a Polish space obeys the LDP, with the rate function given by the relative entropy rate.

The usual approach in large deviations theory is to establish an LDP for some comparatively tractable basic case through explicit calculations, and then use the machinery of Section 30.2 to extend it to LDPs for more complicated cases. This chapter applies this strategy to IID sequences.

31.1 Cumulant Generating Functions and Relative Entropy

Suppose the only inequality we knew in probability theory was Markov's inequality, $\mathbb{P}(X \geq a) \leq \mathbf{E}[X]/a$ when $X \geq 0$. How might we extract an exponential probability bound from it? Well, for any real-valued variable, e^{tX} is positive, so we can say that $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \mathbf{E}[e^{tX}]/e^{ta}$. $\mathbf{E}[e^{tX}]$ is of course the moment generating function of X . It has the nice property that addition of independent random variables leads to multiplication of their moment generating functions, as $\mathbf{E}[e^{t(X_1+X_2)}] = \mathbf{E}[e^{tX_1}e^{tX_2}] = \mathbf{E}[e^{tX_1}]\mathbf{E}[e^{tX_2}]$ if $X_1 \perp\!\!\!\perp X_2$. If X_1, X_2, \dots are IID, then we can get a deviation bound for their sample mean \bar{X}_n through the moment generating function:

$$\begin{aligned} \mathbb{P}(\bar{X}_n \geq a) &= \mathbb{P}\left(\sum_{i=1}^n X_i \geq na\right) \\ \mathbb{P}(\bar{X}_n \geq a) &\leq e^{-nta} (\mathbf{E}[e^{tX_1}])^n \\ \frac{1}{n} \log \mathbb{P}(\bar{X}_n \geq a) &\leq -ta + \log \mathbf{E}[e^{tX_1}] \\ &\leq \inf_t -ta + \log \mathbf{E}[e^{tX_1}] \\ &\leq -\sup_t ta - \log \mathbf{E}[e^{tX_1}] \end{aligned}$$

This suggests that the functions $\log \mathbf{E}[e^{tX}]$ and $\sup_t ta - \log \mathbf{E}[e^{tX}]$ will be useful to us. Accordingly, we encapsulate them in a pair of definitions.

Definition 464 (Cumulant Generating Function) *The cumulant generating function of a random variable X in \mathbb{R}^d is a function $\Lambda : \mathbb{R}^d \mapsto \mathbb{R}$,*

$$\Lambda(t) \equiv \log \mathbf{E}[e^{t \cdot X}] \quad (31.1)$$

Definition 465 (Legendre Transform) *The Legendre transform of a real-valued function f on \mathbb{R}^d is another real-valued function on \mathbb{R}^d ,*

$$f^*(x) \equiv \sup_{t \in \mathbb{R}^d} t \cdot x - f(t) \quad (31.2)$$

The definition of cumulant generating functions and their Legendre transforms can be extended to arbitrary spaces where some equivalent of the inner product (a real-valued form, bilinear in its two arguments) makes sense; f and f^* then must take arguments from the complementary spaces.

Legendre transforms are particularly important in convex analysis¹, since convexity is preserved by taking Legendre transforms. If f is not convex initially,

¹See Rockafellar (1970), or, more concisely, Ellis (1985, ch. VI).

then f^{**} is (in one dimension) something like the greatest convex lower bound on f ; made precise, this statement even remains true in higher dimensions. I make these remarks because of the following fact:

Lemma 466 *The cumulant generating function $\Lambda(t)$ is convex.*

PROOF: Simple calculation, using Hölder's inequality in one step:

$$\Lambda(at + bu) = \log \mathbf{E} \left[e^{(at+bu)X} \right] \quad (31.3)$$

$$= \log \mathbf{E} \left[e^{atX} e^{buX} \right] \quad (31.4)$$

$$= \log \mathbf{E} \left[(e^{tX})^a (e^{uX})^b \right] \quad (31.5)$$

$$\leq \log (\mathbf{E} [e^{tX}])^a (\mathbf{E} [e^{uX}])^b \quad (31.6)$$

$$= a\Lambda(t) + b\Lambda(u) \quad (31.7)$$

which proves convexity. \square

Our previous result, then, is easily stated: if the X_i are IID in \mathbb{R} , then

$$\mathbb{P}(\bar{X}_n \geq a) \leq \Lambda^*(a) \quad (31.8)$$

where $\Lambda^*(a)$ is the Legendre transform of the cumulant generating function of the X_i . This elementary fact is, surprisingly enough, the foundation of the large deviations principle for empirical means.

The notion of cumulant generating functions can be extended to probability measures, and this will be useful when dealing with large deviations of empirical distributions. The definitions follow the pattern one would expect from the complementarity between probability measures and bounded continuous functions.

Definition 467 (Cumulant Generating Functional) *Let X be a random variable on a metric space Ξ , with distribution μ , and let $C_b(\Xi)$ be the class of all bounded, continuous, real-valued functions on Ξ . Then the cumulant-generating functional $\Lambda : C_b(\Xi) \mapsto \mathbb{R}$ is*

$$\Lambda(f) \equiv \log \mathbf{E} \left[e^{f(X)} \right] \quad (31.9)$$

Definition 468 *The Legendre transform of a real-valued functional F on $C_b(\Xi)$ is*

$$F^*(\nu) \equiv \sup_{f \in C_b(\Xi)} \mathbf{E}_\nu[f] - \Lambda(f) \quad (31.10)$$

where $\nu \in \mathcal{P}(\Xi)$, the set of all probability measures on Ξ .

Lemma 469 (Donsker and Varadhan) *The Legendre transform of the cumulant generating functional is the relative entropy:*

$$\Lambda^*(\nu) = D(\nu \| \mu) \quad (31.11)$$

PROOF: First of all, notice that the supremum in Eq. 31.10 can be taken over all bounded *measurable* functions, not just functions in C_b , since C_b is dense. This will let us use indicator functions and simple functions in the subsequent argument.

If $\nu \not\ll \mu$, then $D(\nu\|\mu) = \infty$. But then there is also a set, call it B , with $\mu(B) = 0$, $\nu(B) > 0$. Take $f_n = n\mathbf{1}_B$. Then $\mathbf{E}_\nu[f_n] - \Lambda(f_n) = n\nu(B) - 0$, which can be made arbitrarily large by taking n arbitrarily large, hence the supremum in Eq. 31.10 is ∞ .

If $\nu \ll \mu$, then show that $D(\nu\|\mu) \leq \Lambda^*(\nu)$ and $D(\nu\|\mu) \geq \Lambda^*(\nu)$, so they must be equal. To get the first inequality, start with the observation then $\frac{d\nu}{d\mu}$ exists, so set $f = \log \frac{d\nu}{d\mu}$, which is measurable. Then $D(\nu\|\mu)$ is $\mathbf{E}_\nu[f] - \log \mathbf{E}_\mu[e^f]$. If f is bounded, this shows that $D(\nu\|\mu) \leq \Lambda^*(\nu)$. If f is not bounded, approximate it by a sequence of bounded, measurable functions f_n with $\mathbf{E}_\mu[e^{f_n}] \rightarrow 1$ and $\mathbf{E}_\nu[f_n] \rightarrow \mathbf{E}_\nu[f]$, again concluding that $D(\nu\|\mu) \leq \Lambda^*(\nu)$.

To go the other way, first consider the special case where \mathcal{X} is finite, and so generated by a partition, with cells B_1, \dots, B_n . Then all measurable functions are simple functions, and $\mathbf{E}_\nu[f] - \Lambda(f)$ is

$$g(f) = \sum_{i=1}^n f_i \nu(B_i) - \log \sum_{i=1}^n e^{f_i} \mu(B_i) \quad (31.12)$$

Now, $g(f)$ is concave on all the f_i , and

$$\frac{\partial g(f)}{\partial f_i} = \nu(B_i) - \frac{1}{\sum_{i=1}^n e^{f_i} \mu(B_i)} \mu(B_i) e^{f_i} \quad (31.13)$$

Setting this equal to zero,

$$\frac{\nu(B_i)}{\mu(B_i)} = \frac{1}{\sum_{i=1}^n \mu(B_i) e^{f_i}} e^{f_i} \quad (31.14)$$

$$\log \frac{\nu(B_i)}{\mu(B_i)} = f_i \quad (31.15)$$

gives the maximum value of $g(f)$. (Remember that $0 \log 0 = 0$.) But then $g(f) = D(\nu\|\mu)$. So $\Lambda^*(\nu) \leq D(\nu\|\mu)$ when the σ -algebra is finite. In the general case, consider the case where f is a simple function. Then $\sigma(f)$ is finite, and $\mathbf{E}_\nu[f] - \log \mathbf{E}_\mu[e^f] \leq D(\nu\|\mu)$ follows by the finite case and smoothing. Finally, if f is not simple, but is bounded and measurable, there is a simple h such that $\mathbf{E}_\nu[f] - \log \mathbf{E}_\mu[e^f] \leq \mathbf{E}_\nu[h] - \log \mathbf{E}_\mu[e^h]$, so

$$\sup_{f \in C_b(\Xi)} \mathbf{E}_\nu[f] - \log \mathbf{E}_\mu[e^f] \leq D(\nu\|\mu) \quad (31.16)$$

which completes the proof. \square

31.2 Large Deviations of the Empirical Mean in \mathbb{R}^d

Historically, the oldest and most important result in large deviations is that the empirical mean of an IID sequence of real-valued random variables obeys a large deviations principle with rate n ; the oldest version of this proposition goes back to Harald Cramér in the 1930s, and so it is known as Cramér's theorem, even though the modern version, which is both more refined technically and works in arbitrary finite-dimensional Euclidean spaces, is due to Varadhan in the 1960s.

Theorem 470 (Cramér's Theorem) *If X_i are IID random variables in \mathbb{R}^d , and $\Lambda(t) < \infty$ for all $t \in \mathbb{R}^d$, then their empirical mean obeys an LDP with rate n and good rate function $\Lambda^*(x)$.*

PROOF: The proof has three parts. First, the upper bound for closed sets; second, the lower bound for open sets, under an additional assumption on $\Lambda(t)$; third and finally, lifting of the assumption on Λ by means of a perturbation argument (related to Lemma 463).

To prove the upper bound for closed sets, we first prove the upper bound for sufficiently small balls around arbitrary points. Then, we take our favorite closed set, and divide it into a compact part close to the origin, which we can cover by a finite number of closed balls, and a remainder which is far from the origin and of low probability.

First the small balls of low probability. Because $\Lambda^*(x) = \sup_u u \cdot x - \Lambda(u)$, for any $\epsilon > 0$, we can find some u such that $u \cdot x - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$. (Otherwise, $\Lambda^*(x)$ would not be the *least* upper bound.) Since $u \cdot x$ is continuous in x , it follows that there exists some open ball B of positive radius, centered on x , within which $u \cdot y - \Lambda(x) > \min 1/\epsilon, \Lambda^*(x) - \epsilon$, or $u \cdot y > \Lambda(x) + \min 1/\epsilon, \Lambda^*(x) - \epsilon$. Now use the exponential Markov inequality to get

$$\mathbb{P}(\bar{X}_n \in B) \leq \mathbf{E} \left[e^{u \cdot n \bar{X}_n - n \inf_{y \in B} u \cdot y} \right] \quad (31.17)$$

$$\leq e^{-n(\min \frac{1}{\epsilon}, \Lambda^*(x) - \epsilon)} \quad (31.18)$$

which is small. To get the compact set near the origin of high probability, use the exponential decay of the probability at large $\|x\|$. Since $\Lambda(t) < \infty$ for all t , $\Lambda^*(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$. So, using (once again) the exponential Markov inequality, for every $\epsilon > 0$, there must exist an $r > 0$ such that

$$\frac{1}{n} \log \mathbb{P}(\|\bar{X}_n\| > r) \leq -\frac{1}{\epsilon} \quad (31.19)$$

for all n .

Now pick your favorite closed measurable set $C \in \mathcal{B}^d$. Then $C \cap \{x : \|x\| \leq r\}$ is compact, and I can cover it by m balls B_1, \dots, B_m , with centers x_1, \dots, x_m , of the sort built in the previous paragraph. So I can apply a union bound to

$\mathbb{P}(\bar{X}_n \in C)$, as follows.

$$\mathbb{P}(\bar{X}_n \in C) \tag{31.20}$$

$$= \mathbb{P}(\bar{X}_n \in C \cap \{x : \|x\| \leq r\}) + \mathbb{P}(\bar{X}_n \in C \cap \{x : \|x\| > r\})$$

$$\leq \mathbb{P}\left(\bar{X}_n \in \bigcup_{i=1}^m B_i\right) + \mathbb{P}(\|\bar{X}_n\| > r) \tag{31.21}$$

$$\leq \left(\sum_{i=1}^m \mathbb{P}(\bar{X}_n \in B_i)\right) + \mathbb{P}(\|\bar{X}_n\| > r) \tag{31.22}$$

$$\leq \left(\sum_{i=1}^m e^{-n(\min \frac{1}{\epsilon}, \Lambda^*(x_i) - \epsilon)}\right) + e^{-n\frac{1}{\epsilon}} \tag{31.23}$$

$$\leq (m+1)e^{-n(\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon)} \tag{31.24}$$

with $\Lambda^*(C) = \inf_{x \in C} \Lambda^*(x)$, as usual. So if I take the log, normalize, and go to the limit, I have

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in C) \leq -\min \frac{1}{\epsilon}, \Lambda^*(C) - \epsilon \tag{31.25}$$

$$\leq -\Lambda^*(C) \tag{31.26}$$

since ϵ was arbitrary to start with, and I've got the upper bound for closed sets.

To get the lower bound for open sets, pick your favorite open set $O \in \mathcal{B}^d$, and your favorite $x \in O$. Suppose, for the moment, that $\Lambda(t)/\|t\| \rightarrow \infty$ as $\|t\| \rightarrow \infty$. (This is the growth condition mentioned earlier, which we will left at the end of the proof.) Then, because $\Lambda(t)$ is smooth, there is some u such that $\nabla \Lambda(u) = x$. (You will find it instructive to draw the geometry here.) Now let Y_i be a sequence of IID random variables, whose probability law is given by

$$\mathbb{P}(Y_i \in B) = \frac{\mathbf{E}[e^{uX} \mathbf{1}_B(X)]}{\mathbf{E}[e^{uX}]} = e^{-\Lambda(u)} \mathbf{E}[e^{uX} \mathbf{1}_B(X)] \tag{31.27}$$

It is not hard to show, by manipulating the cumulant generating functions, that $\Lambda_Y(t) = \Lambda_X(t+u) - \Lambda_X(u)$, and consequently that $\mathbf{E}[Y_i] = x$. I construct these Y to allow me to pull the following trick, which works if $\epsilon > 0$ is sufficiently small that the first inequality holds (and I can always chose small enough ϵ):

$$\mathbb{P}(\bar{X}_n \in O) \geq \mathbb{P}(\|\bar{X}_n - x\| < \epsilon) \tag{31.28}$$

$$= e^{n\Lambda(u)} \mathbf{E}\left[e^{-nu\bar{Y}_n} \mathbf{1}\{y : \|y - x\| < \epsilon\}(\bar{Y}_n)\right] \tag{31.29}$$

$$\geq e^{n\Lambda(u) - nu \cdot x - n\epsilon\|u\|} \mathbb{P}(\|\bar{Y}_n - x\| < \epsilon) \tag{31.30}$$

By the strong law of large numbers, $\mathbb{P}(\|\bar{Y}_n - x\| < \epsilon) \rightarrow 1$ for all ϵ , so

$$\liminf \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in O) \geq \Lambda(u) - u \cdot x - \epsilon \|u\| \quad (31.31)$$

$$\geq -\Lambda^*(x) - \epsilon \|u\| \quad (31.32)$$

$$\geq -\Lambda^*(x) \quad (31.33)$$

$$\geq -\inf_{x \in O} \Lambda^*(x) = -\Lambda(O) \quad (31.34)$$

as required. This proves the LDP, as required, if $\Lambda(t)/\|t\| \rightarrow \infty$ as $\|t\| \rightarrow \infty$.

Finally, to lift the last-named restriction (which, remember, only affected the lower bound for open sets), introduce a sequence Z_i of IID standard Gaussian variables, i.e. $Z_i \sim \mathcal{N}(0, I)$, which are completely independent of the X_i . It is easily calculated that the cumulant generating function of the Z_i is $\|t\|^2/2$, so that \bar{Z}_n satisfies the LDP. Another easy calculation shows that $X_i + \sigma Z_i$ has cumulant generating function $\Lambda_X(t) + \frac{\sigma^2}{2}\|t\|^2$, which again satisfies the previous condition. Since $\Lambda_{X+\sigma Z} \geq \Lambda_X$, $\Lambda_X^* \geq \Lambda_{X+\sigma Z}^*$. Now, once again pick any open set O , and any point $x \in O$, and an ϵ sufficiently small that all points within a distance 2ϵ of x are also in O . Since the LDP applies to $X + \sigma Z$,

$$\mathbb{P}(\|\bar{X}_n + \sigma \bar{Z}_n - x\| \leq \epsilon) \geq -\Lambda_{X+\sigma Z}^*(x) \quad (31.35)$$

$$\geq -\Lambda_X^*(x) \quad (31.36)$$

On the other hand, basic probability manipulations give

$$\mathbb{P}(\|\bar{X}_n + \sigma \bar{Z}_n - x\| \leq \epsilon) \leq \mathbb{P}(\bar{X}_n \in O) + \mathbb{P}(\sigma \|\bar{Z}_n\| \geq \epsilon) \quad (31.37)$$

$$\leq 2 \max \mathbb{P}(\bar{X}_n \in O), \mathbb{P}(\sigma \|\bar{Z}_n\| \geq \epsilon) \quad (31.38)$$

Taking the liminf of the normalized log of both sides,

$$\liminf \frac{1}{n} \log \mathbb{P}(\|\bar{X}_n + \sigma \bar{Z}_n - x\| \leq \epsilon) \quad (31.39)$$

$$\leq \liminf \frac{1}{n} \log (\max \mathbb{P}(\bar{X}_n \in O), \mathbb{P}(\sigma \|\bar{Z}_n\| \geq \epsilon))$$

$$\leq \liminf \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in O) \vee \left(-\frac{\epsilon^2}{2\sigma^2} \right) \quad (31.40)$$

$$(31.41)$$

Since σ was arbitrary, we can let it go to zero, and obtain

$$\liminf \frac{1}{n} \log \mathbb{P}(\bar{X}_n \in O) \geq -\Lambda_X^*(x) \quad (31.42)$$

$$\geq -\Lambda_X^*(O) \quad (31.43)$$

as required. \square

31.3 Large Deviations of the Empirical Measure in Polish Spaces

The Polish space setting is, apparently, more general than \mathbb{R}^d , but we will represent distributions on the Polish space in terms of the expectation of a separating set of functions, and then appeal to the Euclidean result.

Proposition 471 *Any Polish space S can be represented as a Borel subset of a compact metric space, namely $[0, 1]^{\mathbb{N}} \equiv M$.*

PROOF: See, for instance, Appendix A of Kallenberg. \square

Strictly speaking, there should be a function mapping points from S to points in M . However, since this is an embedding, I will silently omit it in what follows.

Proposition 472 *$C_b(M)$ has a countable dense separating set $\mathcal{F} = f_1, f_2, \dots$*

PROOF: See Kallenberg again. \square

Because \mathcal{F} is separating, to specify a probability distribution on K is equivalent to specifying the expectation value of all the functions in \mathcal{F} . Write $f_1^d(X)$ to abbreviate the d -dimensional vector $(f_1(X), f_2(X), \dots, f_d(X))$, and $f_1^\infty(X)$ to abbreviate the corresponding infinite-dimensional vector.

Lemma 473 *Empirical means are expectations with respect to empirical measure. That is, let f be a real-valued measurable function and $Y_i = f(X_i)$. Then $\bar{Y}_n = \mathbf{E}_{\hat{P}_n} [f(X)]$.*

PROOF: Direct calculation.

$$\bar{Y}_n \equiv \frac{1}{n} \sum_{i=1}^n f(X_i) \quad (31.44)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbf{E}_{\delta_{X_i}} [f(X)] \quad (31.45)$$

$$\equiv \mathbf{E}_{\hat{P}_n} [f(X)] \quad (31.46)$$

\square

Lemma 474 *Let X_i be a sequence of IID random variables in a Polish space Ξ . For each d , the sequence of vectors $(\mathbf{E}_{\hat{P}_n} [f_1], \dots, \mathbf{E}_{\hat{P}_n} [f_d])$ obeys the LDP with rate n and good rate function J_d .*

PROOF: For each d , the sequence of vectors $(f_1(X_i), \dots, f_d(X_i))$ are IID, so, by Cramér's Theorem (470), their empirical mean obeys the LDP with rate n and good rate function

$$J_d(x) = \sup_{t \in \mathbb{R}^d} t \cdot x - \log \mathbf{E} \left[e^{t \cdot f_1^d(X)} \right] \quad (31.47)$$

But, by Lemma 473, the empirical means are expectations over the empirical distributions, so the latter must also obey the LDP, with the same rate and rate function. \square

Notice, incidentally, that the fact that the $f_i \in \mathcal{F}$ isn't relevant for the proof of the lemma; it will however be relevant for the proof of the theorem.

Theorem 475 (Sanov's Theorem) *Let X_i , $i \in \mathbb{N}$, be IID random variables in a Polish space Ξ , with common probability measure μ . Then the empirical distributions \hat{P}_n obey an LDP with rate n and good rate function $J(\nu) = D(\nu \parallel \mu)$.*

PROOF: Combining Lemma 474 and Theorem 461, we see that $\mathbf{E}_{\hat{P}_n} [f_1^\infty(X)]$ obeys the LDP with rate n and good rate function

$$J(x) = \sup_d J_d(\pi_d x) \quad (31.48)$$

$$= \sup_d \sup_{t \in \mathbb{R}^d} t \cdot \pi_d x - \log \mathbf{E} \left[e^{t \cdot f_1^d(X)} \right] \quad (31.49)$$

Since $\mathcal{P}(M)$ is compact (so all random sequences in it are exponentially tight), and the mapping from $\nu \in \mathcal{P}(M)$ to $\mathbf{E}_\nu [f_1^\infty] \in \mathbb{R}^\mathbb{N}$ is continuous, apply the inverse contraction principle (Theorem 459) to get that \hat{P}_n satisfies the LDP with good rate function

$$J(\nu) = J(\mathbf{E}_\nu [f_1^\infty]) \quad (31.50)$$

$$= \sup_d \sup_{t \in \mathbb{R}^d} t \cdot \mathbf{E}_\nu [f_1^d] - \log \mathbf{E}_\mu \left[e^{t \cdot f_1^d(X)} \right] \quad (31.51)$$

$$= \sup_{f \in \text{span} \mathcal{F}} \mathbf{E}_\nu [f] - \Lambda(f) \quad (31.52)$$

$$= \sup_{f \in C_b(M)} \mathbf{E}_\nu [f] - \Lambda(f) \quad (31.53)$$

$$= D(\nu \parallel \mu) \quad (31.54)$$

Notice however that this is an LDP in the space $\mathcal{P}(M)$, not in $\mathcal{P}(\Xi)$. However, the embedding taking $\mathcal{P}(\Xi)$ to $\mathcal{P}(M)$ is continuous, and it is easily verified (see Lemma 27.17 in Kallenberg) that \hat{P}_n is exponentially tight in $\mathcal{P}(\Xi)$, so another application of the inverse contraction principle says that \hat{P}_n must obey the LDP in the restricted space $\mathcal{P}(\Xi)$, and with the same rate. \square

31.4 Large Deviations of the Empirical Process in Polish Spaces

A fairly straightforward modification of the proof for Sanov's theorem establishes a large deviations principle for the finite-dimensional empirical distributions of an IID sequence.

Corollary 476 *Let X_i be an IID sequence in a Polish space Ξ , with common measure μ . Then, for every finite positive integer k , the k -dimensional empirical distribution \hat{P}_n^k , obeys an LDP with rate n and good rate function $J_k(\nu) = D(\nu \| \pi_{k-1}\nu \otimes \mu)$ if $\nu \in \mathcal{P}(\Xi^k)$ is shift invariant, and $J(\nu) = \infty$ otherwise.*

This leads to the following important generalization.

Theorem 477 *If X_i are IID in a Polish space, with a common measure μ , then the empirical process distribution \hat{P}_n^∞ obeys an LDP with rate n and good rate function $J_\infty(\nu) = d(\nu \| \mu^\infty)$, the relative entropy rate, if ν is a shift-invariant probability measure, and $= \infty$ otherwise.*

PROOF: By Corollary 476 and the projective limit theorem 461, \hat{P}_n^∞ obeys an LDP with rate n and good rate function

$$J_\infty(\nu) = \sup_k J_k(\pi_k \nu) = \sup_k D(\pi_k \nu \| \pi_{k-1} \nu \otimes \mu) \quad (31.55)$$

But, applying the chain rule for relative entropy (Lemma 404),

$$D(\pi_n \nu \| \mu^n) = D(\pi_n \nu \| \pi_{n-1} \nu \otimes \mu) + D(\pi_{n-1} \nu \| \mu^{n-1}) \quad (31.56)$$

$$= \sum_{k=1}^n D(\pi_k \nu \| \pi_{k-1} \nu \otimes \mu) \quad (31.57)$$

$$\lim \frac{1}{n} D(\pi_n \nu \| \mu^n) = \lim \frac{1}{n} \sum_{k=1}^n D(\pi_k \nu \| \pi_{k-1} \nu \otimes \mu) \quad (31.58)$$

$$= \sup_k D(\pi_k \nu \| \pi_{k-1} \nu \otimes \mu) \quad (31.59)$$

But $\lim n^{-1} D(\pi_n \nu \| \mu^n)$ is the relative entropy rate, $d(\nu \| \mu^\infty)$, and we've already identified the right-hand side as the rate function. \square

The strength of Theorem 477 lies in the fact that, via the contraction principle (Theorem 451), it implies that the LDP holds for any continuous function of the empirical process distribution. This in particular includes the finite-dimensional distributions, the empirical mean, functions of finite-length trajectories, etc. Moreover, Theorem 451 also provides a means to calculate the rate function for all these quantities.

Chapter 32

Large Deviations for Markov Sequences

This chapter establishes large deviations principles for Markov sequences as natural consequences of the large deviations principles for IID sequences in Chapter 31. (LDPs for continuous-time Markov processes will be treated in the chapter on Freidlin-Wentzell theory.)

Section 32.1 uses the exponential-family representation of Markov sequences to establish an LDP for the two-dimensional empirical distribution (“pair measure”). The rate function is a relative entropy.

Section 32.2 extends the results of Section 32.1 to other observables for Markov sequences, such as the empirical process and time averages of functions of the state.

For the whole of this chapter, let X_1, X_2, \dots be a homogeneous Markov sequence, taking values in a Polish space Ξ , with transition probability kernel μ , and initial distribution ν and invariant distribution ρ . If Ξ is not discrete, we will assume that ν and ρ have densities n and r with respect to some reference measure, and that $\mu(x, dy)$ has density $m(x, y)$ with respect to that same reference measure, for all x . (LDPs can be proved for Markov sequences without such density assumptions — see, e.g., Ellis (1988) — but the argument is more complicated.)

32.1 Large Deviations for Pair Measure of Markov Sequences

It is perhaps not sufficiently appreciated that Markov sequences form exponential families (Billingsley, 1961; Küchler and Sørensen, 1997). Suppose Ξ is

discrete. Then

$$\mathbb{P}(X_1^n = x_1^t) = \nu(x_1) \prod_{i=1}^{t-1} \mu(x_i, x_{i+1}) \quad (32.1)$$

$$= \nu(x_1) e^{\sum_{i=1}^{t-1} \log \mu(x_i, x_{i+1})} \quad (32.2)$$

$$= \nu(x_1) e^{\sum_{x,y \in \Xi^2} T_{x,y}(x_1^t) \log \mu(x,y)} \quad (32.3)$$

where $T_{x,y}(x_1^t)$ counts the number of times the state y follows the state x in the sequence x_1^t , i.e., it gives the *transition counts*. What we have just established is that the Markov chains on Ξ with a given initial distribution form an exponential family, whose natural sufficient statistics are the transition counts, and whose natural parameters are the logarithms of the transition probabilities.

(If Ξ is not continuous, but we make the density assumptions mentioned at the beginning of this chapter, we can write

$$p_{X_1^t}(x_1^t) = n(x_1) \prod_{i=1}^{t-1} m(x_i, x_{i+1}) \quad (32.4)$$

$$= n(x_1) e^{\int_{\Xi^2} dT(x_1^t) \log m(x,y)} \quad (32.5)$$

where now $T(x_1^t)$ puts probability mass $\frac{1}{n-1}$ at x, y for every i such that $x_i = x$, $x_{i+1} = y$.)

We can use this exponential family representation to establish the following basic theorem.

Theorem 478 *Let X_i be a Markov sequence obeying the assumptions set out at the beginning of this chapter, and furthermore that $\mu(x, y)/\rho(y)$ is bounded above (in the discrete-state case) or that $m(x, y)/r(y)$ is bounded above (in the continuous-state case). Then the two-dimensional empirical distribution (“pair measure”) \hat{P}_t^2 obeys an LDP with rate n and with rate function $J_2(\psi) = D(\psi \| \pi_1 \psi \times \mu)$ if ν is shift-invariant, $J(\nu) = \infty$ otherwise.*

PROOF: I will just give the proof for the discrete case, since the modifications for the continuous case are straightforward (given the assumptions made about densities), largely a matter of substituting Roman letters for Greek ones.

First, modify the representation of the probabilities in Eq. 32.3 slightly, so that it refers directly to \hat{P}_t^2 (as laid down in Definition 454), rather than to the transition counts.

$$\mathbb{P}(X_1^t = x_1^t) = \frac{\nu(x_1)}{\mu(x_t, x_1)} e^{t \sum_{x,y \in \Xi} \hat{P}_t^2(x,y) \log \mu(x,y)} \quad (32.6)$$

$$= \frac{\nu(x_1)}{\mu(x_t, x_1)} e^{n \mathbf{E}_{\hat{P}_t^2} [\log \mu(X, Y)]} \quad (32.7)$$

Now construct a sequence of IID variables Y_i , all distributed according to ρ , the invariant measure of the Markov chain:

$$\mathbb{P}(Y_1^t = y_1^t) = e^{n \mathbf{E}_{\hat{P}_t^2} [\log \rho(Y)]} \quad (32.8)$$

The ratio of these probabilities is the Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_X}{d\mathbb{P}_Y}(x_1) = \frac{\nu(x_1)}{\mu(x_1, x_1)} e^{t\mathbf{E}_{\hat{P}_n^2}[t \log \frac{\mu(X, Y)}{\rho(Y)}]} \quad (32.9)$$

(In the continuous- Ξ case, the derivative is the ratio of the densities with respect to the common reference measure, and the principle is the same.) Introducing the functional $F(\nu) = \mathbf{E}_\nu \left[\log \frac{\mu(X, Y)}{\rho(Y)} \right]$, the derivative is equal to $O(1)e^{tF(\hat{P}_t^2)}$, and our initial assumption amounts to saying that F is not just continuous (which it must be) but bounded from above.

Now introduce $Q_{t,X}$, the distribution of the empirical pair measure \hat{P}_t^2 under the Markov process, and $Q_{t,Y}$, the distribution of \hat{P}_t^2 for the IID samples produced by Y_i . From Eq. 32.9,

$$\frac{1}{t} \log \mathbb{P}(\hat{P}_t^2 \in B) = \frac{1}{t} \log \int_B dQ_{t,X}(\psi) \quad (32.10)$$

$$= \frac{1}{t} \log \int_B \frac{dQ_{t,X}}{dQ_{t,Y}} dQ_{t,Y}(\psi) \quad (32.11)$$

$$= O\left(\frac{1}{t}\right) + \frac{1}{t} \log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) \quad (32.12)$$

It is thus clear that

$$\liminf \frac{1}{t} \log \mathbb{P}(\hat{P}_t^2 \in B) = \liminf \frac{1}{t} \log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) \quad (32.13)$$

$$\limsup \frac{1}{t} \log \mathbb{P}(\hat{P}_t^2 \in B) = \limsup \frac{1}{t} \log \int_B e^{tF(\psi)} dQ_{t,Y}(\psi) \quad (32.14)$$

Introduce a (final) proxy random sequence, also taking values in $\mathcal{P}(\Xi^2)$, call it Z_t , with $\mathbb{P}(Z_t \in B) = \int_B e^{tF(\psi)} dQ_{t,Y}(\psi)$. We know (Corollary 476) that, under $Q_{t,Y}$, the empirical pair measure satisfies an LDP with rate t and good rate function $J_Y = D(\psi \| \pi_1 \psi \otimes \rho)$, so by Corollary 457, Z_t satisfies an LDP with rate t and good rate function

$$J_F(\psi) = -(F(\psi) - J_Y(\psi)) + \sup_{\zeta \in \mathcal{P}(\Xi^2)} F(\zeta) - J_Y(\zeta) \quad (32.15)$$

A little manipulation turns this into

$$J_F(\psi) = D(\psi \| \pi_1 \psi \otimes \mu) - \inf_{\zeta \in \mathcal{P}(\Xi^2)} D(\zeta \| \pi_1 \zeta \otimes \mu) \quad (32.16)$$

and the infimum is clearly zero. Since this is the rate function Z_t , in view of Eqs. 32.13 and 32.14 it is also the rate function for \hat{P}_n^2 , which we have agreed to call J_2 . \square

Remark 1: The key to making this work is the assumption that F is bounded from above. This can fail if, for instance, the process is not ergodic, although usually in that case one can rescue the general idea by some kind of ergodic decomposition.

Remark 2: The LDP for the pair measure of an IID sequence can now be seen to be a special case of the LDP for the pair measure of a Markov sequence. The same is true, generally speaking, of all the other LDPs for IID and Markov sequences. Calculations are almost always easier for the IID case, however, which permits us to give explicit formulae for the rate functions of empirical means and empirical distributions unavailable (generally speaking) in the Markovian case.

Corollary 479 *The minima of the rate function J_2 are the invariant distributions.*

PROOF: The rate function is $D(\psi \| \pi_1 \psi \otimes \mu)$. Since relative entropy is ≥ 0 , and equal to zero iff the two distributions are equal (Lemma 401), we get a minimum of zero in the rate function iff $\psi = \pi_1 \psi \otimes \mu$, or $\psi = \rho^2$, for some $\rho \in \mathcal{P}(\Xi)$ such that $\rho\mu = \rho$. Conversely, if ψ is of this form, then $J_2(\psi) = 0$. \square

Corollary 480 *The empirical distribution \hat{P}_t obeys an LDP with rate t and good rate function*

$$J_1(\psi) = \inf_{\zeta \in \mathcal{P}(\Xi^2): \pi_1 \zeta = \psi} D(\zeta \| \pi_1 \zeta \otimes \mu) \quad (32.17)$$

PROOF: This is a direct application of the Contraction Principle (Theorem 451), as in Corollary 456. \square

Remark: Observe that if ψ is invariant under the action of the Markov chain, then $J_1(\psi) = 0$ by a combination of the preceding corollaries. This is good, because we know from ergodic theory that the empirical distribution converges on the invariant distribution for an ergodic Markov chain. In fact, in view of Lemma 402, which says that $D(\psi \| \rho) \geq \frac{1}{2 \ln 2} \|\psi - \rho\|_1^2$, the probability that the empirical distribution differs from the invariant distribution ρ by more than δ , in total variation distance, goes down like $O(e^{-t\delta^2/2})$.

Corollary 481 *If Theorem 478 holds, then time averages of observables, $A_t f$, obey a large deviations principle with rate function*

$$J_0(a) = \inf_{\zeta \in \mathcal{P}(\Xi^2): \mathbf{E}_{\pi_1 \zeta}[f(X)] = a} D(\zeta \| \pi_1 \zeta \otimes \mu) \quad (32.18)$$

PROOF: Another application the Contraction Principle, as in Corollary 456. \square

Remark: Observe that if $a = \mathbf{E}_\rho[f(X)]$, with ρ invariant, then the $J_0(a) = 0$. Again, it is reassuring to see that large deviations theory is compatible with ergodic theory, which tells us to expect the almost-sure convergence of $A_t f$ on $\mathbf{E}_\rho[f(X)]$.

Corollary 482 *If X_i are from a Markov sequence of order $k+1$, then, under conditions analogous to Theorem 478, the $k+1$ -dimensional empirical distribution \hat{P}_t^{k+1} obeys an LDP with rate t and good rate function*

$$D(\nu \| \pi_{k-1} \nu \otimes \mu) \quad (32.19)$$

PROOF: An obvious extension of the argument for Theorem 478, using the appropriate exponential-family representation of the higher-order process. \square

Whether all exponential-family stochastic processes (Küchler and Sørensen, 1997) obey LDPs is an interesting question; I'm not sure if anyone knows the answer.

32.2 Higher LDPs for Markov Sequences

In this section, I assume without further comment that the Markov sequence X obeys the LDP of Theorem 478.

Theorem 483 *For all $k \geq 2$, the finite-dimensional empirical distribution \hat{P}_t^k obeys an LDP with rate t and good rate function $J_k(\psi) = D(\psi \| \pi_{k-1} \psi \otimes \mu)$, if $\psi \in \mathcal{P}(\Xi^k)$ is shift-invariant, and $= \infty$ otherwise.*

PROOF: The case $k = 2$ is just Theorem 478. However, if $k \geq 3$, the argument preceding that theorem shows that $\mathbb{P}(\hat{P}_t^k \in B)$ depends only on $\pi_2 \hat{P}_t^k$, the pair measure implied by the k -dimensional distribution, so the proof of that theorem can be adapted to apply to \hat{P}_t^k , in conjunction with Corollary 476, establishing the LDP for finite-dimensional distributions of IID sequences. The identification of the rate function follows the same argument, too. \square

Theorem 484 *The empirical process distribution obeys an LDP with rate t and good rate function $J_\infty(\psi) = d(\psi \| \rho)$, with ρ here standing for the stationary process distribution of the Markov sequence.*

PROOF: Entirely parallel to the proof of Theorem 477, with Theorem 483 substituting for Corollary 476. \square

Consequently, any continuous function of the empirical process distribution has an LDP.

Chapter 33

Large Deviations for Weakly Dependent Sequences via the Gärtner-Ellis Theorem

This chapter proves the Gärtner-Ellis theorem, establishing an LDP for not-too-dependent processes taking values in topological vector spaces. Most of our earlier LDP results can be seen as consequences of this theorem.

33.1 The Gärtner-Ellis Theorem

The Gärtner-Ellis theorem is a powerful result which establishes the existence of a large deviation principle for processes where the cumulant generating function tends towards a well-behaved limit, implying not-too-strong dependence between successive values. (Exercise 33.5 clarifies the meaning of “too strong”.) It will imply our LDPs for IID and Markovian sequences. I could have started with it, but its proof, as you’ll see, is pretty technical, and so it seemed better to use the more elementary arguments of the preceding chapters.

To fix notation, Ξ will be a real topological vector space, and Ξ^* will be its dual space, of continuous linear functions $\Xi \mapsto \mathbb{R}$. (If $\Xi = \mathbb{R}^d$, we can identify Ξ and Ξ^* by means of the inner product. In differential geometry, on the other hand, Ξ might be a space of tangent vectors, and Ξ^* the corresponding one-forms.) X_ϵ will be a family of Ξ -valued random variables, parameterized by $\epsilon > 0$. Refer to Definitions 464 and 465 in Section 31.1 for the definition of the cumulant generating function and its Legendre transform (respectively), which I will denote by $\Lambda_\epsilon : \Xi^* \mapsto \mathbb{R}$ and $\Lambda_\epsilon^* : \Xi \mapsto \overline{\mathbb{R}}$.

The proof of the Gärtner-Ellis theorem goes through a number of lemmas.

Basically, the upper large deviation bound holds under substantially weaker conditions than the lower bound does, and it's worthwhile having the partial results available to use in estimates even if the full large deviations principle does not apply.

Definition 485 *The upper-limiting cumulant generating function is*

$$\bar{\Lambda}(t) \equiv \limsup_{\epsilon \rightarrow 0} \epsilon \Lambda_{\epsilon}(t/\epsilon) \quad (33.1)$$

and its Legendre transform is written $\bar{\Lambda}^(x)$.*

The point of this is that the limsup always exists, whereas the limit doesn't, necessarily. But we can show that the limsup has some reasonable properties, and in fact it's enough to give us an upper bound.

Lemma 486 *$\bar{\Lambda}(t)$ is convex, and $\bar{\Lambda}^*(x)$ is a convex rate function.*

PROOF: The proof of the convexity of $\Lambda_{\epsilon}(t)$ follows the proof in Lemma 466, and the convexity of $\bar{\Lambda}(t)$ by passing to the limit. To establish $\bar{\Lambda}^*(x)$ as a rate function, we need it to be non-negative and lower semi-continuous. Since $\Lambda_{\epsilon}(0) = 0$ for all ϵ , $\bar{\Lambda}(0) = 0$. This in turn implies that $\bar{\Lambda}^*(x) \geq 0$. Since the latter is the supremum of a class of continuous functions, namely $t(x) - \bar{\Lambda}(t)$, it must be lower semi-continuous. Finally, its convexity is implied by its being a Legendre transform. \square

Lemma 487 (Upper Bound in Gärtner-Ellis Theorem: Compact Sets)

For any compact set $K \subset \Xi$,

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_{\epsilon} \in K) \leq -\bar{\Lambda}^*(K) \quad (33.2)$$

PROOF: Entirely parallel to the proof of the upper bound in Cramér's Theorem (470), up through the point where closed sets are divided into a compact part and a remainder far from the origin, of exponentially-small probability. Because K is compact, we can proceed as though the remainder is empty. \square

Lemma 488 (Upper Bound in Gärtner-Ellis Theorem: Closed Sets) *If the family of distributions $\mathcal{L}(X_{\epsilon})$ are exponentially tight, then for all closed $C \subset \Xi$,*

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_{\epsilon} \in C) \leq -\bar{\Lambda}^*(C) \quad (33.3)$$

PROOF: Exponential tightness, by definition (458), will let us repeat Theorem 470 trick of dividing closed sets into a compact part, and an exponentially-vanishing non-compact part. \square

Definition 489 The limiting cumulant generating function is

$$\Lambda(t) \equiv \lim_{\epsilon \rightarrow 0} \epsilon \Lambda_\epsilon(t/\epsilon) \quad (33.4)$$

when the limit exists. Its domain of finiteness $D \equiv \{t \in \Xi^* : \Lambda(t) < \text{infty}\}$. Its limiting Legendre transform is Λ^* , with domain of finiteness D^* .

Lemma 490 If $\Lambda(t)$ exists, then it is convex, $\Lambda^*(x)$ is a convex rate function, and Eq. 33.2 applies to the latter. If in addition the process is exponentially tight, then Eq. 33.3 holds for $\Lambda^*(x)$.

PROOF: Because, if a limit exists, it is equal to the limsup. \square

Lemma 491 If $\Xi = \mathbb{R}^d$, then $0 \in \text{int} D$ is sufficient for exponential tightness.

PROOF: Exercise. \square

Unfortunately, this is *not* good enough to get exponential tightness in arbitrary vector spaces.

Definition 492 (Exposed Point) A point $x \in \Xi$ is exposed for $\bar{\Lambda}^*(\cdot)$ when there is a $t \in \Xi^*$ such that $\bar{\Lambda}^*(y) - \bar{\Lambda}^*(x) > t(y - x)$ for all $y \neq x$. t is the exposing hyper-plane for x .

In \mathbb{R}^1 , a point x is exposed if the curve $\bar{\Lambda}^*(y)$ lies strictly above the line of slope t through the point $(x, \bar{\Lambda}^*(x))$. Similarly, in \mathbb{R}^d , the $\bar{\Lambda}^*(y)$ surface must lie strictly above the hyper-plane passing through $(x, \bar{\Lambda}^*(x))$ with surface normal t . Since $\bar{\Lambda}^*(y)$ is convex, we could generally arrange this by making this the tangent hyper-plane, but we do not, yet, have any reason to think that the tangent is well-defined. — Obviously, if $\Lambda(t)$ exists, we can replace $\bar{\Lambda}^*(\cdot)$ by $\Lambda^*(\cdot)$ in the definition and the rest of this paragraph.

Definition 493 An exposed point $x \in \Xi$ with exposing hyper-plane t is nice, $x \in N$, if

$$\lim_{\epsilon \rightarrow 0} \epsilon \Lambda_\epsilon(t/\epsilon) \quad (33.5)$$

exists, and, for some $r > 1$,

$$\bar{\Lambda}(rt) < \infty \quad (33.6)$$

Note: Most books on large deviations do not give this property any particular name.

Lemma 494 (Lower Bound in Gärtner-Ellis Theorem) If the X_ϵ are exponentially tight, then, for any open set $O \subseteq \Xi$,

$$\liminf_{\epsilon \rightarrow 0} \epsilon \log \mathbb{P}(X_\epsilon \in O) \geq -\bar{\Lambda}^*(O \cap N) \quad (33.7)$$

PROOF: If we pick any nice, exposed point $x \in O$, we can repeat the proof of the lower bound from Cramér's Theorem (470). In fact, the point of Definition 493 is to ensure this. Taking the supremum over all such x gives the lemma. \square

Theorem 495 (Abstract Gärtner-Ellis Theorem) *If the X_ϵ are exponentially tight, and $\bar{\Lambda}^*(O \cap E) = \bar{\Lambda}^*(O)$ for all open sets $O \subseteq \Xi$, then X_ϵ obey an LDP with rate $1/\epsilon$ and good rate function $\bar{\Lambda}^*(x)$.*

PROOF: The large deviations upper bound is Lemma 488. The large deviations lower bound is implied by Lemma 494 and the additional hypothesis of the theorem. \square

Matters can be made a bit more concrete in Euclidean space (the original home of the theorem), using, however, one or two more bits of convex analysis.

Definition 496 (Relative Interior) *The relative interior of a non-empty and convex set A is*

$$\text{rint}A \equiv \{x \in A : \forall y \in A, \exists \delta > 0, x - \delta(y - x) \in A\} \quad (33.8)$$

Notice that $\text{int}A \subseteq \text{rint}A$, since the latter, in some sense, doesn't care about points outside of A .

Definition 497 (Essentially Smooth) *Λ is essentially smooth if (i) $\text{int}D \neq \emptyset$, (ii) Λ is differentiable on $\text{int}D$, and (iii) Λ is "steep", meaning that if D has a boundary, ∂D , then $\lim_{t \rightarrow \partial D} \|\nabla \Lambda(t)\| = \infty$.*

This definition was introduced to exploit the following theorem of convex analysis.

Proposition 498 *If Λ is essentially smooth, then Λ^* is continuous on $\text{rint}D^*$, and $\text{rint}D^* \subseteq N$, the set of exposed, nice points.*

PROOF: D^* is non-empty, because there is at least one point x_0 where $\Lambda^*(x_0) = 0$. Moreover, it is a convex set, because Λ^* is a convex function (Lemma 490), so it has a relative interior. Now appeal to Rockafellar (1970, Corollary 26.4.1). \square

Remark: You might want to try to prove that Λ^* is continuous on $\text{rint}D^*$.

Theorem 499 (Euclidean Gärtner-Ellis Theorem) *Suppose X_ϵ , taking values in \mathbb{R}^d , are exponentially tight, and that $\Lambda(t)$ exists and is essentially smooth. Then X_ϵ obey an LDP with rate $1/\epsilon$ and good rate function $\Lambda^*(x)$.*

PROOF: From the abstract Gärtner-Ellis Theorem 495, it is enough to show that $\Lambda^*(O \cap N) = \Lambda^*(O)$, for any open set O . That is, we want

$$\inf_{x \in O \cap N} \Lambda^*(x) = \inf_{x \in O} \Lambda^*(x) \quad (33.9)$$

Since it's automatic that

$$\inf_{x \in O \cap N} \Lambda^*(x) \geq \inf_{x \in O} \Lambda^*(x) \quad (33.10)$$

what we need to show is that

$$\inf_{x \in O \cap N} \Lambda^*(x) \leq \inf_{x \in O} \Lambda^*(x) \quad (33.11)$$

In view of Proposition 498, it's really just enough to show that $\Lambda^*(O \cap \text{rint} D^*) \leq \Lambda^*(O)$. This is trivial when the intersection $O \cap D^*$ is empty, so assume it isn't, and pick any x in that intersection. Because O is open, and because of the definition of the relative interior, we can pick any point $y \in \text{rint} D^*$, and, for sufficiently small δ , $\delta y + (1 - \delta)x \in O \cap \text{rint} D^*$. Since Λ^* is convex, $\Lambda^*(\delta y + (1 - \delta)x) \leq \delta \Lambda^*(y) + (1 - \delta)\Lambda^*(x)$. Taking the limit as $\delta \rightarrow 0$,

$$\Lambda^*(O \cap \text{rint} D^*) \leq \Lambda^*(x) \quad (33.12)$$

and the claim follows by taking the infimum over x .

33.2 Exercises

Exercise 33.1 Show that Cramér's Theorem (470), is a special case of the Euclidean Gärtner-Ellis Theorem (499).

Exercise 33.2 Let Z_1, Z_2, \dots be IID random variables in a discrete space, and let X_1, X_2, \dots be the empirical distributions they generate. Use the Gärtner-Ellis Theorem to re-prove Sanov's Theorem (475). Can you extend this to the case where the Z_i take values in an arbitrary Polish space?

Exercise 33.3 Let Z_1, Z_2, \dots be values from a stationary ergodic Markov chain (i.e. the state space is discrete). Repeat the previous exercise for the pair measure. Again, can the result be extended to arbitrary Polish state spaces?

Exercise 33.4 Let Z_i be real-valued mean-zero stationary Gaussian variables, with $\sum_{i=-\infty}^{\infty} |\text{cov}(Z_0, Z_i)| < \infty$. Let $X_t = t^{-1} \sum_{i=1}^t Z_i$. Show that these time averages obey an LDP with rate t and rate function $x^2/2\Gamma$, where $\Gamma = \sum_{i=-\infty}^{\infty} \text{cov}(Z_0, Z_i)$. (Cf. the mean-square ergodic theorem 293 of chapter 22.) If Z_i are not Gaussian but are weakly stationary, find an additional hypothesis such that the time averages still obey an LDP.

Exercise 33.5 (Too-Strong Dependence) Let $X_i = X$, a random variable in \mathbb{R} , for all i . Show that the Gärtner-Ellis Theorem fails, unless X is degenerate.

Chapter 34

Large Deviations for Stochastic Differential Equations

This last chapter revisits large deviations for stochastic differential equations in the small-noise limit, first raised in Chapter 21.

Section 34.1 establishes the LDP for the Wiener process (Schilder's Theorem).

Section 34.2 proves the LDP for stochastic differential equations where the driving noise is independent of the state of the process.

Section 34.3 states the corresponding result for SDEs when the noise is state-dependent, and gestures in the direction of the proof.

In Chapter 21, we looked at how the diffusions X_ϵ which solve the SDE

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon dW, \quad X_\epsilon(0) = x_0 \quad (34.1)$$

converge on the trajectory $x_0(t)$ solving the ODE

$$\frac{dx}{dt} = a(x(t)), \quad x(0) = x_0 \quad (34.2)$$

in the “small noise” limit, $\epsilon \rightarrow 0$. Specifically, Theorem 270 gave a (fairly crude) upper bound on the probability of deviations:

$$\lim_{\epsilon \rightarrow 0} \epsilon^2 \log \mathbb{P} \left(\sup_{0 \leq t \leq T} \Delta_\epsilon(t) > \delta \right) \leq -\delta^2 e^{-2K_a T} \quad (34.3)$$

where K_a depends on the Lipschitz coefficient of the drift function a . The theory of large deviations for stochastic differential equations, known as Freidlin-Wentzell theory for its original developers, shows that, using the metric implicit

in the left-hand side of Eq. 34.3, the family of processes X_ϵ obey a large deviations principle with rate ϵ^{-2} , and a good rate function.

(The full Freidlin-Wentzell theory actually goes somewhat further than just SDEs, to consider small-noise perturbations of dynamical systems of many sorts, perturbations by Markov processes (rather than just white noise), etc. Time does not allow us to consider the full theory (Freidlin and Wentzell, 1998), or its many applications to nonparametric estimation (Ibragimov and Has'minskii, 1979/1981), systems analysis and signal processing (Kushner, 1984), statistical mechanics (Olivieri and Vares, 2005), etc.)

As in Chapter 31, the strategy is to first prove a large deviations principle for a comparatively simple case, and then transfer it to more subtle processes which can be represented as appropriate functionals of the basic case. Here, the basic case is the Wiener process $W(t)$, with t restricted to the unit interval $[0, 1]$.

34.1 Large Deviations of the Wiener Process

We start with a standard d -dimensional Wiener process W , and consider its dilation by a factor ϵ , $X_\epsilon(t) = \epsilon W(t)$. There are a number of ways of establishing that X_ϵ obeys a large deviation principle as $\epsilon \rightarrow 0$. One approach (see Dembo and Zeitouni (1998, ch. 5) starts with establishing an LDP for continuous-time random walks, ultimately based on the Gärtner-Ellis Theorem, and then showing that the convergence of such processes to the Wiener process (the Functional Central Limit Theorem, Theorem 215 of Chapter 16) is sufficiently fast that the LDP carries over. However, this approach involves a number of surprisingly tricky topological issues, so I will avoid it, in favor of a more probabilistic path, marked out by Freidlin and Wentzell (Freidlin and Wentzell, 1998, sec. 3.2).

Until further notice, $\|w\|_\infty$ will denote the supremum norm in the space of continuous curves over the unit interval, $\mathbf{C}([0, 1], \mathbb{R}^d)$.

Definition 500 (Cameron-Martin Spaces) *The Cameron-Martin space H_T consists of all continuous sample paths $x \in \mathbf{C}([0, T], \mathbb{R}^d)$ where $x(0) = 0$, x is absolutely continuous, and its Radon-Nikodym derivative \dot{x} is square-integrable.*

Lemma 501 (Cameron-Martin Spaces are Hilbert) *Cameron-Martin spaces are Hilbert spaces, with norm $\|x\|_{CM} = \int_0^T |\dot{x}(t)|^2 dt$.*

PROOF: An exercise (34.1) in verifying that the axioms of a Hilbert space are satisfied. \square

Definition 502 (Effective Wiener Action) *The effective Wiener action of an continuous function $x \in \mathbf{C}([0, t], \mathbb{R}^d)$ is*

$$J_T(x) \equiv \frac{1}{2} \|x\|_{CM}^2 \quad (34.4)$$

if $x \in H_T$, and ∞ otherwise. In particular,

$$J_1(x) \equiv \frac{1}{2} \int_0^1 |\dot{x}(t)|^2 dt \quad (34.5)$$

For every $j > 0$, let $L_T(j) = \{x : J_T(x) \leq j\}$.

Proposition 503 (Girsanov Formula for Deterministic Drift) *Fix a function $f \in H_1$, and let $Y_\epsilon = X_\epsilon - f$. Then $\mathcal{L}(Y_\epsilon) = \nu_\epsilon$ is absolutely continuous with respect to $\mathcal{L}(X_\epsilon) = \mu_\epsilon$, and the Radon-Nikodym derivative is*

$$\frac{d\nu_\epsilon}{d\mu_\epsilon}(\epsilon w) = \exp \left\{ -\frac{1}{\epsilon} \int_0^1 \dot{w}(t) \cdot dW - \frac{1}{2\epsilon^2} \int_0^1 |\dot{w}(t)|^2 dt \right\} \quad (34.6)$$

PROOF: This is a special case of Girsanov's Theorem. See Corollary 18.25 on p. 365 of Kallenberg, or, more transparently perhaps, the relevant parts of Liptser and Shiryaev (2001, vol. I). \square

Lemma 504 (Exponential Bound on the Probability of Tubes Around Given Trajectories) *For any $\delta, \gamma, K > 0$, there exists an $\epsilon_0 > 0$ such that, if $\epsilon < \epsilon_0$,*

$$\mathbb{P}(\|X_\epsilon - x\|_\infty \leq \delta) \geq e^{-\frac{J_1(x) + \gamma}{\epsilon^2}} \quad (34.7)$$

provided $x(0) = 0$ and $J_1(x) < K$.

PROOF: Using Proposition 503,

$$\mathbb{P}(\|X_\epsilon - x\|_\infty \leq \delta) = \mathbb{P}(\|Y_\epsilon - 0\|_\infty \leq \delta) \quad (34.8)$$

$$= \int_{\|\epsilon w\|_\infty < \delta} \frac{d\nu_\epsilon}{d\mu_\epsilon}(\epsilon w) d\mu_\epsilon(\epsilon w) \quad (34.9)$$

$$= e^{-\frac{J_1(x)}{\epsilon^2}} \int_{\|\epsilon w\|_\infty < \delta} e^{-\frac{1}{\epsilon} \int_0^1 \dot{x} \cdot dW} d\mu_\epsilon(w) \quad (34.10)$$

From Lemma 268 in Chapter 21, we can see that $\mathbb{P}(\|\epsilon W\|_\infty < \delta) \rightarrow 1$ as $\epsilon \rightarrow 0$. So, if ϵ is sufficiently small, $\mathbb{P}(\|\epsilon W\|_\infty < \delta) \geq 3/4$. Now, applying Chebyshev's inequality to the integrand,

$$\mathbb{P} \left(-\frac{1}{\epsilon} \int_0^1 \dot{x} \cdot dW \leq -\frac{2\sqrt{2}}{\epsilon} \sqrt{J_1(x)} \right) \quad (34.11)$$

$$\leq \mathbb{P} \left(\left| \frac{1}{\epsilon} \int_0^1 \dot{x} \cdot dW \right| \leq \frac{2\sqrt{2}}{\epsilon} \sqrt{J_1(x)} \right) \quad (34.12)$$

$$\leq \frac{\epsilon^2 \mathbf{E} \left[\left(\int_0^1 \dot{x} \cdot dW \right)^2 \right]}{8\epsilon^2 J_1(x)} \quad (34.13)$$

$$= \frac{\int_0^1 |\dot{x}|^2 dt}{8J_1(x)} = \frac{1}{4} \quad (34.14)$$

using the Itô isometry (Corollary 239). Thus

$$\mathbb{P} \left(e^{-\frac{1}{\epsilon} \int_0^1 \dot{x} \cdot dW} \geq e^{-\frac{2\sqrt{2}}{\epsilon} \sqrt{J_1(x)}} \right) \geq \frac{3}{4} \quad (34.15)$$

$$\int_{\|\epsilon w\|_\infty < \delta} e^{-\frac{1}{\epsilon} \int_0^1 \dot{x} \cdot dW} d\mu_\epsilon(w) > \frac{1}{2} e^{-\frac{2\sqrt{2}}{\epsilon} \sqrt{J_1(x)}} \quad (34.16)$$

$$\mathbb{P}(\|X_\epsilon - x\|_\infty \leq \delta) > \frac{1}{2} e^{-\frac{J_1(x)}{\epsilon^2} - \frac{2\sqrt{2}}{\epsilon} \sqrt{J_1(x)}} \quad (34.17)$$

where the second term in the exponent can be made less than any desired γ by taking ϵ small enough. \square

Lemma 505 (Trajectories are Rarely Far from Action Minima) *For every $j > 0$, $\delta > 0$, let $U(j, \delta)$ be the open δ neighborhood of $L_1(j)$, i.e., all the trajectories coming within δ of a trajectory whose action is less than or equal to j . Then for any $\gamma > 0$, there is an $\epsilon_0 > 0$ such that, if $\epsilon < \epsilon_0$ and*

$$\mathbb{P}(X_\epsilon \notin U(j, \delta)) \leq e^{-\frac{\gamma}{\epsilon^2}} \quad (34.18)$$

PROOF: Basically, approximating the Wiener process by a continuous piecewise-linear function, and showing that the approximation is sufficiently fine-grained. Chose a natural number n , and let $Y_{n,\epsilon}(t)$ be the piecewise linear random function which coincides with X_ϵ at times $0, 1/n, 2/n, \dots, 1$, i.e.,

$$Y_{n,\epsilon}(t) = X_\epsilon([tn]/n) + \left(t - \frac{[tn]}{n}\right) X_\epsilon([tn+1]/n) \quad (34.19)$$

We will see that, for large enough n , this is exponentially close to X_ϵ . First, though, let's bound the probability in Eq. 34.18.

$$\begin{aligned} \mathbb{P}(X_\epsilon \notin U(j, \delta)) &= \mathbb{P}(X_\epsilon \notin U(j, \delta), \|X_\epsilon - Y_{n,\epsilon}\|_\infty < \delta) \\ &\quad + \mathbb{P}(X_\epsilon \notin U(j, \delta), \|X_\epsilon - Y_{n,\epsilon}\|_\infty \geq \delta) \end{aligned} \quad (34.20)$$

$$\leq \mathbb{P}(X_\epsilon \notin U(j, \delta), \|X_\epsilon - Y_{n,\epsilon}\|_\infty < \delta) + \mathbb{P}(\|X_\epsilon - Y_{n,\epsilon}\|_\infty \geq \delta) \quad (34.21)$$

$$\leq \mathbb{P}(J_1(Y_{n,\epsilon}) > j) + \mathbb{P}(\|X_\epsilon - Y_{n,\epsilon}\|_\infty \geq \delta) \quad (34.22)$$

$J_1(Y_{n,\epsilon})$ can be gotten at from the increments of the Wiener process:

$$J_1(Y_{n,\epsilon}) = n \frac{\epsilon^2}{2} \sum_{i=1}^n |W(i/n) - W((i-1)/n)|^2 \quad (34.23)$$

$$= \frac{\epsilon^2}{2} \sum_{i=1}^{dn} \xi_i \quad (34.24)$$

where the ξ_i have the χ^2 distribution with one degree of freedom. Using our results on such distributions and their sums in Ch. 21, it is not hard to show that, for sufficiently small ϵ ,

$$\mathbb{P}(J_1(Y_{n,\epsilon}) > j) \leq \frac{1}{2}e^{-\frac{j-\gamma}{\epsilon^2}} \quad (34.25)$$

To estimate the probability that the distance between X_ϵ and $Y_{n,\epsilon}$ reaches or exceeds δ , start with the independent-increments property of X_ϵ , and the fact that the two processes coincide when $t = i/n$.

$$\begin{aligned} \mathbb{P}(\|X_\epsilon - Y_{n,\epsilon}\|_\infty \geq \delta) \\ \leq \sum_{i=1}^n \mathbb{P}\left(\max_{(i-1)/n \leq t \leq i/n} |X_\epsilon(t) - Y_{n,\epsilon}(t)| \geq \delta\right) \end{aligned} \quad (34.26)$$

$$= n\mathbb{P}\left(\max_{0 \leq t \leq 1/n} |X_\epsilon(t) - Y_{n,\epsilon}(t)| \geq \delta\right) \quad (34.27)$$

$$= n\mathbb{P}\left(\max_{0 \leq t \leq 1/n} |\epsilon W(t) - n\epsilon W(1/n)| \geq \delta\right) \quad (34.28)$$

$$\leq n\mathbb{P}\left(\max_{0 \leq t \leq 1/n} |\epsilon W(t)| \geq \frac{\delta}{2}\right) \quad (34.29)$$

$$\leq 4dn\mathbb{P}\left(W_1(1/n) \geq \frac{\delta}{2d\epsilon}\right) \quad (34.30)$$

$$\leq 4dn \frac{2d\epsilon}{\delta\sqrt{2\pi n}} e^{-\frac{n\delta^2}{8d^2\epsilon^2}} \quad (34.31)$$

again freely using our calculations from Ch. 21. If $n > 4d^2j/\delta^2$, then $\mathbb{P}(\|X_\epsilon - Y_{n,\epsilon}\|_\infty) \leq \frac{1}{2}e^{-\frac{j-\gamma}{\epsilon^2}}$, and we have overall

$$\mathbb{P}(X_\epsilon \notin U(j, \delta)) \leq e^{-\frac{j-\gamma}{\epsilon^2}} \quad (34.32)$$

as required. \square

Proposition 506 (Compact Level Sets of the Wiener Action) *The Cameron-Martin norm has compact level sets.*

PROOF: See Kallenberg, Lemma 27.7, p. 543. \square

Theorem 507 (Schilder's Theorem on Large Deviations of the Wiener Process on the Unit Interval) *If W is a d -dimensional Wiener process on the unit interval, then $X_\epsilon = \epsilon W$ obeys an LDP on $\mathbf{C}([0, 1], \mathbb{R}^d)$, with rate ϵ^{-2} and good rate function $J_1(x)$, the effective Wiener action over $[0, 1]$.*

PROOF: It is easy to show that Lemma 504 implies the large deviation lower bound for open sets. (Exercise 34.2.) The tricky part is the upper bound. Pick any closed set C and any $\gamma > 0$. Let $s = J_1(C) - \gamma$. By Lemma 506, the set

$K = L_1(s) = \{x : J_1(x) \leq s\}$ is compact. By construction, $C \cap K = \emptyset$. So $\delta = \inf_{x \in C, y \in K} \|x - y\|_\infty > 0$. Let U be the closed δ -neighborhood of K . Then use Lemma 505

$$\mathbb{P}(X_\epsilon \in C) \leq \mathbb{P}(X_\epsilon \notin U) \quad (34.33)$$

$$\leq e^{-\frac{s-\gamma}{\epsilon^2}} \quad (34.34)$$

$$\leq e^{-\frac{J_1(C)-2\gamma}{\epsilon^2}} \quad (34.35)$$

$$\log \mathbb{P}(X_\epsilon \in C) \leq -\frac{J_1(C) - 2\gamma}{\epsilon^2} \quad (34.36)$$

$$\epsilon^2 \log \mathbb{P}(X_\epsilon \in C) \leq -J_1(C) - 2\gamma \quad (34.37)$$

$$\limsup_{\epsilon \rightarrow 0} \epsilon^2 \log \mathbb{P}(X_\epsilon \in C) \leq -J_1(C) - 2\gamma \quad (34.38)$$

Since γ was arbitrary, this completes the proof. \square

Remark: The trick used here, about establishing results like Lemmas 505 and 504, and then using compact level sets to prove large deviations, works more generally. See Theorem 3.3 in Freidlin and Wentzell (1998, sec. 3.3).

Corollary 508 (Extension of Schilder's Theorem to $[0, T]$) *Schilder's theorem remains true for Wiener processes on $[0, T]$, for all $T > 0$, with rate function J_T , the effective Wiener action on $[0, T]$.*

PROOF: If W is a Wiener process on $[0, 1]$, then, for every T , $S(W) = \sqrt{T}W(t/T)$ is a Wiener process on $[0, T]$. (Show this!) Since the mapping S is continuous from $\mathbf{C}([0, 1], \mathbb{R}^d)$ to $\mathbf{C}([0, T], \mathbb{R}^d)$, by the Contraction Principle (Theorem 451) the family $\epsilon S(W)$ obey an LDP with rate ϵ^{-2} and good rate function $J_T(x) = J_1(S^{-1}(x))$. (Notice that S is invertible, so $S^{-1}(x)$ is a function, not a set of functions.) Since $x \in H_1$ iff $y = S(x) \in H_T$, it's easy to check that for such, $\dot{y}(t) = T^{-1/2}\dot{x}(t/T)$, meaning that

$$\|y\|_{CM} = \int_0^T |\dot{y}(t)|^2 dt = \int_0^T |\dot{x}(t/T)|^2 \frac{dt}{T} = \|x\|_{CM} \quad (34.39)$$

which completes the proof. \square

Corollary 509 (Schilder's Theorem on \mathbb{R}^+) *Schilder's theorem remains true for Wiener processes on \mathbb{R}^+ , with good rate function J_∞ given by the effective Wiener action on \mathbb{R}^+ ,*

$$J_\infty(x) \equiv \frac{1}{2} \int_0^\infty |\dot{x}(t)|^2 dt \quad (34.40)$$

if $x \in H_\infty$, $J_\infty(x) = \infty$ otherwise.

PROOF: For each natural number n , let $\pi_n x$ be the restriction of x to the interval $[0, n]$. By Corollary 508, each of them obeys an LDP with rate function $\frac{1}{2} \int_0^n |\dot{x}(t)|^2 dt$. Now apply the projective limit theorem (461) to get that $J_\infty(x) = \sup_n J_n(x)$, which is clearly Eq. 34.40, as the integrand is non-negative. \square

34.2 Large Deviations for SDEs with State-Independent Noise

Having established an LDP for the Wiener process, it is fairly straightforward to get an LDP for stochastic differential equations where the driving noise is *independent* of the state of the diffusion process.

Definition 510 (SDE with Small State-Independent Noise) *An SDE with small state-independent noise is a stochastic differential equation of the form*

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon dW \quad (34.41)$$

$$X_\epsilon(0) = 0 \quad (34.42)$$

where $a : \mathbb{R}^d \mapsto \mathbb{R}^d$ is uniformly Lipschitz continuous.

Notice that any non-random initial condition x_0 can be handled by a simple change of coordinates.

Definition 511 (Effective Action under State-Independent Noise) *The effective action of a trajectory $x \in H_\infty$ is*

$$J(x) \equiv \frac{1}{2} \int_0^t |\dot{x}(t) - a(x(t))|^2 dt \quad (34.43)$$

and $= \infty$ if $x \in \mathbf{C} \setminus H_\infty$.

Lemma 512 (Continuous Mapping from Wiener Process to SDE Solutions) *The map $F : \mathbf{C}(\mathbb{R}^+, \mathbb{R}^d) \mapsto \mathbf{C}(\mathbb{R}^+, \mathbb{R}^d)$ given by*

$$x(t) = w(t) + \int_0^t a(x(s))ds \quad (34.44)$$

when $x = F(w)$ is continuous.

PROOF: This goes rather in the same manner as the proof of existence and uniqueness for SDEs (Theorem 260). For any $w_1, w_2 \in \mathbf{C}(\mathbb{R}^+, \mathbb{R}^d)$, set $x_1 = F(w_1)$, $x_2 = F(w_2)$. From the Lipschitz property of a ,

$$|x_1(t) - x_2(t)| \leq \|w_1 - w_2\| + K_a \int_0^t |x_1(s) - x_2(s)| ds \quad (34.45)$$

(writing $|y(t)|$ for the norm of Euclidean vectors y , and $\|x\|$ for the supremum norm of continuous curves). By Gronwall's Inequality (Lemma 258), then,

$$\|x_1 - x_2\| \leq \|w_1 - w_2\| e^{K_a T} \quad (34.46)$$

on every interval $[0, T]$. So we can make sure that $\|x_1 - x_2\|$ is less than any desired amount by making sure that $\|w_1 - w_2\|$ is sufficiently small, and so F is continuous. \square

Lemma 513 (Mapping Cameron-Martin Spaces Into Themselves) *If $w \in H_\infty$, then $x = F(w)$ is in H_∞ .*

PROOF: Exercise 34.3. \square

Theorem 514 (Freidlin-Wentzell Theorem with State-Independent Noise) *The Itô processes X_ϵ of Definition 510 obey the large deviations principle with rate ϵ^{-2} and good rate function given by the effective action $J(x)$.*

PROOF: For every ϵ , $X_\epsilon = F(\epsilon W)$. Corollary 509 tells us that ϵW obeys the large deviation principle with rate ϵ^{-2} and good rate function J_∞ . Since (Lemma 512) F is continuous, by the Contraction Principle (Theorem 451) X_ϵ also obeys the LDP, with rate given by $J_\infty(F^{-1}(x))$. If $F^{-1}(x) \cap H_\infty = \emptyset$, this is ∞ . On the other hand, if $F^{-1}(x)$ does contain curves in H_∞ , then $J_\infty(F^{-1}(x)) = J_\infty(F^{-1}(x) \cap H_\infty)$. By Lemma 513, this implies that $x \in H_\infty$, too. For any curve $w \in F^{-1}(x) \cap H_\infty$, $\dot{x} = \dot{w} + a(x)$, or $\dot{w} = \dot{x} - a(x)$. $J_\infty(w) = \int_0^\infty |\dot{x} - a(x)|^2 dt$ is however the effective action of the trajectory x (Definition 511). \square

34.3 Large Deviations for State-Dependent Noise

If the diffusion term in the SDE does depend on the state of the process, one obtains a very similar LDP to the results in the previous section. However, the approach must be modified: the mapping from W to X_ϵ , while still measurable, is no longer necessarily continuous, so we can't use the contraction principle as before.

Definition 515 (SDE with Small State-Dependent Noise) *An SDE with small state-dependent noise is a stochastic differential equation of the form*

$$dX_\epsilon = a(X_\epsilon)dt + \epsilon b(X_\epsilon)dW \quad (34.47)$$

$$X_\epsilon(0) = 0 \quad (34.48)$$

where a and b are uniformly Lipschitz continuous, and b is non-singular.

Definition 516 (Effective Action under State-Dependent Noise) *The effective action of a trajectory $x \in H_\infty$ is given by*

$$J(x) \equiv \int_0^\infty L(x(t), \dot{x}(t)) dt \quad (34.49)$$

where

$$L(q, p) = \frac{1}{2} (p_i - a_i(q)) B_{ij}^{-1}(q) (p_j - a_j(q)) \quad (34.50)$$

and

$$B(q) = b(q)b^T(q) \quad (34.51)$$

with $J(x) = \infty$ if $x \in \mathbf{C} \setminus H_\infty$.

Theorem 517 (Freidlin-Wentzell Theorem for State-Dependent Noise)

The processes X_ϵ obey a large deviations principle with rate ϵ^{-2} and good rate function equal to the effective action.

PROOF: Considerably more complicated. (See, e.g., Dembo and Zeitouni (1998, sec. 5.6, pp. 213–220).) The essence, however, is to consider an approximating Itô process X_n , where $a(X_t)$ and $b(X_t)$ are replaced in Eq. 34.47 by $a(X_n([tn]/n))$ and $b(X_n([tn]/n))$. Here the mapping from W to X_n is continuous, so it's not too hard to show that the latter obey an LDP with a reasonable rate function, and also that they're exponentially equivalent (in n) to X_ϵ . \square

34.4 Exercises

Exercise 34.1 (Cameron-Martin Spaces are Hilbert Spaces) *Prove Lemma 501.*

Exercise 34.2 (Lower Bound in Schilder's Theorem) *Prove that Lemma 504 implies the large deviations lower bound for open sets.*

Exercise 34.3 (Mapping Cameron-Martin Spaces Into Themselves) *Prove Lemma 513. Hint: Use Gronwall's Inequality (Lemma 258) again to show that F maps H_T into H_T , and then show that $H_\infty = \bigcap_{n=1}^\infty H_n$.*

Bibliography

- Abramowitz, Milton and Irene A. Stegun (eds.) (1964). *Handbook of Mathematical Functions*. Washington, D.C.: National Bureau of Standards. URL <http://www.math.sfu.ca/~cbm/aands/>.
- Algoet, Paul (1992). “Universal Schemes for Prediction, Gambling and Portfolio Selection.” *The Annals of Probability*, **20**: 901–941. URL <http://links.jstor.org/sici?sici=0091-1798%28199204%2920%3A2%3C901%3AUSFPGA%3E2.0.CO%3B2-J>. Correction, *The Annals of Probability*, **23** (1995): 474–478, JSTOR.
- Algoet, Paul H. and Thomas M. Cover (1988). “A Sandwich Proof of the Shannon-McMillan-Breiman Theorem.” *The Annals of Probability*, **16**: 899–909. URL <http://links.jstor.org/sici?sici=0091-1798%28198804%2916%3A2%3C899%3AASPOTS%3E2.0.CO%3B2-M>.
- Amari, Shun-ichi and Hiroshi Nagaoka (1993/2000). *Methods of Information Geometry*. Providence, Rhode Island: American Mathematical Society. Translated by Daishi Harada. As *Joho Kika no Hoho*, Tokyo: Iwanami Shoten Publishers.
- Arnol’d, V. I. (1973). *Ordinary Differential Equations*. Cambridge, Massachusetts: MIT Press. Trans. Richard A. Silverman from *Obyknovennye differentsial’nye Uravneniya*.
- (1978). *Mathematical Methods of Classical Mechanics*. Berlin: Springer-Verlag. First published as *Matematicheskie metody klassicheskoi mekhaniki*, Moscow: Nauka, 1974.
- Arnol’d, V. I. and A. Avez (1968). *Ergodic Problems of Classical Mechanics*. Mathematical Physics Monograph Series. New York: W. A. Benjamin.
- Badino, Massimiliano (2004). “An Application of Information Theory to the Problem of the Scientific Experiment.” *Synthese*, **140**: 355–389. URL <http://philsci-archive.pitt.edu/archive/00001830/>.
- Baldi, Pierre and Søren Brunak (2001). *Bioinformatics: The Machine Learning Approach*. Cambridge, Massachusetts: MIT Press, 2nd edn.

- Banks, J., J. Brooks, G. Cairns, G. Davis and P. Stacy (1992). "On Devaney's Definition of Chaos." *American Mathematical Monthly*, **99**: 332–334.
- Bartlett, M. S. (1955). *An Introduction to Stochastic Processes, with Special Reference to Methods and Applications*. Cambridge, England: Cambridge University Press.
- Basharin, Gely P., Amy N. Langville and Valeriy A. Naumov (2004). "The Life and Work of A. A. Markov." *Linear Algebra and its Applications*, **386**: 3–26. URL <http://decision.csl.uiuc.edu/~meyn/pages/Markov-Work-and-life.pdf>.
- Biggers, Earl Derr (1928). *Behind That Curtain*. New York: Grosset and Dunlap.
- Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.
- Blackwell, David and M. A. Girshick (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.
- Bosq, Denis (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Berlin: Springer-Verlag, 2nd edn.
- Caires, S. and J. A. Ferreira (2005). "On the Non-parametric Prediction of Conditionally Stationary Sequences." *Statistical Inference for Stochastic Processes*, **8**: 151–184. Correction, vol. 9 (2006), pp. 109–110.
- Charniak, Eugene (1993). *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press.
- Chomsky, Noam (1957). *Syntactic Structures*. The Hague: Mouton.
- Courant, Richard and David Hilbert (1953). *Methods of Mathematical Physics*. New York: Wiley.
- Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.
- Cramér, Harald (1945). *Mathematical Methods of Statistics*. Uppsala: Almqvist and Wiksells.
- Dembo, Amir and Ofer Zeitouni (1998). *Large Deviations Techniques and Applications*. New York: Springer Verlag, 2nd edn.
- Devaney, Robert L. (1992). *A First Course in Chaotic Dynamical Systems: Theory and Experiment*. Reading, Mass.: Addison-Wesley.
- Doob, Joseph L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. New York: Wiley.

- Doukhan, Paul (1995). *Mixing: Properties and Examples*. New York: Springer-Verlag.
- Durbin, James and Siem Jam Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
- Durrett, Richard (1991). *Probability: Theory and Examples*. Belmont, California: Duxbury.
- Dynkin, E. B. (1978). “Sufficient statistics and extreme points.” *Annals of Probability*, **6**: 705–730. URL <http://links.jstor.org/sici?sici=0091-1798%28197810%296%3A5%3C705%3ASSAEP%3E2.0.CO%3B2-D>.
- Eckmann, Jean-Pierre and David Ruelle (1985). “Ergodic Theory of Chaos and Strange Attractors.” *Reviews of Modern Physics*, **57**: 617–656.
- Eichinger, L., J. A. Pachebat, G. Glockner *et al.* (2005). “The genome of the social amoeba *Dictyostelium discoideum*.” *Nature*, **435**: 43–57.
- Ellis, Richard S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*. Berlin: Springer-Verlag.
- (1988). “Large Deviations for the Empirical Measure of a Markov Chain with an Application to the Multivariate Empirical Measure.” *The Annals of Probability*, **16**: 1496–1508. URL <http://links.jstor.org/sici?sici=0091-1798%28198810%2916%3A4%3C1496%3ALDFTEM%3E2.0.CO%3B2-S>.
- Embrechts, Paul and Makoto Maejima (2002). *Selfsimilar processes*. Princeton, New Jersey: Princeton University Press.
- Ethier, Stewart N. and Thomas G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. New York: Wiley.
- Eyink, Gregory L. (1996). “Action principle in nonequilibrium statistical dynamics.” *Physical Review E*, **54**: 3419–3435.
- Fisher, Ronald Aylmer (1958). *The Genetical Theory of Natural Selection*. New York: Dover, 2nd edn. First edition published Oxford: Clarendon Press, 1930.
- Forster, Dieter (1975). *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*. Reading, Massachusetts: Benjamin Cummings.
- Freidlin, M. I. and A. D. Wentzell (1998). *Random Perturbations of Dynamical Systems*. Berlin: Springer-Verlag, 2nd edn. First edition first published as *Fluktuatsii v dinamicheskikh sistemakh pod deistviem malykh sluchainykh vozmushchenii*, Moscow: Nauka, 1979.
- Frisch, Uriel (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge, England: Cambridge University Press.

- Fristedt, Bert and Lawrence Gray (1997). *A Modern Approach to Probability Theory*. Probability Theory and Its Applications. Boston: Birkhäuser.
- Gikhman, I. I. and A. V. Skorokhod (1965/1969). *Introduction to the Theory of Random Processes*. Philadelphia: W. B. Saunders. Trans. Richard Silverman from *Vvedenie v teoriu sluchainikh protsessov*, Moscow: Nauka; reprinted Mineola, New York: Dover, 1996.
- Gillespie, John H. (1998). *Population Genetics: A Concise Guide*. Baltimore: Johns Hopkins University Press.
- Gnedenko, B. V. and A. N. Kolmogorov (1954). *Limit Distributions for Sums of Independent Random Variables*. Cambridge, Massachusetts: Addison-Wesley. Translated from the Russian and annotated by K. L. Chung, with an Appendix by J. L. Doob.
- Gray, Robert M. (1988). *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag. URL <http://ee-www.stanford.edu/~gray/arp.html>.
- (1990). *Entropy and Information Theory*. New York: Springer-Verlag. URL <http://www-ee.stanford.edu/~gray/it.html>.
- Haldane, J. B. S. (1932). *The Causes of Evolution*. Longmans, Green and Co. Reprinted Princeton: Princeton University Press, 1990.
- Halmos, Paul R. (1950). *Measure Theory*. New York: Van Nostrand.
- Hille, Einer (1948). *Functional Analysis and Semi-Groups*. New York: American Mathematical Society.
- Howard, Ronald A. (1971). *Markov Models*, vol. 1 of *Dynamic Probabilistic Systems*. New York: Wiley.
- Ibragimov, I. A. and R. Z. Has'minskii (1979/1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag. Translated by Samuel Kotz from *Asimptoticheskaya teoriya otsenivaniya*, Moscow: Nauka.
- Jakubowski, A. and Z. S. Szewczak (1990). "A normal convergence criterion for strongly mixing stationary sequences." In *Limit Theorems in Probability and Statistics* (I. Berkes and Endre Csáki and Pal Révész, eds.), vol. 57 of *Colloquia mathematica Societatis Janos Bolyai*, pp. 281–292. New York: North-Holland. Proceedings of the Third Hungarian Colloquium on Limit Theorems in Probability and Statistics, held in Pécs, Hungary, July 3–7, 1989 and sponsored by the Bolyai János Mathematical Society.
- Kac, Mark (1947). "On the Notion of Recurrence in Discrete Stochastic Processes." *Bulletin of the American Mathematical Society*, **53**: 1002–1010. Reprinted in Kac (1979), pp. 231–239.

- (1979). *Probability, Number Theory, and Statistical Physics: Selected Papers*. Cambridge, Massachusetts: MIT Press. Edited by K. Baclawski and M. D. Donsker.
- Kallenberg, Olav (2002). *Foundations of Modern Probability*. New York: Springer-Verlag, 2nd edn.
- Kass, Robert E. and Paul W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. New York: Wiley.
- Katznelson, I. and B. Weiss (1982). “A simple proof of some ergodic theorems.” *Israel Journal of Mathematics*, **42**: 291–296.
- Keizer, Joel (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag.
- Khinchin, Aleksandr Iakovlevich (1949). *Mathematical Foundations of Statistical Mechanics*. New York: Dover Publications. Translated from the Russian by G. Gamow.
- Knight, Frank B. (1975). “A Predictive View of Continuous Time Processes.” *Annals of Probability*, **3**: 573–596. URL <http://links.jstor.org/sici?sici=0091-1798%28197508%293%3A4%3C573%3AAPVOCT%3E2.0.CO%3B2-N>.
- (1992). *Foundations of the Prediction Process*. Oxford: Clarendon Press.
- Kolmogorov, A. N. and S. V. Fomin (1975). *Introductory Real Analysis*. New York: Dover. Translated and edited by Richard A. Silverman.
- Kontoyiannis, I., P. H. Algoet, Yu. M. Suhov and A. J. Wyner (1998). “Non-parametric entropy estimation for stationary processes and random fields, with applications to English text.” *IEEE Transactions on Information Theory*, **44**: 1319–1327. URL <http://www.dam.brown.edu/people/yiannis/PAPERS/suhov2.pdf>.
- Küchler, Uwe and Michael Sørensen (1997). *Exponential Families of Stochastic Processes*. Berlin: Springer-Verlag.
- Kulhavý, Rudolf (1996). *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. Berlin: Springer-Verlag.
- Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn.
- Kurtz, Thomas G. (1970). “Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes.” *Journal of Applied Probability*, **7**: 49–58. URL <http://links.jstor.org/sici?sici=0021-9002%28197004%297%3A1%3C49%3AS00DEA%3E2.0.CO%3B2-P>.

- (1971). “Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes.” *Journal of Applied Probability*, **8**: 344–356. URL <http://links.jstor.org/sici?sici=0021-9002%28197106%298%3A2%3C344%3ALTFS0J%3E2.0.CO%3B2-M>.
- (1975). “Semigroups of Conditioned Shifts and Approximation of Markov Processes.” *The Annals of Probability*, **3**: 618–642. URL <http://links.jstor.org/sici?sici=0091-1798%28197508%293%3A4%3C618%3ASOCSAA%3E2.0.CO%3B2-C>.
- Kushner, Harold J. (1984). *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*. Cambridge, Massachusetts: MIT Press.
- Lamperti, John (1962). “Semi-Stable Stochastic Processes.” *Transactions of the American Mathematical Society*, **104**: 62–78. URL <http://links.jstor.org/sici?sici=0002-9947%28196207%29104%3A1%3C62%3ASSP%3E2.0.CO%3B2-I>.
- Lasota, Andrzej and Michael C. Mackey (1994). *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. Berlin: Springer-Verlag. First edition, *Probabilistic Properties of Deterministic Systems*, Cambridge University Press, 1985.
- Lehmann, E. L. and George Casella (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Berlin: Springer-Verlag, 2nd edn.
- Liptser, Robert S. and Albert N. Shiryaev (2001). *Statistics of Random Processes*. Berlin: Springer-Verlag, 2nd edn. Two volumes. Trans. A. B. Aries. First English edition 1977–1978. First published as *Statistika sluchainykh protessov*, Moscow: Nauka, 1974.
- Loève, Michel (1955). *Probability Theory*. New York: D. Van Nostrand Company, 1st edn.
- Mackey, Michael C. (1992). *Time’s Arrow: The Origins of Thermodynamic Behavior*. Berlin: Springer-Verlag.
- Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.
- Øksendal, Bernt (1995). *Stochastic Differential Equations: An Introduction with Applications*. Berlin: Springer-Verlag, 4th edn.
- Olivieri, Enzo and Maria Eulália Vares (2005). *Large Deviations and Metastability*. Cambridge, England: Cambridge University Press.
- Ornstein, Donald S. and Benjamin Weiss (1990). “How Sampling Reveals a Process.” *Annals of Probability*, **18**: 905–930. URL <http://links.jstor.org/sici?sici=0091-1798%28199007%2918%3A3%3C905%3AHSRAP%3E2.0.CO%3B2-J>.

- Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems*. New York: Morgan Kaufmann.
- Peres, Yuval and Paul Shields (2005). “Two New Markov Order Estimators.” E-print, arxiv.org. URL <http://arxiv.org/abs/math.ST/0506080>.
- Pollard, David (2002). *A User’s Guide to Measure Theoretic Probability*. Cambridge, England: Cambridge University Press.
- Pullum, Geoffrey K. (1991). *The Great Eskimo Vocabulary Hoax, and Other Irreverent Essays on the Study of Language*. Chicago: University of Chicago Press.
- Rabiner, Lawrence R. (1989). “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE*, **77**: 257–286. URL <http://dx.doi.org/10.1109/5.18626>.
- Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press.
- Rogers, L. C. G. and David Williams (1994). *Foundations*, vol. 1 of *Diffusions, Markov Processes and Martingales*. New York: John Wiley, 2nd edn. Reprinted Cambridge, England: Cambridge University Press, 2000.
- (2000). *Itô Calculus*, vol. 2 of *Diffusions, Markov Processes and Martingales*. Cambridge, England: Cambridge University Press, 2nd edn.
- Rosenblatt, Murray (1956). “A Central Limit Theorem and a Strong Mixing Condition.” *Proceedings of the National Academy of Sciences (USA)*, **42**: 43–47. URL <http://www.pnas.org/cgi/reprint/42/1/43>.
- (1971). *Markov Processes. Structure and Asymptotic Behavior*. Berlin: Springer-Verlag.
- Schroeder, Manfred (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. San Francisco: W. H. Freeman.
- Selmeçzi, David, Simon F. Tolic-Norrelykke, Erik Schaeffer, Peter H. Hagedorn, Stephan Mosler, Kirstine Berg-Sorensen, Niels B. Larsen and Henrik Flyvbjerg (2006). “Brownian Motion after Einstein: Some new applications and new experiments.” In *Controlled Nanoscale Motion in Biological and Artificial Systems* (H. Linke and A. Mansson, eds.), vol. 131 of *Proceedings of the Nobel Symposium*, p. forthcoming. Berlin: Springer. URL <http://arxiv.org/abs/physics/0603142>.
- Shalizi, Cosma Rohilla and James P. Crutchfield (2001). “Computational Mechanics: Pattern and Prediction, Structure and Simplicity.” *Journal of Statistical Physics*, **104**: 817–879. URL <http://arxiv.org/abs/cond-mat/9907176>.

- Shannon, Claude E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**: 379–423. URL <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>. Reprinted in Shannon and Weaver (1963).
- Shannon, Claude E. and Warren Weaver (1963). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.
- Shiryaev, Albert N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*. Singapore: World Scientific. Trans. N. Kruzhilin.
- Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Cambridge, Massachusetts: MIT Press, 2nd edn.
- Taniguchi, Masanobu and Yoshihide Kakizawa (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Berlin: Springer-Verlag.
- Tuckwell, Henry C. (1989). *Stochastic Processes in the Neurosciences*. Philadelphia: SIAM.
- Tyran-Kamińska, Marta (2005). "An Invariance Principle for Maps with Polynomial Decay of Correlations." *Communications in Mathematical Physics*, **260**: 1–15. URL <http://arxiv.org/abs/math.DS/0408185>.
- von Plato, Jan (1994). *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective*. Cambridge, England: Cambridge University Press.
- Wax, Nelson (ed.) (1954). *Selected Papers on Noise and Stochastic Processes*. Dover.
- Wiener, Norbert (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.
- (1958). *Nonlinear Problems in Random Theory*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.
- (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.
- Wu, Wei Biao (2005). "Nonlinear system theory: Another look at dependence." *Proceedings of the National Academy of Sciences (USA)*, **102**: 14150–14154.