

# Contents

*List of Figures*      xix

*Series Foreword*      xxi

*Foreword*      xxiii

*Preface*      xxv

## **I Introductory Material      1**

### **1 *Learning, Regularity, and Compression*      3**

- 1.1 Regularity and Learning      4
- 1.2 Regularity and Compression      4
- 1.3 Solomonoff's Breakthrough – Kolmogorov Complexity      8
- 1.4 Making the Idea Applicable      10
- 1.5 Crude MDL, Refined MDL and Universal Coding      12
  - 1.5.1 From Crude to Refined MDL      14
  - 1.5.2 Universal Coding and Refined MDL      17
  - 1.5.3 Refined MDL for Model Selection      18
  - 1.5.4 Refined MDL for Prediction and Hypothesis Selection      20
- 1.6 Some Remarks on Model Selection      23
  - 1.6.1 Model Selection among Non-Nested Models      23
  - 1.6.2 Goals of Model vs. Point Hypothesis Selection      25
- 1.7 The MDL Philosophy      26
- 1.8 MDL, Occam's Razor, and the "True Model"      29
  - 1.8.1 Answer to Criticism No. 1      30

1.8.2	Answer to Criticism No. 2	32
1.9	History and Forms of MDL	36
1.9.1	What Is MDL?	37
1.9.2	MDL Literature	38
1.10	Summary and Outlook	40
<b>2</b>	<b><i>Probabilistic and Statistical Preliminaries</i></b>	<b>41</b>
2.1	General Mathematical Preliminaries	41
2.2	Probabilistic Preliminaries	46
2.2.1	Definitions; Notational Conventions	46
2.2.2	Probabilistic Sources	53
2.2.3	Limit Theorems and Statements	55
2.2.4	Probabilistic Models	57
2.2.5	Probabilistic Model Classes	60
2.3	Kinds of Probabilistic Models*	62
2.4	Terminological Preliminaries	69
2.5	Modeling Preliminaries:	
	Goals and Methods for Inductive Inference	71
2.5.1	Consistency	71
2.5.2	Basic Concepts of Bayesian Statistics	74
2.6	Summary and Outlook	78
<b>3</b>	<b><i>Information-Theoretic Preliminaries</i></b>	<b>79</b>
3.1	Coding Preliminaries	79
3.1.1	Restriction to Prefix Coding Systems; Descriptions as Messages	83
3.1.2	Different Kinds of Codes	86
3.1.3	Assessing the Efficiency of Description Methods	90
3.2	The Most Important Section of This Book: Probabilities and Code Lengths	90
3.2.1	The Kraft Inequality	91
3.2.2	Code Lengths “Are” Probabilities	95
3.2.3	Immediate Insights and Consequences	99
3.3	Probabilities and Code Lengths, Part II	101
3.3.1	(Relative) Entropy and the Information Inequality	103
3.3.2	Uniform Codes, Maximum Entropy, and Minimax Codelength	106
3.4	Summary, Outlook, Further Reading	106
<b>4</b>	<b><i>Information-Theoretic Properties of Statistical Models</i></b>	<b>109</b>

4.1	Introduction	109
4.2	Likelihood and <i>Observed</i> Fisher Information	111
4.3	KL Divergence and <i>Expected</i> Fisher Information	117
4.4	Maximum Likelihood: Data vs. Parameters	124
4.5	Summary and Outlook	130
<b>5</b>	<b><i>Crude Two-Part Code MDL</i></b>	<b>131</b>
5.1	Introduction: Making Two-Part MDL Precise	132
5.2	Two-Part Code MDL for Markov Chain Selection	133
5.2.1	The Code $C_2$	135
5.2.2	The Code $C_1$	137
5.2.3	Crude Two-Part Code MDL for Markov Chains	138
5.3	Simplistic Two-Part Code MDL Hypothesis Selection	139
5.4	Two-Part MDL for Tasks Other Than Hypothesis Selection	141
5.5	Behavior of Two-Part Code MDL	142
5.6	Two-Part Code MDL and Maximum Likelihood	144
5.6.1	The Maximum Likelihood <i>Principle</i>	144
5.6.2	MDL vs. ML	147
5.6.3	MDL as a <i>Maximum Probability Principle</i>	148
5.7	Computing and Approximating Two-Part MDL in Practice	150
5.8	Justifying Crude MDL: Consistency and Code Design	152
5.8.1	A General Consistency Result	153
5.8.2	Code Design for Two-Part Code MDL	157
5.9	Summary and Outlook	163
5.A	Appendix: Proof of Theorem 5.1	163
<b>II</b>	<b>Universal Coding</b>	<b>165</b>
<b>6</b>	<b><i>Universal Coding with Countable Models</i></b>	<b>171</b>
6.1	Universal Coding: The Basic Idea	172
6.1.1	Two-Part Codes as Simple Universal Codes	174
6.1.2	From Universal Codes to Universal Models	175
6.1.3	Formal Definition of Universality	177
6.2	The Finite Case	178
6.2.1	Minimax Regret and Normalized ML	179
6.2.2	NML vs. Two-Part vs. Bayes	182
6.3	The Countably Infinite Case	184
6.3.1	The Two-Part and Bayesian Codes	184

6.3.2	The NML Code	187	
6.4	Prequential Universal Models	190	
6.4.1	Distributions as Prediction Strategies	190	
6.4.2	Bayes Is Prequential; NML and Two-part Are Not	193	
6.4.3	The Prequential Plug-In Model	197	
6.5	Individual vs. Stochastic Universality*	199	
6.5.1	Stochastic Redundancy	199	
6.5.2	Uniformly Universal Models	201	
6.6	Summary, Outlook and Further Reading	204	
<b>7</b>	<b><i>Parametric Models: Normalized Maximum Likelihood</i></b>	<b>207</b>	
7.1	Introduction	207	
7.1.1	Preliminaries	208	
7.2	Asymptotic Expansion of Parametric Complexity	211	
7.3	The Meaning of $\int_{\Theta} \sqrt{\det I(\theta)} d\theta$	216	
7.3.1	Complexity and Functional Form	217	
7.3.2	KL Divergence and Distinguishability	219	
7.3.3	Complexity and Volume	222	
7.3.4	Complexity and the Number of Distinguishable Distributions*	224	
7.4	Explicit and Simplified Computations	226	
<b>8</b>	<b><i>Parametric Models: Bayes</i></b>	<b>231</b>	
8.1	The Bayesian Regret	231	
8.1.1	Basic Interpretation of Theorem 8.1	233	
8.2	Bayes Meets Minimax – Jeffreys’ Prior	234	
8.2.1	Jeffreys’ Prior and the Boundary	237	
8.3	How to Prove the Bayesian and NML Regret Theorems	239	
8.3.1	Proof Sketch of Theorem 8.1	239	
8.3.2	Beyond Exponential Families	241	
8.3.3	Proof Sketch of Theorem 7.1	243	
8.4	Stochastic Universality*	244	
8.A	Appendix: Proofs of Theorem 8.1 and Theorem 8.2	248	
<b>9</b>	<b><i>Parametric Models: Prequential Plug-in</i></b>	<b>257</b>	
9.1	Prequential Plug-in for Exponential Families	257	
9.2	The Plug-in vs. the Bayes Universal Model	262	
9.3	More Precise Asymptotics	265	
9.4	Summary	269	

<b>10</b>	<b><i>Parametric Models: Two-Part</i></b>	<b>271</b>
10.1	The Ordinary Two-Part Universal Model	271
10.1.1	Derivation of the Two-Part Code Regret	274
10.1.2	Proof Sketch of Theorem 10.1	277
10.1.3	Discussion	282
10.2	The Conditional Two-Part Universal Code*	284
10.2.1	Conditional Two-Part Codes for Discrete Exponential Families	286
10.2.2	Distinguishability and the Phase Transition*	290
10.3	Summary and Outlook	293
<b>11</b>	<b><i>NML With Infinite Complexity</i></b>	<b>295</b>
11.1	Introduction	295
11.1.1	Examples of Undefined NML Distribution	298
11.1.2	Examples of Undefined Jeffreys' Prior	299
11.2	Metauniversal Codes	301
11.2.1	Constrained Parametric Complexity	302
11.2.2	Meta-Two-Part Coding	303
11.2.3	Renormalized Maximum Likelihood*	306
11.3	NML with Luckiness	308
11.3.1	Asymptotic Expansion of LNML	312
11.4	Conditional Universal Models	316
11.4.1	Bayesian Approach with Jeffreys' Prior	317
11.4.2	Conditional NML	320
11.4.3	Liang and Barron's Approach	325
11.5	Summary and Remarks	329
11.A	Appendix: Proof of Theorem 11.4	329
<b>12</b>	<b><i>Linear Regression</i></b>	<b>335</b>
12.1	Introduction	336
12.1.1	Prelude: The Normal Location Family	338
12.2	Least-Squares Estimation	340
12.2.1	The Normal Equations	342
12.2.2	Composition of Experiments	345
12.2.3	Penalized Least-Squares	346
12.3	The Linear Model	348
12.3.1	Bayesian Linear Model $\mathcal{M}^{\mathbf{X}}$ with Gaussian Prior	354
12.3.2	Bayesian Linear Models $\mathcal{M}^{\mathbf{X}}$ and $\mathcal{S}^{\mathbf{X}}$ with Noninformative Priors	359

12.4	Universal Models for Linear Regression	363
12.4.1	NML	363
12.4.2	Bayes and LNML	364
12.4.3	Bayes-Jeffreys and CNML	365
<b>13</b>	<b><i>Beyond Parametrics</i></b>	<b>369</b>
13.1	Introduction	370
13.2	CUP: Unions of Parametric Models	372
13.2.1	CUP vs. Parametric Models	375
13.3	Universal Codes Based on Histograms	376
13.3.1	Redundancy of Universal CUP Histogram Codes	380
13.4	Nonparametric Redundancy	383
13.4.1	Standard CUP Universal Codes	384
13.4.2	Minimax Nonparametric Redundancy	387
13.5	Gaussian Process Regression*	390
13.5.1	Kernelization of Bayesian Linear Regression	390
13.5.2	Gaussian Processes	394
13.5.3	Gaussian Processes as Universal Models	396
13.6	Conclusion and Further Reading	402
<b>III</b>	<b>Refined MDL</b>	<b>403</b>
<b>14</b>	<b><i>MDL Model Selection</i></b>	<b>409</b>
14.1	Introduction	409
14.2	Simple Refined MDL Model Selection	411
14.2.1	Compression Interpretation	415
14.2.2	Counting Interpretation	416
14.2.3	Bayesian Interpretation	418
14.2.4	Prequential Interpretation	419
14.3	General Parametric Model Selection	420
14.3.1	Models with Infinite Complexities	420
14.3.2	Comparing Many or Infinitely Many Models	422
14.3.3	The General Picture	425
14.4	Practical Issues in MDL Model Selection	428
14.4.1	Calculating Universal Codelengths	428
14.4.2	Computational Efficiency and Practical Quality of Non-NML Universal Codes	429

14.4.3	Model Selection with Conditional NML and Plug-in Codes	431
14.4.4	General Warnings about Model Selection	435
14.5	MDL Model Selection for Linear Regression	438
14.5.1	Rissanen's RNML Approach	439
14.5.2	Hansen and Yu's gMDL Approach	443
14.5.3	Liang and Barron's Approach	446
14.5.4	Discussion	448
14.6	Worst Case vs. Average Case*	451
<b>15</b>	<b><i>MDL Prediction and Estimation</i></b>	<b>459</b>
15.1	Introduction	459
15.2	MDL for Prediction and Predictive Estimation	460
15.2.1	Prequential MDL Estimators	461
15.2.2	Prequential MDL Estimators Are Consistent	465
15.2.3	Parametric and Nonparametric Examples	469
15.2.4	Césaro KL consistency vs. KL consistency*	472
15.3	Two-Part Code MDL for Point Hypothesis Selection	476
15.3.1	Discussion of Two-Part Consistency Theorem	478
15.4	MDL Parameter Estimation	483
15.4.1	MDL Estimators vs. Luckiness ML Estimators	487
15.4.2	What Estimator To Use?	491
15.4.3	Comparison to Bayesian Estimators*	493
15.5	Summary and Outlook	498
15.A	Appendix: Proof of Theorem 15.3	499
<b>16</b>	<b><i>MDL Consistency and Convergence</i></b>	<b>501</b>
16.1	Introduction	501
16.1.1	The Scenarios Considered	501
16.2	Consistency: Prequential and Two-Part MDL Estimators	502
16.3	Consistency: MDL Model Selection	505
16.3.1	Selection between a Union of Parametric Models	505
16.3.2	Nonparametric Model Selection Based on CUP Model Class	508
16.4	MDL Consistency Peculiarities	511
16.5	Risks and Rates	515
16.5.1	Relations between Divergences and Risk Measures	517
16.5.2	Minimax Rates	519

16.6	MDL Rates of Convergence	520
16.6.1	Prequential and Two-Part MDL Estimators	520
16.6.2	MDL Model Selection	522
<b>17</b>	<b>MDL in Context</b>	<b>523</b>
17.1	MDL and Frequentist Paradigms	524
17.1.1	Sanity Check or Design Principle?	525
17.1.2	The Weak Prequential Principle	528
17.1.3	MDL vs. Frequentist Principles: Remaining Issues	529
17.2	MDL and Bayesian Inference	531
17.2.1	Luckiness Functions vs. Prior Distributions	534
17.2.2	MDL, Bayes, and Occam	539
17.2.3	MDL and Brands of Bayesian Statistics	544
17.2.4	Conclusion: a Common Future after All?	548
17.3	MDL, AIC and BIC	549
17.3.1	BIC	549
17.3.2	AIC	550
17.3.3	Combining the Best of AIC and BIC	552
17.4	MDL and MML	555
17.4.1	Strict Minimum Message Length	556
17.4.2	Comparison to MDL	558
17.4.3	The Wallace-Freeman Estimator	560
17.5	MDL and Prequential Analysis	562
17.6	MDL and Cross-Validation	565
17.7	MDL and Maximum Entropy	567
17.8	Kolmogorov Complexity and Structure Function	570
17.9	MDL and Individual Sequence Prediction	573
17.10	MDL and Statistical Learning Theory	579
17.10.1	Structural Risk Minimization	581
17.10.2	PAC-Bayesian Approaches	585
17.10.3	PAC-Bayes and MDL	588
17.11	The Road Ahead	592
<b>IV</b>	<b>Additional Background</b>	<b>597</b>
<b>18</b>	<b>The Exponential or "Maximum Entropy" Families</b>	<b>599</b>
18.1	Introduction	600
18.2	Definition and Overview	601

18.3	Basic Properties	605
18.4	Mean-Value, Canonical, and Other Parameterizations	609
18.4.1	The Mean Value Parameterization	609
18.4.2	Other Parameterizations	611
18.4.3	Relating Mean-Value and Canonical Parameters**	613
18.5	Exponential Families of General Probabilistic Sources*	617
18.6	Fisher Information Definitions and Characterizations*	619
<b>19</b>	<b>Information-Theoretic Properties of Exponential Families</b>	<b>623</b>
19.1	Introduction	624
19.2	Robustness of Exponential Family Codes	624
19.2.1	If $\Theta_{\text{mean}}$ Does Not Contain the Mean**	627
19.3	Behavior <i>at</i> the ML Estimate $\hat{\beta}$	629
19.4	Behavior <i>of</i> the ML Estimate $\hat{\beta}$	632
19.4.1	Central Limit Theorem	633
19.4.2	Large Deviations	634
19.5	Maximum Entropy and Minimax Codelength	637
19.5.1	Exponential Families and Maximum Entropy	638
19.5.2	Exponential Families and Minimax Codelength	641
19.5.3	The Compression Game	643
19.6	Likelihood Ratio Families and Rényi Divergences*	645
19.6.1	The Likelihood Ratio Family	647
19.7	Summary	650
	<b>References</b>	<b>651</b>
	<b>Index</b>	<b>674</b>
	List of Symbols	675
	Subject Index	679