



Information Theory

From Coding to Learning

FIRST EDITION

Yury Polyanskiy

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Yihong Wu

Department of Statistics and Data Science
Yale University





Dedicated to

Names



Contents

<i>Preface</i>	<i>page</i> xv
<i>Introduction</i>	xvi
Frequently used notation	1
Part I Information measures	3
1 Entropy	6
1.1 Entropy and conditional entropy	6
1.2 Axiomatic characterization	11
1.3 History of entropy	11
1.4* Submodularity	13
1.5* Han's inequality and Shearer's Lemma	14
2 Divergence	17
2.1 Divergence and Radon-Nikodym derivatives	17
2.2 Divergence: main inequality and equivalent expressions	21
2.3 Differential entropy	23
2.4 Markov kernels	25
2.5 Conditional divergence, chain rule, data-processing inequality	27
2.6* Local behavior of divergence and Fisher information	32
2.6.1* Local behavior of divergence for mixtures	32
2.6.2* Parametrized family	34
3 Mutual information	37
3.1 Mutual information	37
3.2 Mutual information as difference of entropies	39
3.3 Examples of computing mutual information	42
3.4 Conditional mutual information and conditional independence	45
3.5 Sufficient statistics and data processing	48
4 Variational characterizations and continuity of D and I	50
4.1 Geometric interpretation of mutual information	51
4.2 Variational characterizations of divergence: Gelfand-Yaglom-Perez	54
4.3 Variational characterizations of divergence: Donsker-Varadhan	55

Contents vii

4.4 Continuity of divergence	57
4.5* Continuity under monotone limits of σ -algebras	58
4.6 Variational characterizations and continuity of mutual information	61
5 Extremization of mutual information: capacity saddle point	64
5.1 Convexity of information measures	64
5.2 Extremization of mutual information	65
5.3 Capacity as information radius	69
5.4 Existence of capacity-achieving output distribution (general case)	70
5.5 Gaussian saddle point	73
5.6 Iterative algorithms: Blahut-Arimoto, Expectation-Maximization, Sinkhorn	75
6 Tensorization. Fano's inequality. Entropy rate.	78
6.1 Tensorization (single-letterization) of mutual information	78
6.2* Gaussian capacity via orthogonal symmetry	79
6.3 Information measures and probability of error	80
6.4 Entropy rate	83
6.5 Entropy and symbol (bit) error rate	84
6.6 Entropy and contiguity	85
6.7 Mutual information rate	86
7 <i>f</i>-divergences	88
7.1 Definition and basic properties of <i>f</i> -divergences	88
7.2 Data-processing inequality; approximation by finite partitions	91
7.3 Total variation and Hellinger distance in hypothesis testing	95
7.4 Inequalities between <i>f</i> -divergences and joint range	98
7.5 Examples of computing joint range	102
7.5.1 Hellinger distance versus total variation	102
7.5.2 KL divergence versus total variation	103
7.5.3 Chi-squared versus total variation	103
7.6 A selection of inequalities between various divergences	104
7.7 Divergences between Gaussians	105
7.8 Mutual information based on <i>f</i> -divergence	106
7.9 Empirical distribution and χ^2 -information	108
7.10 Most <i>f</i> -divergences are locally χ^2 -like	110
7.11 <i>f</i> -divergences in parametric families: Fisher information	111
7.12 Rényi divergences and tensorization	116
7.13 Variational representation of <i>f</i> -divergences	118
7.14*Technical proofs: convexity, local expansions and variational representations	122

viii **Contents**

8 Entropy method in combinatorics and geometry	128
8.1 Binary vectors of average weights	129
8.2 Shearer's lemma & counting subgraphs	130
8.3 Brégman's Theorem	132
8.4 Euclidean geometry: Bollobás-Thomason and Loomis-Whitney	134
9 Random number generators	136
9.1 Setup	136
9.2 Converse	137
9.3 Elias' construction from data compression	138
9.4 Peres' iterated von Neumann's scheme	139
9.5 Bernoulli factory	141
Exercises for Part I	144
Part II Lossless data compression	159
10 Variable-length lossless compression	163
10.1 Variable-length lossless compression	163
10.2 Mandelbrot's argument for universality of Zipf's (power) law	168
10.3 Uniquely decodable codes, prefix codes and Huffman codes	171
11 Fixed-length (almost lossless) compression. Slepian-Wolf.	177
11.1 Fixed-length almost lossless code. Asymptotic Equipartition Property (AEP).	177
11.2 Linear Compression	182
11.3 Compression with side information at both compressor and decompressor	184
11.4 Slepian-Wolf (Compression with side information at decompressor only)	185
11.5 Multi-terminal Slepian Wolf	187
11.6*Source-coding with a helper (Ahlswede-Körner-Wyner)	189
12 Compressing stationary ergodic sources	192
12.1 Bits of ergodic theory	193
12.2 Proof of the Shannon-McMillan Theorem	196
12.3*Proof of the Birkhoff-Khintchine Theorem	198
12.4*Sinaï's generator theorem	200
13 Universal compression	205
13.1 Arithmetic coding	206
13.2 Combinatorial construction of Fitingof	207

13.2.1	Universal compressor for all finite-order Markov chains	208
13.3	Optimal compressors for a class of sources. Redundancy.	208
13.4*	Approximate minimax solution: Jeffreys prior	210
13.5	Sequential probability assignment: Krichevsky-Trofimov	213
13.6	Individual sequence and universal prediction	214
13.7	Lempel-Ziv compressor	217
Exercises for Part II		220
Part III Binary hypothesis testing		227
14 Neyman-Pearson lemma		230
14.1	Neyman-Pearson formulation	230
14.2	Likelihood ratio tests	233
14.3	Converse bounds on $\mathcal{R}(P, Q)$	235
14.4	Achievability bounds on $\mathcal{R}(P, Q)$	236
14.5	Stein's regime	239
14.6	Chernoff regime: preview	242
15 Information projection and large deviations		244
15.1	Basics of large deviations theory	244
15.1.1	Log MGF and rate function	245
15.1.2	Tilted distribution	249
15.2	Large-deviations exponents and KL divergence	251
15.3	Information Projection	254
15.4	Interpretation of Information Projection	258
15.5	Generalization: Sanov's theorem	258
16 Hypothesis testing: error exponents		260
16.1	(E_0, E_1) -Tradeoff	260
16.2	Equivalent forms of Theorem 16.1	263
16.3*	Sequential Hypothesis Testing	266
16.4	Composite, robust and goodness-of-fit hypothesis testing	270
Exercises for Part III		272
Part IV Channel coding		281
17 Error correcting codes		284
17.1	Codes and probability of error	284
17.2	Coding for Binary Symmetric Channels	286

x **Contents**

17.3 Optimal decoder	288
17.4 Weak converse bound	289
18 Random and maximal coding	291
18.1 Information density	291
18.2 Shannon's random coding bound	294
18.3 Dependence-testing bound	296
18.4 Feinstein's maximal coding bound	298
18.5 RCU and Gallager's bound	300
18.6 Linear codes	302
18.7 Why random and maximal coding work well?	307
19 Channel capacity	311
19.1 Channels and channel capacity	311
19.2 Shannon's noisy channel coding theorem	316
19.3 Examples of computing capacity	320
19.4*Symmetric channels	321
19.5*Information Stability	325
19.6 Capacity under bit error rate	329
19.7 Joint Source Channel Coding	331
20 Channels with input constraints. Gaussian channels.	335
20.1 Channel coding with input constraints	335
20.2 Channel capacity under separable cost constraints	338
20.3 Stationary AWGN channel	340
20.4 Parallel AWGN channel	342
20.5*Non-stationary AWGN	344
20.6*Additive colored Gaussian noise channel	345
20.7*Additive White Gaussian Noise channel with Intersymbol Interference	347
20.8*Gaussian channels with amplitude constraints	348
20.9*Gaussian channels with fading	348
21 Energy-per-bit, continuous-time channels	351
21.1 Energy per bit	351
21.2 Capacity per unit cost	354
21.3 Energy-per-bit for the fading channel	357
21.4 Capacity of the continuous-time AWGN channel	358
21.5*Capacity of the continuous-time band-limited AWGN channel	360

22 Strong converse. Channel dispersion and error exponents. Finite Blocklength Bounds.	363
22.1 Strong Converse	363
22.2 Stationary memoryless channel without strong converse	368
22.3 Meta-converse	369
22.4*Error exponents	371
22.5 Channel dispersion	375
22.6 Finite blocklength bounds and normal approximation	378
22.7 Normalized Rate	380
23 Channel coding with feedback	382
23.1 Feedback does not increase capacity for stationary memoryless channels	382
23.2*Alternative proof of Theorem 23.3 and Massey's directed information	386
23.3 When is feedback really useful?	388
23.3.1 Code with very small (e.g. zero) error probability	389
23.3.2 Code with variable length	392
23.3.3 Code with variable power	393
Exercises for Part IV	396
Part V Rate-distortion theory and metric entropy	407
24 Rate-distortion theory	410
24.1 Scalar and vector quantization	410
24.1.1 Scalar Uniform Quantization	410
24.1.2 Scalar Non-uniform Quantization	411
24.1.3 Optimal Scalar Quantizers	412
24.1.4 Fine quantization	414
24.1.5 Fine quantization and variable rate	415
24.2 Information-theoretic formulation	416
24.3 Converse bounds	418
24.4*Converting excess distortion to average	420
25 Rate distortion: achievability bounds	422
25.1 Shannon's rate-distortion theorem	422
25.1.1 Intuition	423
25.1.2 Proof of Theorem 25.1	425
25.2*Covering lemma	429
25.3*Wyner's common information	431
25.4*Approximation of output statistics and the soft-covering lemma	433

xii Contents

26 Evaluating rate-distortion function. Lossy Source-Channel separation.	436
26.1 Evaluation of $R(D)$	436
26.1.1 Bernoulli Source	436
26.1.2 Gaussian Source	437
26.2*Analog of saddle-point property in rate-distortion	440
26.3 Lossy joint source-channel coding	443
26.3.1 Converse	444
26.3.2 Achievability via separation	445
26.4 What is lacking in classical lossy compression?	448
27 Metric entropy	450
27.1 Covering and packing	450
27.2 Finite-dimensional space and volume bound	453
27.3 Beyond the volume bound	456
27.3.1 Sudakov minorization	458
27.3.2 Maurey's empirical method	460
27.3.3 Duality of metric entropy	461
27.4 Infinite-dimensional space: smooth functions	462
27.5 Hilbert ball has metric entropy $\frac{1}{\epsilon^2}$	465
27.6 Metric entropy and small-ball probability	467
27.7 Metric entropy and rate-distortion theory	469
Exercises for Part V	473
Part VI Statistical applications	477
28 Basics of statistical decision theory	480
28.1 Basic setting	480
28.2 Gaussian Location Model (GLM)	482
28.3 Bayes risk, minimax risk, and the minimax theorem	483
28.3.1 Bayes risk	484
28.3.2 Minimax risk	485
28.3.3 Minimax and Bayes risk: a duality perspective	487
28.3.4 Minimax theorem	488
28.4 Multiple observations and sample complexity	489
28.5 Tensor product of experiments	490
28.6 Log-concavity, Anderson's lemma and exact minimax risk in GLM	492
29 Classical large-sample asymptotics	496
29.1 Statistical lower bound from data processing	496

Contents **xiii**

29.1.1	Hammersley-Chapman-Robbins (HCR) lower bound	496
29.1.2	Bayesian Cramér-Rao lower bound	498
29.2	Bayesian Cramér-Rao lower bounds	499
29.3	Maximum Likelihood Estimator and asymptotic efficiency	502
29.4	Application: Estimating discrete distributions and entropy	504
30	Mutual information method	506
30.1	GLM revisited and Shannon lower bound	507
30.2	GLM with sparse means	510
30.3	Community detection	512
30.4	Estimation better than chance	513
31	Lower bounds via reduction to hypothesis testing	515
31.1	Le Cam's two-point method	515
31.2	Assouad's Lemma	518
31.2.1	Assouad's lemma from the Mutual Information Method	519
31.3	Fano's method	520
32	Entropic upper bound for statistical estimation	523
32.1	Yang-Barron's construction	523
32.1.1	Bayes risk as conditional mutual information and capacity bound	526
32.1.2	Capacity upper bound via KL covering numbers	529
32.1.3	Capacity lower bound via Hellinger packing number	530
32.1.4	General bounds between cumulative and individual (one-step) risks	531
32.2	Pairwise comparison à la Le Cam-Birgé	532
32.2.1	Composite hypothesis testing and Hellinger distance	532
32.2.2	Hellinger guarantee on Le Cam-Birgé's pairwise comparison estimator	533
32.2.3	Refinement using local entropy	535
32.2.4	Lower bound using local Hellinger packing	538
32.3	Yatracos' class and minimum distance estimator	540
32.4	Application: Estimating smooth densities	542
33	Strong data processing inequality	544
33.1	Computing a boolean function with noisy gates	544
33.2	Strong Data Processing Inequality	547
33.3	Directed Information Percolation	551
33.4	Input-dependent SDPI	555
33.4.1	Application: Broadcasting and coloring on trees	556
33.4.2	Application: distributed correlation estimation	559
33.5	Channel comparison: degradation, less noisy, more capable	560

xiv **Contents**

33.6 Undirected information percolation	562
33.6.1 Application: Spiked Wigner model	565
33.7 Strong data post-processing inequality (Post-SDPI)	566
33.7.1 Application: Distributed Mean Estimation	570
Exercises for Part VI	572
<i>References</i>	584

Preface

This book is a modern introduction to the field of information theory. In the last two decades, information theory has evolved from a discipline primarily dealing with problems of information storage and transmission (“coding”) to one focusing increasingly on information extraction and denoising (“learning”). This transformation is reflected in the title and content of this book.

The content grew out of the lecture notes accumulated over a decade of the authors’ teaching regular courses at MIT, University of Illinois, and Yale, as well as short courses at EPFL (Switzerland) and ENSAE (France). Our intention is to use this manuscript as a textbook for a first course on information theory for graduate (and advanced undergraduate) students, or for a second (topics) course delving deeper into specific areas. A significant part of the book is devoted to the exposition of information-theoretic methods which have found influential applications in other fields such as statistical learning and computer science. (Specifically, we cover Kolmogorov’s metric entropy, strong data processing inequalities, and entropic upper bounds for statistical estimation). We also include some lesser known classical material (for example, connections to ergodicity) along with the latest developments, which are often covered by the exercises (following the style of Csiszár and Körner [80]).

It is hard to mention everyone, who helped us start and finish this work, but some stand out especially. First and foremost, we owe our debt to Sergio Verdú, whose course at Princeton is responsible for our life-long admiration of the subject. Furthermore, some technical choices, such as the “one-shot” approach to coding theorems and simultaneous treatment of discrete and continuous alphabets, reflect the style we learned from his courses. Next, we were fortunate to have many bright students contribute to typing the lecture notes (precursor of this book), as well as to correcting and extending the content. Among them, we especially thank Ganesh Ajjanagadde, Austin Collins, Yuzhou Gu, Richard Guo, Qingqing Huang, Yunus Inan, Reka Inovan, Jason Klusowski, Anuran Makur, Pierre Quinton, Aolin Xu, Sheng Xu, Pengkun Yang, Junhui Zhang.

Y. Polyanskiy <yp@mit.edu>

MIT

Y. Wu <yihong.wu@yale.edu>

Yale

Introduction

What is information?

The Oxford English Dictionary lists 18 definitions of the word *information*, while the Merriam-Webster Dictionary lists 17. This emphasizes the diversity of meaning and domains in which the word information may appear. This book, however, is only concerned with a precise mathematical understanding of information, independent of the application context.

How can we measure something that we cannot even define well? Among the earliest attempts of quantifying information we can list R.A. Fisher's works on the uncertainty of statistical estimates ("confidence intervals") and R. Hartley's definition of information as the logarithm of the number of possibilities. Around the same time, Fisher [125] and others identified connection between information and thermodynamic *entropy*. This line of thinking culminated in Claude Shannon's magnum opus [268], where he formalized the concept of (what we call today the) Shannon information and forever changed the human language by accepting John Tukey's word *bit* as the unit of its measurement. In addition to possessing a number of elegant properties, Shannon information turned out to also answer certain rigorous mathematical questions (such as the optimal rate of data compression and data transmission). This singled out Shannon's definition as the right way of quantifying information. Classical information theory, as taught in [75, 80, 130], focuses exclusively on this point of view.

In this book, however, we take a slightly more general point of view. To introduce it, let us quote an eminent physicist L. Brillouin [52]:

We must start with a precise definition of the word "information". We consider a problem involving a certain number of possible answers, if we have no special information on the actual situation. When we happen to be in possession of some information on the problem, the number of possible answers is reduced, and complete information may even leave us with only one possible answer. Information is a function of the ratio of the number of possible answers before and after, and we choose a logarithmic law in order to insure additivity of the information contained in independent situations.

Note that only the last sentence specializes the more general term information to the Shannon's special version. In this book, we think of information without that last sentence. Namely, for us information is a measure of *difference between two beliefs about the system state*. For example, it could be the amount of *change* in our worldview following an observation or an event. Specifically, suppose that initially the probability distribution P describes our understanding of the world (e.g., P allows us to answer questions such as how likely it is to rain today). Following an observation our distribution changes to Q (e.g., upon observing clouds or a clear sky). The amount of information in the observation is the *dissimilarity* between P and Q . How to quantify dissimilarity depends on the particular context. As argued by Shannon, in many cases the right choice is the Kullback-Leibler

(KL) divergence $D(Q\|P)$, see Definition 2.1. Indeed, if the prior belief is described by a probability mass function $P = (p_1, \dots, p_k)$ on the set of k possible outcomes, then the observation of the first outcome results in the new (posterior) belief vector $Q = (1, 0, \dots, 0)$ giving $D(Q\|P) = \log \frac{1}{p_1}$, and similarly for other outcomes. Since the outcome i happens with probability p_i we see that the average dissimilarity between the prior and posterior beliefs is

$$\sum_{i=1}^k p_i \log \frac{1}{p_i},$$

which is precisely the Shannon entropy, cf. Definition 1.1.

However, it is our conviction that measures of dissimilarity (or “information measures”) other than the KL divergence are needed for applying information theory beyond the classical realms. For example, the concepts of total variation, Hellinger distance and χ^2 -divergence (both prominent members of the f -divergence family) have found deep and fruitful applications in the theory of statistical estimation and probability, as well as contemporary topics in theoretical computer science such as communication complexity, estimation with communication constraints, property testing (we discuss these in detail in Part VI). Therefore, when we talk about information measures in Part I of this book we do not exclusively focus on those of Shannon type, although the latter are justly given a premium treatment.

What is information theory?

Similarly to *information*, the subject of *information theory* does not have a precise definition. In the narrowest sense, it is a scientific discipline concerned with optimal methods of transmitting and storing data. The highlights of this part of the subject are so called “coding theorems” showing existence of algorithms for compressing and communicating information across noisy channels. Classical results, such as Shannon’s noisy channel coding theorem (Theorem 19.8), not only show existence of algorithms, but also quantify their performance and show that such performance is best possible. This part is, thus, concerned with identifying fundamental limits of practically relevant (engineering) problems. Consequently, this branch is sometimes called “IEEE¹-style information theory”, and it influenced or revolutionized much of information technology we witness today: digital communication, wireless (cellular and WiFi) networks, cryptography (Diffie-Hellman), data compression (Lempel-Ziv family of algorithms), and a lot more.

However, the true scope of the field is much broader. Indeed, the Hilbert’s 13th problem (for smooth functions) was illuminated and resolved by Arnold and Kolmogorov via the idea of metric entropy that Kolmogorov introduced following Shannon’s rate-distortion theory [?]. The (non-)isomorphism problem for Bernoulli shifts in ergodic theory has been solved by introducing the Kolmogorov-Sinai entropy. In physics, the Landauer principle and other works on Maxwell demon

¹ For Institute of Electrical and Electronics Engineers; pronounced “Eye-triple-E”.

xviii Introduction

have been heavily influenced by the information theory. Many more topics ranging from biology, neuroscience and thermodynamics to pattern recognition, artificial intelligence and control theory all regularly appear in information-theoretic conferences and journals.

It seems that objectively circumscribing the territory claimed by information theory is futile. Instead, we highlight what we believe to be the most interesting developments of late.

First, information processing systems of today are much more varied compared to those of last century. A modern controller (robot) is not just reacting to a few-dimensional vector of observations, modeled as a linear time-invariant system. Instead, it has million-dimensional inputs (e.g., a rasterized image), delayed and quantized, which also need to be communicated across noisy links. The target of statistical inference is no longer a low-dimensional parameter, but rather a high-dimensional (possibly discrete) object with structure (e.g. a sparse matrix, or a graph between communities). Furthermore, observations arrive to a statistician from spatially or temporally separated sources, which need to be transmitted cognizant of rate limitations. Recognizing these new challenges, multiple communities simultaneously started re-investigating classical results (Chapter 29) on the optimality of maximum-likelihood and the (optimal) variance bounds given by the Fisher information. These developments in high-dimensional statistics, computer science and statistical learning depend on the mastery of the f -divergences (Chapter 7), the mutual-information method (Chapter 30), and the strong version of the data-processing inequality (Chapter 33).

Second, since the 1990s technological advances have brought about a slew of new noisy channel models. While classical theory addresses the so-called memoryless channels, the modern channels, such as in flash storage, or urban wireless (multi-path, multi-antenna) communication, are far from memoryless. In order to analyze these, the classical “asymptotic i.i.d.” theory is insufficient. The resolution is the so-called “one-shot” approach to information theory, in which all main results are developed while treating the channel inputs and outputs as abstract [298]. Only at the last step those inputs are given the structure of long sequences and the asymptotic values are calculated. This new “one-shot” approach has additional relevance to anyone willing to learn quantum information theory, where it is in fact necessary.

Third, and perhaps the most important, is the explosion in the interest of understanding the methods and limits of machine learning from data. Information-theoretic methods were instrumental for several discoveries in this area. As examples, we recall the concept of metric entropy (Chapter 27) that is a cornerstone of Vapnik’s approach to supervised learning (known as empirical risk minimization). In addition, metric entropy turns out to govern the fundamental limits of, and suggest algorithms for, the problem of density estimation, the canonical building block of unsupervised learning (Chapter 32). Another fascinating connection is that the optimal prediction performance of online-learning algorithms is given by the maximum of the mutual information. This is shown through a deep connection between prediction and universal compression (Chapter 13), which lead to the multiplicative weight update algorithms [317, 73]. Finally, there is a common information-theoretic method for solving a series of problems in distributed estimation, community detection (in graphs), and computation with noisy logic gates. This method is a strong version of the classical data-processing inequality (see Chapter 33), and is being actively developed and applied.

Why another book on information theory?

In short, we think that the three important developments of late – the f -divergences, the one-shot point of view, the connections with statistical learning – are not covered adequately in existing textbooks. At the same time these topics are future-looking: their significance will only grow with time. Hence, studying them along with the classical information theory is a good investment of students' effort.

There are two great classical textbooks that are unlikely to become irrelevant any time soon: [75] and [80] (and the revised edition of the latter [83]). The former has been a primary textbook for the majority of undergraduate courses on information theory in the world. It manages to rigorously introduce the concepts of entropy, information and divergence and prove all the main results of the field. Furthermore, [75] touches upon non-standard topics, such as universal compression, gambling and portfolio theory.

The [80] spearheaded the combinatorial point of view on information theory, known as “the method of types”. While more mathematically demanding than [75], [80] manages to introduce stronger results such as sharp estimates of error exponents and, especially, rate regions in multi-terminal communication systems. However, both books are almost exclusively focused on asymptotics and Shannon-type information measures.

Many more specialized treatments are available as well. For a communication-oriented reader, the classical [130] is still indispensable. The one-shot point of view is taken in [298]. Connections to statistical learning theory and learning on graphs (belief propagation) is beautifully covered in [200]. Ergodic theory is the central subject in [142]. Quantum information theory – a burgeoning field – is treated in the recent [320]. The only textbook dedicated to the connection between information theory and statistics is by Kullback [184], though restricted to large-sample asymptotics in hypothesis testing. In nonparametric statistics, application of information-theoretic methods is briefly (but elegantly) covered in [304].

Nevertheless, it is not possible to quilt this textbook from chapters of these excellent predecessors. A number of important topics are treated exclusively here, such as those in Chapters 7 (f -divergences), 18 (one-shot coding theorems), 22 (finite blocklength), 27 (metric entropy), 30 (mutual information method), 32 (entropic bounds on estimation), and 33 (strong data-processing inequalities). Furthermore, building up to these chapters requires numerous small innovations across the rest of the textbook and are not available elsewhere.

Going to omissions, this book completely skips the topic of multi-terminal information theory. This difficult subject captivated much of the effort in the post-Shannon “IEEE-style” theory. We refer to the classics [83] and the recent excellent textbook [110] containing an encyclopedic coverage of this area.

Another unfortunate omission is the connection between information theory and functional inequalities [75, Chapter 17]. This topic has seen a flurry of recent activity, especially in logarithmic Sobolev inequalities, isoperimetry, concentration of measure, Brascamp-Lieb inequalities, (Marton-Talagrand) information-transportation inequalities and others; see the monograph [249].

A note to statisticians

The interplay between information theory and statistics is a constant theme in the development of both fields. Since its inception, information theory has been indispensable for understanding the fundamental limits of statistical estimation. The prominent role of information-theoretic quantities, such as mutual information, f -divergence, metric entropy, and capacity, in establishing the minimax rates of estimation has long been recognized since the seminal work of Le Cam [188], Ibragimov and Khas'minski [158], Pinsker [228], Birgé [34], Haussler and Opper [153], Yang and Barron [329], among many others. In Part VI of this book we give an exposition to some of the most influential information-theoretic ideas and their applications in statistics. Of course, this is not meant to be a thorough treatment of decision theory or mathematical statistics; for that purpose, we refer to the classics [158, 192, 43, 304] and the more recent monographs [54, 256, 137, 318] focusing on high dimensions. Instead, we apply the theory developed in previous Parts I–V of this book to several concrete and carefully chosen examples of determining the minimax risk in both classical (fixed-dimensional, large-sample asymptotic) and modern (high-dimensional, non-asymptotic) settings.

At a high level, the connection between information theory (in particular, data transmission) and statistical inference is that both problems are defined by a conditional distribution $P_{Y|X}$, which is referred to as the *channel* for the former and the *statistical model* or *experiment* for the latter. In data transmission we optimize the encoder, which maps messages to codewords, chosen in a way that permits the decoder to reconstruct the message based on the noisy observation Y . In statistical settings, Y is still the observation while X plays the role of the parameter which determines the distribution of Y via $P_{Y|X}$; the major distinction is that here we no longer have the freedom to preselect X and the only task is to smartly estimate X (in either the average or the worst case) on the basis of the data Y . Despite this key difference, many information-theoretic ideas still have influential and fruitful applications for statistical problems, as we shall see next.

In Chapter 29 we show how the data processing inequality can be used to deduce classical lower bounds (Hammersley-Chapman-Robbins, Cramér-Rao, van Trees). In Chapter 30 we introduce *the mutual information method*, based on the reasoning in joint source-channel coding. Namely, by comparing the amount of information contained in the data and the amount of information required for achieving a given estimation accuracy, both measured in bits, this method allows us to apply the theory of capacity and rate-distortion function developed in Parts IV and V to lower bound the statistical risk. Besides being principled, this approach also unifies the three popular methods for proving minimax lower bounds due to Le Cam, Assouad, and Fano respectively (Chapter 31).

It is a common misconception that information theory only supplies techniques for proving negative results in statistics. In Chapter 32 we present three *upper bounds* on statistical estimation risk based on *metric entropy*: Yang-Barron's construction inspired by universal compression, Le Cam-Birgé's tournament based on pairwise hypothesis testing, and Yatracos' minimum-distance approach. These powerful methods are responsible for some of the strongest and most general results in statistics and applicable for both high-dimensional and nonparametric problems. Finally, in Chapter 33 we introduce the method based on strong data processing inequalities and apply it to resolve an array of contemporary problems including community detection on graphs, distributed

estimation with communication constraints and generating random tree colorings. These problems are increasingly captivating the minds of computer scientists as well.

How to use this textbook

An introductory class on information theory aiming at advanced undergraduate or graduate students can proceed with the following sequence:

- Part I: Chapters 1–3, Sections 4.1, 5.1–5.3, 6.1, and 6.3, focusing only on discrete probability space and ignoring Radon-Nikodym derivatives. Some mention of applications in combinatorics and cryptography (Chapters 8, 9) is recommended.
- Part II: Chapter 10, Sections 11.1–11.4.
- Part III: Chapter 14, Sections 15.1–15.3, and 16.1.
- Part IV: Chapters 17–18, Sections 19.1–19.3, 19.7, 20.1–20.2, 23.1.
- Part V: Sections 24.1–24.3, 25.1, 26.1, and 26.3.
- Conclude with a few applications of information theory outside the classical domain (Chapters 30 and 33).

A graduate-level class on information theory with a traditional focus on communication and compression can proceed faster through Part I (omitting f -divergences and other non-essential chapters), but then cover II–V in depth, including strong converse, finite-blocklength regime, and communication with feedback, but omitting Chapter 27. It is important to work through exercises at the end of Part IV for this kind of class.

For a graduate-level class on information theory with an emphasis on statistical learning, start with Part I (especially Chapter 7), followed by Part II (especially Chapter 13) and Part III, from Part IV limit coverage to Chapters 17–19, and from Part V to Chapter 27 (especially, Sections 27.1–27.4). This should leave more than half of the semester for carefully working through Part VI. For example, for a good pace we suggest leaving at least 5–6 lectures for Chapters 32 and 33. These last chapters contain some bleeding-edge research results and open problems, hopefully welcoming students to work on them. For that we also recommend going over the exercises at the end of Parts I and VI.

Frequently used notation

General conventions

- The symbol \triangleq reads *defined as* and \equiv *abbreviated as*.
- The set of real numbers and integers are denoted by \mathbb{R} and \mathbb{Z} . Let $\mathbb{N} = \{1, 2, \dots\}$, $\mathbb{Z}_+ = \{0, 1, \dots\}$, $\mathbb{R}_+ = \{x : x \geq 0\}$.
- For $n \in \mathbb{N}$, let $[n] = \{1, \dots, n\}$.
- Throughout the book, $x^n \triangleq (x_1, \dots, x_n)$ denotes an n -dimensional vector, $x_i^j \triangleq (x_i, \dots, x_j)$ for $1 \leq i < j \leq n$ and $x_S \triangleq \{x_i : i \in S\}$ for $S \subset [n]$.
- Unless explicitly specified, the logarithm \log and exponential \exp are with respect to a generic common base. The natural logarithm is denoted by $\ln = \log_e$.
- $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.
- For $p \in [0, 1]$, $\bar{p} \triangleq 1 - p$.
- $x^+ = \max\{x, 0\}$.
- $w_H(x)$ denotes the Hamming weight (number of ones) of a binary vector x . $d_H(x, y) = \sum_{i=1}^n 1_{\{x_i \neq y_i\}}$ denotes the Hamming distance between vectors x and y of length n .
- Standard big O notations are used throughout the book: e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ if there is an absolute constant $c > 0$ such that $a_n \leq cb_n$; $a_n = \Omega(b_n)$ if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ if both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$, we also write $a_n \asymp b_n$ in these cases; $a_n = o(b_n)$ or $b_n = \omega(a_n)$ if $a_n \leq \epsilon_n b_n$ for some $\epsilon_n \rightarrow 0$.

Analysis

- Let $\text{int}(E)$ and $\text{cl}(E)$ denote the interior and closure of a set E , namely, the largest open set contained in and smallest closed set containing E , respectively.
- Let $\text{co}(E)$ denote the convex hull of E (without topology), namely, the smallest convex set containing E , given by $\text{co}(E) = \{\sum_{i=1}^n \alpha_i x_i : \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1, x_i \in E, n \in \mathbb{N}\}$.
- For subsets A, B of a real vector space and $\lambda \in \mathbb{R}$, denote the dilation $\lambda A = \{\lambda a : a \in A\}$ and the Minkowski sum $A + B = \{a + b : a \in A, B \in B\}$.
- For a metric space (\mathcal{X}, d) , a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is called C -Lipschitz if $|f(x) - f(y)| \leq Cd(x, y)$ for all $x, y \in \mathcal{X}$. We set $\|f\|_{\text{Lip}(\mathcal{X})} = \inf\{C : f \text{ is } C\text{-Lipschitz}\}$.

Measure theory and probability

- The Lebesgue measure on Euclidean spaces is denoted by Leb and also by vol (volume).
- Throughout the book, all measurable spaces $(\mathcal{X}, \mathcal{E})$ are standard Borel spaces. Unless explicitly needed, we suppress the underlying σ -algebra \mathcal{E} .

2 Frequently used notation

- The collection of all probability measures on \mathcal{X} is denoted by $\Delta(\mathcal{X})$. For finite spaces we abbreviate $\Delta_k \equiv \Delta([k])$, a $(k - 1)$ -dimensional simplex.
- For measures P and Q , their product measure is denoted by $P \times Q$ or $P \otimes Q$. The n -fold product of P is denoted by P^n or $P^{\otimes n}$.
- Let P be absolutely continuous with respect to Q , denoted by $P \ll Q$. The Radon-Nikodym derivative of P with respect to Q is denoted by $\frac{dP}{dQ}$. For a probability measure P , if $Q = \text{Leb}$, $\frac{dP}{dQ}$ is referred to the probability density function (pdf); if Q is the counting measure on a countable \mathcal{X} , $\frac{dP}{dQ}$ is the probability mass function (pmf).
- Let $P \perp Q$ denote their mutual singularity, namely, $P(A) = 0$ and $Q(A) = 1$ for some A .
- The *support* of a probability measure P , denoted by $\text{supp}(P)$, is the smallest closed set C such that $P(C) = 1$. An *atom* x of P is such that $P(\{x\}) > 0$. A distribution P is *discrete* if $\text{supp}(P)$ is a countable set (consisting of its atoms).
- Let X be a random variable taking values on \mathcal{X} , which is referred to as the *alphabet* of X . Typically upper case, lower case, and script case are reserved for random variables, realizations, and alphabets. Oftentimes \mathcal{X} and \mathcal{Y} are automatically assumed to be the alphabet of X and Y , etc.
- Let P_X denote the distribution (law) of the random variable X , $P_{X,Y}$ the joint distribution of X and Y , and $P_{Y|X}$ the conditional distribution of Y given X .
- The independence of random variables X and Y is denoted by $X \perp\!\!\!\perp Y$, in which case $P_{X,Y} = P_X \times P_Y$. Similarly, $X \perp\!\!\!\perp Y|Z$ denotes their conditional independence given Z , in which case $P_{X,Y|Z} = P_{X|Z} \times P_{Y|Z}$.
- Throughout the book, $X^n \equiv X_1^n \triangleq (X_1, \dots, X_n)$ denotes an n -dimensional random vector. We write $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$ if they are independently and identically distributed (iid) as P , in which case $P_{X^n} = P^n$.
- The empirical distribution of a sequence x_1, \dots, x_n denoted by \hat{P}_{x^n} ; empirical distribution of a random sample X_1, \dots, X_n denoted by $\hat{P}_n \equiv \hat{P}_{X^n}$.
- Let $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\mathbb{P}}$, \xrightarrow{d} denote convergence almost surely, in probability, and in distribution (law), respectively. Let $\stackrel{d}{=}$ denote equality in distribution.
- Some commonly used distributions are as follows:
 - $\text{Ber}(p)$: Bernoulli distribution with mean p .
 - $\text{Bin}(n, p)$: Binomial distribution with n trials and success probability p .
 - $\text{Poisson}(\lambda)$: Poisson distribution with mean λ .
 - Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian (normal) distribution on \mathbb{R} with mean μ and σ^2 and $\mathcal{N}(\mu, \Sigma)$ the Gaussian distribution on \mathbb{R}^d with mean μ and covariance matrix Σ . Denote the standard normal density by $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, the CDF and complementary CDF by $\Phi(t) = \int_{-\infty}^t \varphi(x) dx$ and $Q(t) = \Phi^c(t) = 1 - \Phi(t)$. The inverse of Q is denoted by $Q^{-1}(\epsilon)$.
 - $Z \sim \mathcal{N}_c(\mu, \sigma^2)$ denotes the complex-valued circular symmetric normal distribution with expectation $\mathbb{E}[Z] = \mu \in \mathbb{C}$ and $\mathbb{E}[|Z - \mu|^2] = \sigma^2$.
 - For a compact subset \mathcal{X} of \mathbb{R}^d with non-empty interior, $\text{Unif}(\mathcal{X})$ denotes the uniform distribution on \mathcal{X} , with $\text{Unif}(a, b) \equiv \text{Unif}([a, b])$ for interval $[a, b]$. We also use $\text{Unif}(\mathcal{X})$ to denote the uniform (equiprobable) distribution on a finite set \mathcal{X} .

Part I

Information measures



Information measures form the backbone of information theory. The first part of this book is devoted to an in-depth study of various information measures, notably, entropy, divergence, mutual information, as well as their conditional versions (Chapters 1–3). In addition to basic definitions illustrated through concrete examples, we will also study various aspects including regularity, tensorization, variational representation, local expansion, convexity and optimization properties, as well as the data processing principle (Chapters 4–6). These information measures will be imbued with operational meaning when we proceed to classical topics in information theory such as data compression and transmission, in subsequent parts of the book.

In addition to the classical (Shannon) information measures, Chapter 7 provides a systematic treatment of f -divergences, a generalization of (Shannon) measures introduced by Csíkszár that plays an important role in many statistical problems (see Parts III and VI). Finally, towards the end of this part we will discuss two operational topics: random number generators in Chapter 9 and the application of entropy method to combinatorics and geometry Chapter 8.

1

Entropy

This chapter introduces the first information measure – Shannon entropy. After studying its standard properties (chain rule, conditioning), we will briefly describe how one could arrive at its definition. We discuss axiomatic characterization, the historical development in statistical mechanics, as well as the underlying combinatorial foundation (“method of types”). We close the chapter with Han’s and Shearer’s inequalities, that both exploit submodularity of entropy. After this Chapter, the reader is welcome to consult the applications in combinatorics (Chapter 8) and random number generation (Chapter 9), which are independent of the rest of this Part.

1.1 Entropy and conditional entropy

Definition 1.1 (Entropy). Let X be a discrete random variable with probability mass function $P_X(x), x \in \mathcal{X}$. The *entropy* (or *Shannon entropy*) of X is

$$\begin{aligned} H(X) &= \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \end{aligned}$$

When computing the sum, we agree that (by continuity of $x \mapsto x \log \frac{1}{x}$)

$$0 \log \frac{1}{0} = 0. \quad (1.1)$$

Since entropy only depends on the distribution of a random variable, it is customary in information theory to also write $H(P_X)$ in place of $H(X)$, which we will do freely in this book. The basis of the logarithm in Definition 1.1 determines the units of the entropy:

$$\begin{aligned} \log_2 &\leftrightarrow \text{bits} \\ \log_e &\leftrightarrow \text{nats} \\ \log_{256} &\leftrightarrow \text{bytes} \\ \log &\leftrightarrow \text{arbitrary units, base always matches exp} \end{aligned}$$

Different units will be convenient in different cases and so most of the general results in this book are stated with “baseless” log/exp.

1.1 Entropy and conditional entropy

Definition 1.2 (Joint entropy). The *joint entropy* of n discrete random variables $X^n \triangleq (X_1, X_2, \dots, X_n)$ is

$$H(X^n) = H(X_1, \dots, X_n) = \mathbb{E} \left[\log \frac{1}{P_{X_1, \dots, X_n}(X_1, \dots, X_n)} \right].$$

Note that joint entropy is a special case of Definition 1.1 applied to the random vector $X^n = (X_1, X_2, \dots, X_n)$ taking values in the product space.

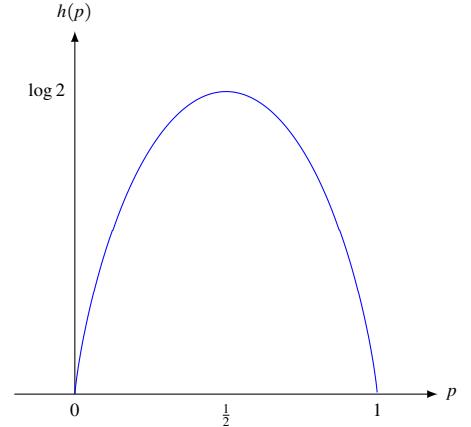
Remark 1.1. The name “entropy” originates from thermodynamics – see Section 1.3, which also provides combinatorial justification for this definition. Another common justification is to derive $H(X)$ as a consequence of natural axioms for any measure of “information content” – see Section 1.2. There are also natural experiments suggesting that $H(X)$ is indeed the amount of “information content” in X . For example, one can measure time it takes for ant scouts to describe the location of the food to ants-workers. It was found that when nest is placed at the root of a full binary tree of depth d and food at one of the leaves, the time was proportional to the entropy of a random variable describing the food location [253]. (It was also estimated that ants communicate with about 0.7–1 bit/min and that communication time reduces if there are some regularities in path-description: paths like “left,right,left,right,left,right” are described by scouts faster).

Entropy measures the intrinsic randomness or uncertainty of a random variable. In the simple setting where X takes values uniformly over a finite set \mathcal{X} , the entropy is simply given by log-cardinality: $H(X) = \log |\mathcal{X}|$. In general, the more spread out (resp. concentrated) a probability mass function is, the higher (resp. lower) is its entropy, as demonstrated by the following example.

Example 1.1 (Bernoulli). Let $X \sim \text{Ber}(p)$, with $P_X(1) = p$ and $P_X(0) = \bar{p} \triangleq 1 - p$. Then

$$H(X) = h(p) \triangleq p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}}.$$

Here $h(\cdot)$ is called the *binary entropy function*, which is continuous, concave on $[0, 1]$, symmetric around $\frac{1}{2}$, and satisfies $h'(p) = \log \frac{\bar{p}}{p}$, with infinite slope at 0 and 1. The highest entropy is achieved at $p = \frac{1}{2}$ (uniform), while the lowest entropy is achieved at $p = 0$ or 1 (deterministic). It is instructive to compare the plot of the binary entropy function with the variance $p(1-p)$.



Example 1.2 (Geometric). Let X be geometrically distributed, with $P_X(i) = p\bar{p}^i$, $i = 0, 1, \dots$. Then $\mathbb{E}[X] = \frac{\bar{p}}{p}$ and

$$H(X) = \mathbb{E}[\log \frac{1}{p\bar{p}^X}] = \log \frac{1}{p} + \mathbb{E}[X] \log \frac{1}{\bar{p}} = \frac{h(p)}{p}.$$

Example 1.3 (Infinite entropy). Is it possible that $H(X) = +\infty$? Yes, for example, $\mathbb{P}[X = k] \propto \frac{1}{k \ln^2 k}$, $k = 2, 3, \dots$.

Many commonly used information measures have their conditional counterparts, defined by applying the original definition to a conditional probability measure followed by a further averaging. For entropy this is defined as follows.

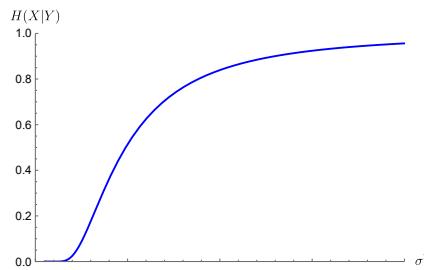
Definition 1.3 (Conditional entropy). Let X be a discrete random variable and Y arbitrary. Denote by $P_{X|Y=y}(\cdot)$ or $P_{X|Y}(\cdot|y)$ the conditional distribution of X given $Y = y$. The conditional entropy of X given Y is

$$H(X|Y) = \mathbb{E}_{y \sim P_Y}[H(P_{X|Y=y})] = \mathbb{E}\left[\log \frac{1}{P_{X|Y}(X|Y)}\right],$$

Similar to entropy, conditional entropy measures the remaining randomness of a random variable when another is revealed. As such, $H(X|Y) = H(X)$ whenever Y is independent of X . But when Y depends on X , observing Y does lower the entropy of X . Before formalizing this in the next theorem, here is a concrete example.

Example 1.4 (Conditional entropy and noisy channel). Let Y be a noisy observation of $X \sim \text{Ber}(\frac{1}{2})$ as follows.

- 1 $Y = X \oplus Z$, where \oplus denotes binary addition (XOR) and $Z \sim \text{Ber}(\delta)$ independently of X . In other words, Y agrees with X with probability δ and disagrees with probability $\bar{\delta}$. Then $P_{X|Y=0} = \text{Ber}(\delta)$ and $P_{X|Y=1} = \text{Ber}(\bar{\delta})$. Since $h(\delta) = h(\bar{\delta})$, $H(X|Y) = h(\delta)$. Note that when $\delta = \frac{1}{2}$, Y is independent of X and $H(X|Y) = H(X) = 1$ bits; when $\delta = 0$ or 1, X is completely determined by Y and hence $H(X|Y) = 0$.
- 2 $Y = X + Z$ be real-valued, where $Z \sim \mathcal{N}(0, \sigma^2)$. Then $H(X|Y) = \mathbb{E}[h(\mathbb{P}[X = 1|Y])]$, where $\mathbb{P}[X = 1|Y = y] = \frac{\varphi(\frac{y-1}{\sigma})}{\varphi(\frac{y}{\sigma}) + \varphi(\frac{y-1}{\sigma})}$ and $Y \sim \frac{1}{2}(\mathcal{N}(0, \sigma^2) + \mathcal{N}(1, \sigma^2))$. Below is a numerical plot of $H(X|Y)$ as a function of σ^2 which can be shown to be monotonically increasing from 0 to 1 bit. (Hint: Theorem 1.5(d).)



Before discussing various properties of entropy and conditional entropy, let us first review some relevant facts from convex analysis, which will be used extensively throughout the book.

1.1 Entropy and conditional entropy

9

Review: Convexity

- **Convex set:** A subset S of some vector space is *convex* if $x, y \in S \Rightarrow \alpha x + \bar{\alpha}y \in S$ for all $\alpha \in [0, 1]$. (Recall: $\bar{\alpha} \triangleq 1 - \alpha$)

Examples: unit interval $[0, 1]$; $S = \{\text{probability distributions on } \mathcal{X}\}$; $S = \{P_X : \mathbb{E}[X] = 0\}$.

- **Convex function:** $f: S \rightarrow \mathbb{R}$ is

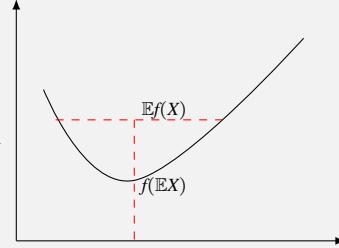
- convex if $f(\alpha x + \bar{\alpha}y) \leq \alpha f(x) + \bar{\alpha}f(y)$ for all $x, y \in S, \alpha \in [0, 1]$.
- strictly convex if $f(\alpha x + \bar{\alpha}y) < \alpha f(x) + \bar{\alpha}f(y)$ for all $x \neq y \in S, \alpha \in (0, 1)$.
- (strictly) concave if $-f$ is (strictly) convex.

Examples: $x \mapsto x \log x$ is strictly convex; the mean $P \mapsto \int x dP$ is convex but not strictly convex, variance is concave (Question: is it strictly concave? Think of zero-mean distributions.).

- **Jensen's inequality:**

For any S -valued random variable X ,

- f is convex $\Rightarrow f(\mathbb{E}X) \leq \mathbb{E}f(X)$
- f is strictly convex $\Rightarrow f(\mathbb{E}X) < \mathbb{E}f(X)$, unless X is a constant ($X = \mathbb{E}X$ a.s.)

**Theorem 1.4** (Properties of entropy).

- (Positivity) $H(X) \geq 0$ with equality, iff X is a constant (no randomness).
- (Uniform distribution maximizes entropy) For finite \mathcal{X} , $H(X) \leq \log |\mathcal{X}|$, with equality iff X is uniform on \mathcal{X} .
- (Invariance under relabeling) $H(X) = H(f(X))$ for any bijective f .
- (Conditioning reduces entropy) $H(X|Y) \leq H(X)$, with equality iff X and Y are independent.
- (Simple chain rule)

$$H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y). \quad (1.2)$$

- (Entropy under deterministic transformation) $H(X) = H(X, f(X)) \geq H(f(X))$ with equality iff f is one-to-one on the support of P_X .
- (Full chain rule)

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i|X^{i-1}) \leq \sum_{i=1}^n H(X_i), \quad (1.3)$$

with equality iff X_1, \dots, X_n are mutually independent.

Proof. (a) Since $\log \frac{1}{P_X(X)}$ is a positive random variable, its expectation $H(X)$ is also positive, with $H(X) = 0$ if and only if $\log \frac{1}{P_X(X)} = 0$ almost surely, namely, P_X is a point mass.

(b) Apply Jensen's inequality to the strictly concave function $x \mapsto \log x$:

$$H(X) = \mathbb{E} \left[\log \frac{1}{P_X(X)} \right] \leq \log \mathbb{E} \left[\frac{1}{P_X(X)} \right] = \log |\mathcal{X}|.$$

(c) $H(X)$ as a summation only depends on the values of P_X , not locations.

(d) Abbreviate $P(x) \equiv P_X(x)$ and $P(x|y) \equiv P_{X|Y}(x|y)$. Using $P(x) = \mathbb{E}_Y[P(x|Y)]$ and applying Jensen's inequality to the strictly concave function $x \mapsto x \log \frac{1}{x}$,

$$H(X|Y) = \sum_{x \in \mathcal{X}} \mathbb{E}_Y \left[P(x|Y) \log \frac{1}{P(x|Y)} \right] \leq \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)} = H(X).$$

Additionally, this also follows from (and is equivalent to) Corollary 3.1 in Chapter 3 or Theorem 5.4 in Chapter 5.

(e) Telescoping $P_{X,Y}(X, Y) = P_{Y|X}(Y|X)P_X(X)$ and noting that both sides are positive $P_{X,Y}$ -almost surely, we have

$$\mathbb{E}[\log \frac{1}{P_{X,Y}(X, Y)}] = \mathbb{E} \left[\log \frac{1}{P_X(X) \cdot P_{Y|X}(Y|X)} \right] = \underbrace{\mathbb{E} \left[\log \frac{1}{P_X(X)} \right]}_{H(X)} + \underbrace{\mathbb{E} \left[\log \frac{1}{P_{Y|X}(Y|X)} \right]}_{H(Y|X)}$$

(f) The intuition is that $(X, f(X))$ contains the same amount of information as X . Indeed, $x \mapsto (x, f(x))$ is one-to-one. Thus by (c) and (e):

$$H(X) = H(X, f(X)) = H(f(X)) + H(X|f(X)) \geq H(f(X))$$

The bound is attained iff $H(X|f(X)) = 0$ which in turn happens iff X is a *constant* given $f(X)$.

(g) Similar to (e), telescoping

$$P_{X_1 X_2 \dots X_n} = P_{X_1} P_{X_2 | X_1} \cdots P_{X_n | X^{n-1}}$$

and taking the logarithm prove the equality. The inequality follows from (d), with the case of equality occurring if and only if $P_{X_i | X^{i-1}} = P_{X_i}$ for $i = 1, \dots, n$, namely, $P_{X^n} = \prod_{i=1}^n P_{X_i}$. \square

To give a preview of the *operational meaning* of entropy, let us play the game of *20 Questions*. We are allowed to make queries about some unknown discrete RV X by asking *yes-no* questions. The objective of the game is to guess the realized value of the RV X . For example, $X \in \{a, b, c, d\}$ with $\mathbb{P}[X = a] = 1/2$, $\mathbb{P}[X = b] = 1/4$, and $\mathbb{P}[X = c] = \mathbb{P}[X = d] = 1/8$. In this case, we can ask “ $X = a?$ ”. If not, proceed by asking “ $X = b?$ ”. If not, ask “ $X = c?$ ”, after which we will know for sure the realization of X . The resulting average number of questions is $1/2 + 1/4 \times 2 + 1/8 \times 3 + 1/8 \times 3 = 1.75$, which equals $H(X)$ in bits. An alternative strategy is to ask “ $X = a, b$ or c, d ” in the first round then proceeds to determine the value in the second round, which always requires two questions and does worse on average.

It turns out (Section 10.3) that the minimal average number of yes-no questions to pin down the value of X is always between $H(X)$ bits and $H(X) + 1$ bits. In this special case the above scheme is optimal because (intuitively) it always splits the probability in half.

1.2 Axiomatic characterization

One might wonder why entropy is defined as $H(P) = \sum p_i \log \frac{1}{p_i}$ and if there are other definitions. Indeed, the information-theoretic definition of entropy is related to entropy in statistical physics. Also, it arises as answers to specific operational problems, e.g., the minimum average number of bits to describe a random variable as discussed above. Therefore it is fair to say that it is not pulled out of thin air.

Shannon in 1948 paper has also showed that entropy can be defined *axiomatically*, as a function satisfying several natural conditions. Denote a probability distribution on m letters by $P = (p_1, \dots, p_m)$ and consider a functional $H_m(p_1, \dots, p_m)$. If H_m obeys the following axioms:

- (a) Permutation invariance
- (b) Expansible: $H_m(p_1, \dots, p_{m-1}, 0) = H_{m-1}(p_1, \dots, p_{m-1})$.
- (c) Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = \log 2$.
- (d) Subadditivity: $H(X, Y) \leq H(X) + H(Y)$. Equivalently, $H_{mn}(r_{11}, \dots, r_{mn}) \leq H_m(p_1, \dots, p_m) + H_n(q_1, \dots, q_n)$ whenever $\sum_{j=1}^n r_{ij} = p_i$ and $\sum_{i=1}^m r_{ij} = q_j$.
- (e) Additivity: $H(X, Y) = H(X) + H(Y)$ if $X \perp\!\!\!\perp Y$. Equivalently, $H_{mn}(p_1 q_1, \dots, p_m q_n) = H_m(p_1, \dots, p_m) + H_n(q_1, \dots, q_n)$.
- (f) Continuity: $H_2(p, 1-p) \rightarrow 0$ as $p \rightarrow 0$.

then $H_m(p_1, \dots, p_m) = \sum_{i=1}^m p_i \log \frac{1}{p_i}$ is the only possibility. The interested reader is referred to [75, p. 53] and the references therein.

1.3 History of entropy

In the early days of industrial age, engineers wondered if it is possible to construct a perpetual motion machine. After many failed attempts, a law of conservation of energy was postulated: a machine cannot produce more work than the amount of energy it consumed from the ambient world. (This is also called the *first law* of thermodynamics.) The next round of attempts was then to construct a machine that would draw energy in the form of heat from a warm body and convert it to equal (or approximately equal) amount of work. An example would be a steam engine. However, again it was observed that all such machines were highly inefficient. That is, the amount of work produced by absorbing heat Q was far less than Q . The remainder of energy was dissipated to the ambient world in the form of heat. Again after many rounds of attempting various designs Clausius and Kelvin proposed another law:

Second law of thermodynamics: There does not exist a machine that operates in a cycle (i.e. returns to its original state periodically), produces useful work and whose only other effect on the outside world is drawing heat from a warm body. (That is, every such machine, should expend some amount of heat to some cold body too!)¹

¹ Note that the reverse effect (that is converting work into heat) is rather easy: friction is an example.

Equivalent formulation is as follows: “There does not exist a cyclic process that transfers heat from a cold body to a warm body”. That is, every such process needs to be helped by expending some amount of external work; for example, the air conditioners, sadly, will always need to use some electricity.

Notice that there is something annoying about the second law as compared to the first law. In the first law there is a quantity that is conserved, and this is somehow logically easy to accept. The second law seems a bit harder to believe in (and some engineers did not, and only their recurrent failures to circumvent it finally convinced them). So Clausius, building on an ingenious work of S. Carnot, figured out that there is an “explanation” to why any cyclic machine should expend heat. He proposed that there must be some hidden quantity associated to the machine, entropy of it (initially described as “transformative content” or *Verwandlungsinhalt* in German), whose value must return to its original state. Furthermore, under any reversible (i.e. quasi-stationary, or “very slow”) process operated on this machine the change of entropy is proportional to the ratio of absorbed heat and the temperature of the machine:

$$\Delta S = \frac{\Delta Q}{T}. \quad (1.4)$$

If heat Q is absorbed at temperature T_{hot} then to return to the original state, one must return some amount of heat Q' , where Q' can be significantly smaller than Q but never zero if Q' is returned at temperature $0 < T_{\text{cold}} < T_{\text{hot}}$. Further logical arguments can convince one that for irreversible cyclic process the change of entropy at the end of the cycle can only be positive, and hence *entropy cannot reduce*.

There were great many experimentally verified consequences that second law produced. However, what is surprising is that the mysterious entropy did not have any formula for it (unlike, say, energy), and thus had to be computed indirectly on the basis of relation (1.4). This was changed with the revolutionary work of Boltzmann and Gibbs, who provided a microscopic explanation of the second law based on statistical physics principles and showed that, e.g., for a system of n independent particles (as in ideal gas) the entropy of a given macro-state can be computed as

$$S = kn \sum_{j=1}^{\ell} p_j \log \frac{1}{p_j}, \quad (1.5)$$

where k is the Boltzmann constant, and we assumed that each particle can only be in one of ℓ molecular states (e.g. spin up/down, or if we quantize the phase volume into ℓ subcubes) and p_j is the fraction of particles in j -th molecular state.

More explicitly, their innovation was two-fold. First, they separated the concept of a micro-state (which in our example above corresponds to a tuple of n states, one for each particle) and the macro-state (a list $\{p_j\}$ of proportions of particles in each state). Second, they postulated that for experimental observations only the macro-state matters, but the multiplicity of the macro-state (number of micro-states that correspond to a given macro-state) is precisely the (exponential of the) entropy. The formula (1.5) then follows from the following explicit result connecting combinatorics and entropy.

Proposition 1.5 (Method of types). *Let n_1, \dots, n_k be non-negative integers with $\sum_{i=1}^k n_i = n$, and denote the distribution $P = (p_1, \dots, p_k)$, $p_i = \frac{n_i}{n}$. Then the multinomial coefficient $\binom{n}{n_1, \dots, n_k} \triangleq \frac{n!}{n_1! \cdots n_k!}$ satisfies*

$$\frac{1}{(1+n)^{k-1}} \exp\{nH(P)\} \leq \binom{n}{n_1, \dots, n_k} \leq \exp\{nH(P)\}.$$

Proof. For the upper bound, let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$ and let $N_i = \sum_{j=1}^n 1_{\{X_j=i\}}$ denote the number of occurrences of i . Then (N_1, \dots, N_k) has a multinomial distribution:

$$\mathbb{P}[N_1 = n'_1, \dots, N_k = n'_k] = \binom{n}{n'_1, \dots, n'_k} \prod_{i=1}^k p_i^{n'_i},$$

for any nonnegative integers n'_i such that $n'_1 + \cdots + n'_k = n$. Recalling that $p_i = n_i/n$, the upper bound follows from $\mathbb{P}[N_1 = n_1, \dots, N_k = n_k] \leq 1$. In addition, since (N_1, \dots, N_k) takes at most $(n+1)^{k-1}$ values, the lower bound follows if we can show that (n_1, \dots, n_k) is its mode. Indeed, for any n'_i with $n'_1 + \cdots + n'_k = n$, defining $\Delta_i = n'_i - n_i$ we have

$$\frac{\mathbb{P}[N_1 = n'_1, \dots, N_k = n'_k]}{\mathbb{P}[N_1 = n_1, \dots, N_k = n_k]} = \prod_{i=1}^k \frac{n_i!}{(n_i + \Delta_i)!} p_i^{\Delta_i} \leq \prod_{i=1}^k n_i^{-\Delta_i} p_i^{\Delta_i} = 1,$$

where the inequality follows from $\frac{m!}{(m+\Delta)!} \leq m^{-\Delta}$ and the last equality follows from $\sum_{i=1}^k \Delta_i = 0$. \square

Proposition 1.6 shows that the multinomial coefficient can be approximated up to a polynomial (in n) term by $\exp(nH(P))$. More refined estimates can be obtained; see Ex. I.2. In particular, the binomial coefficient can be approximated using the binary entropy function as follows: Provided that $p = \frac{k}{n} \in (0, 1)$,

$$e^{-1/6} \leq \frac{\binom{n}{k}}{\sqrt{2\pi np(1-p)}} e^{nh(p)} \leq 1. \quad (1.6)$$

For more on combinatorics and entropy, see Ex. I.1, I.3 and Chapter 8.

1.4* Submodularity

Recall that $[n]$ denotes a set $\{1, \dots, n\}$, $\binom{S}{k}$ denotes subsets of S of size k and 2^S denotes all subsets of S . A set function $f: 2^S \rightarrow \mathbb{R}$ is called *submodular* if for any $T_1, T_2 \subset S$

$$f(T_1 \cup T_2) + f(T_1 \cap T_2) \leq f(T_1) + f(T_2) \quad (1.7)$$

Submodularity is similar to concavity, in the sense that “adding elements gives diminishing returns”. Indeed consider $T' \subset T$ and $b \notin T$. Then

$$f(T \cup b) - f(T) \leq f(T' \cup b) - f(T').$$

Theorem 1.6. Let X^n be discrete RV. Then $T \mapsto H(X_T)$ is submodular.

Proof. Let $A = X_{T_1 \setminus T_2}$, $B = X_{T_1 \cap T_2}$, $C = X_{T_2 \setminus T_1}$. Then we need to show

$$H(A, B, C) + H(B) \leq H(A, B) + H(B, C).$$

This follows from a simple chain

$$H(A, B, C) + H(B) = H(A, C|B) + 2H(B) \quad (1.8)$$

$$\leq H(A|B) + H(C|B) + 2H(B) \quad (1.9)$$

$$= H(A, B) + H(B, C) \quad (1.10)$$

□

Note that entropy is not only submodular, but also monotone:

$$T_1 \subset T_2 \implies H(X_{T_1}) \leq H(X_{T_2}).$$

So fixing n , let us denote by Γ_n the set of all non-negative, monotone, submodular set-functions on $[n]$. Note that via an obvious enumeration of all non-empty subsets of $[n]$, Γ_n is a closed convex cone in $\mathbb{R}_+^{2^n-1}$. Similarly, let us denote by Γ_n^* the set of all set-functions corresponding to distributions on X^n . Let us also denote $\bar{\Gamma}_n^*$ the closure of Γ_n^* . It is not hard to show, cf. [335], that $\bar{\Gamma}_n^*$ is also a closed convex cone and that

$$\Gamma_n^* \subset \bar{\Gamma}_n^* \subset \Gamma_n.$$

The astonishing result of [336] is that

$$\Gamma_2^* = \bar{\Gamma}_2^* = \Gamma_2 \quad (1.11)$$

$$\Gamma_3^* \subsetneq \bar{\Gamma}_3^* = \Gamma_3 \quad (1.12)$$

$$\Gamma_n^* \subsetneq \bar{\Gamma}_n^* \subsetneq \Gamma_n \quad n \geq 4. \quad (1.13)$$

This follows from the fundamental new information inequality not implied by the submodularity of entropy (and thus called *non-Shannon inequality*). Namely, [336] showed that for any 4-tuple of discrete random variables:

$$I(X_3; X_4) - I(X_3; X_4|X_1) - I(X_3; X_4|X_2) \leq \frac{1}{2}I(X_1; X_2) + \frac{1}{4}I(X_1; X_3, X_4) + \frac{1}{4}I(X_2; X_3, X_4).$$

(This can be restated in the form of an entropy inequality using Theorem 3.4 but the resulting expression is too cumbersome).

1.5* Han's inequality and Shearer's Lemma

Theorem 1.7 (Han's inequality). Let X^n be discrete n -dimensional RV and denote $\bar{H}_k(X^n) = \frac{1}{\binom{n}{k}} \sum_{T \in \binom{[n]}{k}} H(X_T)$ the average entropy of a k -subset of coordinates. Then $\frac{\bar{H}_k}{k}$ is decreasing in k :

$$\frac{1}{n}\bar{H}_n \leq \dots \leq \frac{1}{k}\bar{H}_k \dots \leq \bar{H}_1. \quad (1.14)$$

1.5* Han's inequality and Shearer's Lemma 15

Furthermore, the sequence \bar{H}_k is increasing and concave in the sense of decreasing slope:

$$\bar{H}_{k+1} - \bar{H}_k \leq \bar{H}_k - \bar{H}_{k-1}. \quad (1.15)$$

Proof. Denote for convenience $\bar{H}_0 = 0$. Note that $\frac{\bar{H}_m}{m}$ is an average of differences:

$$\frac{1}{m}\bar{H}_m = \frac{1}{m} \sum_{k=1}^m (\bar{H}_k - \bar{H}_{k-1})$$

Thus, it is clear that (1.15) implies (1.14) since increasing m by one adds a smaller element to the average. To prove (1.15) observe that from submodularity

$$H(X_1, \dots, X_{k+1}) + H(X_1, \dots, X_{k-1}) \leq H(X_1, \dots, X_k) + H(X_1, \dots, X_{k-1}, X_{k+1}).$$

Now average this inequality over all $n!$ permutations of indices $\{1, \dots, n\}$ to get

$$\bar{H}_{k+1} + \bar{H}_{k-1} \leq 2\bar{H}_k$$

as claimed by (1.15).

Alternative proof: Notice that by “conditioning decreases entropy” we have

$$H(X_{k+1}|X_1, \dots, X_k) \leq H(X_{k+1}|X_2, \dots, X_k).$$

Averaging this inequality over all permutations of indices yields (1.15). \square

Theorem 1.8 (Shearer’s Lemma). *Let X^n be discrete n -dimensional RV and let $S \subset [n]$ be a random variable independent of X^n and taking values in subsets of $[n]$. Then*

$$H(X_S|S) \geq H(X^n) \cdot \min_{i \in [n]} \mathbb{P}[i \in S]. \quad (1.16)$$

Remark 1.2. In the special case where S is uniform over all subsets of cardinality k , (1.16) reduces to Han’s inequality $\frac{1}{n}H(X^n) \leq \frac{1}{k}\bar{H}_k$. The case of $n = 3$ and $k = 2$ can be used to give an entropy proof of the following well-known geometry result that relates the size of 3-D object to those of its 2-D projections: Place N points in \mathbb{R}^3 arbitrarily. Let N_1, N_2, N_3 denote the number of distinct points projected onto the xy , xz and yz -plane, respectively. Then $N_1N_2N_3 \geq N^2$. For another application, see Section 8.2.

Proof. We will prove an equivalent (by taking a suitable limit) version: If $\mathcal{C} = (S_1, \dots, S_M)$ is a list (possibly with repetitions) of subsets of $[n]$ then

$$\sum_j H(X_{S_j}) \geq H(X^n) \cdot \min_i \deg(i), \quad (1.17)$$

where $\deg(i) \triangleq \#\{j : i \in S_j\}$. Let us call \mathcal{C} a *chain* if all subsets can be rearranged so that $S_1 \subseteq S_2 \dots \subseteq S_M$. For a chain, (1.17) is trivial, since the minimum on the right-hand side is either

zero (if $S_M \neq [n]$) or equals multiplicity of S_M in \mathcal{C} ,² in which case we have

$$\sum_j H(X_{S_j}) \geq H(X_{S_M}) \# \{j : S_j = S_M\} = H(X^n) \cdot \min_i \deg(i).$$

For the case of \mathcal{C} not a chain, consider a pair of sets S_1, S_2 that are not related by inclusion and replace them in the collection with $S_1 \cap S_2, S_1 \cup S_2$. Submodularity (1.7) implies that the sum on the left-hand side of (1.17) does not increase under this replacement, values $\deg(i)$ are not changed. Notice that the total number of pairs that are not related by inclusion strictly decreases by this replacement: if T was related by inclusion to S_1 then it will also be related to at least one of $S_1 \cup S_2$ or $S_1 \cap S_2$; if T was related to both S_1, S_2 then it will be related to both of the new sets as well. Therefore, by applying this operation we must eventually arrive to a chain, for which (1.17) has already been shown. \square

Remark 1.3. Han's inequality (1.15) holds for any submodular set-function. For Han's inequality (1.14) we also need $f(\emptyset) = 0$ (this can be achieved by adding a constant to all values of f). Shearer's lemma holds for any submodular set-function that is also non-negative.

Example 1.5 (Non-entropy submodular function). Another submodular set-function is

$$S \mapsto I(X_S; X_{S^c}).$$

Han's inequality for this one reads

$$0 = \frac{1}{n} I_n \leq \dots \leq \frac{1}{k} I_k \dots \leq I_1,$$

where $I_k = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} I(X_S; X_{S^c})$ measures the amount of k -subset coupling in the random vector X^n .

² Note that, consequently, for X^n without constant coordinates, and if \mathcal{C} is a chain, (1.17) is only tight if \mathcal{C} consists of only \emptyset and $[n]$ (with multiplicities). Thus if degrees $\deg(i)$ are known and non-constant, then (1.17) can be improved, cf. [201].

2 Divergence

In this chapter we study divergence $D(P\|Q)$ (also known as information divergence, Kullback-Leibler (KL) divergence, relative entropy), which is the first example of dissimilarity (information) measure between a pair of distributions P and Q . As we will see later in Chapter 7, KL divergence is a special case of f -divergences. Defining KL divergence and its conditional version in full generality requires some measure-theoretic acrobatics (Radon-Nikodym derivatives and Markov kernels), that we spend some time on. (We stress again that all this abstraction can be ignored if one is willing to only work with finite or countably-infinite alphabets.)

Besides definitions we prove the “main inequality” showing that KL-divergence is non-negative. Coupled with the chain rule for divergence, this inequality implies the **data-processing inequality**, which is arguably the central pillar of information theory and this book. We conclude the chapter by studying local behavior of divergence when P and Q are close. In the special case when P and Q belong to a parametric family, we will see that divergence is locally quadratic with Hessian being the Fisher information, explaining the fundamental role of the latter in classical statistics.

2.1 Divergence and Radon-Nikodym derivatives

Review: Measurability

For an exposition of measure-theoretic preliminaries, see [58, Chapters I and IV]. We emphasize two aspects. *First*, in this book we understand Lebesgue integration $\int_{\mathcal{X}} f d\mu$ as defined for measurable functions that are extended real-valued, i.e. $f: \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$. In particular, for negligible set E , i.e. $\mu[E] = 0$, we have $\int_{\mathcal{X}} 1_E f d\mu = 0$ regardless of (possibly infinite) values of f on E , cf. [58, Chapter I, Prop. 4.13]. *Second*, we almost always assume that alphabets are standard Borel spaces. Some of the nice properties of standard Borel spaces:

- All complete separable metric spaces, endowed with Borel σ -algebras are standard Borel. In particular, countable alphabets and \mathbb{R}^n and \mathbb{R}^∞ (space of sequences) are standard Borel.
- If $\mathcal{X}_i, i = 1, \dots$ are standard Borel, then so is $\prod_{i=1}^{\infty} \mathcal{X}_i$.
- Singletons $\{x\}$ are measurable sets.
- The diagonal $\{(x, x) : x \in \mathcal{X}\}$ is measurable in $\mathcal{X} \times \mathcal{X}$.

We now need to define the second central concept of this book: the *relative entropy*, or *Kullback-Leibler divergence*. Before giving the formal definition, we start with special cases. For that we fix some alphabet \mathcal{A} . The relative entropy from between distributions P and Q on \mathcal{X} is denoted by $D(P\|Q)$, defined as follows.

- Suppose \mathcal{A} is a discrete (finite or countably infinite) alphabet. Then

$$D(P\|Q) \triangleq \begin{cases} \sum_{a \in \mathcal{A}: P(a), Q(a) > 0} P(a) \log \frac{P(a)}{Q(a)}, & \text{supp}(P) \subset \text{supp}(Q) \\ +\infty, & \text{otherwise} \end{cases} \quad (2.1)$$

- Suppose $\mathcal{A} = \mathbb{R}^k$, P and Q have densities (pdfs) p and q with respect to the Lebesgue measure. Then

$$D(P\|Q) = \begin{cases} \int_{\{p>0, q>0\}} p(x) \log \frac{p(x)}{q(x)} dx & \text{Leb}\{p > 0, q = 0\} = 0 \\ +\infty & \text{otherwise} \end{cases} \quad (2.2)$$

These two special cases cover a vast majority of all cases that we encounter in this book. However, mathematically it is not very satisfying to restrict to these two special cases. For example, it is not clear how to compute $D(P\|Q)$ when P and Q are two measures on a manifold (such as a unit sphere) embedded in \mathbb{R}^k . Another problematic case is computing $D(P\|Q)$ between measures on the space of sequences (stochastic processes). To address these cases we need to recall the concepts of *Radon-Nikodym derivative* and *absolute continuity*.

Recall that for two measures P and Q , we say P is absolutely continuous w.r.t. Q (denoted by $P \ll Q$) if $Q(E) = 0$ implies $P(E) = 0$ for all measurable E . If $P \ll Q$, then Radon-Nikodym theorem show that there exists a function $f: \mathcal{X} \rightarrow \mathbb{R}_+$ such that for any measurable set E ,

$$P(E) = \int_E f dQ. \quad [\text{change of measure}] \quad (2.3)$$

Such f is called a *relative density* or a Radon-Nikodym derivative of P w.r.t. Q , denoted by $\frac{dP}{dQ}$. Note that $\frac{dP}{dQ}$ may not be unique. In the simple cases, $\frac{dP}{dQ}$ is just the familiar *likelihood ratio*:

- For discrete distributions, we can just take $\frac{dP}{dQ}(x)$ to be the ratio of pmfs.
- For continuous distributions, we can take $\frac{dP}{dQ}(x)$ to be the ratio of pdfs.

We can see that the two special cases of $D(P\|Q)$ were both computing $\mathbb{E}_P[\log \frac{dP}{dQ}]$. This turns out to be the most general definition that we are looking for. However, we will state it slightly differently, following the tradition.

Definition 2.1 (Kullback-Leibler (KL) Divergence). Let P, Q be distributions on \mathcal{A} , with Q called the reference measure. The divergence (or relative entropy) between P and Q is

$$D(P\|Q) = \begin{cases} \mathbb{E}_Q[\frac{dP}{dQ} \log \frac{dP}{dQ}] & P \ll Q \\ +\infty & \text{otherwise} \end{cases} \quad (2.4)$$

2.1 Divergence and Radon-Nikodym derivatives 19

adopting again the convention from (1.1), namely, $0 \log 0 = 0$.

Below we will show (Lemma 2.4) that the expectation in (2.4) is well-defined (but possibly infinite) and coincides with $\mathbb{E}_P[\log \frac{dP}{dQ}]$ whenever $P \ll Q$.

To demonstrate the general definition in the case not covered by discrete/continuous specializations, consider the situation in which both P and Q are given as densities with respect to a common dominating measure μ , written as $dP = f_P d\mu$ and $dQ = f_Q d\mu$ for some non-negative f_P, f_Q . (In other words, $P \ll \mu$ and $f_P = \frac{dP}{d\mu}$.) For example, taking $\mu = P + Q$ always allows one to specify P and Q in this form. In this case, we have the following expression for divergence:

$$D(P\|Q) = \begin{cases} \int_{f_Q > 0, f_P > 0} d\mu f_P \log \frac{f_P}{f_Q} & \mu(\{f_Q = 0, f_P > 0\}) = 0, \\ +\infty & \text{otherwise} \end{cases} \quad (2.5)$$

Indeed, first note that, under the assumption of $P \ll \mu$ and $Q \ll \mu$, we have $P \ll Q$ iff $\mu(\{f_Q = 0, f_P > 0\}) = 0$. Furthermore, if $P \ll Q$, then $\frac{dP}{dQ} = \frac{f_P}{f_Q}$ Q -a.e., in which case applying (2.3) and (1.1) reduces (2.5) to (2.4). Namely, $D(P\|Q) = \mathbb{E}_Q[\frac{dP}{dQ} \log \frac{dP}{dQ}] = \mathbb{E}_Q[\frac{f_P}{f_Q} \log \frac{f_P}{f_Q}] = \int d\mu f_P \log \frac{f_P}{f_Q} 1_{\{f_Q > 0\}} = \int d\mu f_P \log \frac{f_P}{f_Q} 1_{\{f_Q > 0, f_P > 0\}}$.

Note that $D(P\|Q)$ was defined to be $+\infty$ if $P \not\ll Q$. However, it can also be $+\infty$ even when $P \ll Q$. For example, $D(\text{Cauchy}\|\text{Gaussian}) = \infty$. However, it does not mean that there are somehow two different ways in which D can be infinite. Indeed, what can be shown is that in both cases there exists a sequence of (finer and finer) finite partitions Π of the space \mathcal{A} such that evaluating KL divergence between the induced discrete distributions $P|_\Pi$ and $Q|_\Pi$ grows without a bound. This will be subject of Theorem 4.6 below.

Our next observation is that, generally, $D(P\|Q) \neq D(Q\|P)$ and, therefore, divergence is not a distance. We will see later, that this is natural in many cases; for example it reflects the inherent *asymmetry* of hypothesis testing (see Part III and, in particular, Section 14.5). Consider the example of coin tossing where under P the coin is fair and under Q the coin always lands on the head. Upon observing HHHHHHH, one tends to believe it is Q but can never be absolutely sure; upon observing HHT, one knows for sure it is P . Indeed, $D(P\|Q) = \infty$, $D(Q\|P) = 1 \text{ bit}$.

Having made these remarks we proceed to some examples. First, we show that D is unsurprisingly a generalization of entropy.

Theorem 2.2 (H v.s. D). *If distribution P is supported on a finite set \mathcal{A} , then*

$$H(P) = \log |\mathcal{A}| - D(P\|U_{\mathcal{A}}),$$

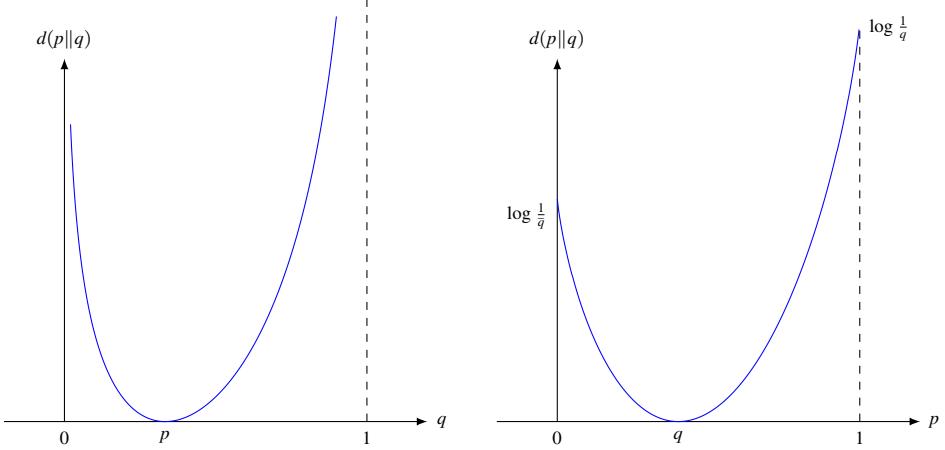
where $U_{\mathcal{A}}$ is a uniform distribution on \mathcal{A} .

Proof. $D(P\|U_{\mathcal{A}}) = \mathbb{E}_P[\log \frac{P(X)}{1/|\mathcal{A}|}] = \log |\mathcal{A}| - H(P)$. \square

Example 2.1 (Binary divergence). Consider $P = \text{Ber}(p)$ and $Q = \text{Ber}(q)$ on $\mathcal{A} = \{0, 1\}$. Then

$$D(P\|Q) = d(p\|q) \triangleq p \log \frac{p}{q} + \bar{p} \log \frac{\bar{p}}{\bar{q}}. \quad (2.6)$$

Here is how $d(p\|q)$ depends on p and q :



The following quadratic lower bound is easily checked:

$$d(p\|q) \geq 2(p - q)^2 \log e.$$

In fact, this is a special case of a famous Pinsker's inequality (Theorem 7.17).

Example 2.2 (Real Gaussian). For two Gaussians on $\mathcal{A} = \mathbb{R}$,

$$D(\mathcal{N}(m_1, \sigma_1^2) \| \mathcal{N}(m_0, \sigma_0^2)) = \frac{\log e}{2} \frac{(m_1 - m_0)^2}{\sigma_0^2} + \frac{1}{2} \left[\log \frac{\sigma_0^2}{\sigma_1^2} + \left(\frac{\sigma_1^2}{\sigma_0^2} - 1 \right) \log e \right]. \quad (2.7)$$

Here, the first and second term compares the means and the variances, respectively.

Similarly, in the vector case of $\mathcal{A} = \mathbb{R}^k$ and assuming $\det \Sigma_0 \neq 0$, we have

$$\begin{aligned} & D(\mathcal{N}(m_1, \Sigma_1) \| \mathcal{N}(m_0, \Sigma_0)) \\ &= \frac{\log e}{2} (m_1 - m_0)^\top \Sigma_0^{-1} (m_1 - m_0) + \frac{1}{2} \left(\log \det \Sigma_0 - \log \det \Sigma_1 + \text{tr}(\Sigma_0^{-1} \Sigma_1 - I) \log e \right). \end{aligned} \quad (2.8)$$

Example 2.3 (Complex Gaussian). The complex Gaussian distribution $\mathcal{N}_c(m, \sigma^2)$ with mean $m \in \mathbb{C}$ and variance σ^2 has a density $\frac{1}{\pi\sigma^2} e^{-|z-m|^2/\sigma^2}$ for $z \in \mathbb{C}$. In other words, the real and imaginary parts are independent real Gaussians:

$$\mathcal{N}_c(m, \sigma^2) = \mathcal{N} \left(\begin{bmatrix} \text{Re}(m) & \text{Im}(m) \end{bmatrix}, \begin{bmatrix} \sigma^2/2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix} \right)$$

Then

$$D(\mathcal{N}_c(m_1, \sigma_1^2) \| \mathcal{N}_c(m_0, \sigma_0^2)) = \frac{\log e}{2} \frac{|m_1 - m_0|^2}{\sigma_0^2} + \log \frac{\sigma_0^2}{\sigma_1^2} + \left(\frac{\sigma_1^2}{\sigma_0^2} - 1 \right) \log e. \quad (2.9)$$

which follows from (2.8). More generally, for complex Gaussian vectors on \mathbb{C}^k , assuming $\det \Sigma_0 \neq 0$,

$$\begin{aligned} D(\mathcal{N}_c(m_1, \Sigma_1) \| \mathcal{N}_c(m_0, \Sigma_0)) &= (m_1 - m_0)^\top \Sigma_0^{-1} (m_1 - m_0) \log e \\ &\quad + \log \det \Sigma_0 - \log \det \Sigma_1 + \text{tr}(\Sigma_0^{-1} \Sigma_1 - I) \log e \end{aligned}$$

2.2 Divergence: main inequality and equivalent expressions

Many inequalities in information can be attributed to the following fundamental result, namely, the nonnegativity of divergence.

Theorem 2.3 (Information Inequality).

$$D(P\|Q) \geq 0,$$

with equality iff $P = Q$.

Proof. In view of the definition (2.4), it suffices to consider $P \ll Q$. Let $\varphi(x) \triangleq x \log x$, which is strictly convex on \mathbb{R}_+ . Applying Jensen's Inequality:

$$D(P\|Q) = \mathbb{E}_Q \left[\varphi \left(\frac{dP}{dQ} \right) \right] \geq \varphi \left(\mathbb{E}_Q \left[\frac{dP}{dQ} \right] \right) = \varphi(1) = 0,$$

with equality iff $\frac{dP}{dQ} = 1$ Q -a.e., namely, $P = Q$. \square

The above proof explains the reason for defining $D(P\|Q) = \mathbb{E}_Q[\frac{dP}{dQ} \log \frac{dP}{dQ}]$ as opposed to $D(P\|Q) = \mathbb{E}_P[\log \frac{dP}{dQ}]$; nevertheless, the two definitions are equivalent. Furthermore, the next result unifies the two cases ($P \ll Q$ vs $P \not\ll Q$) in Definition 2.1.

Lemma 2.4. Let $P, Q, R \ll \mu$ and f_P, f_Q, f_R denote their densities relative to μ . Define a bivariate function $\text{Log}_{\frac{a}{b}} : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{\pm\infty\}$ by

$$\text{Log}_{\frac{a}{b}} = \begin{cases} -\infty & a = 0, b > 0 \\ +\infty & a > 0, b = 0 \\ 0 & a = 0, b = 0 \\ \log \frac{a}{b} & a > 0, b > 0. \end{cases} \quad (2.10)$$

Then the following results hold:

- First,

$$\mathbb{E}_P \left[\text{Log}_{\frac{f_R}{f_Q}} \right] = D(P\|Q) - D(P\|R), \quad (2.11)$$

provided at least one of the divergences is finite.

- Second, the expectation $\mathbb{E}_P \left[\text{Log}_{\frac{f_P}{f_Q}} \right]$ is well-defined (but possibly infinite) and, furthermore,

$$D(P\|Q) = \mathbb{E}_P \left[\text{Log}_{\frac{f_P}{f_Q}} \right]. \quad (2.12)$$

In particular, when $P \ll Q$ we have

$$D(P\|Q) = \mathbb{E}_P \left[\log \frac{dP}{dQ} \right]. \quad (2.13)$$

Remark 2.1. Note that ignoring the issue of dividing by or taking a log of 0, the proof of (2.12) is just the simple identity $\log \frac{dR}{dQ} = \log \frac{dRdP}{dQdP} = \log \frac{dP}{dQ} - \log \frac{dP}{dR}$. What permits us to handle zeros is the Log function, which satisfies several natural properties of the log: for every $a, b \in \mathbb{R}_+$

$$\text{Log} \frac{a}{b} = -\text{Log} \frac{b}{a}$$

and for every $c > 0$ we have

$$\text{Log} \frac{a}{b} = \text{Log} \frac{a}{c} + \text{Log} \frac{c}{b} = \text{Log} \frac{ac}{b} - \log(c)$$

except for the case $a = b = 0$.

Proof. First, suppose $D(P||Q) = \infty$ and $D(P||R) < \infty$. Then $P[f_R(Y) = 0] = 0$, and hence in computation of the expectation in (2.11) only the second part of convention (2.10) can possibly apply. Since also $f_P > 0$ P -almost surely, we have

$$\text{Log} \frac{f_R}{f_Q} = \text{Log} \frac{f_R}{f_P} + \text{Log} \frac{f_P}{f_Q}, \quad (2.14)$$

with both logarithms evaluated according to (2.10). Taking expectation over P we see that the first term, equal to $-D(P||R)$, is finite, whereas the second term is infinite. Thus, the expectation in (2.11) is well-defined and equal to $+\infty$, as is the LHS of (2.11).

Now consider $D(P||Q) < \infty$. This implies that $P[f_Q(Y) = 0] = 0$ and this time in (2.11) only the first part of convention (2.10) can apply. Thus, again we have identity (2.14). Since the P -expectation of the second term is finite, and of the first term non-negative, we again conclude that expectation in (2.11) is well-defined, equals the LHS of (2.11) (and both sides are possibly equal to $-\infty$).

For the second part, we first show that

$$\mathbb{E}_P \left[\min \left(\text{Log} \frac{f_P}{f_Q}, 0 \right) \right] \geq -\frac{\log e}{e}. \quad (2.15)$$

Let $g(x) = \min(x \log x, 0)$. It is clear $-\frac{\log e}{e} \leq g(x) \leq 0$ for all x . Since $f_P(Y) > 0$ for P -almost all Y , in convention (2.10) only the $\frac{1}{0}$ case is possible, which is excluded by the $\min(\cdot, 0)$ from the expectation in (2.15). Thus, the LHS in (2.15) equals

$$\begin{aligned} \int_{\{f_P > f_Q > 0\}} f_P(y) \log \frac{f_P(y)}{f_Q(y)} d\mu &= \int_{\{f_P > f_Q > 0\}} f_Q(y) \frac{f_P(y)}{f_Q(y)} \log \frac{f_P(y)}{f_Q(y)} d\mu \\ &= \int_{\{f_Q > 0\}} f_Q(y) g \left(\frac{f_P(y)}{f_Q(y)} \right) d\mu \\ &\geq -\frac{\log e}{e}. \end{aligned}$$

Since the negative part of $\mathbb{E}_P \left[\text{Log} \frac{f_P}{f_Q} \right]$ is bounded, the expectation $\mathbb{E}_P \left[\text{Log} \frac{f_P}{f_Q} \right]$ is well-defined. If $P[f_Q = 0] > 0$ then it is clearly $+\infty$, as is $D(P||Q)$ (since $P \not\ll Q$). Otherwise, let $E = \{f_P >$

2.3 Differential entropy 23

$0, f_Q > 0\}$. Then $P[E] = 1$ and on E we have $f_P = f_Q \cdot \frac{f_P}{f_Q}$. Thus, we obtain

$$\mathbb{E}_P \left[\log \frac{f_P}{f_Q} \right] = \int_E d\mu f_P \log \frac{f_P}{f_Q} = \int_E d\mu f_Q \varphi \left(\frac{f_P}{f_Q} \right) = \mathbb{E}_Q \left[1_E \varphi \left(\frac{f_P}{f_Q} \right) \right].$$

From here, we notice that $Q[f_Q > 0] = 1$ and on $\{f_P = 0, f_Q > 0\}$ we have $\varphi(\frac{f_P}{f_Q}) = 0$. Thus, the term 1_E can be dropped and we obtain the desired (2.12).

The final statement of the Lemma follows from taking $\mu = Q$ and noticing that P -almost surely we have

$$\log \frac{\frac{dP}{dQ}}{1} = \log \frac{dP}{dQ}.$$

□

2.3 Differential entropy

The definition of $D(P||Q)$ extends verbatim to measures P and Q (not necessarily probability measures), in which case $D(P||Q)$ can be negative. A sufficient condition for $D(P||Q) \geq 0$ is that P is a probability measure and Q is a sub-probability measure, i.e., $\int dQ \leq 1 = \int dP$. The notion of *differential entropy* is simply the divergence with respect to the Lebesgue measure:

Definition 2.5. The differential entropy of a random vector X is

$$h(X) = h(P_X) \triangleq -D(P_X||\text{Leb}). \quad (2.16)$$

In particular, if X has probability density function (pdf) p , then $h(X) = \mathbb{E} \log \frac{1}{p(X)}$; otherwise $h(X) = -\infty$. The conditional differential entropy is $h(X|Y) \triangleq \mathbb{E} \log \frac{1}{p_{X|Y}(X|Y)}$ where $p_{X|Y}$ is a conditional pdf.

Example 2.4 (Gaussian). For $X \sim N(\mu, \sigma^2)$,

$$h(X) = \frac{1}{2} \log(2\pi e \sigma^2) \quad (2.17)$$

More generally, for $X \sim N(\mu, \Sigma)$ in \mathbb{R}^d ,

$$h(X) = \frac{1}{2} \log((2\pi e)^d \det \Sigma) \quad (2.18)$$

Warning: Even for continuous random variable X , $h(X)$ can be positive, negative, take values of $\pm\infty$ or even undefined.¹ There are many crucial differences between the Shannon entropy and the differential entropy. For example, from Theorem 1.5 we know that deterministic processing cannot increase the Shannon entropy, i.e. $H(f(X)) \leq H(X)$ for any discrete X , which is intuitively clear. However, this fails completely for differential entropy (e.g. consider scaling). Furthermore,

¹ For an example, consider a piecewise-constant pdf taking value $e^{(-1)^n n}$ on the n -th interval of width $\Delta_n = \frac{c}{n^2} e^{(-1)^n n}$.

for sums of independent random variables, for integer-valued X and Y , $H(X+Y)$ is finite whenever $H(X)$ and $H(Y)$ are, because $H(X+Y) \leq H(X, Y) = H(X) + H(Y)$. This again fails for differential entropy. In fact, there exists real-valued X with finite $h(X)$ such that $h(X+Y) = \infty$ for any independent Y such that $h(Y) > -\infty$; there also exist X and Y with finite differential entropy such that $h(X+Y)$ does not exist (cf. [41, Section V]).

Nevertheless, differential entropy shares many functional properties with the usual Shannon entropy. For a short application to Euclidean geometry see Section 8.4.

Theorem 2.6 (Properties of differential entropy). *Assume that all differential entropies appearing below exist and are finite (in particular all RVs have pdfs and conditional pdfs).*

- (a) (Uniform distribution maximizes differential entropy) If $\mathbb{P}[X^n \in S] = 1$ then $h(X^n) \leq \log \text{Leb}(S)$, with equality iff X^n is uniform on S .
- (b) (Scaling and shifting) $h(X^n + x) = h(X^n)$, $h(\alpha X^n) = h(X^n) + k \log |\alpha|$ and for an invertible matrix A , $h(AX^n) = h(X^n) + \log |\det A|$.
- (c) (Conditioning reduces differential entropy) $h(X|Y) \leq h(X)$. (Here Y is arbitrary.)
- (d) (Chain rule) Let X^n has a joint probability density function. Then

$$h(X^n) = \sum_{k=1}^n h(X_k | X^{k-1}).$$

- (e) (Submodularity) The set-function $T \mapsto h(X_T)$ is submodular.
- (f) (Han's inequality) The function $k \mapsto \frac{1}{k \binom{n}{k}} \sum_{T \in \binom{[n]}{k}} h(X_T)$ is decreasing in k .

Proof. Parts (a), (c), and (d) follow from the similar argument in the proof (b), (d), and (g) of Theorem 1.5. Part (b) is by a change of variable in the density. Finally, (e) and (f) are analogous to Theorems 1.7 and 1.8. \square

Interestingly, the first property is robust to small additive perturbations, cf. Ex. I.6. Regarding maximizing entropy under quadratic constraints, we have the following characterization of Gaussians.

Theorem 2.7. *Let $\text{Cov}(X) = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top$ denote the covariance matrix of a random vector X . For any $d \times d$ positive definite matrix Σ ,*

$$\max_{P_X: \text{Cov}(X) \preceq \Sigma} h(X) = h(N(0, \Sigma)) = \frac{1}{2} \log((2\pi e)^d \det \Sigma) \quad (2.19)$$

Furthermore, for any $a > 0$,

$$\max_{P_X: \mathbb{E}[\|X\|^2] \leq a} h(X) = h\left(N\left(0, \frac{a}{d} I_d\right)\right) = \frac{d}{2} \log \frac{2\pi e a}{d}. \quad (2.20)$$

Proof. To show (2.19), without loss of generality, assume that $\mathbb{E}[X] = 0$. By comparing to Gaussian, we have

$$\begin{aligned} 0 &\leq D(P_X \| N(0, \Sigma)) \\ &= -h(X) + \frac{1}{2} \log((2\pi)^d \det(\Sigma)) + \frac{\log e}{2} \mathbb{E}[X^\top \Sigma^{-1} X] \\ &\leq -h(X) + h(N(0, \Sigma)), \end{aligned}$$

where in the last step we apply $\mathbb{E}[X^\top \Sigma^{-1} X] = \text{Tr}(\mathbb{E}[XX^\top] \Sigma^{-1}) \leq \text{Tr}(I)$ due to the constraint $\text{Cov}(X) \preceq \Sigma$ and the formula (2.18). The inequality (2.20) follows analogously by choosing the reference measure to be $N(0, \frac{a}{d} I_d)$. \square

Finally, let us mention a connection between the differential entropy and the Shannon entropy. Let X be a continuous random vector in \mathbb{R}^d . Denote its discretized version by $X_m = \frac{1}{m} \lfloor mX \rfloor$ for $m \in \mathbb{N}$, where $\lfloor \cdot \rfloor$ is taken componentwise. Rényi showed that [252, Theorem 1] provided $H(\lfloor X \rfloor) < \infty$ and $h(X)$ is defined, we have

$$H(X_m) = d \log m + h(X) + o(1), \quad m \rightarrow \infty. \quad (2.21)$$

To interpret this result, consider, for simplicity, $d = 1, m = 2^k$ and assume that X takes values in the unit interval, in which case X_{2^k} is the k -bit uniform quantization of X . Then (2.21) suggests that for large k , the quantized bits behave as independent fair coin flips. The underlying reason is that for “nice” density functions, the restriction to small intervals is approximately uniform. For more on quantization see Section 24.1 (notably Section 24.1.5) in Chapter 24.

2.4 Markov kernels

The main objects in this book are random variables and probability distributions. The main operation for creating new random variables, as well as for defining relations between random variables, is that of a *Markov kernel* (also known as a *transition probability kernel*).

Definition 2.8. A Markov kernel $K : \mathcal{X} \rightarrow \mathcal{Y}$ is a bivariate function $K(\cdot | \cdot)$, whose first argument is a measurable subset of \mathcal{Y} and the second is an element of \mathcal{X} , such that:

- 1 For any $x \in \mathcal{X}$: $K(\cdot | x)$ is a probability measure on \mathcal{Y}
- 2 For any measurable set A : $x \mapsto K(A|x)$ is a measurable function on \mathcal{X} .

The kernel K can be viewed as a random transformation acting from \mathcal{X} to \mathcal{Y} , which draws Y from a distribution depending on the realization of X , including deterministic transformations as special cases. For this reason, we write $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ and also $X \xrightarrow{P_{Y|X}} Y$. In information-theoretic context, we also refer to $P_{Y|X}$ as a *channel*, where X and Y are the channel input and output respectively. There are two ways of obtaining Markov kernels. The first way is defining them explicitly. Here are some examples of that:

- 1 Deterministic system: $Y = f(X)$. This corresponds to setting $P_{Y|X=x} = \delta_{f(x)}$.
- 2 Decoupled system: $Y \perp\!\!\!\perp X$. Here we set $P_{Y|X=x} = P_Y$.
- 3 Additive noise (convolution): $Y = X + Z$ with $Z \perp\!\!\!\perp X$. This time we choose $P_{Y|X=x}(\cdot) = P_Z(\cdot - x)$. The term convolution corresponds to the fact that the resulting marginal distribution $P_Y = P_X * P_Z$ is a convolution of measures.

The second way is to *disintegrate* a joint distribution $P_{X,Y}$ by conditioning on X , which is denoted simply by $P_{Y|X}$. Specifically, we have the following result [58, Chapter IV, Theorem 2.18]:

Theorem 2.9 (Disintegration). *Suppose $P_{X,Y}$ is a distribution on $\mathcal{X} \times \mathcal{Y}$ with \mathcal{Y} being standard Borel. Then there exists a Markov kernel $K : \mathcal{X} \rightarrow \mathcal{Y}$ so that for any measurable $E \subset \mathcal{X} \times \mathcal{Y}$ and any integrable f we have*

$$\begin{aligned} P_{X,Y}[E] &= \int_{\mathcal{X}} P_X(dx) K(E^x|x), \quad E^x \triangleq \{y : (x,y) \in E\} \quad (2.22) \\ \int_{\mathcal{X} \times \mathcal{Y}} f(x,y) P_{X,Y}(dx dy) &= \int_{\mathcal{X}} P_X(dx) \int_{\mathcal{Y}} f(x,y) K(dy|x). \end{aligned}$$

Note that above we have implicitly used the facts that the slices E^x of E are measurable subsets of \mathcal{Y} for each x and that the function $x \mapsto K(E^x|x)$ is measurable (cf. [58, Chapter I, Prop. 6.8 and 6.9], respectively). We also notice that one joint distribution $P_{X,Y}$ can have many different *versions* of $P_{Y|X}$ differing on a measure-zero set of x 's.

The operation of combining an input distribution on \mathcal{X} and a kernel $K : \mathcal{X} \rightarrow \mathcal{Y}$ as we did in (2.22) is going to appear extensively in this book. We will usually denote it as *multiplication*: Given P_X and kernel $P_{Y|X}$ we can multiply them to obtain $P_{X,Y} \triangleq P_X P_{Y|X}$, which in the discrete case simply means that the joint PMF factorizes as product of marginal and conditional PMFs:

$$P_{X,Y}(x,y) = P_{Y|X}(y|x) P_X(x),$$

and more generally is given by (2.22) with $K = P_{Y|X}$.

Another useful operation will be that of *composition* (marginalization), which we denote by $P_{Y|X} \circ P_X \triangleq P_Y$. In words, this means forming a distribution $P_{X,Y} = P_X P_{Y|X}$ and then computing the marginal P_Y , or, explicitly,

$$P_Y[E] = \int_{\mathcal{X}} P_X(dx) P_{Y|X}(E|x).$$

To denote this (linear) relation between the input P_X and the output P_Y we sometimes also write $P_X \xrightarrow{P_{Y|X}} P_Y$.

We must remark that technical assumptions such as restricting to standard Borel spaces are really necessary for constructing any sensible theory of disintegration/conditioning and multiplication. To emphasize this point we consider a (cautionary!) example involving a pathological measurable space \mathcal{Y} .

2.5 Conditional divergence, chain rule, data-processing inequality 27

Example 2.5 ($X \perp\!\!\!\perp Y$ but $P_{Y|X=x} \ll P_Y$ for all x). Consider \mathcal{X} a unit interval with Borel σ -algebra and \mathcal{Y} a unit interval with the σ -algebra $\sigma\mathcal{Y}$ consisting of all sets which are either countable or have a countable complement. Clearly $\sigma\mathcal{Y}$ is a sub- σ -algebra of Borel one. We define the following kernel $K : \mathcal{X} \rightarrow \mathcal{Y}$:

$$K(A|x) \triangleq 1\{x \in A\}.$$

This is simply saying that Y is produced from X by setting $Y = X$. It should be clear that for every $A \in \sigma\mathcal{Y}$ the map $x \mapsto K(A|x)$ is measurable, and thus K is a valid Markov kernel. Letting $X \sim \text{Unif}(0, 1)$ and using formula (2.22) we can define a joint distribution $P_{X,Y}$. But what is the conditional distribution $P_{Y|X}$? On one hand, clearly we can set $P_{Y|X}(A|x) = K(A|x)$, since this was how $P_{X,Y}$ was constructed. On the other hand, we will show that $P_{X,Y} = P_X P_Y$, i.e. $X \perp\!\!\!\perp Y$ and $X = Y$ at the same time! Indeed, consider any set $E = B \times C \subset \mathcal{X} \times \mathcal{Y}$. We always have $P_{X,Y}[B \times C] = P_X[B \cap C]$. Thus if C is countable then $P_{X,Y}[E] = 0$ and so is $P_X P_Y[E] = 0$. On the other hand, if C^c is countable then $P_X[C] = P_Y[C] = 1$ and $P_{X,Y}[E] = P_X P_Y[E]$ again. Thus, both $P_{Y|X} = K$ and $P_{Y|X} = P_Y$ are valid conditional distributions. But notice that since $P_Y[\{x\}] = 0$, we have $K(\cdot|x) \ll P_Y$ for every $x \in \mathcal{X}$. In particular, the value of $D(P_{Y|X=x} \| P_Y)$ can either be 0 or $+\infty$ for every x depending on the choice of the version of $P_{Y|X}$. It is, thus, advisable to stay within the realm of standard Borel spaces.

We will also need to use the following result extensively. We remind that a σ -algebra is called separable if it is generated by a countable collection of sets. Any standard Borel space's σ -algebra is separable. The following is another useful result about Markov kernels, cf. [58, Chapter 5, Theorem 4.44]:

Theorem 2.10 (Doob's version of Radon-Nikodym Theorem). *Assume that \mathcal{Y} is a measurable space with a separable σ -algebra. Let $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ and $R_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ be two Markov kernels. Suppose that for every x we have $P_{Y|X=x} \ll R_{Y|X=x}$. Then there exists a measurable function $(x, y) \mapsto f(y|x) \geq 0$ such that for every $x \in \mathcal{X}$ and every measurable subset E of \mathcal{Y} ,*

$$P_{Y|X}(E|x) = \int_E f(y|x) R_{Y|X}(dy|x).$$

The meaning of this theorem is that the Radon-Nikodym derivative $\frac{dP_{Y|X=x}}{dR_{Y|X=x}}$ can be made jointly measurable with respect to (x, y) .

2.5 Conditional divergence, chain rule, data-processing inequality

We aim to define the conditional divergence between two Markov kernels. Throughout this chapter we fix a pair of Markov kernels $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$, and also a probability measure P_X on \mathcal{X} . First, let us consider the case of discrete \mathcal{X} . We define the conditional divergence as

$$D(P_{Y|X} \| Q_{Y|X}|P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X=x} \| Q_{Y|X=x}).$$

In order to extend the above definition to more general \mathcal{X} , we need to first understand whether the map $x \mapsto D(P_{Y|X=x} \| Q_{Y|X=x})$ is even measurable.

Lemma 2.11. *Suppose that \mathcal{Y} is standard Borel. The set $A_0 \triangleq \{x : P_{Y|X=x} \ll Q_{Y|X=x}\}$ and the function*

$$x \mapsto D(P_{Y|X=x} \| Q_{Y|X=x})$$

are both measurable.

Proof. Take $R_{Y|X} = \frac{1}{2}P_{Y|X} + \frac{1}{2}Q_{Y|X}$ and define $f_P(y|x) \triangleq \frac{dP_{Y|X=x}}{dR_{Y|X=x}}(y)$ and $f_Q(y|x) \triangleq \frac{dQ_{Y|X=x}}{dR_{Y|X=x}}(y)$. By Theorem 2.11 these can be chosen to be jointly measurable on $\mathcal{X} \times \mathcal{Y}$. Let us define $B_0 \triangleq \{(x, y) : f_P(y|x) > 0, f_Q(y|x) = 0\}$ and its slice $B_0^x = \{y : (x, y) \in B_0\}$. Then note that $P_{Y|X=x} \ll Q_{Y|X=x}$ iff $R_{Y|X=x}[B_0^x] = 0$. In other words, $x \in A_0$ iff $R_{Y|X=x}[B_0^x] = 0$. The measurability of B_0 implies that of $x \mapsto R_{Y|X=x}[B_0^x]$ and thus that of A_0 . Finally, from (2.12) we get that

$$D(P_{Y|X=x} \| Q_{Y|X=x}) = \mathbb{E}_{Y \sim P_{Y|X=x}} \left[\log \frac{f_P(Y|x)}{f_Q(Y|x)} \right], \quad (2.23)$$

which is measurable, e.g. [58, Chapter 1, Prop. 6.9]. \square

With this preparation we can give the following definition.

Definition 2.12 (Conditional divergence). Assuming \mathcal{Y} is standard Borel, define

$$D(P_{Y|X} \| Q_{Y|X}|P_X) \triangleq \mathbb{E}_{x \sim P_X} [D(P_{Y|X=x} \| Q_{Y|X=x})]$$

We observe that as usual in Lebesgue integration it is possible that a conditional divergence is finite even though $D(P_{Y|X=x} \| Q_{Y|X=x}) = \infty$ for some (P_X -negligible set of) x .

Theorem 2.13 (Chain rule). *For any pair of measures $P_{X,Y}$ and $Q_{X,Y}$ we have*

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{Y|X} \| Q_{Y|X}|P_X) + D(P_X \| Q_X), \quad (2.24)$$

regardless of the versions of conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ one chooses.

Proof. First, let us consider the simplest case: \mathcal{X}, \mathcal{Y} are discrete and $Q_{X,Y}(x, y) > 0$ for all x, y . Letting $(X, Y) \sim P_{X,Y}$ we get

$$\begin{aligned} D(P_{X,Y} \| Q_{X,Y}) &= \mathbb{E} \left[\log \frac{P_{X,Y}(X, Y)}{Q_{X,Y}(X, Y)} \right] = \mathbb{E} \left[\log \frac{P_X(X)P_{Y|X}(Y|X)}{Q_X(X)Q_{Y|X}(Y|X)} \right] \\ &= \mathbb{E} \left[\log \frac{P_{Y|X}(Y|X)}{Q_{Y|X}(Y|X)} \right] + \mathbb{E} \left[\log \frac{P_X(X)}{Q_X(X)} \right] \end{aligned}$$

completing the proof.

2.5 Conditional divergence, chain rule, data-processing inequality 29

Next, let us address the general case. If $P_X \ll Q_X$ then $P_{X,Y} \ll Q_{X,Y}$ and both sides of (2.24) are infinity. Thus, we assume $P_X \ll Q_X$ and set $\lambda_P(x) \triangleq \frac{dP_X}{dQ_X}(x)$. Define $f_P(y|x)$, $f_Q(y|x)$ and $R_{Y|X}$ as in the proof of Lemma 2.12. Then we have $P_{X,Y}, Q_{X,Y} \ll R_{X,Y} \triangleq Q_X R_{Y|X}$, and for any measurable E

$$P_{X,Y}[E] = \int_E \lambda_P(x)f_P(y|x)R_{X,Y}(dx dy), \quad Q_{X,Y}[E] = \int_E f_Q(y|x)R_{X,Y}(dx dy).$$

Then from (2.12) we have

$$D(P_{X,Y}\|Q_{X,Y}) = \mathbb{E}_{P_{X,Y}} \left[\text{Log} \frac{f_P(Y|X)\lambda_P(X)}{f_Q(Y|X)} \right]. \quad (2.25)$$

Note the following property of Log: For any $c > 0$

$$\text{Log} \frac{ac}{b} = \log(c) + \text{Log} \frac{a}{b}$$

unless $a = b = 0$. Now, since $P_{X,Y}[f_P(Y|X) > 0, \lambda_P(X) > 0] = 1$, we conclude that $P_{X,Y}$ -almost surely

$$\text{Log} \frac{f_P(Y|X)\lambda_P(X)}{f_Q(Y|X)} = \log \lambda_P(X) + \text{Log} \frac{f_P(Y|X)}{f_Q(Y|X)}.$$

We aim to take the expectation of both sides over $P_{X,Y}$ and invoke linearity of expectation. To ensure that the issue of $\infty - \infty$ does not arise, we notice that the negative part of each term has finite expectation by (2.15). Overall, continuing (2.25) and invoking linearity we obtain

$$D(P_{X,Y}\|Q_{X,Y}) = \mathbb{E}_{P_{X,Y}}[\log \lambda_P(X)] + \mathbb{E}_{P_{X,Y}} \left[\text{Log} \frac{f_P(Y|X)}{f_Q(Y|X)} \right],$$

where the first term equals $D(P_X\|Q_X)$ by (2.12) and the second $D(P_{Y|X}\|Q_{Y|X}|P_X)$ by (2.23) and the definition of conditional divergence. \square

The chain rule has a number of useful corollaries, which we summarize below.

Theorem 2.14 (Properties of Divergence). *Assume that \mathcal{X} and \mathcal{Y} are standard Borel. Then*

(a) *Conditional divergence can be expressed unconditionally:*

$$D(P_{Y|X}\|Q_{Y|X}|P_X) = D(P_X P_{Y|X}\|P_X Q_{Y|X}).$$

(b) *(Monotonicity) $D(P_{X,Y}\|Q_{X,Y}) \geq D(P_Y\|Q_Y)$.*

(c) *(Full chain rule)*

$$D(P_{X_1 \dots X_n}\|Q_{X_1 \dots X_n}) = \sum_{i=1}^n D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}}).$$

In the special case of $Q_{X^n} = \prod_{i=1}^n Q_{X_i}$,

$$D(P_{X_1 \dots X_n}\|Q_{X_1} \dots Q_{X_n}) = D(P_{X_1 \dots X_n}\|P_{X_1} \dots P_{X_n}) + \sum_{i=1}^n D(P_{X_i}\|Q_{X_i})$$

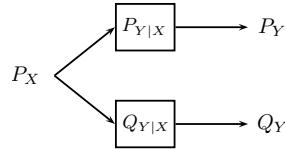
$$\geq \sum_{i=1}^n D(P_{X_i} \| Q_{X_i}), \quad (2.26)$$

where the inequality holds with equality if and only if $P_{X^n} = \prod_{j=1}^n P_{X_j}$.

(d) (Tensorization)

$$D\left(\prod_{j=1}^n P_{X_j} \middle\| \prod_{j=1}^n Q_{X_j}\right) = \sum_{j=1}^n D(P_{X_j} \| Q_{X_j}).$$

(e) (Conditioning increases divergence) Given $P_{Y|X}$, $Q_{Y|X}$ and P_X , let $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = Q_{Y|X} \circ P_X$, as represented by the diagram:



Then $D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_{Y|X} | P_X)$, with equality iff $D(P_{X|Y} \| Q_{X|Y} | P_Y) = 0$.

We remark that as before without the standard Borel assumption even the first property can fail. For example, Example 2.5 shows an example where $P_X P_{Y|X} = P_X Q_{Y|X}$ but $P_{Y|X} \neq Q_{Y|X}$ and $D(P_{Y|X} \| Q_{Y|X} | P_X) = \infty$.

Proof. (a) This follows from the chain rule (2.24) since $P_X = Q_X$.

(b) Apply (2.24), with X and Y interchanged and use the fact that conditional divergence is non-negative.

(c) By telescoping $P_{X^n} = \prod_{i=1}^n P_{X_i|X^{i-1}}$ and $Q_{X^n} = \prod_{i=1}^n Q_{X_i|X^{i-1}}$.

(d) Apply (c).

(e) The inequality follows from (a) and (b). To get conditions for equality, notice that by the chain rule for D :

$$\begin{aligned} D(P_{X,Y} \| Q_{X,Y}) &= D(P_{Y|X} \| Q_{Y|X} | P_X) + \underbrace{D(P_X \| P_X)}_{=0} \\ &= D(P_{X|Y} \| Q_{X|Y} | P_Y) + D(P_Y \| Q_Y). \end{aligned} \quad \square$$

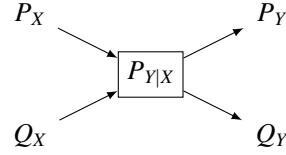
Some remarks are in order:

- There is a nice interpretation of the full chain rule as a decomposition of the “distance” from P_{X^n} to Q_{X^n} as a sum of “distances” between intermediate distributions, cf. Ex. I.33.
- In general, $D(P_{X,Y} \| Q_{X,Y})$ and $D(P_X \| Q_X) + D(P_Y \| Q_Y)$ are incomparable. For example, if $X = Y$ under P and Q , then $D(P_{X,Y} \| Q_{X,Y}) = D(P_X \| Q_X) < 2D(P_X \| Q_X)$. Conversely, if $P_X = Q_X$ and $P_Y = Q_Y$ but $P_{X,Y} \neq Q_{X,Y}$ we have $D(P_{X,Y} \| Q_{X,Y}) > 0 = D(P_X \| Q_X) + D(P_Y \| Q_Y)$.

2.5 Conditional divergence, chain rule, data-processing inequality 31

The following result, known as the *Data-Processing Inequality (DPI)*, is an important principle in all of information theory. In many ways, it underpins the whole concept of information. The intuitive interpretation is that it is easier to distinguish two distributions using clean (resp. full) data as opposed to noisy (resp. partial) data. DPI is a recurring theme in this book, and later we will study DPI for other information measures such as those for mutual information and f -divergences.

Theorem 2.15 (DPI for KL divergence). *Let $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$, as represented by the diagram:*



Then

$$D(P_Y\|Q_Y) \leq D(P_X\|Q_X), \quad (2.27)$$

with equality if and only if $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$.

Proof. This follows from either the chain rule or monotonicity:

$$\begin{aligned} D(P_{X,Y}\|Q_{X,Y}) &= \underbrace{D(P_{Y|X}\|Q_{Y|X}|P_X)}_{=0} + D(P_X\|Q_X) \\ &= D(P_{X|Y}\|Q_{X|Y}|P_Y) + D(P_Y\|Q_Y) \end{aligned}$$

□

Corollary 2.16 (Divergence under deterministic transformation). *Let $Y = f(X)$. Then $D(P_Y\|Q_Y) \leq D(P_X\|Q_X)$, with equality iff f is one-to-one.*

Note that $D(P_{f(X)}\|Q_{f(X)}) = D(P_X\|Q_X)$ does not imply that f is one-to-one; as an example, consider $P_X = \text{Gaussian}$, $Q_X = \text{Laplace}$, $Y = |X|$. In fact, the equality happens precisely when $f(X)$ is a *sufficient statistic* for testing P against Q ; in other words, there is no loss of information in summarizing X into $f(X)$ as far as testing these two hypotheses is concerned. See Example 3.8 for details.

A particular useful application of Corollary 2.1 is when we take f to be an indicator function:

Corollary 2.17 (Large deviations estimate). *For any subset $E \subset \mathcal{X}$ we have*

$$d(P_X[E]\|Q_X[E]) \leq D(P_X\|Q_X),$$

where $d(\cdot\|\cdot)$ is the binary divergence function in (2.6).

Proof. Consider $Y = 1_{\{X \in E\}}$. □

This method will be highly useful in large deviations theory which studies rare events (Section 14.5 and Section 15.2), where we apply Corollary 2.2 to an event E which is highly likely under P but highly unlikely under Q .

2.6* Local behavior of divergence and Fisher information

As we shall see in Section 4.4, KL divergence is in general not continuous. Nevertheless, it is reasonable to expect that the functional $D(P\|Q)$ vanishes when P approaches Q “smoothly”. Due to the smoothness and strict convexity of $x \log x$ at $x = 1$, it is then also natural to expect that this functional decays “quadratically”. In this section we examine this question first along the linear interpolation between P and Q , then, more generally, in smooth parametrized families of distributions. These properties will be extended to more general divergences later in Sections 7.10 and 7.11.

2.6.1* Local behavior of divergence for mixtures

Let $0 \leq \lambda \leq 1$ and consider $D(\lambda P + \bar{\lambda} Q \| Q)$, which vanishes as $\lambda \rightarrow 0$. Next, we show that this decay is always sublinear.

Proposition 2.18. *When $D(P\|Q) < \infty$, the one-sided derivative at $\lambda = 0$ vanishes:*

$$\frac{d}{d\lambda} \Big|_{\lambda=0} D(\lambda P + \bar{\lambda} Q \| Q) = 0$$

If we exchange the arguments, the criterion is even simpler:

$$\frac{d}{d\lambda} \Big|_{\lambda=0} D(Q \| \lambda P + \bar{\lambda} Q) = 0 \iff P \ll Q \quad (2.28)$$

Proof.

$$\frac{1}{\lambda} D(\lambda P + \bar{\lambda} Q \| Q) = \mathbb{E}_Q \left[\frac{1}{\lambda} (\lambda f + \bar{\lambda}) \log(\lambda f + \bar{\lambda}) \right]$$

where $f = \frac{dP}{dQ}$. As $\lambda \rightarrow 0$ the function under expectation decreases to $(f - 1) \log e$ monotonically. Indeed, the function

$$\lambda \mapsto g(\lambda) \triangleq (\lambda f + \bar{\lambda}) \log(\lambda f + \bar{\lambda})$$

is convex and equals zero at $\lambda = 0$. Thus $\frac{g(\lambda)}{\lambda}$ is increasing in λ . Moreover, by the convexity of $x \mapsto x \log x$:

$$\frac{1}{\lambda} (\lambda f + \bar{\lambda}) (\log(\lambda f + \bar{\lambda})) \leq \frac{1}{\lambda} (\lambda f \log f + \bar{\lambda} 1 \log 1) = f \log f$$

2.6* Local behavior of divergence and Fisher information 33

and by assumption $f \log f$ is Q -integrable. Thus the Monotone Convergence Theorem applies.

To prove (2.28) first notice that if $P \ll Q$ then there is a set E with $p = P[E] > 0 = Q[E]$. Applying data-processing for divergence to $X \mapsto 1_E(X)$, we get

$$D(Q\|\lambda P + \bar{\lambda}Q) \geq d(0\|\lambda p) = \log \frac{1}{1 - \lambda p}$$

and derivative is non-zero. If $P \ll Q$, then let $f = \frac{dP}{dQ}$ and notice simple inequalities

$$\log \bar{\lambda} \leq \log(\bar{\lambda} + \lambda f) \leq \lambda(f - 1) \log e.$$

Dividing by λ and assuming $\lambda < \frac{1}{2}$ we get for some absolute constants c_1, c_2 :

$$\left| \frac{1}{\lambda} \log(\bar{\lambda} + \lambda f) \right| \leq c_1 f + c_2.$$

Thus, by the dominated convergence theorem we get

$$\frac{1}{\lambda} D(Q\|\lambda P + \bar{\lambda}Q) = - \int dQ \left(\frac{1}{\lambda} \log(\bar{\lambda} + \lambda f) \right) \xrightarrow{\lambda \rightarrow 0} \int dQ(1 - f) = 0.$$

□

Remark 2.2. More generally, under suitable technical conditions,

$$\frac{d}{d\lambda} \Big|_{\lambda=0} D(\lambda P + \bar{\lambda}Q\|R) = \mathbb{E}_P \left[\log \frac{dQ}{dR} \right] - D(Q\|R)$$

and

$$\frac{d}{d\lambda} \Big|_{\lambda=0} D(\bar{\lambda}P_1 + \lambda Q_1\|\bar{\lambda}P_0 + \lambda Q_0) = \mathbb{E}_{Q_1} \left[\log \frac{dP_1}{dP_0} \right] - D(P_1\|P_0) + \mathbb{E}_{P_1} \left[1 - \frac{dQ_0}{dP_0} \right] \log e.$$

The main message of Proposition 2.17 is that the function

$$\lambda \mapsto D(\lambda P + \bar{\lambda}Q\|Q),$$

is $o(\lambda)$ as $\lambda \rightarrow 0$. In fact, in most cases it is quadratic in λ . To make a precise statement, we need to define the concept of χ^2 -divergence – a version of f -divergence (see Chapter 7):

$$\chi^2(P\|Q) \triangleq \int dQ \left(\frac{dP}{dQ} - 1 \right)^2.$$

This is a popular dissimilarity measure between P and Q , frequently used in statistics. It has many important properties, but we will only mention that χ^2 dominates KL-divergence (cf. (7.31)):

$$D(P\|Q) \leq \log(1 + \chi^2(P\|Q)).$$

Our second result about the local behavior of KL-divergence is the following (see Section 7.10 for generalizations):

Proposition 2.19 (KL is locally χ^2 -like). *We have*

$$\liminf_{\lambda \rightarrow 0} \frac{1}{\lambda^2} D(\lambda P + \bar{\lambda} Q \| Q) = \frac{\log e}{2} \chi^2(P \| Q), \quad (2.29)$$

where both sides are finite or infinite simultaneously.

Proof. First, we assume that $\chi^2(P \| Q) < \infty$ and prove

$$D(\lambda P + \bar{\lambda} Q \| Q) = \frac{\lambda^2 \log e}{2} \chi^2(P \| Q) + o(\lambda^2), \quad \lambda \rightarrow 0.$$

To that end notice that

$$D(P \| Q) = \mathbb{E}_Q \left[g \left(\frac{dP}{dQ} \right) \right],$$

where

$$g(x) \triangleq x \log x - (x - 1) \log e.$$

Note that $x \mapsto \frac{g(x)}{(x-1)^2 \log e} = \int_0^1 \frac{sds}{x(1-s)+s}$ is decreasing in x on $(0, \infty)$. Therefore

$$0 \leq g(x) \leq (x - 1)^2 \log e,$$

and hence

$$0 \leq \frac{1}{\lambda^2} g \left(\bar{\lambda} + \lambda \frac{dP}{dQ} \right) \leq \left(\frac{dP}{dQ} - 1 \right)^2 \log e.$$

By the dominated convergence theorem (which is applicable since $\chi^2(P \| Q) < \infty$) we have

$$\lim_{\lambda \rightarrow 0} \frac{1}{\lambda^2} \mathbb{E}_Q \left[g \left(\bar{\lambda} + \lambda \frac{dP}{dQ} \right) \right] = \frac{g''(1)}{2} \mathbb{E}_Q \left[\left(\frac{dP}{dQ} - 1 \right)^2 \right] = \frac{\log e}{2} \chi^2(P \| Q).$$

Second, we show that unconditionally

$$\liminf_{\lambda \rightarrow 0} \frac{1}{\lambda^2} D(\lambda P + \bar{\lambda} Q \| Q) \geq \frac{\log e}{2} \chi^2(P \| Q). \quad (2.30)$$

Indeed, this follows from Fatou's lemma:

$$\liminf_{\lambda \rightarrow 0} \mathbb{E}_Q \left[\frac{1}{\lambda^2} g \left(\bar{\lambda} + \lambda \frac{dP}{dQ} \right) \right] \geq \mathbb{E}_Q \left[\liminf_{\lambda \rightarrow 0} g \left(\bar{\lambda} + \lambda \frac{dP}{dQ} \right) \right] = \frac{\log e}{2} \chi^2(P \| Q).$$

Therefore, from (2.30) we conclude that if $\chi^2(P \| Q) = \infty$ then so is the LHS of (2.29). \square

2.6.2* Parametrized family

Extending the setting of Section 2.6.1*, consider a parametrized set of distributions $\{P_\theta : \theta \in \Theta\}$ where the parameter space Θ is an open subset of \mathbb{R}^d . Furthermore, suppose that distribution P_θ are all given in the form of

$$P_\theta(dx) = p_\theta(x)\mu(dx),$$

2.6* Local behavior of divergence and Fisher information 35

where μ is some common dominating measure (e.g. Lebesgue or counting measure). If for each fixed x , the density $p_\theta(x)$ depends smoothly on θ , one can define the *Fisher information matrix* with respect to the parameter θ as

$$J_F(\theta) \triangleq \mathbb{E}_\theta [VV^\top], \quad V \triangleq \nabla_\theta \ln p_\theta(X), \quad (2.31)$$

where \mathbb{E}_θ is with respect to $X \sim P_\theta$. In particular, V is known as the *score*.

Under suitable regularity conditions, we have the identity

$$\mathbb{E}_\theta[V] = 0 \quad (2.32)$$

and several equivalent expressions for the Fisher information matrix:

$$\begin{aligned} J_F(\theta) &= \operatorname{cov}_\theta(V) \\ &= 4 \int \mu(dx) (\nabla_\theta \sqrt{p_\theta(x)}) (\nabla_\theta \sqrt{p_\theta(x)})^\top \\ &= -\mathbb{E}_\theta[\operatorname{Hess}_\theta(\ln p_\theta(X))], \end{aligned}$$

where the last identity is obtained by differentiating (2.32) with respect to each θ_j .

The significance of Fisher information matrix arises from the fact that it gauges the local behaviour of divergence for smooth parametric families. Namely, we have (again under suitable technical conditions):²

$$D(P_{\theta_0} \| P_{\theta_0 + \xi}) = \frac{\log e}{2} \xi^\top J_F(\theta_0) \xi + o(\|\xi\|^2), \quad (2.33)$$

which is obtained by integrating the Taylor expansion:

$$\ln p_{\theta_0 + \xi}(x) = \ln p_{\theta_0}(x) + \xi^\top \nabla_\theta \ln p_{\theta_0}(x) + \frac{1}{2} \xi^\top \operatorname{Hess}_\theta(\ln p_{\theta_0}(x)) \xi + o(\|\xi\|^2).$$

We will establish this fact rigorously later in Section 7.11. Property (2.33) is of paramount importance in statistics. We should remember it as: *Divergence is locally quadratic on the parameter space, with Hessian given by the Fisher information matrix*. Note that for the Gaussian location model $P_\theta = \mathcal{N}(\theta, \Sigma)$, (2.33) is in fact exact with $J_F(\theta) \equiv \Sigma^{-1}$ – cf. Example 2.2.

As another example, note that Proposition 2.19 is a special case of (2.33) by considering $P_\lambda = \bar{\lambda}Q + \lambda P$ parametrized by $\lambda \in [0, 1]$. In this case, the Fisher information at $\lambda = 0$ is simply $\chi^2(P \| Q)$. Nevertheless, Proposition 2.19 is completely general while the asymptotic expansion (2.33) is not without regularity conditions (see Section 7.11).

Remark 2.3. Some useful properties of Fisher information are as follows:

² To illustrate the subtlety here, consider a scalar location family, i.e. $p_\theta(x) = f_0(x - \theta)$ for some density f_0 . In this case Fisher information $J_F(\theta_0) = \int \frac{(f'_0)^2}{f_0}$ does not depend on θ_0 and is well-defined even for compactly supported f_0 , provided f'_0 vanishes at the endpoints sufficiently fast. But at the same time the left-hand side of (2.33) is infinite for any $\xi > 0$. In such cases, a better interpretation for Fisher information is as the coefficient of the expansion $D(P_{\theta_0} \| \frac{1}{2}P_{\theta_0} + \frac{1}{2}P_{\theta_0 + \xi}) = \frac{\xi^2}{8} J_F(\theta_0) + o(\xi^2)$. We will discuss this in more detail in Section 7.11.

- Reparametrization: It can be seen that if one introduces another parametrization $\tilde{\theta} \in \tilde{\Theta}$ by means of a smooth invertible map $\tilde{\Theta} \rightarrow \Theta$, then Fisher information matrix changes as

$$J_F(\tilde{\theta}) = A^\top J_F(\theta) A, \quad (2.34)$$

where $A = \frac{d\theta}{d\tilde{\theta}}$ is the Jacobian of the map. So we can see that J_F transforms similarly to the metric tensor in Riemannian geometry. This idea can be used to define a Riemannian metric on the parameter space Θ , called the Fisher-Rao metric. This is explored in a field known as information geometry [11].

- Additivity: Suppose we are given a sample of n iid observations $X^n \stackrel{\text{i.i.d.}}{\sim} P_\theta$. As such, consider the parametrized family of product distributions $\{P_\theta^{\otimes n} : \theta \in \Theta\}$, whose Fisher information matrix is denoted by $J_F^{\otimes n}(\theta)$. In this case, the score is an iid sum. Applying (2.31) and (2.32) yields

$$J_F^{\otimes n}(\theta) = n J_F(\theta). \quad (2.35)$$

Example 2.6. Let $P_\theta = (\theta_0, \dots, \theta_d)$ be a probability distribution on the finite alphabet $\{0, \dots, d\}$. We will take $\theta = (\theta_1, \dots, \theta_d)$ as the free parameter and set $\theta_0 = 1 - \sum_{i=1}^d \theta_i$. So all derivatives are with respect to $\theta_1, \dots, \theta_d$ only. Then we have

$$p_\theta(i) = \begin{cases} \theta_i, & i = 1, \dots, d \\ 1 - \sum_{i=1}^d \theta_i, & i = 0 \end{cases}$$

and for Fisher information matrix we get

$$J_F(\theta) = \text{diag}\left(\frac{1}{\theta_1}, \dots, \frac{1}{\theta_d}\right) + \frac{1}{1 - \sum_{i=1}^d \theta_i} \mathbf{1} \mathbf{1}^\top, \quad (2.36)$$

where $\mathbf{1}$ is the $d \times 1$ vector of all ones. For future references (see Sections 29.4 and 13.4*), we also compute the inverse and determinant of $J_F(\theta)$. By the matrix inversion lemma $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$, we have

$$J_F^{-1}(\theta) = \text{diag}(\theta) - \theta \theta^\top. \quad (2.37)$$

For the determinant, notice that $\det(A + xy^\top) = \det A \cdot \det(I + A^{-1}xy^\top) = \det A \cdot (1 + y^\top A^{-1}x)$, where we used the identity $\det(I + AB) = \det(I + BA)$. Thus, we have

$$\det J_F(\theta) = \prod_{i=0}^d \frac{1}{\theta_i}. \quad (2.38)$$

3 Mutual information

After technical preparations in previous chapters we define perhaps the most famous concept in the entire field of information theory, the *mutual information*. It was originally defined by Shannon, although the name was coined later by Robert Fano¹. It has two equivalent expressions (as a KL divergence and as difference of entropies), both having its merits. In this chapter, we prove first properties of mutual information (non-negativity, chain rule and the data-processing inequality). While defining conditional information, we also introduce the language of *directed graphical models*, and connect the equality case in the data-processing inequality with Fisher's concept of sufficient statistics.

3.1 Mutual information

Mutual information was first defined by Shannon to measure the decrease in entropy of a random quantity following the observation of another (correlated) random quantity. Unlike the concept of entropy itself, which was well-known by then in statistical mechanics, the mutual information was new and revolutionary and had no analogs in science. Today, however, it is preferred to define mutual information in a different form (proposed in [268, Appendix 7]).

Definition 3.1 (Mutual information). For a pair of random variables X and Y we define

$$I(X; Y) = D(P_{X,Y} \| P_X P_Y).$$

The intuitive interpretation of mutual information is that $I(X; Y)$ measures the dependency between X and Y by comparing their joint distribution to the product of the marginals in the KL divergence, which, as we show next, is also equivalent to comparing the conditional distribution to the unconditional.

The way we defined $I(X; Y)$ it is a functional of the joint distribution $P_{X,Y}$. However, it is also rather fruitful to look at it as a functional of the pair $(P_X, P_{Y|X})$ – more on this in Section 5.1.

In general, the divergence $D(P_{X,Y} \| P_X P_Y)$ should be evaluated using the general definition (2.4). Note that $P_{X,Y} \ll P_X P_Y$ need not always hold. Let us consider the following examples, though.

¹ Professor of electrical engineering at MIT, who developed the first course on information theory and as part of it formalized and rigorized much of Shannon's ideas. Most famously, he showed the “converse part” of the noisy channel coding theorem, see Section 17.4.

Example 3.1. If $X = Y \sim N(0, 1)$ then $P_{X,Y} \ll P_X P_Y$ and $I(X; Y) = \infty$. This reflects our intuition that X contains an “infinite” amount of information requiring infinitely many bits to describe. On the other hand, if even one of X or Y is discrete, then we *always* have $P_{X,Y} \ll P_X P_Y$. Indeed, consider any $E \subset \mathcal{X} \times \mathcal{Y}$ measurable in the product sigma algebra with $P_{X,Y}(E) > 0$. Since $P_{X,Y}(E) = \sum_{x \in S} \mathbb{P}[(X, Y) \in E, X = x]$, there exists some $x_0 \in S$ such that $P_Y(E^{x_0}) \geq \mathbb{P}[X = x_0, Y \in E^{x_0}] > 0$, where $E^{x_0} \triangleq \{y : (x_0, y) \in E\}$ is a section of E (measurable for every x_0). But then $P_X P_Y(E) \geq P_X P_Y(\{x_0\} \times E^{x_0}) = P_X(\{x_0\}) P_Y(E^{x_0}) > 0$, implying that $P_{X,Y} \ll P_X P_Y$.

Theorem 3.2 (Properties of mutual information).

(a) (*Mutual information as conditional divergence*) Whenever \mathcal{Y} is standard Borel,

$$I(X; Y) = D(P_{Y|X} \| P_Y | P_X). \quad (3.1)$$

(b) (*Symmetry*) $I(X; Y) = I(Y; X)$

(c) (*Positivity*) $I(X; Y) \geq 0$; $I(X; Y) = 0$ iff $X \perp\!\!\!\perp Y$

(d) For any function f , $I(f(X); Y) \leq I(X; Y)$. If f is one-to-one (with a measurable inverse), then $I(f(X); Y) = I(X; Y)$.

(e) (*More data \Rightarrow More information*) $I(X_1, X_2; Z) \geq I(X_1; Z)$

Proof. (a) This follows from Theorem 2.15(a) with $Q_{Y|X} = P_Y$.

(b) Consider a Markov kernel K sending $(x, y) \mapsto (y, x)$. This kernel sends measure $P_{X,Y} \xrightarrow{K} P_{Y,X}$ and $P_X P_Y \xrightarrow{K} P_Y P_X$. Therefore, from the DPI Theorem 2.16 applied to this kernel we get

$$D(P_{X,Y} \| P_X P_Y) \geq D(P_{Y,X} \| P_Y P_X).$$

Applying this argument again, shows that inequality is in fact equality.

(c) This is just $D \geq 0$ from Theorem 2.3.

(d) Consider a Markov kernel K sending $(x, y) \mapsto (f(x), y)$. This kernel sends measure $P_{X,Y} \xrightarrow{K} P_{f(X),Y}$ and $P_X P_Y \xrightarrow{K} P_{f(X)} P_Y$. Therefore, from the DPI Theorem 2.16 applied to this kernel we get

$$D(P_{X,Y} \| P_X P_Y) \geq D(P_{f(X),Y} \| P_{f(X)} P_Y).$$

It is clear that the two sides correspond to the two mutual informations. For bijective f , simply apply the inequality to f and f^{-1} .

(e) Apply (d) with $f(X_1, X_2) = X_1$. □

Proof. (a) $I(X; Y) = \mathbb{E} \log \frac{P_{X,Y}}{P_X P_Y} = \mathbb{E} \log \frac{P_{Y|X}}{P_Y} = \mathbb{E} \log \frac{P_{X|Y}}{P_X}$.

(b) Apply data-processing inequality twice to the map $(x, y) \rightarrow (y, x)$ to get $D(P_{X,Y} \| P_X P_Y) = D(P_{Y,X} \| P_Y P_X)$.

(c) By definition and Theorem 2.3.

3.1 Mutual information 39

- (d) We will use the data-processing inequality of mutual information (to be proved shortly in Theorem 3.7(c)). For bijective f , consider the chain of data processing: $(x, y) \mapsto (f(x), y) \mapsto (f^{-1}(f(x)), y)$. Then $I(X; Y) \geq I(f(X); Y) \geq I(f^{-1}(f(X)); Y) = I(X; Y)$.
- (e) Apply (d) with $f(X_1, X_2) = X_1$. \square

Of the results above, the one we will use the most is (3.1). Note that it implies that $D(P_{X,Y} \| P_X P_Y) < \infty$ if and only if

$$x \mapsto D(P_{Y|X=x} \| P_Y)$$

is P_X -integrable. This property has a counterpart in terms of absolute continuity, as follows.

Lemma 3.3. *Let \mathcal{Y} be standard Borel. Then*

$$P_{X,Y} \ll P_X P_Y \iff P_{Y|X=x} \ll P_Y \text{ for } P_X\text{-a.e. } x$$

Proof. Suppose $P_{X,Y} \ll P_X P_Y$. We need to prove that *any* version of the conditional probability satisfies $P_{Y|X=x} \ll P_Y$ for almost every x . Note, however, that if we prove this for *some* version $\tilde{P}_{Y|X}$ then the statement for any version follows, since $P_{Y|X=x} = \tilde{P}_{Y|X=x}$ for P_X -a.e. x . (This measure-theoretic fact can be derived from the chain rule (2.24): since $P_X \tilde{P}_{Y|X} = P_{X,Y} = P_X P_{Y|X}$ we must have $0 = D(P_{X,Y} \| P_X P_Y) = D(\tilde{P}_{Y|X} \| P_{Y|X} | P_X) = \mathbb{E}_{x \sim P_X}[D(\tilde{P}_{Y|X=x} \| P_{Y|X=x})]$, implying the stated fact.) So let $g(x, y) = \frac{dP_{X,Y}}{dP_X P_Y}(x, y)$ and $\rho(x) \triangleq \int_{\mathcal{Y}} g(x, y) P_Y(dy)$. Fix any set $E \subset \mathcal{X}$ and notice

$$P_X[E] = \int_{\mathcal{X} \times \mathcal{Y}} 1_E(x) g(x, y) P_X(dx) P_Y(dy) = \int_{\mathcal{X}} 1_E(x) \rho(x) P_X(dx).$$

On the other hand, we also have $P_X[E] = \int 1_E dP_X$, which implies $\rho(x) = 1$ for P_X -a.e. x . Now define

$$\tilde{P}_{Y|X}(dy|x) = \begin{cases} g(x, y) P_Y(dy), & \rho(x) = 1 \\ P_Y(dy), & \rho(x) \neq 1. \end{cases}$$

Directly plugging $\tilde{P}_{Y|X}$ into (2.22) shows that $\tilde{P}_{Y|X}$ does define a valid version of the conditional probability of Y given X . Since by construction $\tilde{P}_{Y|X=x} \ll P_Y$ for every x , the result follows.

Conversely, let $P_{Y|X}$ be a kernel such that $P_X[E] = 1$, where $E = \{x : P_{Y|X=x} \ll P_Y\}$ (recall that E is measurable by Lemma 2.12). Define $\tilde{P}_{Y|X=x} = P_{Y|X=x}$ if $x \in E$ and $\tilde{P}_{Y|X=x} = P_Y$, otherwise. By construction $P_X \tilde{P}_{Y|X} = P_X P_{Y|X} = P_{X,Y}$ and $\tilde{P}_{Y|X=x} \ll P_Y$ for every x . Thus, by Theorem 2.11 there exists a jointly measurable $f(y|x)$ such that

$$\tilde{P}_{Y|X}(dy|x) = f(y|x) P_Y(dy),$$

and, thus, by (2.22)

$$P_{X,Y}[E] = \int_E f(y|x) P_Y(dy) P_X(dx),$$

implying that $P_{X,Y} \ll P_X P_Y$. \square

3.2 Mutual information as difference of entropies

As promised, we next introduce a different point of view on $I(X; Y)$, namely as a difference of entropies. This (conditional entropy) point of view of Shannon emphasizes that $I(X; Y)$ is also measuring the change in the spread or uncertainty of the distribution of X following the observation of Y .

Theorem 3.4.

$$(a) I(X; X) = \begin{cases} H(X) & X \text{ discrete} \\ +\infty & \text{otherwise.} \end{cases}$$

(b) If X is discrete, then

$$I(X; Y) + H(X|Y) = H(X). \quad (3.2)$$

Consequently, either $H(X|Y) = H(X) = \infty$,² or $H(X|Y) < \infty$ and

$$I(X; Y) = H(X) - H(X|Y). \quad (3.3)$$

(c) If both X and Y are discrete, then

$$I(X; Y) + H(X, Y) = H(X) + H(Y),$$

so that whenever $H(X, Y) < \infty$ we have

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

(d) Similarly, if X, Y are real-valued random vectors with a joint PDF, then

$$I(X; Y) = h(X) + h(Y) - h(X, Y)$$

provided that $h(X, Y) < \infty$. If X has a marginal PDF p_X and a conditional PDF $p_{X|Y}(x|y)$, then

$$I(X; Y) = h(X) - h(X|Y),$$

provided $h(X|Y) < \infty$.

(e) If X or Y are discrete then $I(X; Y) \leq \min(H(X), H(Y))$, with equality iff $H(X|Y) = 0$ or $H(Y|X) = 0$, or, equivalently, iff one is a deterministic function of the other.

Proof. (a) By Theorem 3.2.(a), $I(X; X) = D(P_{X|X}\|P_X|P_X) = \mathbb{E}_{x \sim X} D(\delta_x\|P_X)$. If P_X is discrete, then $D(\delta_x\|P_X) = \log \frac{1}{P_X(x)}$ and $I(X; X) = H(X)$. If P_X is not discrete, let $\mathcal{A} = \{x : P_X(x) > 0\}$ denote the set of atoms of P_X . Let $\Delta = \{(x, x) : x \notin \mathcal{A}\} \subset \mathcal{X} \times \mathcal{X}$. (Δ is measurable since it's

² This is indeed possible if one takes $Y = 0$ (constant) and X from Example 1.3, demonstrating that (3.3) does not always hold.

3.2 Mutual information as difference of entropies 41

the intersection of $\mathcal{A}^c \times \mathcal{A}^c$ with the diagonal $\{(x, x) : x \in \mathcal{X}\}$.) Then $P_{X,X}(\Delta) = P_X(\mathcal{A}^c) > 0$ but since

$$(P_X \times P_X)(E) \triangleq \int_{\mathcal{X}} P_X(dx_1) \int_{\mathcal{X}} P_X(dx_2) \mathbf{1}\{(x_1, x_2) \in E\}$$

we have by taking $E = \Delta$ that $(P_X \times P_X)(\Delta) = 0$. Thus $P_{X,X} \ll P_X \times P_X$ and thus by definition

$$I(X; X) = D(P_{X,X} \| P_X P_X) = +\infty.$$

- (b) Since X is discrete there exists a countable set S such that $\mathbb{P}[X \in S] = 1$, and for any $x_0 \in S$ we have $\mathbb{P}[X = x_0] > 0$. Let λ be a counting measure on S and let $\mu = \lambda \times P_Y$, so that $P_X P_Y \ll \mu$. As shown in Example 3.1 we also have $P_{X,Y} \ll \mu$. Furthermore, $f_P(x, y) \triangleq \frac{dP_{X,Y}}{d\mu}(x, y) = p_{X|Y}(x|y)$, where the latter denotes conditional pmf of X given Y (which is a proper pmf for almost every y , since $\mathbb{P}[X \in S | Y = y] = 1$ for a.e. y). We also have $f_Q(x, y) = \frac{dP_X P_Y}{d\mu}(x, y) = \frac{dP_X}{d\lambda}(x) = p_X(x)$, where the latter is an unconditional pmf of X . Note that by definition of Radon-Nikodym derivatives we have

$$\mathbb{E}[p_{X|Y}(x_0 | Y)] = p_X(x_0). \quad (3.4)$$

Next, according to (2.12) we have

$$I(X; Y) = \mathbb{E} \left[\log \frac{f_P(X, Y)}{f_Q(X, Y)} \right] = \mathbb{E}_{y \sim P_Y} \sum_{x \in S} \left[p_{X|Y}(x|y) \log \frac{p_{X|Y}(x|y)}{p_X(x)} \right].$$

Note that $P_{X,Y}$ -almost surely both $p_{X|Y}(X|Y) > 0$ and $p_X(x) > 0$, so we can replace Log with log in the above. On the other hand,

$$H(X|Y) = \mathbb{E}_{y \sim P_Y} \sum_{x \in S} \left[p_{X|Y}(x|y) \log \frac{1}{p_{X|Y}(x|y)} \right].$$

Adding these two expressions, we obtain

$$\begin{aligned} I(X; Y) + H(X|Y) &\stackrel{(a)}{=} \mathbb{E}_{y \sim P_Y} \sum_{x \in S} \left[p_{X|Y}(x|y) \log \frac{1}{p_X(x)} \right] \\ &\stackrel{(b)}{=} \sum_{x \in S} \mathbb{E}_{y \sim P_Y} [p_{X|Y}(x|y)] \log \frac{1}{p_X(x)} \stackrel{(c)}{=} \mathbb{E} \left[\log \frac{1}{P_X(X)} \right] \triangleq H(X), \end{aligned}$$

where in (a) we used linearity of Lebesgue integral $\mathbb{E}_{P_Y} \sum_x$, in (b) we interchange \mathbb{E} and \sum via Fubini; and (c) holds due to (3.4).

- (c) Simply add $H(Y)$ to both sides of (3.2) and use the chain rule for H from (1.2).
- (d) These arguments are similar to discrete case, except that counting measure is replaced with Lebesgue. We leave the details as an exercise.
- (e) Follows from (b). \square

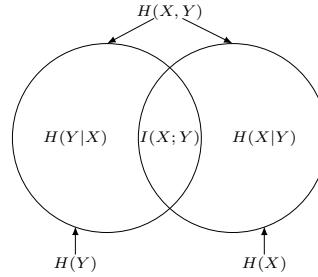
From (3.2) we deduce the following result, which was previously shown in Theorem 1.5(d).

Corollary 3.5 (Conditioning reduces entropy). *For discrete X , $H(X|Y) \leq H(X)$, with equality iff $X \perp\!\!\!\perp Y$.*

Proof. If $H(X) = \infty$ then there is nothing to prove. Otherwise, apply (3.2). \square

Thus, the intuition behind the last corollary (and an important innovation of Shannon) is to give meaning to the amount of entropy reduction (mutual information). It is important to note that conditioning reduces entropy *on average*, not per realization. Indeed, take $X = U \text{ OR } Y$, where $U, Y \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. Then $X \sim \text{Ber}(\frac{3}{4})$ and $H(X) = h(\frac{1}{4}) < 1 \text{ bit} = H(X|Y=0)$, i.e., conditioning on $Y=0$ increases entropy. But *on average*, $H(X|Y) = \mathbb{P}[Y=0]H(X|Y=0) + \mathbb{P}[Y=1]H(X|Y=1) = \frac{1}{2} \text{ bits} < H(X)$, by the strong concavity of $h(\cdot)$.

Remark 3.1 (Information, entropy, and Venn diagrams). For discrete random variables, the following Venn diagram illustrates the relationship between entropy, conditional entropy, joint entropy, and mutual information from Theorem 3.4(b) and (c).



Applying analogously the inclusion-exclusion principle to three variables X_1, X_2, X_3 , we see that the triple intersection corresponds to

$$H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_1, X_3) + H(X_1, X_2, X_3) \quad (3.5)$$

which is sometimes denoted by $I(X_1; X_2; X_3)$. It can be both positive and negative (why?).

In general, one can treat random variables as sets (so that the X_i corresponds to set E_i and the pair (X_1, X_2) corresponds to $E_1 \cup E_2$). Then we can define a unique signed measure μ on the finite algebra generated by these sets so that every information quantity is found by replacing

$$I/H \rightarrow \mu \quad ; \rightarrow \cap \quad , \rightarrow \cup \quad | \rightarrow \setminus .$$

As an example, we have

$$H(X_1|X_2, X_3) = \mu(E_1 \setminus (E_2 \cup E_3)), \quad (3.6)$$

$$I(X_1, X_2; X_3|X_4) = \mu(((E_1 \cup E_2) \cap E_3) \setminus E_4). \quad (3.7)$$

By inclusion-exclusion, the quantity in (3.5) corresponds to $\mu(E_1 \cap E_2 \cap E_3)$, which explains why μ is not necessarily a positive measure. For an extensive discussion, see [79, Chapter 1.3].

3.3 Examples of computing mutual information

Below we demonstrate how to compute I in both continuous and discrete settings.

3.3 Examples of computing mutual information 43

Example 3.2 (Bivariate Gaussian). Let X, Y be jointly Gaussian. Then

$$I(X; Y) = \frac{1}{2} \log \frac{1}{1 - \rho_{X,Y}^2} \quad (3.8)$$

where $\rho_{X,Y} \triangleq \frac{\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\sigma_X \sigma_Y} \in [-1, 1]$ is the correlation coefficient; see Fig. 3.1 for a plot. To

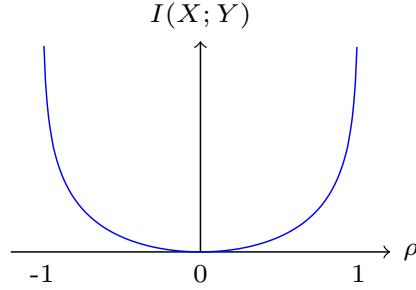


Figure 3.1 Mutual information between correlated Gaussians.

show (3.8), by shifting and scaling if necessary, we can assume without loss of generality that $\mathbb{E}X = \mathbb{E}Y = 0$ and $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$. Then $\rho = \mathbb{E}XY$. By joint Gaussianity, $Y = \rho X + Z$ for some $Z \sim \mathcal{N}(0, 1 - \rho^2) \perp\!\!\!\perp X$. Then using the divergence formula for Gaussians (2.7), we get

$$\begin{aligned} I(X; Y) &= D(P_{Y|X} \| P_Y | P_X) \\ &= \mathbb{E}D(\mathcal{N}(\rho X, 1 - \rho^2) \| \mathcal{N}(0, 1)) \\ &= \mathbb{E}\left[\frac{1}{2} \log \frac{1}{1 - \rho^2} + \frac{\log e}{2} ((\rho X)^2 + 1 - \rho^2 - 1)\right] \\ &= \frac{1}{2} \log \frac{1}{1 - \rho^2}. \end{aligned}$$

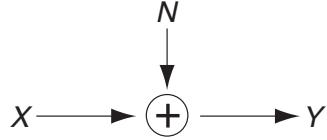
Alternatively, we can use the differential entropy representation in Theorem 3.4(d) and the entropy formula (2.17) for Gaussians:

$$\begin{aligned} I(X; Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z) \\ &= \frac{1}{2} \log(2\pi e) - \frac{1}{2} \log(2\pi e(1 - \rho^2)) = \frac{1}{2} \log \frac{1}{1 - \rho^2}. \end{aligned}$$

where the second equality follows $h(Y|X) = h(Y - X|X) = h(Z|X) = h(Z)$ applying the shift-invariance of h and the independence between X and Z .

Similar to the role of mutual information, the correlation coefficient also measures the dependency between random variables which are real-valued (more generally, on an inner-product space) in a certain sense. In contrast, mutual information is invariant to bijections and thus more general: it can be defined not just for numerical but for arbitrary random variables.

Example 3.3 (AWGN channel). Let $X \perp\!\!\!\perp N$ be independent Gaussian. Consider the additive white Gaussian noise (AWGN) channel: $Y = X + N$; pictorially,



Then

$$I(X; Y) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2} \right),$$

where $\frac{\sigma_X^2}{\sigma_N^2}$ is frequently referred to as the *signal-to-noise ratio (SNR)*.

Example 3.4 (Gaussian vectors). Let $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$ be jointly Gaussian. Then

$$I(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \log \frac{\det \Sigma_{\mathbf{X}} \det \Sigma_{\mathbf{Y}}}{\det \Sigma_{[\mathbf{X}, \mathbf{Y}]}}$$

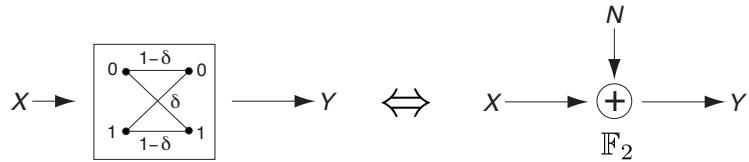
where $\Sigma_{\mathbf{X}} \triangleq \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top]$ denotes the covariance matrix of $\mathbf{X} \in \mathbb{R}^m$, and $\Sigma_{[\mathbf{X}, \mathbf{Y}]}$ denotes the covariance matrix of the random vector $[\mathbf{X}, \mathbf{Y}] \in \mathbb{R}^{m+n}$.

In the special case of additive noise: $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ for $\mathbf{N} \perp\!\!\!\perp \mathbf{X}$, we have

$$I(\mathbf{X}; \mathbf{X} + \mathbf{N}) = \frac{1}{2} \log \frac{\det(\Sigma_{\mathbf{X}} + \Sigma_{\mathbf{N}})}{\det \Sigma_{\mathbf{N}}}$$

since $\det \Sigma_{[\mathbf{X}, \mathbf{X} + \mathbf{N}]} = \det \begin{pmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} \\ \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} + \Sigma_{\mathbf{N}} \end{pmatrix} \stackrel{\text{why?}}{=} \det \Sigma_{\mathbf{X}} \det \Sigma_{\mathbf{N}}$.

Example 3.5 (Binary symmetric channel). Recall the setting in Example 1.4(1). Let $X \sim \text{Ber}(\frac{1}{2})$ and $N \sim \text{Ber}(\delta)$ be independent. Let $Y = X \oplus N$; or equivalently, Y is obtained by flipping X with probability δ .



As shown in Example 1.4(1), $H(X|Y) = H(N) = h(\delta)$ and hence

$$I(X; Y) = \log 2 - h(\delta).$$

The channel $P_{Y|X}$, called the binary symmetric channel with parameter δ and denoted by BSC_δ , will be encountered frequently in this book.

3.4 Conditional mutual information and conditional independence 45

Example 3.6 (Addition over finite groups). Generalizing Example 3.5, let X and Z take values on a finite group G . If X is uniform on G and independent of Z , then

$$I(X; X + Z) = \log |G| - H(Z),$$

which simply follows from that $X + Z$ is uniform on G regardless of the distribution of Z .

3.4 Conditional mutual information and conditional independence

Definition 3.6 (Conditional mutual information). If \mathcal{X} and \mathcal{Y} are standard Borel, then we define

$$I(X; Y|Z) \triangleq D(P_{X,Y|Z} \| P_{X|Z}P_{Y|Z}|P_Z) \quad (3.9)$$

$$= \mathbb{E}_{z \sim P_Z}[I(X; Y|Z = z)]. \quad (3.10)$$

where the product $P_{X|Z}P_{Y|Z}$ is a conditional distribution such that $(P_{X|Z}P_{Y|Z})(A \times B|z) = P_{X|Z}(A|z)P_{Y|Z}(B|z)$, under which X and Y are independent conditioned on Z .

Denoting $I(X; Y)$ as a functional $I(P_{X,Y})$ of the joint distribution $P_{X,Y}$, we have $I(X; Y|Z) = \mathbb{E}_{z \sim P_Z}[I(P_{X,Y|Z=z})]$. As such, $I(X; Y|Z)$ is a linear functional in P_Z . Measurability of the map $z \mapsto I(P_{X,Y|Z=z})$ is not obvious, but follows from Lemma 2.12.

To further discuss properties of the conditional mutual information, let us first introduce the notation for conditional independence. A family of joint distributions can be represented by a directed acyclic graph encoding the dependency structure of the underlying random variables. A simple example is a Markov chain (path graph) $X \rightarrow Y \rightarrow Z$, which represents distributions that factor as $\{P_{X,Y,Z} : P_{X,Y,Z} = P_X P_{Y|X} P_{Z|Y}\}$. We have the following equivalent descriptions:

$$\begin{aligned} X \rightarrow Y \rightarrow Z &\Leftrightarrow P_{X,Z|Y} = P_{X|Y} \cdot P_{Z|Y} \\ &\Leftrightarrow P_{Z|X,Y} = P_{Z|Y} \\ &\Leftrightarrow P_{X,Y,Z} = P_X \cdot P_{Y|X} \cdot P_{Z|Y} \\ &\Leftrightarrow X, Y, Z \text{ form a Markov chain} \\ &\Leftrightarrow X \perp\!\!\!\perp Z|Y \\ &\Leftrightarrow X \leftarrow Y \rightarrow Z, P_{X,Y,Z} = P_Y \cdot P_{X|Y} \cdot P_{Z|Y} \\ &\Leftrightarrow Z \rightarrow Y \rightarrow X \end{aligned}$$

Theorem 3.7 (Further properties of mutual information). *Suppose that all random variables are valued in standard Borel spaces. Then:*

- (a) $I(X; Z|Y) \geq 0$, with equality iff $X \rightarrow Y \rightarrow Z$.

(b) (Simple chain rule)³

$$\begin{aligned} I(X, Y; Z) &= I(X; Z) + I(Y; Z|X) \\ &= I(Y; Z) + I(X; Z|Y). \end{aligned}$$

(c) (DPI for mutual information) If $X \rightarrow Y \rightarrow Z$, then

- i) $I(X; Z) \leq I(X; Y)$, with equality iff $X \rightarrow Z \rightarrow Y$.
- ii) $I(X; Y|Z) \leq I(X; Y)$, with equality iff $X \perp\!\!\!\perp Z$.

(d) If $X \rightarrow Y \rightarrow Z \rightarrow W$, then $I(X; W) \leq I(Y; Z)$

(e) (Full chain rule)

$$I(X^n; Y) = \sum_{k=1}^n I(X_k; Y|X^{k-1})$$

(f) (Permutation invariance) If f and g are one-to-one (with measurable inverses), then

$$I(f(X); g(Y)) = I(X; Y).$$

Proof. (a) By definition and Theorem 3.2(c).

(b) First, notice that from (3.1) we have (with a self-evident notation):

$$I(Y; Z|X = x) = D(P_{Y|Z, X=x} \| P_{Y|X=x} P_{Z|X=x}).$$

Taking expectation over X here we get

$$I(Y; Z|X) \stackrel{(a)}{=} D(P_{Y|X, Z} \| P_{Y|X} P_{Z|X}).$$

On the other hand, from the chain rule for D , (2.24), we have

$$D(P_{X,Y,Z} \| P_{X,Y} P_Z) \stackrel{(b)}{=} D(P_{X,Z} \| P_X P_Z) + D(P_{Y|X,Z} \| P_{Y|X} P_{Z|X}),$$

where in the second term we noticed that conditioning on X, Z under the measure $P_{X,Y} P_Z$ results in $P_{Y|X}$ (independent of Z). Putting (a) and (b) together completes the proof.

(c) Apply Kolmogorov identity to $I(Y, Z; X)$:

$$\begin{aligned} I(Y, Z; X) &= I(X; Y) + \underbrace{I(X; Z|Y)}_{=0} \\ &= I(X; Z) + I(X; Y|Z) \end{aligned}$$

(d) Several applications of the DPI: $I(X; W) \leq I(X; Z) \leq I(Y; Z)$

(e) Recursive application of Kolmogorov identity. □

Remark 3.2. In general, $I(X; Y|Z)$ and $I(X; Y)$ are incomparable. Indeed, consider the following examples:

³ Also known as “Kolmogorov identities”.

3.4 Conditional mutual information and conditional independence 47

- $I(X; Y|Z) > I(X; Y)$: We need to find an example of X, Y, Z , which do not form a Markov chain. To that end notice that there is only one directed acyclic graph non-isomorphic to $X \rightarrow Y \rightarrow Z$, namely $X \rightarrow Y \leftarrow Z$. With this idea in mind, we construct $X, Z \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2})$ and $Y = X \oplus Z$. Then $I(X; Y) = 0$ since $X \perp\!\!\!\perp Y$; however, $I(X; Y|Z) = I(X; X \oplus Z|Z) = H(X) = 1 \text{ bit}$.
- $I(X; Y|Z) < I(X; Y)$: Simply take X, Y, Z to be any random variables on finite alphabets and $Z = Y$. Then $I(X; Y|Z) = I(X; Y|Y) = H(Y|Y) - H(Y|X, Y) = 0$ by a conditional version of (3.3).

Remark 3.3 (Chain rule for $I \Rightarrow$ Chain rule for H). Set $Y = X^n$. Then $H(X^n) = I(X^n; X^n) = \sum_{k=1}^n I(X_k; X^n|X^{k-1}) = \sum_{k=1}^n H(X_k|X^{k-1})$, since $H(X_k|X^n, X^{k-1}) = 0$.

Remark 3.4 (DPI for divergence \implies DPI for mutual information). We proved DPI for mutual information in Theorem 3.7 using Kolmogorov's identity. In fact, DPI for mutual information is implied by that for divergence in Theorem 2.16:

$$I(X; Z) = D(P_{Z|X}\|P_Z|P_X) \leq D(P_{Y|X}\|P_Y|P_X) = I(X; Y),$$

where note that for each x , we have $P_{Y|X=x} \xrightarrow{P_{Z|Y}} P_{Z|X=x}$ and $P_Y \xrightarrow{P_{Z|Y}} P_Z$. Therefore if we have a bi-variate functional of distributions $\mathcal{D}(P\|Q)$ which satisfies DPI, then we can define a “mutual information-like” quantity via $I_{\mathcal{D}}(X; Y) \triangleq \mathcal{D}(P_{Y|X}\|P_Y|P_X) \triangleq \mathbb{E}_{x \sim P_X} \mathcal{D}(P_{Y|X=x}\|P_Y)$ which will satisfy DPI on Markov chains. A rich class of examples arises by taking $\mathcal{D} = D_f$ (an f -divergence – see Chapter 7).

Remark 3.5 (Strong data-processing inequalities). For many channels $P_{Y|X}$, it is possible to strengthen the data-processing inequality (2.27) as follows: For any P_X, Q_X we have

$$D(P_Y\|Q_Y) \leq \eta_{KL} D(P_X\|Q_X),$$

where $\eta_{KL} < 1$ and depends on the channel $P_{Y|X}$ only. Similarly, this gives an improvement in the data-processing inequality for mutual information in Theorem 3.7(c): For any $P_{U,X}$ we have

$$U \rightarrow X \rightarrow Y \implies I(U; Y) \leq \eta_{KL} I(U; X).$$

For example, for $P_{Y|X} = \text{BSC}_\delta$ we have $\eta_{KL} = (1 - 2\delta)^2$. Strong data-processing inequalities (SDPIs) quantify the intuitive observation that noise intrinsic in the channel $P_{Y|X}$ must reduce the information that Y carries about the data U , regardless of how we optimize the encoding $U \mapsto X$. We explore SDPI further in Chapter 33 as well as their ramifications in statistics.

In addition to the case of strict inequality in DPI, the case of equality is also worth taking a closer look. If $U \rightarrow X \rightarrow Y$ and $I(U; X) = I(U; Y)$, intuitively it means that, as far as U is concerned, there is no loss of information in summarizing X into Y . In statistical parlance, we say that Y is a sufficient statistic of X for U . This is the topic for the next section.

3.5 Sufficient statistics and data processing

Much later in the book we will be interested in estimating parameters θ of probability distributions of X . To that end, one often first tries to remove unnecessary information contained in X . Let us formalize the setting as follows:

- Let P_X^θ be a collection of distributions of X parameterized by $\theta \in \Theta$;
- Let $P_{T|X}$ be some Markov kernel. Let $P_T^\theta \triangleq P_{T|X} \circ P_X^\theta$ be the induced distribution on T for each θ .

Definition 3.8 (Sufficient statistic). We say that T is a *sufficient statistic* of X for θ if there exists a transition probability kernel $P_{X|T}$ so that $P_X^\theta P_{T|X} = P_T^\theta P_{X|T}$, i.e., $P_{X|T}$ can be chosen to not depend on θ .

The intuitive interpretation of T being sufficient is that, with T at hand, one can ignore X ; in other words, T contains all the relevant information to infer about θ . This is because X can be simulated on the sole basis of T without knowing θ . As such, X provides no extra information for identification of θ . Any one-to-one transformation of X is sufficient, however, this is not the interesting case. In the interesting cases dimensionality of T will be much smaller (typically equal to that of θ) than that of X . See examples below.

Observe also that the parameter θ need not be a random variable, as Definition 3.12 does not involve any distribution (prior) on θ . This is a so-called *frequentist* point of view on the problem of parameter estimation.

Theorem 3.9. Let θ, X, T be as in the setting above. Then the following are equivalent

- T is a sufficient statistic of X for θ .
- $\forall P_\theta, \theta \rightarrow T \rightarrow X$.
- $\forall P_\theta, I(\theta; X|T) = 0$.
- $\forall P_\theta, I(\theta; X) = I(\theta; T)$, i.e., the data processing inequality for mutual information holds with equality.

Proof. We omit the details, which amount to either restating the conditions in terms of conditional independence, or invoking equality cases in the properties stated in Theorem 3.7. \square

The following result of Fisher provides a criterion for verifying sufficiency:

Theorem 3.10 (Fisher's factorization theorem). For all $\theta \in \Theta$, let P_X^θ have a density p_θ with respect to a common dominating measure μ . Let $T = T(X)$ be a deterministic function of X . Then T is a sufficient statistic of X for θ iff

$$p_\theta(x) = g_\theta(T(x))h(x)$$

for some measurable functions g_θ and h and all $\theta \in \Theta$.

3.5 Sufficient statistics and data processing 49

Proof. We only give the proof in the discrete case where p_θ represents the PMF. (The argument for the general case is similar replacing \sum by $\int d\mu$). Let $t = T(x)$.

“ \Rightarrow ”: Suppose T is a sufficient statistic of X for θ . Then $p_\theta(x) = P_\theta(X = x) = P_\theta(X = x, T = t) = P_\theta(X = x|T = t)P_\theta(T = t) = \underbrace{P(X = x|T = T(x))}_{h(x)} \underbrace{P_\theta(T = T(x))}_{g_\theta(T(x))}$

“ \Leftarrow ”: Suppose the factorization holds. Then

$$P_\theta(X = x|T = t) = \frac{p_\theta(x)}{\sum_x 1_{\{T(x)=t\}} p_\theta(x)} = \frac{g_\theta(t)h(x)}{\sum_x 1_{\{T(x)=t\}} g_\theta(t)h(x)} = \frac{h(x)}{\sum_x 1_{\{T(x)=t\}} h(x)},$$

free of θ . \square

Example 3.7 (Independent observations). In the following examples, a parametrized distribution generates an independent sample of size n , which can be summarized into a scalar-valued sufficient statistic. These can be verified by checking the factorization of the n -fold product distribution and applying Theorem 3.14.

- *Normal mean model.* Let $\theta \in \mathbb{R}$ and observations $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$. Then the sample mean $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ is a sufficient statistic of X^n for θ .
- *Coin flips.* Let $B_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$. Then $\sum_{i=1}^n B_i$ is a sufficient statistic of B^n for θ .
- *Uniform distribution.* Let $U_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$. Then $\max_{i \in [n]} U_i$ is a sufficient statistic of U^n for θ .

Example 3.8 (Sufficient statistic for hypothesis testing). Let $\Theta = \{0, 1\}$. Given $\theta = 0$ or 1 , $X \sim P_X$ or Q_X , respectively. Then Y – the output of $P_{Y|X}$ – is a sufficient statistic of X for θ iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, i.e., $P_{X|Y} = Q_{X|Y}$ holds P_Y -a.s. Indeed, the latter means that for kernel $Q_{X|Y}$ we have

$$P_X P_{Y|X} = P_Y Q_{X|Y} \quad \text{and} \quad Q_X P_{Y|X} = Q_Y Q_{X|Y},$$

which is precisely the definition of sufficient statistic when $\theta \in \{0, 1\}$. This example explains the condition for equality in the data-processing for divergence in Theorem 2.16. Then assuming $D(P_Y\|Q_Y) < \infty$ we have:

$$D(P_X\|Q_X) = D(P_Y\|Q_Y) \iff Y \text{ is a sufficient statistic for testing } P_X \text{ vs. } Q_X$$

Proof. Let $Q_{X,Y} = Q_X P_{Y|X}$, $P_{X,Y} = P_X P_{Y|X}$, then

$$\begin{aligned} D(P_{X,Y}\|Q_{X,Y}) &= \underbrace{D(P_{Y|X}\|Q_{Y|X}|P_X)}_{=0} + D(P_X\|Q_X) \\ &= D(P_{X|Y}\|Q_{X|Y}|P_Y) + D(P_Y\|Q_Y) \\ &\geq D(P_Y\|Q_Y) \end{aligned}$$

with equality iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, which is equivalent to Y being a sufficient statistic for testing P_X vs Q_X as desired. \square

4**Variational characterizations and continuity of D and I**

In this chapter we collect some results on variational characterizations. It is a well known method in analysis to study a functional by proving a variational characterization of the form $F(x) = \sup_{\lambda \in \Lambda} f_\lambda(x)$ or $F(x) = \inf_{\mu \in M} g_\mu(x)$. Such representations can be useful for multiple purposes:

- Convexity: pointwise supremum of convex functions is convex.
- Regularity: pointwise supremum of lower semicontinuous (lsc) functions is lsc.
- Bounds: upper/lower bound on F follows by choosing any λ (μ) and evaluating f_λ (g_μ).

We will see in this chapter that divergence has two different sup characterizations (over partitions and over functions). The mutual information is even more special. In addition to inheriting the ones from KL divergence, it possesses two very special ones: an inf over (centroid) measures Q_Y and a sup over Markov kernels.

As the main applications of these variational characterizations, we will first pursue the topic of continuity. In fact, we will discuss several types of continuity.

First, is the continuity in discretization. This is related to the issue of *computation*. For complicated P and Q direct computation of $D(P||Q)$ might be hard. Instead, one may want to discretize the infinite alphabet and compute numerically the finite sum. Is this procedure stable, i.e., as the quantization becomes finer, does this procedure guarantee to converge to the true value? The answer is positive and this continuity with respect to discretization is guaranteed by Theorem 4.6.

Second, is the continuity under change of the distribution. For example, this arises in the problem of *estimating information measures*. In many statistical setups, oftentimes we do not know P or Q , and we estimate the distribution by \hat{P}_n using n iid observations sampled from P (in discrete cases we may set \hat{P}_n to be simply the empirical distribution). Does $D(\hat{P}_n||Q)$ provide a good estimator for $D(P||Q)$? Does $D(\hat{P}_n||Q) \rightarrow D(P||Q)$ if $\hat{P}_n \rightarrow P$? The answer is delicate – see Section 4.4.

Third, there is yet another kind of continuity: continuity “in the σ -algebra”. Despite the scary name, this one is useful even in the most “discrete” situations. For example, imagine that $\theta \sim \text{Unif}(0, 1)$ and $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$. Suppose that you observe a sequence of X_i ’s until the random moment τ equal to the first occurrence of the pattern 0101. How much information about θ did you learn by time τ ? We can encode these observations as

$$Z_j = \begin{cases} X_j, & j \leq \tau, \\ ?, & j > \tau \end{cases},$$

4.1 Geometric interpretation of mutual information 51

where ? designates the fact that we don't know the value of X_j on those times. Then the question we asked above is to compute $I(\theta; Z^\infty)$. We will show in this chapter that

$$I(\theta; Z^\infty) = \lim_{n \rightarrow \infty} I(\theta; Z^n) = \sum_{n=1}^{\infty} I(\theta; Z_n | Z^{n-1}) \quad (4.1)$$

thus reducing computation to evaluating an infinite sum of simpler terms (not involving infinite-dimensional vectors). Thus, even in this simple question about biased coin flips we have to understand how to safely work with infinite-dimensional vectors.

4.1 Geometric interpretation of mutual information

Mutual information can be understood as the weighted “distance” from the conditional distributions to the marginal distribution. Indeed, for discrete X , we have

$$I(X; Y) = D(P_{Y|X} \| P_Y | P_X) = \sum_{x \in \mathcal{X}} D(P_{Y|X=x} \| P_Y) P_X(x).$$

Furthermore, it turns out that P_Y , similar to the center of gravity, minimizes this weighted distance and thus can be thought as the best approximation for the “center” of the collection of distributions $\{P_{Y|X=x} : x \in \mathcal{X}\}$ with weights given by P_X . We formalize these results in this section and start with the proof of a “golden formula”. Its importance is in bridging the two points of view on mutual information: (4.3) is the difference of (relative) entropies in the style of Shannon, while retaining applicability to continuous spaces in the style of Fano.

Theorem 4.1 (Golden formula). *For any Q_Y we have*

$$D(P_{Y|X} \| Q_Y | P_X) = I(X; Y) + D(P_Y \| Q_Y). \quad (4.2)$$

Thus, if $D(P_Y \| Q_Y) < \infty$, then

$$I(X; Y) = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y). \quad (4.3)$$

Proof. In the discrete case and ignoring the possibility of dividing by zero, the argument is really simple. We just need to write

$$I(X; Y) \stackrel{(3.1)}{=} \mathbb{E}_{P_{X,Y}} \left[\log \frac{P_{Y|X}}{P_Y} \right] = \mathbb{E}_{P_{X,Y}} \left[\log \frac{P_{Y|X} Q_Y}{P_Y Q_Y} \right]$$

and then expand $\log \frac{P_{Y|X} Q_Y}{P_Y Q_Y} = \log \frac{P_{Y|X}}{Q_Y} - \log \frac{P_Y}{Q_Y}$. The argument below is a rigorous implementation of this idea.

First, notice that by Theorem 2.15(e) we have $D(P_{Y|X} \| Q_Y | P_X) \geq D(P_Y \| Q_Y)$ and thus if $D(P_Y \| Q_Y) = \infty$ then both sides of (4.2) are infinite. Thus, we assume $D(P_Y \| Q_Y) < \infty$ and in particular $P_Y \ll Q_Y$. Rewriting LHS of (4.2) via the chain rule (2.24) we see that Theorem amounts to proving

$$D(P_{X,Y} \| P_X Q_Y) = D(P_{X,Y} \| P_X P_Y) + D(P_Y \| Q_Y).$$

The case of $D(P_{X,Y}\|P_XQ_Y) = D(P_{X,Y}\|P_XP_Y) = \infty$ is clear. Thus, we can assume at least one of these divergences is finite, and, hence, also $P_{X,Y} \ll P_XQ_Y$.

Let $\lambda(y) = \frac{dP_Y}{dQ_Y}(y)$. Since $\lambda(Y) > 0$ P_Y -a.s., applying the definition of Log in (2.10), we can write

$$\mathbb{E}_{P_Y}[\log \lambda(Y)] = \mathbb{E}_{P_{X,Y}}\left[\log \frac{\lambda(Y)}{1}\right]. \quad (4.4)$$

Notice that the same $\lambda(y)$ is also the density $\frac{dP_XP_Y}{dP_XQ_Y}(x,y)$ of the product measure P_XP_Y with respect to P_XQ_Y . Therefore, the RHS of (4.4) by (2.11) applied with $\mu = P_XQ_Y$ coincides with

$$D(P_{X,Y}\|P_XQ_Y) - D(P_{X,Y}\|P_XP_Y),$$

while the LHS of (4.4) by (2.13) equals $D(P_Y\|Q_Y)$. Thus, we have shown the required

$$D(P_Y\|Q_Y) = D(P_{X,Y}\|P_XQ_Y) - D(P_{X,Y}\|P_XP_Y).$$

□

By dropping the second term in (4.2) we obtain the following result.

Corollary 4.2 (Mutual information as center of gravity). *For any Q_Y we have*

$$I(X; Y) \leq D(P_{Y|X}\|Q_Y|P_X)$$

and, consequently,

$$I(X; Y) = \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X). \quad (4.5)$$

If $I(X; Y) < \infty$, the unique minimizer is $Q_Y = P_Y$.

Remark 4.1. The variational representation (4.5) is useful for upper bounding mutual information by choosing an appropriate Q_Y . Indeed, often each distribution in the collection $P_{Y|X=x}$ is simple, but their mixture, P_Y , is very hard to work with. In these cases, choosing a suitable Q_Y in (4.5) provides a convenient upper bound. As an example, consider the AWGN channel $Y = X + Z$ in Example 3.3, where $\text{Var}(X) = \sigma^2$, $Z \sim \mathcal{N}(0, 1)$. Then, choosing the best possible Gaussian Q and applying the above bound, we have:

$$I(X; Y) \leq \inf_{\mu \in \mathbb{R}, s \geq 0} \mathbb{E}[D(\mathcal{N}(X, 1)\|\mathcal{N}(\mu, s))] = \frac{1}{2} \log(1 + \sigma^2),$$

which is tight when X is Gaussian. For more examples and statistical applications, see Chapter 30.

Theorem 4.3 (Mutual information as distance to product distributions).

$$I(X; Y) = \min_{Q_X, Q_Y} D(P_{X,Y}\|Q_XQ_Y)$$

with the unique minimizer $(Q_X, Q_Y) = (P_X, P_Y)$.

4.1 Geometric interpretation of mutual information 53

Proof. We only need to use the previous corollary and the chain rule (2.24):

$$D(P_{X,Y}\|Q_XQ_Y) \stackrel{(2.24)}{=} D(P_{Y|X}\|Q_Y|P_X) + D(P_X\|Q_X) \geq I(X; Y).$$

□

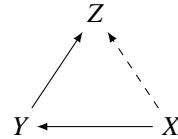
Interestingly, the point of view in the previous result extends to conditional mutual information as follows: We have

$$I(X; Z|Y) = \min_{Q_{X,Y,Z}: X \rightarrow Y \rightarrow Z} D(P_{X,Y,Z}\|Q_{X,Y,Z}), \quad (4.6)$$

where the minimization is over all $Q_{X,Y,Z} = Q_XQ_{Y|X}Q_{Z|Y}$, cf. Section 3.4. Showing this characterization is very similar to the previous theorem. By repeating the same argument as in (4.2) we get

$$\begin{aligned} & D(P_{X,Y,Z}\|Q_XQ_{Y|X}Q_{Z|Y}) \\ &= D(P_{X,Y,Z}\|P_XP_{Y|X}P_{Z|Y}) + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_{Z|Y}\|Q_{Z|Y}|P_Y) \\ &= D(P_{X,Y,Z}\|P_YP_{X|Y}P_{Z|Y}) + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_{Z|Y}\|Q_{Z|Y}|P_Y) \\ &= \underbrace{D(P_{XZ|Y}\|P_{X|Y}P_{Z|Y}|P_Y)}_{I(X; Z|Y)} + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_{Z|Y}\|Q_{Z|Y}|P_Y) \\ &\geq I(X; Z|Y). \end{aligned}$$

Characterization (4.6) can be understood as follows. The most general graphical model for the triplet (X, Y, Z) is a 3-clique (triangle).



What is the information flow on the dashed edge $X \rightarrow Z$? To answer this, notice that removing this edge restricts the joint distribution to a Markov chain $X \rightarrow Y \rightarrow Z$. Thus, it is natural to ask what is the minimum (KL-divergence) distance between a given $P_{X,Y,Z}$ and the set of all distributions $Q_{X,Y,Z}$ satisfying the Markov chain constraint. By the above calculation, optimal $Q_{X,Y,Z} = P_YP_{X|Y}P_{Z|Y}$ and hence the distance is $I(X; Z|Y)$. For this reason, we may interpret $I(X; Z|Y)$ as the amount of information flowing through the $X \rightarrow Z$ edge.

In addition to inf-characterization, mutual information also has a sup-characterization.

Theorem 4.4. *For any Markov kernel $Q_{X|Y}$ such that $Q_{X|Y=y} \ll P_X$ for P_Y -a.e. y we have*

$$I(X; Y) \geq \mathbb{E}_{P_{X,Y}} \left[\log \frac{dQ_{X|Y}}{dP_X} \right].$$

If $I(X; Y) < \infty$ then

$$I(X; Y) = \sup_{Q_{X|Y}} \mathbb{E}_{P_{X,Y}} \left[\log \frac{dQ_{X|Y}}{dP_X} \right], \quad (4.7)$$

where supremum is over Markov kernels $Q_{X|Y}$ as in the first sentence.

Remark 4.2. Similar to how Theorem 4.1 is used to upper-bound $I(X; Y)$ by choosing a good approximation to P_Y , this result is used to lower-bound $I(X; Y)$ by selecting a good (but computable) approximation $Q_{X|Y}$ to usually a very complicated posterior $P_{X|Y}$. See Section 5.6 for applications.

Proof. Since modifying $Q_{X|Y=y}$ on a negligible set of y 's does not change the expectations, we will assume that $Q_{X|Y=y} \ll P_Y$ for every y . If $I(X; Y) = \infty$ then there is nothing to prove. So we assume $I(X; Y) < \infty$, which implies $P_{X,Y} \ll P_X P_Y$. Then by Lemma 3.3 we have that $P_{X|Y=y} \ll P_X$ for almost every y . Choose any such y and apply (2.11) with $\mu = P_X$ and noticing $\text{Log} \frac{dQ_{X|Y=y}/dP_X}{1} = \log \frac{dQ_{X|Y=y}}{dP_X}$ we get

$$\mathbb{E}_{P_{X|Y=y}} \left[\log \frac{dQ_{X|Y=y}}{dP_X} \right] = D(P_{X|Y=y} \| P_X) - D(P_{X|Y=y} \| Q_{X|Y=y}),$$

which is applicable since the first term is finite for a.e. y by (3.1). Taking expectation of the previous identity over y we obtain

$$\mathbb{E}_{P_{X,Y}} \left[\log \frac{dQ_{X|Y}}{dP_X} \right] = I(X; Y) - D(P_{X|Y} \| Q_{X|Y} | P_Y) \leq I(X; Y), \quad (4.8)$$

implying the first part. The equality case in (4.7) follows by taking $Q_{X|Y} = P_{X|Y}$, which satisfies the conditions on Q when $I(X; Y) < \infty$. \square

4.2 Variational characterizations of divergence: Gelfand-Yaglom-Perez

The point of the following theorem is that divergence on general alphabets can be defined via divergence on finite alphabets and discretization. Moreover, as the quantization becomes finer, we approach the value of divergence.

Theorem 4.5 (Gelfand-Yaglom-Perez [133]). *Let P, Q be two probability measures on \mathcal{X} with σ -algebra \mathcal{F} . Then*

$$D(P \| Q) = \sup_{\{E_1, \dots, E_n\}} \sum_{i=1}^n P[E_i] \log \frac{P[E_i]}{Q[E_i]}, \quad (4.9)$$

where the supremum is over all finite \mathcal{F} -measurable partitions: $\bigcup_{j=1}^n E_j = \mathcal{X}$, $E_j \cap E_i = \emptyset$, and $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$ per our usual convention.

4.3 Variational characterizations of divergence: Donsker-Varadhan 55

Remark 4.3. This theorem, in particular, allows us to prove all general identities and inequalities for the cases of discrete random variables and then pass to the limit. In case of mutual information $I(X; Y) = D(P_{X,Y}\|P_X P_Y)$, the partitions over \mathcal{X} and \mathcal{Y} can be chosen separately, see (4.29).

Proof. “ \geq ”: Fix a finite partition E_1, \dots, E_n . Define a function (quantizer/discretizer) $f : \mathcal{X} \rightarrow \{1, \dots, n\}$ as follows: For any x , let $f(x)$ denote the index j of the set E_j to which x belongs. Let X be distributed according to either P or Q and set $Y = f(X)$. Applying data processing inequality for divergence yields

$$\begin{aligned} D(P\|Q) &= D(P_X\|Q_X) \\ &\geq D(P_Y\|Q_Y) \\ &= \sum_i P(E_i) \log \frac{P(E_i)}{Q(E_i)}. \end{aligned} \tag{4.10}$$

“ \leq ”: To show $D(P\|Q)$ is indeed achievable, first note that if $P \ll Q$, then by definition, there exists B such that $Q(B) = 0 < P(B)$. Choosing the partition $E_1 = B$ and $E_2 = B^c$, we have $D(P\|Q) = \infty = \sum_{i=1}^2 P[E_i] \log \frac{P[E_i]}{Q[E_i]}$. In the sequel we assume that $P \ll Q$ and let $X = \frac{dP}{dQ}$. Then $D(P\|Q) = \mathbb{E}_Q[X \log X] = \mathbb{E}_Q[\varphi(X)]$ by (2.4). Note that $\varphi(x) \geq 0$ if and only if $x \geq 1$. By monotone convergence theorem, we have $\mathbb{E}_Q[\varphi(X)1_{\{X < c\}}] \rightarrow D(P\|Q)$ as $c \rightarrow \infty$, regardless of the finiteness of $D(P\|Q)$.

Next, we construct a finite partition. Let $n = c/\epsilon$ be an integer and for $j = 0, \dots, n-1$, let $E_j = \{j\epsilon \leq X(j+1)\epsilon\}$ and $E_n = \{X \geq c\}$. Define $Y = \epsilon[X/\epsilon]$ as the quantized version. Since φ is uniformly continuous on $[0, c]$, for any $x, y \in [0, c]$ such $|x - y| \leq \epsilon$, we have $|\varphi(x) - \varphi(y)| \leq \epsilon'$ for some $\epsilon' = \epsilon'(\epsilon, c)$ such as $\epsilon' \rightarrow 0$ as $\epsilon \rightarrow 0$. Then $\mathbb{E}_Q[\varphi(Y)1_{\{X < c\}}] \geq \mathbb{E}_Q[\varphi(X)1_{\{X < c\}}] - \epsilon'$. Moreover,

$$\begin{aligned} \mathbb{E}_Q[\varphi(Y)1_{\{X < c\}}] &= \sum_{j=0}^{n-1} \varphi(j\epsilon) Q(E_j) \leq \epsilon' + \sum_{j=0}^{n-1} \varphi\left(\frac{P(E_j)}{Q(E_j)}\right) Q(E_j) \\ &\leq \epsilon' + Q(X \geq c) \log e + \sum_{j=0}^n P(E_j) \log \frac{P(E_j)}{Q(E_j)}, \end{aligned}$$

where the first inequality applies the uniform continuity of φ since $j\epsilon \leq \frac{P(E_j)}{Q(E_j)} < (j+1)\epsilon$, and the second applies $\varphi \geq -\log e$. As $Q(X \geq c) \rightarrow 0$ as $c \rightarrow \infty$, the proof is completed by first sending $\epsilon \rightarrow 0$ then $c \rightarrow \infty$. \square

4.3 Variational characterizations of divergence: Donsker-Varadhan

The following is perhaps the most important variational characterization of divergence.

Theorem 4.6 (Donsker-Varadhan [99]). *Let P, Q be probability measures on \mathcal{X} and let \mathcal{C}_Q denote the set of functions $f: \mathcal{X} \rightarrow \mathbb{R}$ such that $\mathbb{E}_Q[\exp\{f(X)\}] < \infty$. We have*

$$D(P\|Q) = \sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}]. \quad (4.11)$$

In particular, if $D(P\|Q) < \infty$ then $\mathbb{E}_P[f(X)]$ is finite for every $f \in \mathcal{C}_Q$. The identity (4.11) holds with \mathcal{C}_Q replaced by the class of all simple functions. If \mathcal{X} is a normal topological space (e.g., a metric space) with Borel σ -algebra, then also

$$D(P\|Q) = \sup_{f \in \mathcal{C}_b} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}], \quad (4.12)$$

where \mathcal{C}_b is a class of bounded continuous functions.

Proof. “ \geq ”: We can assume for this part that $D(P\|Q) < \infty$, since otherwise there is nothing to prove. Then fix $f \in \mathcal{C}_Q$ and define a probability measure Q^f (*tilted version of Q*) via

$$Q^f(dx) = \exp\{f(x) - Z_f\} Q(dx), \quad Z_f \triangleq \log \mathbb{E}_Q[\exp\{f(X)\}].$$

Then, obviously $Q^f \ll Q$ and we have

$$\mathbb{E}_P[f(X)] - Z_f = \mathbb{E}_P \left[\log \frac{dQ^f}{dQ} \right] = \mathbb{E}_P \left[\log \frac{dP dQ^f}{dQ dP} \right] = D(P\|Q) - D(P\|Q^f) \leq D(P\|Q).$$

“ \leq ”: The idea is to just take $f = \log \frac{dP}{dQ}$; however to handle the edge cases we proceed carefully. First, notice that if $P \not\ll Q$ then for some E with $Q[E] = 0 < P[E]$ and $c \rightarrow \infty$ taking $f = c1_E$ shows that both sides of (4.11) are infinite. Thus, we assume $P \ll Q$. For any partition of $\mathcal{X} = \cup_{j=1}^n E_j$ we set $f = \sum_{j=1}^n 1_{E_j} \log \frac{P[E_j]}{Q[E_j]}$. Then the right-hand sides of (4.11) and (4.9) evaluate to the same value and hence by Theorem 4.6 we obtain that supremum over simple functions (and thus over \mathcal{C}_Q) is at least as large as $D(P\|Q)$.

Finally, to show (4.12), we show that for every simple function f there exists a continuous bounded f' such that $\mathbb{E}_P[f'] - \log \mathbb{E}_Q[\exp\{f'\}]$ is arbitrarily close to the same functional evaluated at f . Clearly, for that it is enough to show that for any $a \in \mathbb{R}$ and measurable $A \subset \mathcal{X}$ there exists a sequence of continuous bounded f_n such that

$$\mathbb{E}_P[f_n] \rightarrow aP[A], \quad \text{and} \quad \mathbb{E}_Q[\exp\{f_n\}] \rightarrow \exp\{a\}Q[A] \quad (4.13)$$

hold *simultaneously*. We only consider the case of $a > 0$ below. Let compact F and open U be such that $F \subset A \subset U$ and $\max(P[U] - P[F], Q[U] - Q[F]) \leq \epsilon$. Such F and U exist whenever P and Q are so-called regular measures. Without going into details, we just notice that finite measures on Polish spaces are automatically regular. Then by Urysohn’s lemma there exists a continuous function $f_\epsilon: \mathcal{X} \rightarrow [0, a]$ equal to a on F and 0 on U^c . Then we have

$$\begin{aligned} aP[F] &\leq \mathbb{E}_P[f_\epsilon] \leq aP[U] \\ \exp\{a\}Q[F] &\leq \mathbb{E}_Q[\exp\{f_\epsilon\}] \leq \exp\{a\}Q[U]. \end{aligned}$$

Subtracting $aP[A]$ and $\exp\{a\}Q[A]$ for each of these inequalities, respectively, we see that taking $\epsilon \rightarrow 0$ indeed results in a sequence of functions satisfying (4.13).

4.4 Continuity of divergence 57

□

Remark 4.4. 1 What is the Donsker-Varadhan representation useful for? By setting $f(x) = \epsilon \cdot g(x)$ with $\epsilon \ll 1$ and linearizing exp and log we can see that when $D(P\|Q)$ is small, expectations under P can be approximated by expectations over Q (change of measure): $\mathbb{E}_P[g(X)] \approx \mathbb{E}_Q[g(X)]$. This holds for all functions g with finite exponential moment under Q . Total variation distance provides a similar bound, but for a narrower class of bounded functions:

$$|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| \leq \|g\|_\infty \text{TV}(P, Q).$$

- 2 More formally, the inequality $\mathbb{E}_P[f(X)] \leq \log \mathbb{E}_Q[\exp f(X)] + D(P\|Q)$ is useful in estimating $\mathbb{E}_P[f(X)]$ for complicated distribution P (e.g. over high-dimensional X with weakly dependent coordinates) by making a smart choice of Q (e.g. with iid components).
- 3 In Chapter 5 we will show that $D(P\|Q)$ is convex in P (in fact, in the pair). A general method of obtaining variational formulas like (4.11) is via the Young-Fenchel duality. Indeed, (4.11) is exactly this inequality since the Fenchel-Legendre conjugate of $D(\cdot\|Q)$ is given by a convex map $f \mapsto Z_f$. For more details, see Section 7.13.
- 4 Donsker-Varadhan should also be seen as an “improved version” of the DPI. For example, one of the main applications of the DPI in this book is in obtaining estimates like

$$P[A] \log \frac{1}{Q[A]} \leq D(P\|Q) + \log 2, \quad (4.14)$$

which is the basis of the large deviations theory (Corollary 2.2) and Fano’s inequality (Theorem 6.4). The same estimate can be obtained by applying (4.11) via $f(x) = 1_{\{x \in A\}} \log \frac{1}{Q[A]}$.

4.4 Continuity of divergence

For a finite alphabet \mathcal{X} it is easy to establish the continuity of entropy and divergence:

Proposition 4.7. *Let \mathcal{X} be finite. Fix a distribution Q on \mathcal{X} with $Q(x) > 0$ for all $x \in \mathcal{X}$. Then the map*

$$P \mapsto D(P\|Q)$$

is continuous. In particular,

$$P \mapsto H(P) \quad (4.15)$$

is continuous.

Warning: Divergence is never continuous in the pair, even for finite alphabets. For example, as $n \rightarrow \infty$, $d(\frac{1}{n}\|2^{-n}) \not\rightarrow 0$.

Proof. Notice that

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

and each term is a continuous function of $P(x)$. \square

Our next goal is to study continuity properties of divergence for general alphabets. We start with a negative observation.

Remark 4.5. In general, $D(P\|Q)$ is *not* continuous in either P or Q . For example, let X_1, \dots, X_n be iid and equally likely to be $\{\pm 1\}$. Then by central limit theorem, $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0, 1)$ as $n \rightarrow \infty$. But

$$D(\underbrace{P_{S_n}}_{\text{discrete}} \parallel \underbrace{\mathcal{N}(0, 1)}_{\text{cont's}}) = \infty$$

for all n . Note that this is an example for strict inequality in (4.16).

Nevertheless, there is a very useful semicontinuity property.

Theorem 4.8 (Lower semicontinuity of divergence). *Let \mathcal{X} be a metric space with Borel σ -algebra \mathcal{H} . If P_n and Q_n converge weakly to P and Q , respectively,¹ then*

$$D(P\|Q) \leq \liminf_{n \rightarrow \infty} D(P_n\|Q_n). \quad (4.16)$$

On a general space if $P_n \rightarrow P$ and $Q_n \rightarrow Q$ pointwise² (i.e. $P_n[E] \rightarrow P[E]$ and $Q_n[E] \rightarrow Q[E]$ for every measurable E) then (4.16) also holds.

Proof. This simply follows from (4.12) since $\mathbb{E}_{P_n}[f] \rightarrow \mathbb{E}_P[f]$ and $\mathbb{E}_{Q_n}[\exp\{f\}] \rightarrow \mathbb{E}_Q[\exp\{f\}]$ for every $f \in \mathcal{C}_b$. \square

4.5* Continuity under monotone limits of σ -algebras

Our final and somewhat delicate topic is to understand the (so far neglected) dependence of D and I on the implicit σ -algebra of the space. Indeed, the definition of divergence $D(P\|Q)$ implicitly (via Radon-Nikodym derivative) depends on the σ -algebra \mathcal{F} defining the measurable space $(\mathcal{X}, \mathcal{F})$. To emphasize the dependence on \mathcal{F} we will write in this Section only the underlying σ -algebra explicitly as follows:

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}).$$

Our main results are continuity under monotone limits of σ -algebras

$$\mathcal{F}_n \nearrow \mathcal{F} \implies D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \quad (4.17)$$

$$\mathcal{F}_n \searrow \mathcal{F} \implies D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \searrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \quad (4.18)$$

¹ Recall that sequence of random variables X_n converges in distribution to X if and only if their laws P_{X_n} converge weakly to P_X .

² Pointwise convergence is weaker than convergence in total variation and stronger than weak convergence.

4.5* Continuity under monotone limits of σ -algebras 59

For establishing the first result, it will be convenient to extend the definition of the divergence $D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$ to (a) any *algebra* of sets \mathcal{F} and (b) two positive additive (not necessarily σ -additive) set-functions P, Q on \mathcal{F} .

Definition 4.9 (KL divergence over an algebra). Let P and Q be two positive, additive (not necessarily σ -additive) set-functions defined over an algebra \mathcal{F} of subsets of \mathcal{X} (not necessarily a σ -algebra). We define

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \triangleq \sup_{\{E_1, \dots, E_n\}} \sum_{i=1}^n P[E_i] \log \frac{P[E_i]}{Q[E_i]},$$

where the supremum is over all finite \mathcal{F} -measurable partitions: $\bigcup_{j=1}^n E_j = \mathcal{X}, E_j \cap E_i = \emptyset$, and $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$ per our usual convention.

Note that when \mathcal{F} is not a σ -algebra or P, Q are not σ -additive, we do not have Radon-Nikodym theorem and thus our original definition of KL-divergence is not applicable.

Theorem 4.10 (Measure-theoretic properties of divergence). *Let P, Q be probability measures on the measurable space $(\mathcal{X}, \mathcal{H})$. Assume all algebras below are sub-algebras of \mathcal{H} . Then:*

- (*Monotonicity*) If $\mathcal{F} \subseteq \mathcal{G}$ are nested algebras then

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \leq D(P_{\mathcal{G}}\|Q_{\mathcal{G}}). \quad (4.19)$$

- Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \dots$ be an increasing sequence of algebras and let $\mathcal{F} = \bigcup_n \mathcal{F}_n$ be their limit, then

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}}).$$

- If \mathcal{F} is $(P+Q)$ -dense in \mathcal{G} then³

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) = D(P_{\mathcal{G}}\|Q_{\mathcal{G}}). \quad (4.20)$$

- (*Monotone convergence theorem*) Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \dots$ be an increasing sequence of algebras and let $\mathcal{F} = \bigvee_n \mathcal{F}_n$ be the smallest σ -algebra containing all of \mathcal{F}_n . Then we have

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$$

and, in particular,

$$D(P_{X^\infty}\|Q_{X^\infty}) = \lim_{n \rightarrow \infty} D(P_{X^n}\|Q_{X^n}).$$

Proof. The first two items are straightforward applications of the definition. The third follows from the following fact: if \mathcal{F} is dense in \mathcal{G} then any \mathcal{G} -measurable partition $\{E_1, \dots, E_n\}$ can be approximated by a \mathcal{F} -measurable partition $\{E'_1, \dots, E'_n\}$ with $(P+Q)[E_i \Delta E'_i] \leq \epsilon$. Indeed, first we set E'_1 to be an element of \mathcal{F} with $(P+Q)(E_1 \Delta E'_1) \leq \frac{\epsilon}{2n}$. Then, we set E'_2 to be

³ Recall that \mathcal{F} is μ -dense in \mathcal{G} if $\forall E \in \mathcal{G}, \epsilon > 0 \exists E' \in \mathcal{F}$ s.t. $\mu[E \Delta E'] \leq \epsilon$.

an $\frac{\epsilon}{2n}$ -approximation of $E_2 \setminus E'_1$, etc. Finally, $E'_n = (\cup_{j \leq 1} E'_j)^c$. By taking $\epsilon \rightarrow 0$ we obtain $\sum_i P[E'_i] \log \frac{P[E'_i]}{Q[E'_i]} \rightarrow \sum_i P[E_i] \log \frac{P[E_i]}{Q[E_i]}$.

The last statement follows from the previous one and the fact that any algebra \mathcal{F} is μ -dense in the σ -algebra $\sigma\{\mathcal{F}\}$ it generates for any bounded μ on $(\mathcal{X}, \mathcal{H})$ (cf. [106, Lemma III.7.1]). \square

Finally, we address the continuity under the decreasing σ -algebra, i.e. (4.18).

Proposition 4.11. *Let $\mathcal{F}_n \searrow \mathcal{F}$ be a sequence of decreasing σ -algebras with intersection $\mathcal{F} = \cap_n \mathcal{F}_n$; let P, Q be two probability measures on \mathcal{F}_0 . If $D(P_{\mathcal{F}_0} \| Q_{\mathcal{F}_0}) < \infty$ then we have*

$$D(P_{\mathcal{F}_n} \| Q_{\mathcal{F}_n}) \searrow D(P_{\mathcal{F}} \| Q_{\mathcal{F}}) \quad (4.21)$$

The condition $D(P_{\mathcal{F}_0} \| Q_{\mathcal{F}_0}) < \infty$ can not be dropped, cf. the example after (4.32).

Proof. Let $X_{-n} = \frac{dP}{dQ} \Big|_{\mathcal{F}_n}$. Since $X_{-n} = \mathbb{E}_Q \left[\frac{dP}{dQ} \Big| \mathcal{F}_n \right]$, we have that (\dots, X_{-1}, X_0) is a uniformly integrable martingale. By the martingale convergence theorem in reversed time, cf. [58, Theorem 5.4.17], we have almost surely

$$X_{-n} \rightarrow X_{-\infty} \triangleq \frac{dP}{dQ} \Big|_{\mathcal{F}}. \quad (4.22)$$

We need to prove that

$$\mathbb{E}_Q[X_{-n} \log X_{-n}] \rightarrow \mathbb{E}_Q[X_{-\infty} \log X_{-\infty}].$$

We will do so by decomposing $x \log x$ as follows

$$x \log x = x \log^+ x + x \log^- x,$$

where $\log^+ x = \max(\log x, 0)$ and $\log^- x = \min(\log x, 0)$. Since $x \log^- x$ is bounded, we have from the bounded convergence theorem:

$$\mathbb{E}_Q[X_{-n} \log^- X_{-n}] \rightarrow \mathbb{E}_Q[X_{-\infty} \log^- X_{-\infty}].$$

To prove a similar convergence for \log^+ we need to notice two things. First, the function

$$x \mapsto x \log^+ x$$

is convex. Second, for any non-negative convex function ϕ s.t. $\mathbb{E}[\phi(X_0)] < \infty$ the collection $\{Z_n = \phi(\mathbb{E}[X_0 | \mathcal{F}_n]), n \geq 0\}$ is uniformly integrable. Indeed, we have from Jensen's inequality

$$\mathbb{P}[Z_n > c] \leq \frac{1}{c} \mathbb{E}[\phi(\mathbb{E}[X_0 | \mathcal{F}_n])] \leq \frac{\mathbb{E}[\phi(X_0)]}{c}$$

and thus $\mathbb{P}[Z_n > c] \rightarrow 0$ as $c \rightarrow \infty$. Therefore, we have again by Jensen's

$$\mathbb{E}[Z_n \mathbf{1}\{Z_n > c\}] \leq \mathbb{E}[\phi(X_0) \mathbf{1}\{Z_n > c\}] \rightarrow 0 \quad c \rightarrow \infty.$$

Finally, since $X_{-n} \log^+ X_{-n}$ is uniformly integrable, we have from (4.22)

$$\mathbb{E}_Q[X_{-n} \log^- X_{-n}] \rightarrow \mathbb{E}_Q[X_{-\infty} \log^- X_{-\infty}]$$

and this concludes the proof. \square

4.6 Variational characterizations and continuity of mutual information 61

4.6 Variational characterizations and continuity of mutual information

Again, similarly to Proposition 4.10, it is easy to show that in the case of finite alphabets mutual information is always continuous on finite-dimensional simplices of distributions.⁴

Proposition 4.12. *Let \mathcal{X} and \mathcal{Y} be finite alphabets. Then*

$$P_{X,Y} \mapsto I(X; Y)$$

is continuous. Let \mathcal{X} be finite. Then

$$P_X \mapsto I(X; Y) \tag{4.23}$$

is continuous. Without any assumptions on \mathcal{X} and \mathcal{Y} , let P_X range over the convex hull $\Pi = \text{co}(P_1, \dots, P_n) = \{\sum_{i=1}^n \alpha_i P_i : \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0\}$. If $I(P_j, P_{Y|X}) < \infty$ (using notation $I(P_X, P_{Y|X}) = I(X; Y)$) for all $j \in [n]$, then the map (4.23) is continuous.

Proof. For the first statement, apply representation

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

and (4.15).

For the second statement, take $Q_Y = \frac{1}{|\mathcal{X}|} \sum_x P_{Y|X=x}$. Note that

$$D(P_Y \| Q_Y) = \mathbb{E}_{Q_Y} \left[f \left(\sum_x P_X(x) h_x(Y) \right) \right],$$

where $f(t) = t \log t$ and $h_x(y) = \frac{dP_{Y|X=x}}{dQ_Y}(y)$ are bounded by $|\mathcal{X}|$ and non-negative. Thus, from the bounded-convergence theorem we have that

$$P_X \mapsto D(P_Y \| Q_Y)$$

is continuous. The proof is complete since by the golden formula

$$I(X; Y) = D(P_{Y|X} \| Q_Y | P_X) - D(P_Y \| Q_Y),$$

and the first term is linear in P_X .

For the third statement, form a chain $Z \rightarrow X \rightarrow Y$ with $Z \in [n]$ and $P_{X|Z=j} = P_j$. WLOG assume that P_1, \dots, P_n are distinct extreme points of $\text{co}(P_1, \dots, P_n)$. Then there is a linear bijection between P_Z and $P_X \in \Pi$. Furthermore, $I(X; Y) = I(Z; Y) + I(X; Y|Z)$. The first term is continuous in P_Z by the previous claim, whereas the second one is simply linear in P_Z . Thus, the map $P_Z \mapsto I(X; Y)$ is continuous and so is $P_X \mapsto I(X; Y)$. \square

Further properties of mutual information follow from $I(X; Y) = D(P_{X,Y} \| P_X P_Y)$ and corresponding properties of divergence, e.g.

⁴ Here we only assume that topology on the space of measures is compatible with the linear structure, so that all linear operations on measures are continuous.

1 Donsker-Varadhan and PAC-Bayes:

$$I(X; Y) = \sup_f \mathbb{E}[f(X, Y)] - \log \mathbb{E}[\exp\{f(X, \bar{Y})\}], \quad (4.24)$$

where \bar{Y} is a copy of Y , independent of X and supremum is over bounded, or even bounded continuous functions. Notice, however, that for mutual information we can also get a stronger characterization:⁵

$$I(X; Y) \geq \mathbb{E}[f(X, Y)] - \mathbb{E}[\log \mathbb{E}[\exp\{f(X, \bar{Y})\}|X]], \quad (4.25)$$

from which (4.24) follows by moving the outer expectation inside the log. Both of these can be used to show that $\mathbb{E}[f(X, Y)] \approx \mathbb{E}[f(X, \bar{Y})]$ as long as the dependence between X and Y (as measured by $I(X; Y)$) is weak. For example, suppose that for every x the random variable $h(x, \bar{Y})$ is ϵ -subgaussian, i.e.

$$\log \mathbb{E}[\exp\{\lambda h(x, \bar{Y})\}] \leq \lambda \mathbb{E}[h(x, \bar{Y})] + \frac{1}{2}\epsilon^2\lambda^2.$$

Then plugging $f = \lambda h$ into (4.25) and optimizing λ shows

$$\mathbb{E}[h(X, Y)] - \mathbb{E}[h(X, \bar{Y})] \leq \sqrt{2\epsilon^2 I(X; Y)}. \quad (4.26)$$

This allows one to control expectations of functions of dependent random variables by replacing them with independent pairs at the expense of (square-root of the) mutual information slack [326]. Variant of this idea for bounding deviations with high-probability is a foundation of the PAC-Bayes bounds on generalization of learning algorithms (in there, Y becomes training data, X is the selected hypothesis/predictor, $P_{X|Y}$ the learning algorithm, $\mathbb{E}[h(X, \bar{Y})]$ the test loss, etc); see Ex. I.44 and [57] for more.

- 2 (Uniform convergence and Donsker-Varadhan) There is an interesting other consequence of (4.25). By Theorem 4.1 we have $I(X; Y) \leq D(P_{X|Y} \| Q_X | P_Y)$ for any fixed Q_X . This lets us convert (4.25) into the following inequality: (we denote by \mathbb{E}_Y and $\mathbb{E}_{X|Y}$ the respective unconditional and conditional expectations): For every f , P_Y , Q_X and $P_{X|Y}$ we have

$$\mathbb{E}_Y [\mathbb{E}_{X|Y}[f(X, Y)] - \log \mathbb{E}_{\bar{Y}}[\exp f(X, \bar{Y})] - D(P_{X|Y} \| Q_X)] \leq 0.$$

Now because of the arbitrariness of $P_{X|Y}$, setting measurability issues aside, we get: For every f , P_Y and Q_X

$$\mathbb{E}_Y \left[\sup_{P_X} \{ \mathbb{E}_{X \sim P_X}[f(X, Y)] - \log \mathbb{E}_{\bar{Y}}[\exp f(X, \bar{Y})] - D(P_X \| Q_X) \} \right] \leq 0.$$

As an example, consider a countable collection $\{h_x : x \in \mathcal{X}\}$ of functions on \mathcal{Y} , each of which is ϵ -subgaussian. Then for every P_Y and Q_X , we have for every $\delta > 0$

$$\mathbb{E} \left[\sup_{x \in \mathcal{X}} \left\{ h_x(Y) - \mathbb{E}[h_x(Y)] - \delta \log \frac{1}{Q_X(x)} \right\} \right] \leq \frac{\epsilon^2}{2\delta}.$$

⁵ Just apply Donsker-Varadhan to $D(P_{Y|X=x_0} \| P_Y)$ and average over $x_0 \sim P_X$.

4.6 Variational characterizations and continuity of mutual information 63

For example, taking Q_X to be uniform on N elements recovers the standard bound on the maximum of subgaussian random variables: if H_1, \dots, H_N are ϵ -subgaussian, then

$$\mathbb{E} \left[\max_{1 \leq i \leq N} (H_i - \mathbb{E}[H_i]) \right] \leq \sqrt{2\epsilon^2 \log N}. \quad (4.27)$$

For a generalization see Ex. I.45.

3 If $(X_n, Y_n) \xrightarrow{d} (X, Y)$ converge in distribution, then

$$I(X; Y) \leq \liminf_{n \rightarrow \infty} I(X_n; Y_n). \quad (4.28)$$

- Good example of strict inequality: $X_n = Y_n = \frac{1}{n}Z$. In this case $(X_n, Y_n) \xrightarrow{d} (0, 0)$ but $I(X_n; Y_n) = H(Z) > 0 = I(0; 0)$.
- Even more impressive example: Let (X_p, Y_p) be uniformly distributed on the unit ℓ_p -ball on the plane: $\{x, y : |x|^p + |y|^p \leq 1\}$. Then as $p \rightarrow 0$, $(X_p, Y_p) \xrightarrow{d} (0, 0)$, but $I(X_p; Y_p) \rightarrow \infty$. (See Ex. I.36)

4 Mutual information as supremum over partitions:

$$I(X; Y) = \sup_{\{E_i\} \times \{F_j\}} \sum_{i,j} P_{X,Y}[E_i \times F_j] \log \frac{P_{X,Y}[E_i \times F_j]}{P_X[E_i]P_Y[F_j]}, \quad (4.29)$$

where supremum is over finite partitions of spaces \mathcal{X} and \mathcal{Y} .⁶

5 (Monotone convergence I):

$$I(X^\infty; Y) = \lim_{n \rightarrow \infty} I(X^n; Y) \quad (4.30)$$

$$I(X^\infty; Y^\infty) = \lim_{n \rightarrow \infty} I(X^n; Y^n) \quad (4.31)$$

This implies that the full amount of mutual information between two processes X^∞ and Y^∞ is contained in their finite-dimensional projections, leaving nothing in the tail σ -algebra. Note also that applying the (finite- n) chain rule to (4.30) recovers (4.1).

6 (Monotone convergence II): Let X_{tail} be a random variable such that $\sigma(X_{tail}) = \bigcap_{n \geq 1} \sigma(X_n^\infty)$. Then

$$I(X_{tail}; Y) = \lim_{n \rightarrow \infty} I(X_n^\infty; Y), \quad (4.32)$$

whenever the right-hand side is finite. This is a consequence of Proposition 4.15. Without the finiteness assumption the statement is incorrect. Indeed, consider $X_j \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(1/2)$ and $Y = X_0^\infty$. Then each $I(X_n^\infty; Y) = \infty$, but $X_{tail} = \text{const a.e.}$ by Kolmogorov's 0-1 law, and thus the left-hand side of (4.32) is zero.

⁶ To prove this from (4.9) one needs to notice that algebra of measurable rectangles is dense in the product σ -algebra. See [94, Sec. 2.2].

5

Extremization of mutual information: capacity saddle point

There are four fundamental optimization problems arising in information theory:

- I -projection: Given Q minimize $D(P\|Q)$ over convex class of P . (See Chapter 15.)
- Maximum likelihood: Given P minimize $D(P\|Q)$ over some class of Q . (See Section 29.3.)
- Rate-Distortion: Given P_X minimize $I(X; Y)$ over a convex class of $P_{Y|X}$. (See Chapter 26.)
- Capacity: Given $P_{Y|X}$ maximize $I(X; Y)$ over a convex class of P_X . (This chapter.)

In this chapter we show that all these problems have convex/concave objective functions, discuss iterative algorithms for solving them, and study the capacity problem in more detail.

5.1 Convexity of information measures

Theorem 5.1. $(P, Q) \mapsto D(P\|Q)$ is convex.

Proof. Let $P_X = Q_X = \text{Ber}(\lambda)$ and define two conditional kernels:

$$\begin{aligned} P_{Y|X=0} &= P_0, & P_{Y|X=1} &= P_1 \\ Q_{Y|X=0} &= Q_0, & Q_{Y|X=1} &= Q_1 \end{aligned}$$

An explicit calculation shows that

$$D(P_{X,Y}\|Q_{X,Y}) = \bar{\lambda}D(P_0\|Q_0) + \lambda D(P_1\|Q_1).$$

Therefore, from the DPI (monotonicity) we get:

$$\bar{\lambda}D(P_0\|Q_0) + \lambda D(P_1\|Q_1) = D(P_{X,Y}\|Q_{X,Y}) \geq D(P_Y\|Q_Y) = D(\bar{\lambda}P_0 + \lambda P_1\|\bar{\lambda}Q_0 + \lambda Q_1).$$

□

Remark 5.1. The proof shows that for an arbitrary measure of similarity $\mathcal{D}(P\|Q)$ convexity of $(P, Q) \mapsto \mathcal{D}(P\|Q)$ is equivalent to “conditioning increases divergence” property of \mathcal{D} . Convexity can also be understood as “mixing decreases divergence”.

Remark 5.2. There are a number of alternative arguments possible. For example, $(p, q) \mapsto p \log \frac{p}{q}$ is convex on \mathbb{R}_+^2 , which is a manifestation of a general phenomenon: for a convex $f(\cdot)$ the perspective function $(p, q) \mapsto qf\left(\frac{p}{q}\right)$ is convex too. Yet another way is to invoke the Donsker-Varadhan variational representation Theorem 4.8 and notice that supremum of convex functions is convex.

5.2 Extremization of mutual information 65

Theorem 5.2. *The map $P_X \mapsto H(X)$ is concave. Furthermore, if $P_{Y|X}$ is any channel, then $P_X \mapsto H(X|Y)$ is concave. If \mathcal{X} is finite, then $P_X \mapsto H(X|Y)$ is continuous.*

Proof. For the special case of the first claim, when P_X is on a finite alphabet, the proof is complete by $H(X) = \log |\mathcal{X}| - D(P_X \| U_X)$. More generally, we prove the second claim as follows. Let $f(P_X) = H(X|Y)$. Introduce a random variable $U \sim \text{Ber}(\lambda)$ and define the transformation

$$P_{X|U} = \begin{cases} P_0 & U = 0 \\ P_1 & U = 1 \end{cases}$$

Consider the probability space $U \rightarrow X \rightarrow Y$. Then we have $f(\lambda P_1 + (1 - \lambda)P_0) = H(X|Y)$ and $\lambda f(P_1) + (1 - \lambda)f(P_0) = H(X|Y, U)$. Since $H(X|Y, U) \leq H(X|Y)$, the proof is complete. Continuity follows from Proposition 4.16. \square

Recall that $I(X; Y)$ is a function of $P_{X,Y}$, or equivalently, $(P_X, P_{Y|X})$. Denote $I(P_X, P_{Y|X}) = I(X; Y)$.

Theorem 5.3 (Mutual Information).

- For fixed $P_{Y|X}$, $P_X \mapsto I(P_X, P_{Y|X})$ is concave.
- For fixed P_X , $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ is convex.

Proof. There are several ways to prove the first statement, all having their merits.

- *First proof:* Introduce $\theta \in \text{Ber}(\lambda)$. Define $P_{X|\theta=0} = P_X^0$ and $P_{X|\theta=1} = P_X^1$. Then $\theta \rightarrow X \rightarrow Y$. Then $P_X = \bar{\lambda}P_X^0 + \lambda P_X^1$. $I(X; Y) = I(X, \theta; Y) = I(\theta; Y) + I(X; Y|\theta) \geq I(X; Y|\theta)$, which is our desired $I(\bar{\lambda}P_X^0 + \lambda P_X^1, P_{Y|X}) \geq \bar{\lambda}I(P_X^0, P_{Y|X}) + \lambda I(P_X^1, P_{Y|X})$.
- *Second proof:* $I(X; Y) = \min_Q D(P_{Y|X} \| Q | P_X)$, which is a pointwise minimum of affine functions in P_X and hence concave.
- *Third proof:* Pick a Q and use the golden formula: $I(X; Y) = D(P_{Y|X} \| Q | P_X) - D(P_Y \| Q)$, where $P_X \mapsto D(P_Y \| Q)$ is convex, as the composition of the $P_X \mapsto P_Y$ (affine) and $P_Y \mapsto D(P_Y \| Q)$ (convex).

To prove the second (convexity) statement, simply notice that

$$I(X; Y) = D(P_{Y|X} \| P_Y | P_X).$$

The argument P_Y is a linear function of $P_{Y|X}$ and thus the statement follows from convexity of D in the pair. \square

5.2 Extremization of mutual information

Two problems of interest

- Fix $P_{Y|X} \rightarrow \max_{P_X} I(X; Y)$ — channel coding (Part IV)

Note: This maximum is called “capacity” of a set of distributions $\{P_{Y|X=x}, x \in \mathcal{X}\}$.

- Fix $P_X \rightarrow \min_{P_{Y|X}} I(X; Y)$ — lossy compression (Part V)

Theorem 5.4 (Saddle point). *Let \mathcal{P} be a convex set of distributions on \mathcal{X} . Suppose there exists $P_X^* \in \mathcal{P}$, called a capacity-achieving input distribution, such that*

$$\sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) \triangleq C.$$

Let $P_Y^* = P_X^* P_{Y|X}$, called a capacity-achieving output distribution. Then for all $P_X \in \mathcal{P}$ and for all Q_Y , we have

$$D(P_{Y|X} \| P_Y^* | P_X) \leq D(P_{Y|X} \| P_Y^* | P_X^*) \leq D(P_{Y|X} \| Q_Y | P_X^*). \quad (5.1)$$

Proof. Right inequality: obvious from $C = I(P_X^*, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X^*)$.

Left inequality: If $C = \infty$, then trivial. In the sequel assume that $C < \infty$, hence $I(P_X, P_{Y|X}) < \infty$ for all $P_X \in \mathcal{P}$. Let $P_{X_\lambda} = \lambda P_X + \bar{\lambda} P_X^* \in \mathcal{P}$ and $P_{Y_\lambda} = P_{Y|X} \circ P_{X_\lambda}$. Clearly, $P_{Y_\lambda} = \lambda P_Y + \bar{\lambda} P_Y^*$, where $P_Y = P_{Y|X} \circ P_X$.

We have the following chain then:

$$\begin{aligned} C &\geq I(X_\lambda; Y_\lambda) = D(P_{Y|X} \| P_{Y_\lambda} | P_{X_\lambda}) \\ &= \lambda D(P_{Y|X} \| P_{Y_\lambda} | P_X) + \bar{\lambda} D(P_{Y|X} \| P_{Y_\lambda} | P_X^*) \\ &\geq \lambda D(P_{Y|X} \| P_{Y_\lambda} | P_X) + \bar{\lambda} C \\ &= \lambda D(P_{X,Y} \| P_X P_{Y_\lambda}) + \bar{\lambda} C, \end{aligned}$$

where inequality is by the right part of (5.1) (already shown). Thus, subtracting $\bar{\lambda} C$ and dividing by λ we get

$$D(P_{X,Y} \| P_X P_{Y_\lambda}) \leq C$$

and the proof is completed by taking $\liminf_{\lambda \rightarrow 0}$ and applying the lower semicontinuity of divergence (Theorem 4.12). \square

Corollary 5.5. *In addition to the assumptions of Theorem 5.6, suppose $C < \infty$. Then the capacity-achieving output distribution P_Y^* is unique. It satisfies the property that for any P_Y induced by some $P_X \in \mathcal{P}$ (i.e. $P_Y = P_{Y|X} \circ P_X$) we have*

$$D(P_Y \| P_Y^*) \leq C < \infty \quad (5.2)$$

and in particular $P_Y \ll P_Y^*$.

Proof. The statement is: $I(P_X, P_{Y|X}) = C \Rightarrow P_Y = P_Y^*$. Indeed:

$$\begin{aligned} C &= D(P_{Y|X} \| P_Y | P_X) = D(P_{Y|X} \| P_Y^* | P_X) - D(P_Y \| P_Y^*) \\ &\leq D(P_{Y|X} \| P_Y^* | P_X^*) - D(P_Y \| P_Y^*) \end{aligned}$$

5.2 Extremization of mutual information 67

$$= C - D(P_Y \| P_Y^*) \Rightarrow P_Y = P_Y^*$$

Statement (5.2) follows from the left inequality in (5.1) and “conditioning increases divergence”. \square

Remark 5.3. • The finiteness of C is necessary for Corollary 5.1 to hold. For a counterexample, consider the identity channel $Y = X$, where X takes values on integers. Then any distribution with infinite entropy is a capacity-achieving input (and output) distribution.

- Unlike the output distribution, capacity-achieving input distribution need not be unique. For example, consider $Y_1 = X_1 \oplus Z_1$ and $Y_2 = X_2$ where $Z_1 \sim \text{Ber}(\frac{1}{2})$ is independent of X_1 . Then $\max_{P_{X_1 X_2}} I(X_1, X_2; Y_1, Y_2) = \log 2$, achieved by $P_{X_1 X_2} = \text{Ber}(p) \times \text{Ber}(\frac{1}{2})$ for any p . Note that the capacity-achieving *output* distribution is unique: $P_{Y_1 Y_2}^* = \text{Ber}(\frac{1}{2}) \times \text{Ber}(\frac{1}{2})$.

Review: Minimax and saddlepoint

Suppose we have a bivariate function f . Then we always have the *minimax inequality*:

$$\inf_y \sup_x f(x, y) \geq \sup_x \inf_y f(x, y).$$

When does it hold with equality?

- 1 It turns out minimax equality is implied by the existence of a saddle point (x^*, y^*) , i.e.,

$$f(x, y^*) \leq f(x^*, y^*) \leq f(x^*, y) \quad \forall x, y$$

Furthermore, minimax equality also implies existence of saddle point if inf and sup are achieved c.f. [31, Section 2.6]) for all x, y [Straightforward to check. See proof of corollary below].

- 2 There are a number of known criteria establishing

$$\inf_y \sup_x f(x, y) = \sup_x \inf_y f(x, y)$$

They usually require some continuity of f , compactness of domains and concavity in x and convexity in y . One of the most general version is due to M. Sion [275].

- 3 The mother result of all this minimax theory is a theorem of von Neumann on bilinear functions: Let A and B have finite alphabets, and $g(a, b)$ be arbitrary, then

$$\min_{P_A} \max_{P_B} \mathbb{E}[g(A, B)] = \max_{P_B} \min_{P_A} \mathbb{E}[g(A, B)]$$

Here $(x, y) \leftrightarrow (P_A, P_B)$ and $f(x, y) \leftrightarrow \sum_{a,b} P_A(a)P_B(b)g(a, b)$.

- 4 A more general version is: if \mathcal{X} and \mathcal{Y} are compact convex domains in \mathbb{R}^n , $f(x, y)$ continuous in (x, y) , concave in x and convex in y then

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x, y) = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} f(x, y)$$

Applying Theorem 5.6 to conditional divergence gives the following result.

Corollary 5.6 (Minimax). *Under assumptions of Theorem 5.6, we have*

$$\begin{aligned}\max_{P_X \in \mathcal{P}} I(X; Y) &= \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \\ &= \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X)\end{aligned}$$

Proof. This follows from the saddle-point: Maximizing/minimizing the leftmost/rightmost sides of (5.1) gives

$$\begin{aligned}\min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) &\leq \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) = D(P_{Y|X} \| P_Y^* | P_X^*) \\ &\leq \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X^*) \leq \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X).\end{aligned}$$

but by definition $\min \max \geq \max \min$. Note that we were careful to only use max and min for the cases where we know the optimum is achievable. \square

5.3 Capacity as information radius

Review: Radius and diameter

Let (X, d) be a metric space. Let A be a bounded subset.

- 1 *Radius* (aka Chebyshev radius) of A : the radius of the smallest ball that covers A , i.e.,

$$\text{rad}(A) = \inf_{y \in X} \sup_{x \in A} d(x, y). \quad (5.3)$$

- 2 *Diameter* of A :

$$\text{diam}(A) = \sup_{x, y \in A} d(x, y). \quad (5.4)$$

Note that the radius and the diameter both measure the massiveness/richness of a set.

- 3 From definition and triangle inequality we have

$$\frac{1}{2}\text{diam}(A) \leq \text{rad}(A) \leq \text{diam}(A). \quad (5.5)$$

The lower and upper bounds are achieved when A is, for example, a Euclidean ball and the Hamming space, respectively.

- 4 In many special cases, the upper bound in (5.5) can be improved:

- A result of Bohnenblust [42] shows that in \mathbb{R}^n equipped with any norm we always have $\text{rad}(A) \leq \frac{n}{n+1}\text{diam}(A)$.
- For \mathbb{R}^n with Euclidean distance Jung proved $\text{rad}(A) \leq \sqrt{\frac{n}{2(n+1)}}\text{diam}(A)$, attained by simplex. The best constant is sometimes called the Jung constant of the space.
- For \mathbb{R}^n with ℓ_∞ -norm the situation is even simpler: $\text{rad}(A) = \frac{1}{2}\text{diam}(A)$; such spaces are called centrable.

The next simple corollary shows that capacity is just the radius of a finite collection of distributions $\{P_{Y|X=x} : x \in \mathcal{X}\}$ when distances are measured by divergence (although, we remind, divergence is not a metric).

Corollary 5.7. *For any finite \mathcal{X} and any kernel $P_{Y|X}$, the maximal mutual information over all distributions P_X on \mathcal{X} satisfies*

$$\begin{aligned} \max_{P_X} I(X; Y) &= \max_{x \in \mathcal{X}} D(P_{Y|X=x} \| P_Y^*) \\ &= D(P_{Y|X=x} \| P_Y^*) \quad \forall x : P_X^*(x) > 0. \end{aligned}$$

The last corollary gives a geometric interpretation to capacity: It equals the radius of the smallest divergence-“ball” that encompasses all distributions $\{P_{Y|X=x} : x \in \mathcal{X}\}$. Moreover, the optimal center P_Y^* is a convex combination of some $P_{Y|X=x}$ and is *equidistant* to those.

The following is the information-theoretic version of “radius \leq diameter” (in KL divergence) for arbitrary input space (see Theorem 32.4 for a related representation):

Corollary 5.8. *Let $\{P_{Y|X=x} : x \in \mathcal{X}\}$ be a set of distributions. Then*

$$C = \sup_{P_X} I(X; Y) \leq \underbrace{\inf_Q \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q)}_{\text{radius}} \leq \underbrace{\sup_{x, x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'})}_{\text{diameter}}$$

Proof. By the golden formula Corollary 4.1, we have

$$I(X; Y) = \inf_Q D(P_{Y|X} \| Q | P_X) \leq \inf_Q \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q) \leq \inf_{x' \in \mathcal{X}} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}).$$

□

5.4 Existence of capacity-achieving output distribution (general case)

In the previous section we have shown that the solution to

$$C = \sup_{P_X \in \mathcal{P}} I(X; Y)$$

can be (a) interpreted as a saddle point; (b) written in the minimax form; and (c) that the capacity-achieving output distribution P_Y^* is unique. This was all done under the extra assumption that the supremum over P_X is attainable. It turns out, properties b) and c) can be shown without that extra assumption.

Theorem 5.9 (Kemperman). *For any $P_{Y|X}$ and a convex set of distributions \mathcal{P} such that*

$$C = \sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) < \infty, \quad (5.6)$$

there exists a unique P_Y^ with the property that*

$$C = \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X). \quad (5.7)$$

Furthermore,

$$C = \sup_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \quad (5.8)$$

$$= \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) \quad (5.9)$$

$$= \min_{Q_Y} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q_Y), \quad (\text{if } \mathcal{P} = \{\text{all } P_X\}). \quad (5.10)$$

5.4 Existence of capacity-achieving output distribution (general case) 71

Note that Condition (5.6) is automatically satisfied if there exists a Q_Y such that

$$\sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) < \infty. \quad (5.11)$$

Example 5.1 (Non-existence of capacity-achieving input distribution). Let $Z \sim \mathcal{N}(0, 1)$ and consider the problem

$$C = \sup_{\substack{P_X: \mathbb{E}[X]=0, \mathbb{E}[X^2]=P \\ \mathbb{E}[X^4]=s}} I(X; X + Z). \quad (5.12)$$

Without the constraint $\mathbb{E}[X^4] = s$, the capacity is uniquely achieved at the input distribution $P_X = \mathcal{N}(0, P)$; see Theorem 5.9. When $s \neq 3P^2$, such P_X is no longer feasible. However, for $s > 3P^2$ the maximum

$$C = \frac{1}{2} \log(1 + P)$$

is still attainable. Indeed, we can add a small “bump” to the gaussian distribution as follows:

$$P_X = (1 - p)\mathcal{N}(0, P) + p\delta_x,$$

where $p \rightarrow 0$ and $x \rightarrow \infty$ such that $px^2 \rightarrow 0$ but $px^4 \rightarrow s - 3P^2 > 0$. This shows that for the problem (5.12) with $s > 3P^2$, the capacity-achieving input distribution does not exist, but the capacity-achieving output distribution $P_Y^* = \mathcal{N}(0, 1 + P)$ exists and is unique as Theorem 5.8 shows.

Proof of Theorem 5.8. Let P'_{X_n} be a sequence of input distributions achieving C , i.e., $I(P'_{X_n}, P_{Y|X}) \rightarrow C$. Let \mathcal{P}_n be the convex hull of $\{P'_{X_1}, \dots, P'_{X_n}\}$. Since \mathcal{P}_n is a finite-dimensional simplex, the (concave) function $P_X \mapsto I(P_X, P_{Y|X})$ is continuous (Proposition 4.16) and attains its maximum at some point $P_{X_n} \in \mathcal{P}_n$, i.e.,

$$I_n \triangleq I(P_{X_n}, P_{Y|X}) = \max_{P_X \in \mathcal{P}_n} I(P_X, P_{Y|X}).$$

Denote by P_{Y_n} be the output distribution induced by P_{X_n} . We have then:

$$D(P_{Y_n} \| P_{Y_{n+k}}) = D(P_{Y|X} \| P_{Y_{n+k}} | P_{X_n}) - D(P_{Y|X} \| P_{Y_n} | P_{X_n}) \quad (5.13)$$

$$\leq I(P_{X_{n+k}}, P_{Y|X}) - I(P_{X_n}, P_{Y|X}) \quad (5.14)$$

$$\leq C - I_n, \quad (5.15)$$

where in (5.14) we applied Theorem 5.6 to $(\mathcal{P}_{n+k}, P_{Y_{n+k}})$. The crucial idea is to apply comparison of KL divergence (which is not a distance) with a true distance known as *total variation* defined in (7.3) below. Such comparisons are going to be the topic of Chapter 7. Here we assume for granted validity of Pinsker’s inequality (see Theorem 7.17). According to that inequality and since $I_n \nearrow C$, we conclude that the sequence P_{Y_n} is Cauchy in total variation:

$$\sup_{k \geq 1} \text{TV}(P_{Y_n}, P_{Y_{n+k}}) \rightarrow 0, \quad n \rightarrow \infty.$$

Since the space of all probability distributions on a fixed alphabet is complete in total variation, the sequence must have a limit point $P_{Y_n} \rightarrow P_Y^*$. Convergence in TV implies weak convergence,

and thus by taking a limit as $k \rightarrow \infty$ in (5.15) and applying the lower semicontinuity of divergence (Theorem 4.12) we get

$$D(P_{Y_n} \| P_Y^*) \leq \lim_{k \rightarrow \infty} D(P_{Y_n} \| P_{Y_{n+k}}) \leq C - I_n,$$

and therefore, $P_{Y_n} \rightarrow P_Y^*$ in the (stronger) sense of $D(P_{Y_n} \| P_Y^*) \rightarrow 0$. By Theorem 4.1,

$$D(P_{Y|X} \| P_Y^* | P_{X_n}) = I_n + D(P_{Y_n} \| P_Y^*) \rightarrow C. \quad (5.16)$$

Take any $P_X \in \bigcup_{k \geq 1} \mathcal{P}_k$. Then $P_X \in \mathcal{P}_n$ for all sufficiently large n and thus by Theorem 5.6

$$D(P_{Y|X} \| P_{Y_n} | P_X) \leq I_n \leq C, \quad (5.17)$$

which, by the lower semicontinuity of divergence and Fatou's lemma, implies

$$D(P_{Y|X} \| P_Y^* | P_X) \leq C. \quad (5.18)$$

To prove that (5.18) holds for arbitrary $P_X \in \mathcal{P}$, we may repeat the argument above with \mathcal{P}_n replaced by $\tilde{\mathcal{P}}_n = \text{conv}(\{P_X\} \cup \mathcal{P}_n)$, denoting the resulting sequences by $\tilde{P}_{X_n}, \tilde{P}_{Y_n}$ and the limit point by \tilde{P}_Y^* , and obtain:

$$D(P_{Y_n} \| \tilde{P}_{Y_n}) = D(P_{Y|X} \| \tilde{P}_{Y_n} | P_{X_n}) - D(P_{Y|X} \| P_{Y_n} | P_{X_n}) \quad (5.19)$$

$$\leq C - I_n, \quad (5.20)$$

where (5.20) follows from (5.18) since $P_{X_n} \in \tilde{\mathcal{P}}_n$. Hence taking limit as $n \rightarrow \infty$ we have $\tilde{P}_Y^* = P_Y^*$ and therefore (5.18) holds.

To see the uniqueness of P_Y^* , assuming there exists Q_Y^* that fulfills $C = \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y^* | P_X)$, we show $Q_Y^* = P_Y^*$. Indeed,

$$C \geq D(P_{Y|X} \| Q_Y^* | P_{X_n}) = D(P_{Y|X} \| P_{Y_n} | P_{X_n}) + D(P_{Y_n} \| Q_Y^*) = I_n + D(P_{Y_n} \| Q_Y^*).$$

Since $I_n \rightarrow C$, we have $D(P_{Y_n} \| Q_Y^*) \rightarrow 0$. Since we have already shown that $D(P_{Y_n} \| P_Y^*) \rightarrow 0$, we conclude $P_Y^* = Q_Y^*$ (this can be seen, for example, from Pinsker's inequality and the triangle inequality $\text{TV}(P_Y^*, Q_Y^*) \leq \text{TV}(P_{Y_n}, Q_Y^*) + \text{TV}(P_{Y_n}, P_Y^*) \rightarrow 0$).

Finally, to see (5.9), note that by definition capacity as a max-min is at most the min-max, i.e.,

$$C = \sup_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \leq \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) \leq \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) = C$$

in view of (5.16) and (5.17). \square

Corollary 5.10. *Let \mathcal{X} be countable and \mathcal{P} a convex set of distributions on \mathcal{X} . If $\sup_{P_X \in \mathcal{P}} H(X) < \infty$ then*

$$\sup_{P_X \in \mathcal{P}} H(X) = \min_{Q_X} \sup_{P_X \in \mathcal{P}} \sum_x P_X(x) \log \frac{1}{Q_X(x)} < \infty$$

and the optimizer Q_X^* exists and is unique. If $Q_X^* \in \mathcal{P}$, then it is also the unique maximizer of $H(X)$.

Proof. Just apply Kemperman's Theorem 5.8 to the identity channel $Y = X$. \square

5.5 Gaussian saddle point 73

Example 5.2 (Max entropy). Assume that $f: \mathbb{Z} \rightarrow \mathbb{R}$ is such that $Z(\lambda) \triangleq \sum_{n \in \mathbb{Z}} \exp\{-\lambda f(n)\} < \infty$ for all $\lambda > 0$. Then

$$\max_{X: \mathbb{E}[f(X)] \leq a} H(X) \leq \inf_{\lambda > 0} \{\lambda a + \log Z(\lambda)\}.$$

This follows from taking $Q_X(n) = Z(\lambda)^{-1} \exp\{-\lambda f(n)\}$ in Corollary 5.5. Distributions of this form are known as Gibbs distributions for the energy function f . This bound is often tight and achieved by $P_X(n) = Z(\lambda^*)^{-1} \exp\{-\lambda^* f(n)\}$ with λ^* being the minimizer.

5.5 Gaussian saddle point

For additive noise, there is also a different kind of saddle point between P_X and the distribution of noise:

Theorem 5.11. *Let $X_g \sim \mathcal{N}(0, \sigma_X^2)$, $N_g \sim \mathcal{N}(0, \sigma_N^2)$, $X_g \perp\!\!\!\perp N_g$. Then:*

1. “Gaussian capacity”:

$$C = I(X_g; X_g + N_g) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2} \right)$$

2. “Gaussian input is the best for Gaussian noise”: For all $X \perp\!\!\!\perp N_g$ and $\text{Var } X \leq \sigma_X^2$,

$$I(X; X + N_g) \leq I(X_g; X_g + N_g),$$

with equality iff $X \stackrel{d}{=} X_g$.

3. “Gaussian noise is the worst for Gaussian input”: For all N s.t. $\mathbb{E}[X_g N] = 0$ and $\mathbb{E}N^2 \leq \sigma_N^2$,

$$I(X_g; X_g + N) \geq I(X_g; X_g + N_g),$$

with equality iff $N \stackrel{d}{=} N_g$ and independent of X_g .

Interpretations:

- 1 For AWGN channel, Gaussian input is the most favorable. Indeed, immediately from the second statement we have

$$\max_{X: \text{Var } X \leq \sigma_X^2} I(X; X + N_g) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2} \right)$$

which is the capacity formula for the AWGN channel.

- 2 For Gaussian source, additive Gaussian noise is the worst in the sense that it minimizes the mutual information provided by the noisy version.

Proof. WLOG, assume all random variables have zero mean. Let $Y_g = X_g + N_g$. Define

$$f(x) \triangleq D(P_{Y_g|X_g=x} \| P_{Y_g}) = D(\mathcal{N}(x, \sigma_N^2) \| \mathcal{N}(0, \sigma_X^2 + \sigma_N^2)) = \underbrace{\frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2} \right)}_{=C} + \frac{\log e}{2} \frac{x^2 - \sigma_X^2}{\sigma_X^2 + \sigma_N^2}$$

1. Compute $I(X_g; X_g + N_g) = \mathbb{E}[f(X_g)] = C$
2. Recall the inf-representation (Corollary 4.1): $I(X; Y) = \min_Q D(P_{Y|X} \| Q|P_X)$. Then

$$I(X; X + N_g) \leq D(P_{Y_g|X_g} \| P_{Y_g}|P_X) = \mathbb{E}[f(X)] \leq C < \infty.$$

Furthermore, if $I(X; X + N_g) = C$, then the uniqueness of the capacity-achieving output distribution, cf. Corollary 5.1, implies $P_Y = P_{Y_g}$. But $P_Y = P_X * \mathcal{N}(0, \sigma_N^2)$, where $*$ denotes convolution. Then it must be that $X \sim \mathcal{N}(0, \sigma_X^2)$ simply by considering characteristic functions:

$$\Psi_X(t) \cdot e^{-\frac{1}{2}\sigma_N^2 t^2} = e^{-\frac{1}{2}(\sigma_X^2 + \sigma_N^2)t^2} \Rightarrow \Psi_X(t) = e^{-\frac{1}{2}\sigma_X^2 t^2} \implies X \sim \mathcal{N}(0, \sigma_X^2)$$

3. Let $Y = X_g + N$ and let $P_{Y|X_g}$ be the respective kernel. Note that here we only assume that N is *uncorrelated* with X_g , i.e., $\mathbb{E}[NX_g] = 0$, not necessarily independent. Then

$$I(X_g; X_g + N) \geq \mathbb{E} \log \frac{dP_{X_g|Y_g}(X_g|Y)}{dP_{X_g}(X_g)} \quad (5.21)$$

$$= \mathbb{E} \log \frac{dP_{Y_g|X_g}(Y|X_g)}{dP_{Y_g}(Y)} \quad (5.22)$$

$$= C + \frac{\log e}{2} \mathbb{E} \left[\frac{Y^2}{\sigma_X^2 + \sigma_N^2} - \frac{N^2}{\sigma_N^2} \right] \quad (5.23)$$

$$= C + \frac{\log e}{2} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} \left(1 - \frac{\mathbb{E}N^2}{\sigma_N^2} \right) \quad (5.24)$$

$$\geq C, \quad (5.25)$$

where

- (5.21): follows from (4.7),
- (5.22): $\frac{dP_{X_g|Y_g}}{dP_{X_g}} = \frac{dP_{Y_g|X_g}}{dP_{Y_g}}$
- (5.24): $\mathbb{E}[X_g N] = 0$ and $\mathbb{E}[Y^2] = \mathbb{E}[N^2] + \mathbb{E}[X_g^2]$.
- (5.25): $\mathbb{E}N^2 \leq \sigma_N^2$.

Finally, the conditions for equality in (5.21) (see (4.8)) require

$$D(P_{X_g|Y} \| P_{X_g|Y_g}|P_Y) = 0$$

Thus, $P_{X_g|Y} = P_{X_g|Y_g}$, i.e., X_g is conditionally Gaussian: $P_{X_g|Y=y} = \mathcal{N}(by, c^2)$ for some constants b and c . In other words, under $P_{X_g|Y}$, we have

$$X_g = bY + cZ, \quad Z \sim \text{Gaussian} \perp Y.$$

But then Y must be Gaussian itself by Cramer's Theorem [76] or simply by considering characteristic functions:

$$\Psi_Y(t) \cdot e^{ct^2} = e^{c't^2} \Rightarrow \Psi_Y(t) = e^{c''t^2} \implies Y \text{ is Gaussian}$$

5.6 Iterative algorithms: Blahut-Arimoto, Expectation-Maximization, Sinkhorn 75

Therefore, (X_g, Y) must be jointly Gaussian and hence $N = Y - X_g$ is Gaussian. Thus we conclude that it is only possible to attain $I(X_g; X_g + N) = C$ if N is Gaussian of variance σ_N^2 and independent of X_g . \square

5.6 Iterative algorithms: Blahut-Arimoto, Expectation-Maximization, Sinkhorn

Although the optimization problems that we discussed above are convex (and thus would be considered algorithmically “easy”), there are still clever ideas used to speed up their numerical solutions. The main underlying principle is the following *alternating minimization algorithm*:

- Optimization problem: $\min_t f(t)$.
- Assumption I: $f(t) = \min_s F(t, s)$ (i.e. f can be written as a minimum of some other function F).
- Assumption II: There exist two solvers $t^*(s) = \operatorname{argmin}_t F(t, s)$ and $s^*(t) = \operatorname{argmin}_s F(t, s)$.
- Iterative algorithm:
 - Step 0: Fix some s_0, t_0 .
 - Step $2k - 1$: $s_k = s^*(t_{k-1})$.
 - Step $2k$: $t_k = t^*(s_k)$.

Note that there is a steady improvement at each step (the value $F(s_k, t_k)$ is decreasing), so it can be often proven that the algorithm converges to a local minimum, or even a global minimum under appropriate conditions (e.g. the convexity of f). Below we discuss several applications of this idea, and refer to [81] for proofs of convergence. We need a result, which will be derived in Chapter 15: for any function $c : \mathcal{Y} \rightarrow \mathbb{R}$ and any Q_Y on \mathcal{Y} , under the integrability condition $Z = \int Q_Y(dy) \exp\{-c(y)\} < \infty$,

$$\min_{P_Y} D(P_Y \| Q_Y) + \mathbb{E}_{Y \sim P_Y}[c(Y)] \quad (5.26)$$

is attained at $P_Y^*(dy) = \frac{1}{Z} Q_Y(dy) \exp\{-c(y)\}$.

For simplicity below we only consider the case of discrete alphabets \mathcal{X}, \mathcal{Y} .

Maximizing mutual information (capacity). We have a fixed $P_{Y|X}$ and the optimization problem

$$C = \max_{P_X} I(X; Y) = \max_{P_X} \max_{Q_{X|Y}} \mathbb{E}_{P_{X,Y}} \left[\log \frac{Q_{X|Y}}{P_X} \right].$$

This results in the iterations:

$$\begin{aligned} Q_{X|Y}(x|y) &\leftarrow \frac{1}{Z(y)} P_X(x) P_{Y|X}(y|x) \\ P_X(x) &\leftarrow Q'(x) \triangleq \frac{1}{Z} \exp \left\{ \sum_y P_{Y|X}(y|x) \log Q_{X|Y}(x|y) \right\}, \end{aligned}$$

where $Z(y)$ and Z are normalization constants. To derive this, notice that for a fixed P_X the optimal $Q_{X|Y} = P_{X|Y}$. For a fixed $Q_{X|Y}$, we can see that

$$\mathbb{E}_{P_{X,Y}} \left[\log \frac{Q_{X|Y}}{P_X} \right] = \log Z - D(P_X||Q'),$$

and thus the optimal $P_X = Q'$.

Denoting P_n to be the value of P_X at the n th iteration, we observe that

$$I(P_n, P_{Y|X}) \leq C \leq \sup_x D(P_{Y|X=x}||P_{Y|X} \circ P_n). \quad (5.27)$$

This is useful since at every iteration not only we get an estimate of the optimizer P_n , but also the gap to optimality $C - I(P_n, P_{Y|X}) \leq C - \text{RHS}$. It can be shown, furthermore, that both RHS and LHS in (5.27) monotonically converge to C as $n \rightarrow \infty$, see [81] for details.

Minimizing mutual information (rate-distortion). We have a fixed P_X , a cost function $c(x, y)$ and the optimization problem

$$R = \min_{P_{Y|X}} I(X; Y) + \mathbb{E}[d(X, Y)] = \min_{P_{Y|X}, Q_Y} D(P_{Y|X}||Q_Y|P_X) + \mathbb{E}[d(X, Y)]. \quad (5.28)$$

Using (5.26) we derive the iterations:

$$\begin{aligned} P_{Y|X}(y|x) &\leftarrow \frac{1}{Z(x)} Q_Y(y) \exp\{-d(x, y)\} \\ Q_Y &\leftarrow P_{Y|X} \circ P_X. \end{aligned}$$

A sandwich bound similar to (5.27) holds here, see (5.30), so that one gets two computable sequences converging to R from above and below, as well as $P_{Y|X}$ converging to the argmin in (5.28).

EM algorithm (convex case). Proposed in [87], Expectation-maximization (EM) algorithm is a heuristic for solving the maximal likelihood problem. It is known to converge to the global maximizer for convex problems. Let us consider such a simple case. Given a distribution P_X our goal is to minimize the divergence with respect to the mixture $Q_X = Q_{X|Y} \circ Q_Y$:

$$L = \min_{Q_Y} D(P_X||Q_{X|Y} \circ Q_Y), \quad (5.29)$$

where $Q_{X|Y}$ is a given channel. This is a problem arising in the maximum likelihood estimation for mixture models where Q_Y is the unknown mixing distribution and $P_X = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ is the empirical distribution of the sample (x_1, \dots, x_n) .¹

¹ Note that EM algorithm is also applicable more generally, when $Q_{X|Y}^{(\theta)}$ itself depends on the unknown parameter θ and the goal (see Section 29.3) is to maximize the total log likelihood $\sum_{i=1}^n \log Q_X^{(\theta)}(x_i)$ joint over (Q_Y, θ) . A canonical example (which was one of the original motivations for the EM algorithm) is a k -component Gaussian mixture $Q_X^{(\theta)} = \sum_{j=1}^k w_j \mathcal{N}(\mu_j, 1)$; in other words, $Q_Y = (w_1, \dots, w_k)$, $Q_{X|Y=j}^{(\theta)} = \mathcal{N}(\mu_j, 1)$ and $\theta = (\mu_1, \dots, \mu_k)$. If the centers μ_j 's are known and only the weights w_j 's are to be estimated, then we get the simple convex case in (5.29). In general the log likelihood function is non-convex in the centers and EM iterations may not converge to the global optimum even with infinite sample size (see [165] for an example with $k = 3$).

5.6 Iterative algorithms: Blahut-Arimoto, Expectation-Maximization, Sinkhorn

77

To derive an iterative algorithm for (5.29), we write

$$\min_{Q_Y} D(P_X \| Q_X) = \min_{Q_X} \min_{P_{Y|X}} D(P_{X,Y} \| Q_{X,Y}).$$

(Note that taking $d(x,y) = -\log \frac{dQ_{X|Y}}{dP_X}$ shows that this problem is equivalent to (5.28).) By the chain rule, thus, we find the iterations

$$\begin{aligned} P_{Y|X} &\leftarrow \frac{1}{Z(x)} Q_Y(y) Q_{X|Y}(x|y) \\ Q_Y &\leftarrow P_{Y|X} \circ P_X. \end{aligned}$$

Denote by Q_n the value of $Q_X = Q_{X|Y} \circ Q_Y$ at the n th iteration. Notice that for any n and all Q_X we have from Jensen's inequality

$$D(P_X \| Q_X) - D(P_X \| Q_n) = \mathbb{E}_{X \sim P_X} [\log \mathbb{E}_{Y \sim Q_Y} \frac{dQ_{X|Y}}{dQ_n}] \geq \text{gap}(Q_n),$$

where we defined $\text{gap}(Q_n) = -\log \text{esssup}_y \mathbb{E}_{X \sim P_X} [\frac{dQ_{X|Y=y}}{dQ_n}]$. In all, we get the following sandwich bound:

$$D(P_Y \| Q_n) - \text{gap}(Q_n) \leq L \leq D(P_Y \| Q_n), \quad (5.30)$$

and it can be shown that as $n \rightarrow \infty$ both sides converge to L .

Sinkhorn's algorithm. This algorithm [274] is very similar, but not exactly the same as the ones above. We fix $Q_{X,Y}$, two marginals V_X, V_Y and solve the problem

$$S = \min \{D(P_{X,Y} \| Q_{X,Y}) : P_X = V_X, P_Y = V_Y\}.$$

From the results of Chapter 15 it is clear that the optimal distribution $P_{X,Y}$ is given by

$$P_{X,Y}^* = A(x) Q_{X,Y}(x,y) B(y),$$

for some $A, B \geq 0$. In order to find functions A, B we notice that under a fixed B the value of A that makes $P_X = V_X$ is given by

$$A(x) \leftarrow \frac{V_X(x) Q_{X,Y}(x,y) B(y)}{\sum_y Q_{X,Y}(x,y) B(y)}.$$

Similarly, to fix the Y -marginal we set

$$B(y) \leftarrow \frac{A(x) Q_{X,Y}(x,y) V_Y(y)}{\sum_x A(x) Q_{X,Y}(x,y)}.$$

The Sinkhorn's algorithm alternates the A and B updates until convergence.

The original version in [274] corresponds to $V_X = V_Y = \text{Unif}([n])$, and the goal there was to show that any matrix $\{C_{x,y}\}$ with non-negative entries can be transformed into a doubly-stochastic matrix $\{A(x) C_{x,y} B(y)\}$ by only rescaling rows and columns. The renewed interest in this classical algorithm arose from an observation that taking a jointly Gaussian $Q_{X,Y}(x,y) = c \exp\{-\|x-y\|^2/\epsilon\}$ produces a coupling $P_{X,Y}$ which resembles and approximates (as $\epsilon \rightarrow 0$) the optimal-transport coupling required for computing the Wasserstein distance $W_2(V_X, V_Y)$, see [85] for more.

6**Tensorization. Fano's inequality. Entropy rate.**

In this chapter we start with explaining the important property of mutual information known as tensorization (or single-letterization), which allows one to maximize and minimize mutual information between two high-dimensional vectors. So far in this book we have tacitly failed to give any operational meaning to the value of $I(X; Y)$. In this chapter, we give one fundamental such justification in the form of Fano's inequality. It states that whenever $I(X; Y)$ is small, one should not be able to predict X on the basis of Y with a small probability of error. As such, this inequality will be applied countless times in the rest of the book. We also define concepts of entropy rate (for a stochastic process) and of mutual information rate (for a pair of stochastic processes). For the former, it is shown that two processes that coincide often must have close entropy rates – a fact to be used later in the discussion of ergodicity. For the latter we give a closed form expression for the pair of Gaussian processes in terms of their spectral density.

6.1 Tensorization (single-letterization) of mutual information

For many applications we will have memoryless channels or memoryless sources. The following result is critical for extremizing mutual information in those cases.

Theorem 6.1 (Joint vs. marginal mutual information).

(1) If $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$ then

$$I(X^n; Y^n) \leq \sum_{i=1}^n I(X_i; Y_i) \quad (6.1)$$

with equality iff $P_{Y^n} = \prod P_{Y_i}$. Consequently, the (unconstrained) capacity is additive for memoryless channels:

$$\max_{P_{X^n}} I(X^n; Y^n) = \sum_{i=1}^n \max_{P_{X_i}} I(X_i; Y_i).$$

(2) If $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$ then

$$I(X^n; Y) \geq \sum_{i=1}^n I(X_i; Y) \quad (6.2)$$

6.1 Tensorization (single-letterization) of mutual information 79

with equality iff $P_{X^n|Y} = \prod P_{X_i|Y}$ P_Y -almost surely.¹ Consequently,

$$\min_{P_{Y^n|X^n}} I(X^n; Y^n) = \sum_{i=1}^n \min_{P_{Y_i|X_i}} I(X_i; Y_i).$$

Proof. (1) Use $I(X^n; Y^n) - \sum I(X_i; Y_i) = D(P_{Y^n|X^n} \| \prod P_{Y_i|X_i} | P_{X^n}) - D(P_{Y^n} \| \prod P_{Y_i})$
(2) Reverse the role of X and Y : $I(X^n; Y) - \sum I(X_i; Y) = D(P_{X^n|Y} \| \prod P_{X_i|Y} | P_Y) - D(P_{X^n} \| \prod P_{X_i})$

□

The moral of this result is that

- 1 For product channel, the input maximizing the mutual information is a product distribution.
- 2 For product source, the channel minimizing the mutual information is a product channel.

This type of result is often known as *single-letterization* in information theory. It tremendously simplifies the optimization problem over a high-dimensional (multi-letter) problem to a scalar (single-letter) problem. For example, in the simplest case where X^n, Y^n are binary vectors, optimizing $I(X^n; Y^n)$ over P_{X^n} and $P_{Y^n|X^n}$ entails optimizing over 2^n -dimensional vectors and $2^n \times 2^n$ matrices, whereas optimizing each $I(X_i; Y_i)$ individually is easy. In analysis, the effect when some quantities (or inequalities, such as log-Sobolev [144]) extend additively to tensor powers is called *tensorization*. Since forming a product of channels or distributions is a form of tensor power, the first part of the theorem shows that the capacity tensorizes.

Example 6.1. 1. (6.1) fails for non-product channels. $X_1 \perp\!\!\!\perp X_2 \sim \text{Bern}(1/2)$ on $\{0, 1\} = \mathbb{F}_2$:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 \\ I(X_1; Y_1) &= I(X_2; Y_2) = 0 \text{ but } I(X^2; Y^2) = 2 \text{ bits} \end{aligned}$$

2. Strict inequality in (6.1).

$$\begin{aligned} \forall k \quad Y_k &= X_k = U \sim \text{Bern}(1/2) \Rightarrow I(X_k; Y_k) = 1 \text{ bit} \\ I(X^n; Y^n) &= 1 \text{ bit} < \sum I(X_k; Y_k) = n \text{ bits} \end{aligned}$$

3. Strict inequality in (6.2). $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$

$$\begin{aligned} Y_1 &= X_2, Y_2 = X_3, \dots, Y_n = X_1 \Rightarrow I(X_k; Y_k) = 0 \\ I(X^n; Y^n) &= \sum H(X_i) > 0 = \sum I(X_k; Y_k) \end{aligned}$$

¹ That is, if $P_{X^n, Y} = P_Y \prod_{i=1}^n P_{X_i|Y}$ as joint distributions.

6.2* Gaussian capacity via orthogonal symmetry

Multi-dimensional case (WLOG assume $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$ iid): if Z_1, \dots, Z_n are independent, then

$$\max_{\mathbb{E}[\sum X_k^2] \leq nP} I(X^n; X^n + Z^n) \leq \max_{\mathbb{E}[\sum X_k^2] \leq nP} \sum_{k=1}^n I(X_k; X_k + Z_k)$$

Given a distribution $P_{X_1} \cdots P_{X_n}$ satisfying the constraint, form the “average of marginals” distribution $\bar{P}_X = \frac{1}{n} \sum_{k=1}^n P_{X_k}$, which also satisfies the single letter constraint $\mathbb{E}[X^2] = \frac{1}{n} \sum_{k=1}^n \mathbb{E}[X_k^2] \leq P$. Then from the concavity in P_X of $I(P_X, P_{Y|X})$

$$I(\bar{P}_X; P_{Y|X}) \geq \frac{1}{n} \sum_{k=1}^n I(P_{X_k}, P_{Y|X})$$

So \bar{P}_X gives the same or better mutual information, which shows that the extremization above ought to have the form $nC(P)$ where $C(P)$ is the single letter capacity. Now suppose $Y^n = X^n + Z_G^n$ where $Z_G^n \sim \mathcal{N}(0, \mathbf{I}_n)$. Since an isotropic Gaussian is rotationally symmetric, for any orthogonal transformation $U \in O(n)$, the additive noise has the same distribution $Z_G^n \sim UZ_G^n$, so that $P_{UY^n|UX^n} = P_{Y^n|X^n}$, and

$$I(P_{X^n}, P_{Y^n|X^n}) = I(P_{UX^n}, P_{UY^n|UX^n}) = I(P_{UX^n}, P_{Y^n|X^n})$$

From the “average of marginal” argument above, averaging over many rotations of X^n can only make the mutual information larger. Therefore, the optimal input distribution P_{X^n} can be chosen to be invariant under orthogonal transformations. Consequently, the (unique!) capacity achieving output distribution $P_{Y^n}^*$ must be rotationally invariant. Furthermore, from the conditions for equality in (6.1) we conclude that $P_{Y^n}^*$ must have independent components. Since the only product distribution satisfying the power constraints and having rotational symmetry is an isotropic Gaussian, we conclude that $P_{Y^n} = (P_Y^*)^n$ and $P_Y^* = \mathcal{N}(0, P\mathbf{I}_n)$.

For the other direction in the Gaussian saddle point problem:

$$\min_{P_N: \mathbb{E}[N^2]=1} I(X_G; X_G + N)$$

This uses the same trick, except here the input distribution is automatically invariant under orthogonal transformations.

6.3 Information measures and probability of error

Let W be a random variable and \hat{W} be our prediction. There are three types of problems:

- 1 Random guessing: $W \perp\!\!\!\perp \hat{W}$.
- 2 Guessing with data: $W \rightarrow X \rightarrow \hat{W}$.
- 3 Guessing with noisy data: $W \rightarrow X \rightarrow Y \rightarrow \hat{W}$.

6.3 Information measures and probability of error 81

Our goal is to draw converse statements: for example, if the uncertainty of W is too high or if the information provided by the data is too scarce, then it is difficult to guess the value of W .

Theorem 6.2. Let $|\mathcal{X}| = M < \infty$. Then for any $\hat{X} \perp\!\!\!\perp X$,

$$H(X) \leq F_M(\mathbb{P}[X = \hat{X}]) \quad (6.3)$$

where

$$F_M(x) \triangleq (1-x)\log(M-1) + h(x), x \in [0, 1] \quad (6.4)$$

and $h(x) = x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x}$ is the binary entropy function.

If $P_{\max} \triangleq \max_{x \in \mathcal{X}} P_X(x)$, then recalling that $h(\cdot)$ is a

$$H(X) \leq F_M(P_{\max}) = (1 - P_{\max})\log(M-1) + h(P_{\max}), \quad (6.5)$$

with equality iff $P_X = (P_{\max}, \frac{1-P_{\max}}{M-1}, \dots, \frac{1-P_{\max}}{M-1})$.

The function $F_M(\cdot)$ is shown in Fig. 6.1. Notice that due to its non-monotonicity the statement (6.5) does not imply (6.3), even though $\mathbb{P}[X = \hat{X}] \leq P_{\max}$.

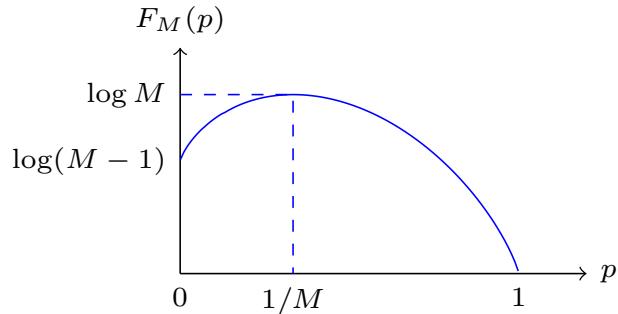


Figure 6.1 The function F_M in (6.4) is concave with maximum $\log M$ at maximizer $1/M$, but not monotone.

Proof. To show (6.3) consider an auxiliary distribution $Q_{X,\hat{X}} = U_X P_{\hat{X}}$, where U_X is uniform on \mathcal{X} . Then $Q[X = \hat{X}] = 1/M$. Denoting $P[X = \hat{X}] \triangleq P_S$, applying the DPI for divergence to the data processor $(X, \hat{X}) \mapsto 1_{\{X=\hat{X}\}}$ yields $d(P_S \| 1/M) \leq D(P_{X\hat{X}} \| Q_{X\hat{X}}) = \log M - H(X)$.

To show the second part, suppose one is trying to guess the value of X without any side information. Then the best bet is obviously the most likely outcome (mode) and the maximal probability of success is

$$\max_{\hat{X} \perp\!\!\!\perp X} \mathbb{P}[X = \hat{X}] = P_{\max} \quad (6.6)$$

Thus, applying (6.3) with \hat{X} being the mode yield (6.5). Finally, suppose that $P = (P_{\max}, P_2, \dots, P_M)$ and introduce $Q = (P_{\max}, \frac{1-P_{\max}}{M-1}, \dots, \frac{1-P_{\max}}{M-1})$. Then the difference of the right and left side of (6.5) equals $D(P \| Q) \geq 0$, with equality iff $P = Q$. \square

Remark 6.1. Let us discuss the unusual proof technique. Instead of studying directly the probability space $P_{X,\hat{X}}$ given to us, we introduced an auxiliary one: $Q_{X,\hat{X}}$. We then drew conclusions about the target metric (probability of error) for the auxiliary problem (the probability of error $= 1 - \frac{1}{M}$). Finally, we used DPI to transport statement about Q to a statement about P : if $D(P||Q)$ is small, then the probabilities of the events (e.g., $\{X \neq \hat{X}\}$) should be small as well. This is a general method, known as *meta-converse*, that we develop in more detail later in this book. For this result, however, there are much more explicit ways to derive it – see Ex. I.42.

Similar to Shannon entropy H , P_{\max} is also a reasonable measure for randomness of P . In fact,

$$H_\infty(P) \triangleq \log \frac{1}{P_{\max}} \quad (6.7)$$

is known as the *Rényi entropy of order ∞* (or the min-entropy in the cryptography literature). Note that $H_\infty(P) = \log M$ iff P is uniform; $H_\infty(P) = 0$ iff P is a point mass. In this regard, the Fano's inequality can be thought of as our first example of a *comparison of information measures*: it compares H and H_∞ .

Theorem 6.3 (Fano's inequality). *Let $|\mathcal{X}| = M < \infty$ and $X \rightarrow Y \rightarrow \hat{X}$. Let $P_e = \mathbb{P}[X \neq \hat{X}]$, then*

$$H(X|Y) \leq F_M(1 - P_e) = P_e \log(M - 1) + h(P_e). \quad (6.8)$$

Furthermore, if $P_{\max} \triangleq \max_{x \in \mathcal{X}} P_X(x) > 0$, then regardless of $|\mathcal{X}|$,

$$I(X; Y) \geq (1 - P_e) \log \frac{1}{P_{\max}} - h(P_e). \quad (6.9)$$

Proof. The benefit of the previous proof is that it trivially generalizes to this new case of (possibly randomized) estimators \hat{X} , which may depend on some observation Y correlated with X . Note that it is clear that the best predictor for X given Y is the maximum posterior (MAP) rule, i.e., posterior mode: $\hat{X}(y) = \operatorname{argmax}_x P_{X|Y}(x|y)$.

To show (6.8) we apply data processing (for divergence) to $P_{X,Y,\hat{X}} = P_X P_{Y|X} P_{\hat{X}|Y}$ vs. $Q_{X,Y,\hat{X}} = U_X P_Y P_{\hat{X}|Y}$ and the data processor (kernel) $(X, Y, \hat{X}) \mapsto 1_{\{X \neq \hat{X}\}}$ (note that $P_{\hat{X}|Y}$ is identical for both).

To show (6.9) we apply data processing (for divergence) to $P_{X,Y,\hat{X}} = P_X P_{Y|X} P_{\hat{X}|Y}$ vs. $Q_{X,Y,\hat{X}} = P_X P_Y P_{\hat{X}|Y}$ and the data processor (kernel) $(X, Y, \hat{X}) \mapsto 1_{\{X \neq \hat{X}\}}$ to obtain:

$$\begin{aligned} I(X; Y) &= D(P_{X,Y,\hat{X}} || Q_{X,Y,\hat{X}}) \geq d(\mathbb{P}[X = \hat{X}] || \mathbb{Q}[X = \hat{X}]) \\ &\geq -h(P_e) + (1 - P_e) \log \frac{1}{\mathbb{Q}[X = \hat{X}]} \geq -h(P_e) - (1 - P_e) \log P_{\max}, \end{aligned}$$

where the last step follows from $\mathbb{Q}[X = \hat{X}] \leq P_{\max}$ since $X \perp\!\!\!\perp \hat{X}$ under \mathbb{Q} . (Again, we refer to Ex. I.42 for a direct proof.) \square

The following corollary of the previous result emphasizes its role in providing converses (or impossibility results) for statistics and data transmission.

6.4 Entropy rate 83

Corollary 6.4 (Lower bound on average probability of error). *Let $W \rightarrow X \rightarrow Y \rightarrow \hat{W}$ and W is uniform on $[M] \triangleq \{1, \dots, M\}$. Then*

$$P_e \triangleq \mathbb{P}[W \neq \hat{W}] \geq 1 - \frac{I(X; Y) + h(P_e)}{\log M} \quad (6.10)$$

$$\geq 1 - \frac{I(X; Y) + \log 2}{\log M}. \quad (6.11)$$

Proof. Apply Theorem 6.4 and the data processing for mutual information: $I(W; \hat{W}) \leq I(X; Y)$.

□

6.4 Entropy rate

Definition 6.5. The entropy rate of a process $\mathbb{X} = (X_1, X_2, \dots)$ is

$$H(\mathbb{X}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n) \quad (6.12)$$

provided the limit exists.

A sufficient condition for the entropy rate to exist is *stationarity*, which essentially means invariance with respect to time shift. Formally, \mathbb{X} is stationary if $(X_{t_1}, \dots, X_{t_n}) \stackrel{d}{=} (X_{t_1+k}, \dots, X_{t_n+k})$ for any $t_1, \dots, t_n, k \in \mathbb{N}$. This definition naturally extends to two-sided processes.

Theorem 6.6. *For any stationary process $\mathbb{X} = (X_1, X_2, \dots)$*

- (a) $H(X_n | X^{n-1}) \leq H(X_{n-1} | X^{n-2})$.
- (b) $\frac{1}{n} H(X^n) \geq H(X_n | X^{n-1})$.
- (c) $\frac{1}{n} H(X^n) \leq \frac{1}{n-1} H(X^{n-1})$.
- (d) $H(\mathbb{X})$ exists and $H(\mathbb{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n) = \lim_{n \rightarrow \infty} H(X_n | X^{n-1})$.
- (e) For a two-sided stationary process $\mathbb{X} = (\dots, X_{-1}, X_0, X_1, X_2, \dots)$, $H(\mathbb{X}) = H(X_1 | X_{-\infty}^0)$ provided that $H(X_1) < \infty$.

Proof.

- (a) Further conditioning + stationarity: $H(X_n | X^{n-1}) \leq H(X_n | X_2^{n-1}) = H(X_{n-1} | X^{n-2})$
- (b) Using chain rule: $\frac{1}{n} H(X^n) = \frac{1}{n} \sum H(X_i | X^{i-1}) \geq H(X_n | X^{n-1})$
- (c) $H(X^n) = H(X^{n-1}) + H(X_n | X^{n-1}) \leq H(X^{n-1}) + \frac{1}{n} H(X^n)$
- (d) $n \mapsto \frac{1}{n} H(X^n)$ is a decreasing sequence and lower bounded by zero, hence has a limit $H(\mathbb{X})$. Moreover by chain rule, $\frac{1}{n} H(X^n) = \frac{1}{n} \sum_{i=1}^n H(X_i | X^{i-1})$. From here we claim that $H(X_n | X^{n-1})$ converges to the same limit $H(\mathbb{X})$. Indeed, from the monotonicity shown in part (a), $\lim_n H(X_n | X^{n-1}) = H'$ exists. Next, recall the following fact from calculus: if $a_n \rightarrow a$, then the Cesàro's mean $\frac{1}{n} \sum_{i=1}^n a_i \rightarrow a$ as well. Thus, $H' = H(\mathbb{X})$.

5. Assuming $H(X_1) < \infty$ we have from (4.30):

$$\lim_{n \rightarrow \infty} H(X_1) - H(X_1 | X_{-n}^0) = \lim_{n \rightarrow \infty} I(X_1; X_{-n}^0) = I(X_1; X_{-\infty}^0) = H(X_1) - H(X_1 | X_{-\infty}^0)$$

□

Example 6.2 (Stationary processes).

- (a) \mathbb{X} – iid source $\Rightarrow H(\mathbb{X}) = H(X_1)$
- (b) \mathbb{X} – mixed sources: Flip a coin with bias p at time $t = 0$, if head, let $\mathbb{X} = \mathbb{Y}$, if tail, let $\mathbb{X} = \mathbb{Z}$. Then $H(\mathbb{X}) = pH(\mathbb{Y}) + \bar{p}H(\mathbb{Z})$.
- (c) \mathbb{X} – stationary Markov chain : $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots$

$$H(X_n | X^{n-1}) = H(X_n | X_{n-1}) \Rightarrow H(\mathbb{X}) = H(X_2 | X_1) = \sum_{a,b} \mu(a) P_{b|a} \log \frac{1}{P_{b|a}}$$

where μ is an invariant measure (possibly non-unique; unique if the chain is ergodic).

- (d) \mathbb{X} – hidden Markov chain : Let $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots$ be a Markov chain. Fix $P_{Y|X}$. Let $X_i \xrightarrow{P_{Y|X}} Y_i$. Then $\mathbb{Y} = (Y_1, \dots)$ is a stationary process. Therefore $H(\mathbb{Y})$ exists but it is very difficult to compute (no closed-form solution known), even if \mathbb{X} is a binary Markov chain and $P_{Y|X}$ is a BSC.

6.5 Entropy and symbol (bit) error rate

In this section we show that the entropy rates of two processes \mathbb{X} and \mathbb{Y} are close whenever they can be “coupled”. Coupling of two processes means defining them on a common probability space so that the average distance between their realizations is small. In the following, we will require that the so-called *symbol error rate* (expected fraction of errors) is small, namely

$$\frac{1}{n} \sum_{j=1}^n \mathbb{P}[X_j \neq Y_j] \leq \epsilon. \quad (6.13)$$

For binary alphabet this quantity is known as the *bit error rate*, which is one of the performance metrics we consider for reliable data transmission in Part IV (see Section 17.1 and Section 19.6). Notice that if we define the Hamming distance as

$$d_H(x^n, y^n) \triangleq \sum_{j=1}^n \mathbb{1}\{x_j \neq y_j\} \quad (6.14)$$

then (6.13) corresponds to requiring $\mathbb{E}[d_H(X^n, Y^n)] \leq n\epsilon$.

Before showing our main result, we show that Fano’s inequality Theorem 6.4 can be tensorized:

Proposition 6.7. *Let X_1, \dots, X_n take values on a finite alphabet \mathcal{X} . Then*

$$H(X^n | Y^n) \leq nF_{|\mathcal{X}|}(1 - \delta) = n(\delta \log(|\mathcal{X}| - 1) + h(\delta)), \quad (6.15)$$

6.6 Entropy and contiguity 85

where the function F_M is defined in (6.4), and

$$\delta = \frac{1}{n} \mathbb{E}[d_H(X^n, Y^n)] = \frac{1}{n} \sum_{j=1}^n \mathbb{P}[X_j \neq Y_j].$$

Proof. For each $j \in [n]$, applying (6.8) to the Markov chain $X_j \rightarrow Y^n \rightarrow Y_j$ yields

$$H(X_j|Y^n) \leq F_M(\mathbb{P}[X_j = Y_j]), \quad (6.16)$$

where we denoted $M = |\mathcal{X}|$. Then, upper-bounding joint entropy by the sum of marginals, cf. (1.3), and combining with (6.16), we get

$$H(X^n|Y^n) \leq \sum_{j=1}^n H(X_j|Y^n) \quad (6.17)$$

$$\leq \sum_{j=1}^n F_M(\mathbb{P}[X_j = Y_j]) \quad (6.18)$$

$$\leq nF_M\left(\frac{1}{n} \sum_{j=1}^n \mathbb{P}[X_j = Y_j]\right) \quad (6.19)$$

where in the last step we used the concavity of F_M and Jensen's inequality. \square

Corollary 6.8. Consider two processes \mathbb{X} and \mathbb{Y} with entropy rates $H(\mathbb{X})$ and $H(\mathbb{Y})$. If

$$\mathbb{P}[X_j \neq Y_j] \leq \epsilon$$

for every j and if \mathbb{X} takes values on a finite alphabet of size M , then

$$H(\mathbb{X}) - H(\mathbb{Y}) \leq F_M(1 - \epsilon).$$

If both processes have alphabets of size M then

$$|H(\mathbb{X}) - H(\mathbb{Y})| \leq \epsilon \log M + h(\epsilon) \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0$$

Proof. There is almost nothing to prove:

$$H(X^n) \leq H(X^n, Y^n) = H(Y^n) + H(X^n|Y^n)$$

and apply (6.15). For the last statement just recall the expression for F_M . \square

6.6 Entropy and contiguity

In this section, we show a related result showing that the entropy rate of a \mathcal{X} -valued process that is “almost iid uniform” is necessarily $\log |\mathcal{X}|$. To quantify “almost” we recall the following concept.

Definition 6.9 (Contiguity). Let $\{P_n\}$ and $\{Q_n\}$ be sequences of probability measures on some Ω_n . We say P_n is *contiguous* with respect to Q_n (denoted by $P_n \triangleleft Q_n$) if for any sequence $\{A_n\}$ of measurable sets, $Q_n(A_n) \rightarrow 0$ implies that $P_n(A_n) \rightarrow 0$. We say P_n and Q_n are *mutually contiguous* (denoted by $P_n \triangleleft\triangleright Q_n$) if $P_n \triangleleft Q_n$ and $Q_n \triangleleft P_n$.

Note that $P_n \triangleleft\triangleright Q_n$ is much weaker than $\text{TV}(P_n, Q_n) \rightarrow 0$. A commonly used sufficient condition for $P_n \triangleleft Q_n$ is bounded second moment $\chi^2(P_n \| Q_n) = O(1)$. Indeed, by the Cauchy-Schwarz inequality, $P_n(A_n) = \mathbb{E}_{P_n}[1_{A_n}] \leq \sqrt{1 + \chi^2(P_n \| Q_n)} \sqrt{Q_n(A_n)}$ which vanishes whenever $Q_n(A_n)$ vanishes. In particular, a sufficient condition for mutual contiguity is the boundedness of likelihood ratio: $c \leq \frac{P_n}{Q_n} \leq C$ for some constants c, C .

Here is the promised result about the entropy rate:

Theorem 6.10. Let \mathcal{X} be a finite set and Q_n the uniform distribution on \mathcal{X}^n . If $P_n \triangleleft Q_n$, then $H(P_n) = H(Q_n) + o(n) = n \log |\mathcal{X}| + o(n)$. Equivalently, $D(P_n \| Q_n) = o(n)$.

Proof. Suppose for the sake of contradiction that $H(P_n) \leq (1 - \epsilon)n \log |\mathcal{X}|$ for some constant ϵ . Let $\epsilon' < \epsilon$ and define $A_n \triangleq \{x^n \in \mathcal{X}^n : P_n(x^n) \geq |\mathcal{X}|^{-(1-\epsilon')n}\}$. Then $|A_n| \leq |\mathcal{X}|^{(1-\epsilon')n}$ and hence $Q_n(A_n) \leq |\mathcal{X}|^{-\epsilon'n}$. Since $P_n \triangleleft Q_n$, we have $P_n(A_n) \rightarrow 0$. On the other hand, $H(P_n) \geq \mathbb{E}_{P_n}[\log \frac{1}{P_n} 1_{A_n^c}] \geq (1 - \epsilon')n \log |\mathcal{X}| P_n(A_n^c)$. Thus $P_n(A_n^c) \leq \frac{1-\epsilon}{1-\epsilon'}$ which is a contradiction. \square

Remark 6.2. It is natural to ask whether Theorem 6.9 holds for non-uniform Q_n , that is, whether $P_n \triangleleft\triangleright Q_n$ implies $H(P_n) = H(Q_n) + o(n)$. This turns out to be false. To see this, choose any μ_n, ν_n and set $P_n \triangleq \frac{1}{2}\mu_n + \frac{1}{2}\nu_n$ and $Q_n \triangleq \frac{1}{3}\mu_n + \frac{2}{3}\nu_n$. Then we always have $P_n \triangleleft\triangleright Q_n$ since $\frac{3}{4} \leq \frac{P_n}{Q_n} \leq \frac{3}{2}$. Using conditional entropy, it is clear that $H(P_n) = \frac{1}{2}H(\mu_n) + \frac{1}{2}H(\nu_n) + O(1)$ and $Q_n = \frac{1}{3}H(\mu_n) + \frac{2}{3}H(\nu_n) + O(1)$. In addition, by data processing, $D(P_n \| Q_n) \leq d(\frac{1}{2} \| \frac{1}{3}) = O(1)$. Choosing, say, $\mu_n = \text{Ber}(\frac{1}{2})^{\otimes n}$ and $\nu_n = \text{Ber}(\frac{1}{3})^{\otimes n}$ leads to $|H(P_n) - H(Q_n)| = \Omega(n)$.

6.7 Mutual information rate

Definition 6.11 (Mutual information rate).

$$I(\mathbb{X}; \mathbb{Y}) = \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n; Y^n)$$

provided the limit exists.

Example 6.3 (Gaussian processes). Consider \mathbb{X}, \mathbb{N} two stationary Gaussian processes, independent of each other. Assume that their auto-covariance functions are absolutely summable and thus there exist continuous power spectral density functions f_X and f_N . Without loss of generality, assume all means are zero. Let $c_X(k) = \mathbb{E}[X_1 X_{k+1}]$. Then f_X is the Fourier transform of the auto-covariance function c_X , i.e., $f_X(\omega) = \sum_{k=-\infty}^{\infty} c_X(k) e^{i\omega k}$. Finally, assume $f_N \geq \delta > 0$. Then recall from Example 3.4:

$$I(X^n; X^n + N^n) = \frac{1}{2} \log \frac{\det(\Sigma_{X^n} + \Sigma_{N^n})}{\det \Sigma_{N^n}}$$

6.7 Mutual information rate 87

$$= \frac{1}{2} \sum_{i=1}^n \log \sigma_i - \frac{1}{2} \sum_{i=1}^n \log \lambda_i,$$

where σ_j, λ_j are the eigenvalues of the covariance matrices $\Sigma_{Y^n} = \Sigma_{X^n} + \Sigma_{N^n}$ and Σ_{N^n} , which are all Toeplitz matrices, e.g., $(\Sigma_{X^n})_{ij} = \mathbb{E}[X_i X_j] = c_X(i-j)$. By Szegö's theorem [143, Sec. 5.2]:

$$\frac{1}{n} \sum_{i=1}^n \log \sigma_i \rightarrow \frac{1}{2\pi} \int_0^{2\pi} \log f_Y(\omega) d\omega \quad (6.20)$$

Note that $c_Y(k) = \mathbb{E}[(X_1 + N_1)(X_{k+1} + N_{k+1})] = c_X(k) + c_N(k)$ and hence $f_Y = f_X + f_N$. Thus, we have

$$\frac{1}{n} I(X^n; X^n + N^n) \rightarrow I(\mathbb{X}; \mathbb{X} + \mathbb{N}) = \frac{1}{4\pi} \int_0^{2\pi} \log \frac{f_X(\omega) + f_N(\omega)}{f_N(\omega)} d\omega.$$

Maximizing this over $f_X(\omega)$ leads to the famous *water-filling* solution $f_X^*(\omega) = |T - f_N(\omega)|^+$.

7

f-divergences

In Chapter 2 we introduced the KL divergence that measures the dissimilarity between two distributions. This turns out to be a special case of the family of *f*-divergence between probability distributions, introduced by Csiszár [78]. Like KL-divergence, *f*-divergences satisfy a number of useful properties:

- operational significance: KL divergence forms a basis of information theory by yielding fundamental answers to questions in channel coding and data compression. Similarly, *f*-divergences such as χ^2 , H^2 and TV have their foundational roles in parameter estimation, high-dimensional statistics and hypothesis testing, respectively.
- invariance to bijective transformations.
- data-processing inequality
- variational representations (à la Donsker-Varadhan)
- local behavior given by χ^2 (in nonparametric cases) or Fisher information (in parametric cases).

The purpose of this chapter is to establish these properties and prepare the ground for applications in subsequent chapters. The important highlight is a *joint range* Theorem of Harremoës and Vajda [151], which gives the sharpest possible comparison inequality between arbitrary *f*-divergences (and puts an end to a long sequence of results starting from Pinsker's inequality – Theorem 7.17). This material can be skimmed on the first reading and referenced later upon need.

7.1 Definition and basic properties of *f*-divergences

Definition 7.1 (*f*-divergence). Let $f: (0, \infty) \rightarrow \mathbb{R}$ be a convex function with $f(1) = 0$. Let P and Q be two probability distributions on a measurable space $(\mathcal{X}, \mathcal{F})$. If $P \ll Q$ then the *f*-divergence is defined as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right) \right] \quad (7.1)$$

where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative and $f(0) \triangleq f(0+)$. More generally, let $f'(\infty) \triangleq \lim_{x \downarrow 0} xf'(1/x)$. Suppose that $Q(dx) = q(x)\mu(dx)$ and $P(dx) = p(x)\mu(dx)$ for some common dominating measure μ , then we have

$$D_f(P\|Q) = \int_{q>0} q(x)f\left(\frac{p(x)}{q(x)}\right) d\mu + f'(\infty)P[q=0] \quad (7.2)$$

7.1 Definition and basic properties of f -divergences 89

with the agreement that if $P[q = 0] = 0$ the last term is taken to be zero regardless of the value of $f'(\infty)$ (which could be infinite).

Remark 7.1. For the discrete case, with $Q(x)$ and $P(x)$ being the respective pmfs, we can also write

$$D_f(P\|Q) = \sum_x Q(x)f\left(\frac{P(x)}{Q(x)}\right)$$

with the understanding that

- $f(0) = f(0+)$,
- $0f(\frac{0}{0}) = 0$, and
- $0f(\frac{a}{0}) = \lim_{x \downarrow 0} xf(\frac{a}{x}) = af'(\infty)$ for $a > 0$.

Remark 7.2. A nice property of $D_f(P\|Q)$ is that the definition is invariant to the choice of the dominating measure μ in (7.2). This is not the case for other dissimilarity measures, e.g., the squared L_2 -distance between the densities $\|p - q\|_{L^2(d\mu)}^2$ which is a popular loss function for density estimation in statistics literature.

The following are common f -divergences:

- **Kullback-Leibler (KL) divergence:** We recover the usual $D(P\|Q)$ in Chapter 2 by taking $f(x) = x \log x$.
- **Total variation:** $f(x) = \frac{1}{2}|x - 1|$,

$$\text{TV}(P, Q) \triangleq \frac{1}{2} \mathbb{E}_Q \left[\left| \frac{dP}{dQ} - 1 \right| \right] = \frac{1}{2} \int |dP - dQ| = 1 - \int d(P \wedge Q). \quad (7.3)$$

Moreover, $\text{TV}(\cdot, \cdot)$ is a metric on the space of probability distributions.¹

- **χ^2 -divergence:** $f(x) = (x - 1)^2$,

$$\chi^2(P\|Q) \triangleq \mathbb{E}_Q \left[\left(\frac{dP}{dQ} - 1 \right)^2 \right] = \int \frac{(dP - dQ)^2}{dQ} = \int \frac{dP^2}{dQ} - 1. \quad (7.4)$$

Note that we can also choose $f(x) = x^2 - 1$. Indeed, f 's differing by a linear term lead to the same f -divergence, cf. Proposition 7.5.

- **Squared Hellinger distance:** $f(x) = (1 - \sqrt{x})^2$,

$$H^2(P, Q) \triangleq \mathbb{E}_Q \left[\left(1 - \sqrt{\frac{dP}{dQ}} \right)^2 \right] = \int (\sqrt{dP} - \sqrt{dQ})^2 = 2 - 2 \int \sqrt{dPdQ}. \quad (7.5)$$

¹ In (7.3), $\int d(P \wedge Q)$ is the usual short-hand for $\int (\frac{dP}{d\mu} \wedge \frac{dQ}{d\mu}) d\mu$ where μ is any dominating measure. The expressions in (7.4) and (7.5) are understood in the similar sense.

Here the quantity $B(P, Q) \triangleq \int \int \sqrt{dP dQ}$ is known as the *Bhattacharyya coefficient* (or Hellinger affinity) [33]. Note that $H(P, Q) = \sqrt{H^2(P, Q)}$ defines a metric on the space of probability distributions: indeed, the triangle inequality follows from that of $L_2(\mu)$ for a common dominating measure. Note, however, that $(P, Q) \mapsto H(P, Q)$ is *not* convex. (This is because metric H is not induced by a Banach norm on the space of measures.)

- **Le Cam distance** [189, p. 47]: $f(x) = \frac{1-x}{2x+2}$,

$$\text{LC}(P, Q) = \frac{1}{2} \int \frac{(dP - dQ)^2}{dP + dQ}. \quad (7.6)$$

Moreover, $\sqrt{\text{LC}(P||Q)}$ is a metric on the space of probability distributions [113].

- **Jensen-Shannon divergence**: $f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$,

$$\text{JS}(P, Q) = D\left(P \parallel \frac{P+Q}{2}\right) + D\left(Q \parallel \frac{P+Q}{2}\right). \quad (7.7)$$

Moreover, $\sqrt{\text{JS}(P||Q)}$ is a metric on the space of probability distributions [113].

Remark 7.3. If $D_f(P||Q)$ is an f -divergence, then it is easy to verify that $D_f(\lambda P + \bar{\lambda} Q||Q)$ and $D_f(P||\lambda P + \bar{\lambda} Q)$ are f -divergences for all $\lambda \in [0, 1]$. In particular, $D_f(Q||P) = D_{\tilde{f}}(P||Q)$ with $\tilde{f}(x) \triangleq xf(\frac{1}{x})$.

We start summarizing some formal observations about the f -divergences

Proposition 7.2 (Basic properties). *The following hold:*

- 1 $D_{f_1+f_2}(P||Q) = D_{f_1}(P||Q) + D_{f_2}(P||Q)$.
- 2 $D_f(P||P) = 0$.
- 3 $D_f(P||Q) = 0$ for all $P \neq Q$ iff $f(x) = c(x-1)$ for some c . For any other f we have $D_f(P||Q) = f(0) + f'(\infty) > 0$ for $P \perp Q$.
- 4 If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$ then the function $x \mapsto D_f(P_{Y|X=x}||Q_{Y|X=x})$ is measurable and

$$D_f(P_{X,Y}||Q_{X,Y}) = \int_X dP_X(x) D_f(P_{Y|X=x}||Q_{Y|X=x}) \triangleq D_f(P_{Y|X}||Q_{Y|X}|P_X), \quad (7.8)$$

the latter referred to as the conditional f -divergence (similar to Definition 2.13 for conditional KL divergence).

- 5 If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$ then

$$D_f(P_{X,Y}||Q_{X,Y}) = D_f(P_X||Q_X). \quad (7.9)$$

In particular,

$$D_f(P_X P_Y||Q_X P_Y) = D_f(P_X||Q_X). \quad (7.10)$$

- 6 Let $f_1(x) = f(x) + c(x-1)$, then

$$D_{f_1}(P||Q) = D_f(P||Q) \quad \forall P, Q.$$

7.2 Data-processing inequality; approximation by finite partitions 91

In particular, we can always assume that $f \geq 0$ and (iff f is differentiable at 1) that $f'(1) = 0$.

Proof. The first and second are clear. For the third property, verify explicitly that $D_f(P\|Q) = 0$ for $f = c(x - 1)$. Next consider general f and observe that for $P \perp Q$, by definition we have

$$D_f(P\|Q) = f(0) + f'(\infty), \quad (7.11)$$

which is well-defined (i.e., $\infty - \infty$ is not possible) since by convexity $f(0) > -\infty$ and $f'(\infty) > -\infty$. So all we need to verify is that $f(0) + f'(\infty) = 0$ if and only if $f = c(x - 1)$ for some $c \in \mathbb{R}$. Indeed, since $f(1) = 0$, the convexity of f implies that $x \mapsto g(x) \triangleq \frac{f(x)}{x-1}$ is non-decreasing. By assumption, we have $g(0+) = g(\infty)$ and hence $g(x)$ is a constant on $x > 0$, as desired.

For property 4, let $R_{Y|X} = \frac{1}{2}P_{Y|X} + \frac{1}{2}Q_{Y|X}$. By Theorem 2.11 there exist jointly measurable $p(y|x)$ and $q(y|x)$ such that $dP_{Y|X=x} = p(y|x)dR_{Y|X=x}$ and $Q_{Y|X} = q(y|x)dR_{Y|X=x}$. We can then take μ in (7.2) to be $\mu = P_X R_{Y|X}$, which gives $dP_{X,Y} = p(y|x)d\mu$ and $dQ_{X,Y} = q(y|x)d\mu$ and thus

$$\begin{aligned} D_f(P_{X,Y}\|Q_{X,Y}) &= \int_{\mathcal{X} \times \mathcal{Y}} d\mu 1\{y : q(y|x) > 0\} q(y|x) f\left(\frac{p(y|x)}{q(y|x)}\right) + f'(\infty) \int_{\mathcal{X} \times \mathcal{Y}} d\mu 1\{y : q(y|x) = 0\} p(y|x) \\ &\stackrel{(7.2)}{=} \int_{\mathcal{X}} dP_X \underbrace{\left\{ \int_{\{y:q(y|x)>0\}} dR_{Y|X=x} q(y|x) f\left(\frac{p(y|x)}{q(y|x)}\right) + f'(\infty) \int_{\{y:q(y|x)=0\}} dR_{Y|X=x} p(y|x) \right\}}_{D_f(P_{Y|X=x}\|Q_{Y|X=x})} \end{aligned}$$

which is the desired (7.8).

Property 5 follows from the observation: if we take $\mu = P_{X,Y} + Q_{X,Y}$ and $\mu_1 = P_X + Q_X$ then $\frac{dP_{X,Y}}{d\mu} = \frac{dP_X}{d\mu_1}$ and similarly for Q .

Property 6 follows from the first and the third. Note also that reducing to $f \geq 0$ is done by taking $c = f'(1)$ (or any subdifferential at $x = 1$ if f is not differentiable). \square

7.2 Data-processing inequality; approximation by finite partitions

Theorem 7.3 (Monotonicity).

$$D_f(P_{X,Y}\|Q_{X,Y}) \geq D_f(P_X\|Q_X). \quad (7.12)$$

Proof. Note that in the case $P_{X,Y} \ll Q_{X,Y}$ (and thus $P_X \ll Q_X$), the proof is a simple application of Jensen's inequality to definition (7.1):

$$\begin{aligned} D_f(P_{X,Y}\|Q_{X,Y}) &= \mathbb{E}_{X \sim Q_X} \mathbb{E}_{Y \sim Q_{Y|X}} \left[f\left(\frac{dP_{Y|X}P_X}{dQ_{Y|X}Q_X}\right) \right] \\ &\geq \mathbb{E}_{X \sim Q_X} \left[f\left(\mathbb{E}_{Y \sim Q_{Y|X}} \left[\frac{dP_{Y|X}P_X}{dQ_{Y|X}Q_X} \right] \right) \right] \\ &= \mathbb{E}_{X \sim Q_X} \left[f\left(\frac{dP_X}{dQ_X}\right) \right]. \end{aligned}$$

To prove the general case we need to be more careful. Let $R_X = \frac{1}{2}(P_X + Q_X)$ and $R_{Y|X} = \frac{1}{2}P_{Y|X} + \frac{1}{2}Q_{Y|X}$. It should be clear that $P_{X,Y}, Q_{X,Y} \ll R_{X,Y} \triangleq R_X R_{Y|X}$ and that for every x : $P_{Y|X=x}, Q_{Y|X=x} \ll R_{Y|X=x}$. By Theorem 2.11 there exist measurable functions p_1, p_2, q_1, q_2 so that

$$dP_{X,Y} = p_1(x)p_2(y|x)dR_{X,Y}, \quad dQ_{X,Y} = q_1(x)q_2(y|x)dR_{X,Y}$$

and $dP_{Y|X=x} = p_2(y|x)dR_{Y|X=x}$, $dQ_{Y|X=x} = q_2(y|x)dR_{Y|X=x}$. We also denote $p(x,y) = p_1(x)p_2(y|x)$, $q(x,y) = q_1(x)q_2(y|x)$.

Fix $t > 0$ and consider a supporting line to f at t with slope μ , so that

$$f(u) \geq f(t) + \mu(u - t), \quad \forall u \geq 0.$$

Thus, $f'(\infty) \geq \mu$ and taking $u = \lambda t$ for any $\lambda \in [0, 1]$ we have shown:

$$f(\lambda t) + \bar{\lambda}t f'(\infty) \geq f(t), \quad \forall t \geq 0, \lambda \in [0, 1]. \quad (7.13)$$

Note that we added $t = 0$ case as well, since for $t = 0$ the statement is obvious (recall, though, that $f(0) \triangleq f(0+)$ can be equal to $+\infty$).

Next, fix some x with $q_1(x) > 0$ and consider the chain

$$\begin{aligned} & \int_{\{y:q_2(y|x)>0\}} dR_{Y|X=x} q_2(y|x) f\left(\frac{p_1(x)p_2(y|x)}{q_1(x)q_2(y|x)}\right) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) = 0] f'(\infty) \\ & \stackrel{(a)}{\geq} f\left(\frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) > 0]\right) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) = 0] f'(\infty) \\ & \stackrel{(b)}{\geq} f\left(\frac{p_1(x)}{q_1(x)}\right) \end{aligned}$$

where (a) is by Jensen's inequality and the convexity of f , and (b) by taking $t = \frac{p_1(x)}{q_1(x)}$ and $\lambda = P_{Y|X=x}[q_2(Y|x) > 0]$ in (7.13). Now multiplying the obtained inequality by $q_1(x)$ and integrating over $\{x : q_1(x) > 0\}$ we get

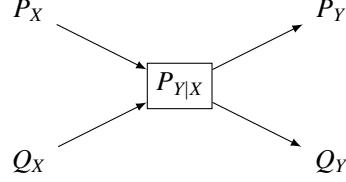
$$\begin{aligned} & \int_{\{q>0\}} dR_{X,Y} q(x,y) f\left(\frac{p(x,y)}{q(x,y)}\right) + f'(\infty) P_{X,Y}[q_1(X) > 0, q_2(Y|X) = 0] \\ & \geq \int_{\{q_1>0\}} dR_X q_1(x) f\left(\frac{p_1(x)}{q_1(x)}\right). \end{aligned}$$

Adding $f'(\infty) P_{X,Y}[q_1(X) = 0]$ to both sides we obtain (7.12) since both sides evaluate to definition (7.2). \square

The following is the main result of this section.

7.2 Data-processing inequality; approximation by finite partitions 93

Theorem 7.4 (Data processing). Consider a channel that produces Y given X based on the conditional law $P_{Y|X}$ (shown below).



Let P_Y (resp. Q_Y) denote the distribution of Y when X is distributed as P_X (resp. Q_X). For any f -divergence $D_f(\cdot \parallel \cdot)$,

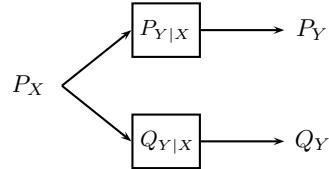
$$D_f(P_Y \parallel Q_Y) \leq D_f(P_X \parallel Q_X).$$

Proof. This follows from the monotonicity (7.12) and (7.9). \square

Next we discuss some of the more useful properties of f -divergence that parallel those of KL divergence in Theorem 2.15:

Theorem 7.5 (Properties of f -divergences).

- (a) Non-negativity: $D_f(P \parallel Q) \geq 0$. If f is strictly convex² at 1, then $D_f(P \parallel Q) = 0$ if and only if $P = Q$.
- (b) Joint convexity: $(P, Q) \mapsto D_f(P \parallel Q)$ is a jointly convex function. Consequently, $P \mapsto D_f(P \parallel Q)$ and $Q \mapsto D_f(P \parallel Q)$ are also convex.
- (c) Conditioning increases f -divergence. Let $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = Q_{Y|X} \circ Q_X$, or, pictorially,



Then

$$D_f(P_Y \parallel Q_Y) \leq D_f(P_{Y|X} \parallel Q_{Y|X} \mid P_X).$$

Proof. (a) Non-negativity follows from monotonicity by taking X to be unary. To show strict positivity, suppose for the sake of contradiction that $D_f(P \parallel Q) = 0$ for some $P \neq Q$. Then there exists some measurable A such that $p = P(A) \neq q = Q(A) > 0$. Applying the data

² By strict convexity at 1, we mean for all $s, t \in [0, \infty)$ and $\alpha \in (0, 1)$ such that $\alpha s + \bar{\alpha}t = 1$, we have $\alpha f(s) + (1 - \alpha)f(t) > f(1)$.

processing inequality (with $Y = 1_{\{X \in A\}}$), we obtain $D_f(\text{Ber}(p) \parallel \text{Ber}(q)) = 0$. Consider two cases

- a $0 < q < 1$: Then $D_f(\text{Ber}(p) \parallel \text{Ber}(q)) = qf\left(\frac{p}{q}\right) + \bar{q}f\left(\frac{\bar{p}}{\bar{q}}\right) = f(1)$;
- b $q = 1$: Then $p < 1$ and $D_f(\text{Ber}(p) \parallel \text{Ber}(q)) = f(p) + \bar{p}f'(\infty) = 0$, i.e. $f'(\infty) = \frac{f(p)}{p-1}$. Since $x \mapsto \frac{f(x)}{x-1}$ is non-decreasing, we conclude that f is affine on $[p, \infty)$.

Both cases contradict the assumed strict convexity of f at 1.

(b) Convexity follows from the DPI as in the proof of Theorem 5.1.

(c) Recall that the conditional divergence was defined in (7.8) and hence the inequality follows from the monotonicity. Another way to see the inequality is as result of applying Jensen's inequality to the jointly convex function $D_f(P \parallel Q)$. \square

Remark 7.4 (Strict convexity). Note that even when f is strictly convex at 1, the map $(P, Q) \mapsto D_f(P \parallel Q)$ may not be strictly convex (e.g. $\text{TV}(\text{Ber}(p), \text{Ber}(q)) = |p - q|$ is piecewise linear). However, if f is strictly convex everywhere on \mathbb{R}_+ then so is D_f . Indeed, if $P \neq Q$, then there exists E such that $P(E) \neq Q(E)$. By the DPI and the strict convexity of f , we have $D_f(P \parallel Q) \geq D_f(\text{Ber}(P(E)) \parallel \text{Ber}(Q(E))) > 0$. Strict convexity of f is also related to other desirable properties of $I_f(X; Y)$, see Ex. I.31.

Remark 7.5 (g -divergences). We note that, more generally, we may call functional $\mathcal{D}(P \parallel Q)$ a “ g -divergence”, or a generalized dissimilarity measure, if it satisfies the following properties: positivity, monotonicity, data processing inequality (DPI), conditioning increases divergence (CID) and convexity in the pair. As we have seen in the proof of Theorem 5.1 the latter two are exactly equivalent. Furthermore, our proof demonstrated that DPI and CID are both implied by monotonicity. If $\mathcal{D}(P \parallel P) = 0$ then monotonicity, as in (7.12), also implies positivity by taking X to be unary. Finally, notice that DPI also implies monotonicity by applying it to the (deterministic) channel $(X, Y) \mapsto X$. Thus, requiring DPI (or monotonicity) for \mathcal{D} automatically implies all the other main properties. We remark also that there exist g -divergences which are not monotone transformations of any f -divergence, cf. [235, Section V]. On the other hand, for finite alphabets, [226] shows that any $\mathcal{D}(P \parallel Q) = \sum_i \phi(P_i, Q_i)$ is a g -divergence iff it is an f -divergence.

The following convenient property, a counterpart of Theorem 4.6, allows us to reduce any general problem about f -divergences to the problem on finite alphabets. The proof is in Section 7.14*.

Theorem 7.6. Let P, Q be two probability measures on \mathcal{X} with σ -algebra \mathcal{F} . Given a finite \mathcal{F} -measurable partitions $\mathcal{E} = \{E_1, \dots, E_n\}$ define the distribution $P_{\mathcal{E}}$ on $[n]$ by $P_{\mathcal{E}}(i) = P[E_i]$ and $Q_{\mathcal{E}}(i) = Q[E_i]$. Then

$$D_f(P \parallel Q) = \sup_{\mathcal{E}} D_f(P_{\mathcal{E}} \parallel Q_{\mathcal{E}}) \tag{7.14}$$

where the supremum is over all finite \mathcal{F} -measurable partitions \mathcal{E} .

7.3 Total variation and Hellinger distance in hypothesis testing

As we will discover throughout the book, different f -divergences have different operational significance. For example, χ^2 -divergence is useful in the study of Markov chains (see Example 33.1 and Exercise VI.18); for estimation the Bayes quadratic risk for a binary prior is determined by Le Cam divergence (7.6). Here we discuss the relation of TV and Hellinger H^2 to the problem of binary hypothesis testing. We will delve deep into this problem in Part III (and return to its composite version in Part VI). In this section, we only introduce some basics for the purpose of illustration.

The *binary hypothesis testing* problem is formulated as follows: one is given an observation (random variable) X , and it is known that either $X \sim P$ (a case referred to as null-hypothesis H_0) or $X \sim Q$ (alternative hypothesis H_1). The goal is to decide, on the basis of X alone, which of the two hypotheses holds. In other words, we want to find a (possibly randomized) decision function $\phi : \mathcal{X} \rightarrow \{0, 1\}$ such that the sum of two types of probabilities of error

$$P[\phi(X) = 1] + Q[\phi(X) = 0] \quad (7.15)$$

is minimized.

In this section we first show that optimization over ϕ naturally leads to the concept of TV. Subsequently, we will see that asymptotic considerations (when P and Q are replaced with $P^{\otimes n}$ and $Q^{\otimes n}$) leads to H^2 . We start with the former case.

Theorem 7.7. (a) *Sup-representation of total variation:*

$$\text{TV}(P, Q) = \sup_E P(E) - Q(E) = \frac{1}{2} \sup_{f \in \mathcal{F}} \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \quad (7.16)$$

where the first supremum is over all measurable sets E , and the second is over $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R}, \|f\|_\infty \leq 1\}$. In particular, the minimal sum of error probabilities in (7.15) is given by

$$\min_{\phi} \{P[\phi(X) = 1] + Q[\phi(X) = 0]\} = 1 - \text{TV}(P, Q), \quad (7.17)$$

where the minimum is over all decision rules $\phi : \mathcal{X} \rightarrow \{0, 1\}$.³

(b) *Inf-representation of TV [288]:* Provided that the diagonal $\{(x, x) : x \in \mathcal{X}\}$ is measurable,

$$\text{TV}(P, Q) = \inf_{P_{X,Y}} \{P_{X,Y}[X \neq Y] : P_X = P, P_Y = Q\}, \quad (7.18)$$

where the set of joint distribution $P_{X,Y}$ with the property $P_X = P$ and $P_Y = Q$ are called couplings of P and Q .

Proof. Let p, q, μ be as in Definition 7.1. Then for any $f \in \mathcal{F}$ we have

$$\int f(x)(p(x) - q(x))d\mu \leq \int |p(x) - q(x)|d\mu = 2\text{TV}(P, Q),$$

³ The extension of (7.17) from simple to composite hypothesis testing is in (32.24).

which establishes that the second supremum in (7.16) lower bounds TV, and hence (by taking $f(x) = 2 \cdot 1_E(x) - 1$) so does the first. For the other direction, let $E = \{x : p(x) > q(x)\}$ and notice

$$0 = \int (p(x) - q(x))d\mu = \int_E + \int_{E^c} (p(x) - q(x))d\mu,$$

implying that $\int_{E^c} (q(x) - p(x))d\mu = \int_E (p(x) - q(x))d\mu$. But the sum of these two integrals precisely equals $2 \cdot \text{TV}$, which implies that this choice of E attains equality in (7.16).

For the inf-representation, we notice that given a coupling $P_{X,Y}$, for any $\|f\|_\infty \leq 1$, we have

$$\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] = \mathbb{E}[f(X) - f(Y)] \leq 2P_{X,Y}[X \neq Y]$$

which, in view of (7.16), shows that the inf-representation is always an upper bound. To show that this bound is tight one constructs X, Y as follows: with probability $\pi \triangleq \int \min(p(x), q(x))d\mu$ we take $X = Y = c$ with c sampled from a distribution with density $r(x) = \frac{1}{\pi} \min(p(x), q(x))$, whereas with probability $1 - \pi$ we take X, Y sampled independently from distributions $p_1(x) = \frac{1}{1-\pi}(p(x) - \min(p(x), q(x)))$ and $q_1(x) = \frac{1}{1-\pi}(q(x) - \min(p(x), q(x)))$ respectively. The result follows upon verifying that this $P_{X,Y}$ indeed defines a coupling of P and Q and applying the last identity of (7.3). \square

Remark 7.6 (Variational representation). The sup-representation (7.16) of the total variation will be extended to general f -divergences in Section 7.13. In turn, the inf-representation (7.18) has no analogs for other f -divergences, with the notable exception of Marton's d_2 , see Remark 7.39. Distances defined via inf-representations over couplings are often called *Wasserstein distances*, and hence we may think of TV as the Wasserstein distance with respect to Hamming distance $d(x, x') = 1\{x \neq x'\}$ on \mathcal{X} . The benefit of variational representations is that choosing a particular coupling in (7.18) gives an upper bound on $\text{TV}(P, Q)$, and choosing a particular f in (7.16) yields a lower bound.

Of particular relevance is the special case of testing with multiple observations, where the data $X = (X_1, \dots, X_n)$ are i.i.d. drawn from either P or Q . In other words, the goal is to test

$$H_0 : X \sim P^{\otimes n} \quad \text{vs} \quad H_1 : X \sim Q^{\otimes n}.$$

By Theorem 7.12, the optimal total probability of error is given by $1 - \text{TV}(P^{\otimes n}, Q^{\otimes n})$. By the data processing inequality, $\text{TV}(P^{\otimes n}, Q^{\otimes n})$ is a non-decreasing sequence in n (and bounded by 1 by definition) and hence converges. One would expect that as $n \rightarrow \infty$, $\text{TV}(P^{\otimes n}, Q^{\otimes n})$ converges to 1 and consequently, the probability of error in the hypothesis test vanishes. It turns out that for fixed distributions $P \neq Q$, large deviation theory (see Chapter 16) shows that $\text{TV}(P^{\otimes n}, Q^{\otimes n})$ indeed converges to one as $n \rightarrow \infty$ and, in fact, exponentially fast:

$$\text{TV}(P^{\otimes n}, Q^{\otimes n}) = 1 - \exp(-nC(P, Q) + o(n)), \tag{7.19}$$

where the exponent $C(P, Q) > 0$ is known as the *Chernoff Information* of P and Q given in (16.2). However, as frequently encountered in high-dimensional statistical problems, if the distributions $P = P_n$ and $Q = Q_n$ depend on n , then the large-deviation asymptotics in (7.19) can no longer be directly applied. Since computing the total variation between two n -fold product distributions is

7.3 Total variation and Hellinger distance in hypothesis testing 97

typically difficult, understanding how a more tractable f -divergence is related to the total variation may give insight on its behavior. It turns out Hellinger distance is precisely suited for this task.

Shortly, we will show the following relation between TV and the Hellinger divergence:

$$\frac{1}{2}H^2(P, Q) \leq \text{TV}(P, Q) \leq H(P, Q)\sqrt{1 - \frac{H^2(P, Q)}{4}} \leq 1. \quad (7.20)$$

Direct consequences of the bound (7.20) are:

- $H^2(P, Q) = 2$, if and only if $\text{TV}(P, Q) = 1$. In this case, the probability of error is zero since essentially P and Q have disjoint supports.
- $H^2(P, Q) = 0$ if and only if $\text{TV}(P, Q) = 0$. In this case, the smallest total probability of error is one, meaning the best test is random guessing.
- Hellinger consistency is equivalent to TV consistency: we have

$$H^2(P_n, Q_n) \rightarrow 0 \iff \text{TV}(P_n, Q_n) \rightarrow 0 \quad (7.21)$$

$$H^2(P_n, Q_n) \rightarrow 2 \iff \text{TV}(P_n, Q_n) \rightarrow 1; \quad (7.22)$$

however, the speed of convergence need not be the same.

Theorem 7.8. *For any sequence of distributions P_n and Q_n , as $n \rightarrow \infty$,*

$$\begin{aligned} \text{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 0 &\iff H^2(P_n, Q_n) = o\left(\frac{1}{n}\right) \\ \text{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 1 &\iff H^2(P_n, Q_n) = \omega\left(\frac{1}{n}\right) \end{aligned}$$

Proof. For convenience, let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} Q_n$. Then

$$\begin{aligned} H^2(P_n^{\otimes n}, Q_n^{\otimes n}) &= 2 - 2\mathbb{E}\left[\sqrt{\prod_{i=1}^n \frac{P_n}{Q_n}(X_i)}\right] \\ &= 2 - 2\prod_{i=1}^n \mathbb{E}\left[\sqrt{\frac{P_n}{Q_n}(X_i)}\right] = 2 - 2\left(\mathbb{E}\left[\sqrt{\frac{P_n}{Q_n}}\right]\right)^n \\ &= 2 - 2\left(1 - \frac{1}{2}H^2(P_n, Q_n)\right)^n. \end{aligned} \quad (7.23)$$

We now use (7.23) to conclude the proof. Recall from (7.21) that $\text{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 0$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 0$, which happens precisely when $H^2(P_n, Q_n) = o(\frac{1}{n})$. Similarly, by (7.22), $\text{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 1$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \rightarrow 2$, which is further equivalent to $H^2(P_n, Q_n) = \omega(\frac{1}{n})$. \square

Remark 7.7. Property (7.23) is known as *tensorization*. More generally, we have

$$H^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) = 2 - 2\prod_{i=1}^n \left(1 - \frac{1}{2}H^2(P_i, Q_i)\right). \quad (7.24)$$

While some other f -divergences also satisfy tensorization, see Section 7.12, the H^2 has the advantage of a sandwich bound (7.20) making it the most convenient tool for checking asymptotic testability of hypotheses.

Remark 7.8 (Kakutani's dichotomy). Let $P = \prod_{i \geq 1} P_i$ and $Q = \prod_{i \geq 1} Q_i$, where $P_i \ll Q_i$. Kakutani's theorem shows the following dichotomy between these two distributions on the infinite sequence space:

- If $\sum_{i \geq 1} H^2(P_i, Q_i) = \infty$, then P and Q are mutually singular.
- If $\sum_{i \geq 1} H^2(P_i, Q_i) < \infty$, then P and Q are equivalent (i.e. absolutely continuous with respect to each other).

In the Gaussian case, say, $P_i = N(\mu_i, 1)$ and $Q_i = N(0, 1)$, the equivalence condition simplifies to $\sum \mu_i^2 < \infty$.

To understand Kakutani's criterion, note that by the tensorization property (7.24), we have

$$H^2(P, Q) = 2 - 2 \prod_{i \geq 1} \left(1 - \frac{H^2(P_i, Q_i)}{2} \right).$$

Thus, if $\prod_{i \geq 1} \left(1 - \frac{H^2(P_i, Q_i)}{2} \right) = 0$, or equivalently, $\sum_{i \geq 1} H^2(P_i, Q_i) = \infty$, then $H^2(P, Q) = 2$, which, by (7.20), is equivalent to $\text{TV}(P, Q) = 0$ and hence $P \perp Q$. If $\sum_{i \geq 1} H^2(P_i, Q_i) < \infty$, then $H^2(P, Q) < 2$. To conclude the equivalence between P and Q , note that the likelihood ratio $\frac{dP}{dQ} = \prod_{i \geq 1} \frac{dP_i}{dQ_i}$ satisfies that either $Q\left(\frac{dP}{dQ} = 0\right) = 0$ or 1 by Kolmogorov's 0-1 law. See [107, Theorem 5.3.5] for details.

7.4 Inequalities between f -divergences and joint range

In this section we study the relationship, in particular, inequalities, between f -divergences. To gain some intuition, we start with the ad hoc approach by proving the *Pinsker's inequality*, which bounds total variation from above in terms of the KL divergence.

Theorem 7.9 (Pinsker's inequality).

$$D(P||Q) \geq (2 \log e) \text{TV}^2(P, Q). \quad (7.25)$$

Proof. It suffices to consider the natural logarithm for the KL divergence. First we show that, by the data processing inequality, it suffices to prove the result for Bernoulli distributions. For any event E , let $Y = 1_{\{X \in E\}}$ which is Bernoulli with parameter $P(E)$ or $Q(E)$. By the DPI, $D(P||Q) \geq d(P(E)||Q(E))$. If Pinsker's inequality holds for all Bernoulli distributions, we have

$$\sqrt{\frac{1}{2} D(P||Q)} \geq \text{TV}(\text{Ber}(P(E)), \text{Ber}(Q(E))) = |P(E) - Q(E)|$$

7.4 Inequalities between f -divergences and joint range 99

Taking the supremum over E gives $\sqrt{\frac{1}{2}D(P\|Q)} \geq \sup_E |P(E) - Q(E)| = \text{TV}(P, Q)$, in view of Theorem 7.12.

The binary case follows easily from a second-order Taylor expansion (with integral remainder form) of $p \mapsto d(p\|q)$:

$$d(p\|q) = \int_q^p \frac{p-t}{t(1-t)} dt \geq 4 \int_q^p (p-t) dt = 2(p-q)^2$$

and $\text{TV}(\text{Ber}(p), \text{Ber}(q)) = |p - q|$. \square

Pinsker's inequality is sharp in the sense that the constant $(2 \log e)$ in (7.25) is not improvable, i.e., there exist $\{P_n, Q_n\}$, e.g., $P_n = \text{Ber}(\frac{1}{2} + \frac{1}{n})$ and $Q_n = \text{Ber}(\frac{1}{2})$, such that $\frac{\text{LHS}}{\text{RHS}} \rightarrow 2$ as $n \rightarrow \infty$. (This is best seen by inspecting the local quadratic behavior in Proposition 2.19.) Nevertheless, this does not mean that the inequality (7.25) is not improvable, as the RHS can be replaced by some other function of $\text{TV}(P, Q)$ with additional higher-order terms. Indeed, several such improvements of Pinsker's inequality are known. But what is the best inequality? In addition, another natural question is the reverse inequality: can we upper-bound $D(P\|Q)$ in terms of $\text{TV}(P, Q)$? Settling these questions rests on characterizing the *joint range* (the set of possible values) of a given pair f -divergences. This systematic approach to comparing f -divergences (as opposed to the ad hoc proof of Theorem 7.17 we presented above) is the subject of this section.

Definition 7.10 (Joint range). Consider two f -divergences $D_f(P\|Q)$ and $D_g(P\|Q)$. Their joint range is a subset of $[0, \infty]^2$ defined by

$$\mathcal{R} \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on some measurable space}\}.$$

In addition, the joint range over all k -ary distributions is defined as

$$\mathcal{R}_k \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on } [k]\}.$$

As an example, Fig. 7.1 gives the joint range \mathcal{R} between the KL divergence and the total variation. By definition, the lower boundary of the region \mathcal{R} gives the optimal refinement of Pinsker's inequality:

$$D(P\|Q) \geq F(\text{TV}(P, Q)), \quad F(\epsilon) \triangleq \inf_{(P, Q) : \text{TV}(P, Q) = \epsilon} D(P\|Q) = \inf\{s : (\epsilon, s) \in \mathcal{R}\}.$$

Also from Fig. 7.1 we see that it is impossible to bound $D(P\|Q)$ from above in terms of $\text{TV}(P, Q)$ due to the lack of upper boundary.

The joint range \mathcal{R} may appear difficult to characterize since we need to consider P, Q over all measurable spaces; on the other hand, the region \mathcal{R}_k for small k is easy to obtain (at least numerically). Revisiting the proof of Pinkser's inequality in Theorem 7.17, we see that the key step is the reduction to Bernoulli distributions. It is natural to ask: to obtain full joint range is it possible to reduce to the binary case? It turns out that it is always sufficient to consider quaternary distributions, or the convex hull of that of binary distributions.

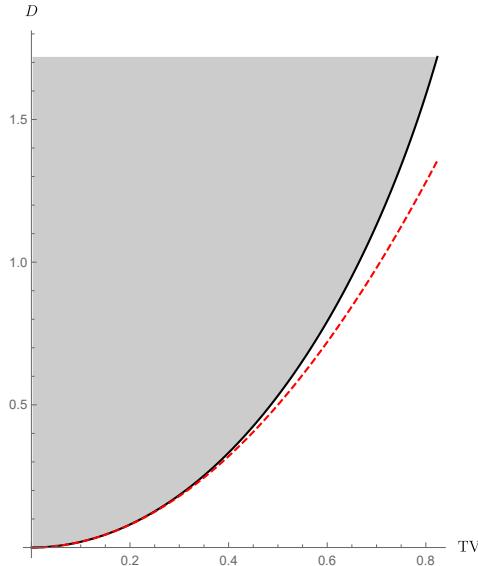


Figure 7.1 Joint range of TV and KL divergence. The dashed line is the quadratic lower bound given by Pinsker’s inequality (7.25).

Theorem 7.11 (Harremoës-Vajda [151]).

$$\mathcal{R} = \text{co}(\mathcal{R}_2) = \mathcal{R}_4.$$

where co denotes the convex hull with a natural extension of convex operations to $[0, \infty]^2$.

We will rely on the following famous result from convex analysis (cf. e.g. [109, Chapter 2, Theorem 18]).

Lemma 7.12 (Fenchel-Eggleston-Carathéodory theorem). *Let $S \subseteq \mathbb{R}^d$ and $x \in \text{co}(S)$. Then there exists a set of $d + 1$ points $S' = \{x_1, x_2, \dots, x_{d+1}\} \in S$ such that $x \in \text{co}(S')$. If S has at most d connected components, then d points are enough.*

Proof. Our proof will consist of three claims:

- *Claim 1:* $\text{co}(\mathcal{R}_2) \subset \mathcal{R}_4$;
- *Claim 2:* $\mathcal{R}_k \subset \text{co}(\mathcal{R}_2)$;
- *Claim 3:* $\mathcal{R} = \mathcal{R}_4$.

Note that Claims 1-2 prove the most interesting part: $\bigcup_{k=1}^{\infty} \mathcal{R}_k = \text{co}(\mathcal{R}_2)$. Claim 3 is more technical and its proof can be found in [151]. However, the approximation result in Theorem 7.11 shows that \mathcal{R} is the closure of $\bigcup_{k=1}^{\infty} \mathcal{R}_k$. Thus for the purpose of obtaining inequalities between D_f and D_g , Claims 1-2 are sufficient.

7.4 Inequalities between f -divergences and joint range 101

We start with Claim 1. Given any two pairs of distributions (P_0, Q_0) and (P_1, Q_1) on some space \mathcal{X} and given any $\alpha \in [0, 1]$, define two joint distributions of the random variables (X, B) where $P_B = Q_B = \text{Ber}(\alpha)$, $P_{X|B=i} = P_i$ and $Q_{X|B=i} = Q_i$ for $i = 0, 1$. Then by (7.8) we get

$$D_f(P_{X,B}\|Q_{X,B}) = \bar{\alpha}D_f(P_0\|Q_0) + \alpha D_f(P_1\|Q_1),$$

and similarly for the D_g . Thus, \mathcal{R} is convex. Next, notice that

$$\mathcal{R}_2 = \tilde{\mathcal{R}}_2 \cup \{(pf'(\infty), pg'(\infty)) : p \in (0, 1]\} \cup \{(qf(0), qg(0)) : q \in (0, 1]\},$$

where $\tilde{\mathcal{R}}_2$ is the image of $(0, 1)^2$ of the continuous map

$$(p, q) \mapsto \left(D_f(\text{Ber}(p)\|\text{Ber}(q)), D_g(\text{Ber}(p)\|\text{Ber}(q)) \right).$$

Since $(0, 0) \in \tilde{\mathcal{R}}_2$, we see that regardless of which $f(0), f'(\infty), g(0), g'(\infty)$ are infinite, the set $\mathcal{R}_2 \cap \mathbb{R}^2$ is connected. Thus, by Lemma 7.20 any point in $\text{co}(\mathcal{R}_2 \cap \mathbb{R}^2)$ is a combination of two points in $\mathcal{R}_2 \cap \mathbb{R}^2$, which, by the argument above, is a subset of \mathcal{R}_4 . Finally, it is not hard to see that $\text{co}(\mathcal{R}_2) \setminus \mathbb{R}^2 \subset \mathcal{R}_4$, which concludes the proof of $\text{co}(\mathcal{R}_2) \subset \mathcal{R}_4$.

Next, we prove Claim 2. Fix P, Q on $[k]$ and denote their PMFs (p_j) and (q_j) , respectively. Note that without changing either $D_f(P\|Q)$ or $D_g(P\|Q)$ (but perhaps, by increasing k by 1), we can make $q_j > 0$ for $j > 1$ and $q_1 = 0$, which we thus assume. Denote $\phi_j = \frac{p_j}{q_j}$ for $j > 1$ and consider the set

$$\mathcal{S} = \left\{ \tilde{Q} = (\tilde{q}_j)_{j \in [k]} : \tilde{q}_j \geq 0, \sum \tilde{q}_j = 1, \tilde{q}_1 = 0, \sum_{j=2}^k \tilde{q}_j \phi_j \leq 1 \right\}.$$

We also define a subset $\mathcal{S}_e \subset \mathcal{S}$ consisting of points \tilde{Q} of two types:

- 1 $\tilde{q}_j = 1$ for some $j \geq 2$ and $\phi_j \leq 1$.
- 2 $\tilde{q}_{j_1} + \tilde{q}_{j_2} = 1$ for some $j_1, j_2 \geq 2$ and $\tilde{q}_{j_1} \phi_{j_1} + \tilde{q}_{j_2} \phi_{j_2} = 1$.

It can be seen that \mathcal{S}_e are precisely all the extreme points of \mathcal{S} . Indeed, any $\tilde{Q} \in \mathcal{S}$ with $\sum_{j \geq 2} \tilde{q}_j \phi_j < 1$ with more than one non-zero atom cannot be extremal (since there is only one active linear constraint $\sum_j \tilde{q}_j = 1$). Similarly, \tilde{Q} with $\sum_{j \geq 2} \tilde{q}_j \phi_j = 1$ can only be extremal if it has one or two non-zero atoms.

We next claim that any point in \mathcal{S} can be written as a convex combination of finitely many points in \mathcal{S}_e . This can be seen as follows. First, we can view \mathcal{S} and \mathcal{S}_e as subsets of \mathbb{R}^{k-1} . Since \mathcal{S} is clearly closed and convex, by the Krein-Milman theorem (see [7, Theorem 7.68]), \mathcal{S} coincides with the closure of the convex hull of its extreme points. Since \mathcal{S}_e is compact (hence closed), so is $\text{co}(\mathcal{S}_e)$ [7, Corollary 5.33]. Thus we have $\mathcal{S} = \text{co}(\mathcal{S}_e)$ and, in particular, there are probability weights $\{\alpha_i : i \in [m]\}$ and extreme points $\tilde{Q}_i \in \mathcal{S}_e$ so that

$$Q = \sum_{i=1}^m \alpha_i \tilde{Q}_i. \tag{7.26}$$

Next, to each \tilde{Q} we associate $\tilde{P} = (\tilde{p}_j)_{j \in [k]}$ as follows:

$$\tilde{p}_j = \begin{cases} \phi_j \tilde{q}_j, & j \in \{2, \dots, k\}, \\ 1 - \sum_{j=2}^k \phi_j \tilde{q}_j, & j = 1 \end{cases}$$

We then have that

$$\tilde{Q} \mapsto D_f(\tilde{P} \parallel \tilde{Q}) = \sum_{j \geq 2} \tilde{q}_j f(\phi_j) + f'(\infty) \tilde{p}_1$$

affinely maps \mathcal{S} to $[0, \infty]$ (note that $f(0)$ or $f'(\infty)$ can equal ∞). In particular, if we denote $\tilde{P}_i = \tilde{P}(\tilde{Q}_i)$ corresponding to \tilde{Q}_i in decomposition (7.26), we get

$$D_f(P \parallel Q) = \sum_{i=1}^m \alpha_i D_f(\tilde{P}_i \parallel \tilde{Q}_i),$$

and similarly for $D_g(P \parallel Q)$. We are left to show that $(\tilde{P}_i, \tilde{Q}_i)$ are supported on at most two points, which verifies that any element of \mathcal{R}_k is a convex combination of k elements of \mathcal{R}_2 . Indeed, for $\tilde{Q} \in \mathcal{S}_e$ the set $\{j \in [k] : \tilde{q}_j > 0 \text{ or } \tilde{p}_j > 0\}$ has cardinality at most two (for the second type extremal points we notice $\tilde{p}_{j_1} + \tilde{p}_{j_2} = 1$ implying $\tilde{p}_1 = 0$). This concludes the proof of Claim 2. \square

7.5 Examples of computing joint range

In this section we show how to apply the method of Harremoës and Vajda for proving the best possible comparison inequalities between various f -divergences.

7.5.1 Hellinger distance versus total variation

The joint range \mathcal{R}_2 of H^2 and TV over binary distributions is simply:

$$\mathcal{R}_2 = \{(2(1 - \sqrt{pq} - \sqrt{\bar{p}\bar{q}}), |p - q|) : 0 \leq p \leq 1, 0 \leq q \leq 1\}.$$

shown as non-convex grey region in Fig. 7.2. By Theorem 7.19, their full joint range \mathcal{R} is the convex hull of \mathcal{R}_2 , which turns out to be exactly described by the sandwich bound (7.20) shown earlier in Section 7.3. This means that (7.20) is not improvable. Indeed, with t ranging from 0 to 1,

- the upper boundary is achieved by $P = \text{Ber}(\frac{1+t}{2}), Q = \text{Ber}(\frac{1-t}{2})$,
- the lower boundary is achieved by $P = (1-t, t, 0), Q = (1-t, 0, t)$.

7.5 Examples of computing joint range 103

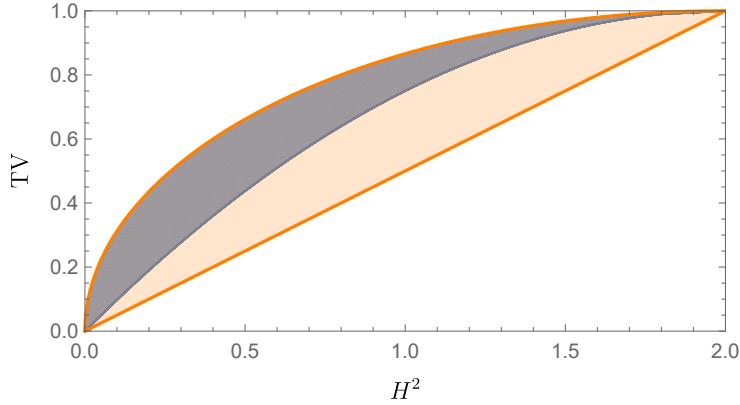


Figure 7.2 The joint range \mathcal{R} of TV and H^2 is characterized by (7.20), which is the convex hull of the grey region \mathcal{R}_2 .

7.5.2 KL divergence versus total variation

The joint range between KL and TV was previously shown in Fig. 7.1. Although there is no known close-form expression, the following parametric formula of the lower boundary (see Fig. 7.1) is known [119, Theorem 1]:

$$\begin{cases} \text{TV}_t = \frac{1}{2}t \left(1 - \left(\coth(t) - \frac{1}{t} \right)^2 \right) \\ D_t = -t^2 \operatorname{csch}^2(t) + t \coth(t) + \log(t \operatorname{csch}(t)) \end{cases}, \quad t \geq 0. \quad (7.27)$$

where we take the natural logarithm. Here is a corollary (weaker bound) due to [307]:

$$D(P\|Q) \geq \log \frac{1 + \text{TV}(P, Q)}{1 - \text{TV}(P, Q)} - \frac{2\text{TV}(P, Q)}{1 + \text{TV}(P, Q)}. \quad (7.28)$$

Both bounds are stronger than Pinsker's inequality (7.25). Note the following consequences:

- $D \rightarrow 0 \Rightarrow \text{TV} \rightarrow 0$, which can be deduced from Pinsker's inequality;
- $\text{TV} \rightarrow 1 \Rightarrow D \rightarrow \infty$ and hence $D = O(1)$ implies that TV is bounded away from one. This can be obtained from (7.27) or (7.28), but not Pinsker's inequality.

7.5.3 Chi-squared versus total variation

Proposition 7.13. *We have the following bound*

$$\chi^2(P\|Q) \geq f(\text{TV}(P, Q)) \geq 4\text{TV}^2(P, Q), \quad f(t) = \begin{cases} 4t^2 & t \leq \frac{1}{2} \\ \frac{t}{1-t} & t \geq \frac{1}{2} \end{cases}, \quad (7.29)$$

where the function f is a convex increasing bijection of $[0, 1]$ onto $[0, \infty)$. Furthermore, for every $s \geq f(t)$ there exists a pair of distributions such that $\chi^2(P\|Q) = s$ and $\text{TV}(P, Q) = t$.

Proof. We claim that the binary joint range is convex. Indeed,

$$\text{TV}(\text{Ber}(p), \text{Ber}(q)) = |p - q| \triangleq t, \quad \chi^2(\text{Ber}(p) \parallel \text{Ber}(q)) = \frac{(p - q)^2}{q(1 - q)} = \frac{t^2}{q(1 - q)}.$$

Given $|p - q| = t$, let us determine the possible range of $q(1 - q)$. The smallest value of $q(1 - q)$ is always 0 by choosing $p = t, q = 0$. The largest value is $1/4$ if $t \leq 1/2$ (by choosing $p = 1/2 - t, q = 1/2$). If $t > 1/2$ then we can at most get $t(1 - t)$ (by setting $p = 0$ and $q = t$). Thus we get $\chi^2(\text{Ber}(p) \parallel \text{Ber}(q)) \geq f(|p - q|)$ as claimed. The convexity of f follows since its derivative is monotonically increasing. Clearly, $f(t) \geq 4t^2$ because $t(1 - t) \leq \frac{1}{4}$. \square

7.6 A selection of inequalities between various divergences

This section presents a collection of useful inequalities. For a more complete treatment, consider [263] and [304, Sec. 2.4]. Most of these inequalities are joint ranges, which means they are tight.

- KL vs TV: see (7.27). For discrete distributions there is partial comparison in the other direction (“reverse Pinsker”, cf. [263, Section VI]):

$$D(P \parallel Q) \leq \log \left(1 + \frac{2}{Q_{\min}} \text{TV}(P, Q)^2 \right) \leq \frac{2 \log e}{Q_{\min}} \text{TV}(P, Q)^2, \quad Q_{\min} = \min_x Q(x)$$

- KL vs Hellinger:

$$D(P \parallel Q) \geq 2 \log \frac{2}{2 - H^2(P, Q)}. \quad (7.30)$$

This is tight at $P = \text{Ber}(0), Q = \text{Ber}(q)$. For a fixed H^2 , in general $D(P \parallel Q)$ has no finite upper bound, as seen from $P = \text{Ber}(p), Q = \text{Ber}(0)$. Therefore (7.30) gives the joint range.

There is a partial result in the opposite direction (log-Sobolev inequality for Bonami-Beckner semigroup, cf. [88, Theorem A.1]):

$$D(P \parallel Q) \leq \frac{\log(\frac{1}{Q_{\min}} - 1)}{1 - 2Q_{\min}} (1 - (1 - H^2(P, Q))^2), \quad Q_{\min} = \min_x Q(x)$$

Another partial result is in Ex. I.48.

- KL vs χ^2 :

$$0 \leq D(P \parallel Q) \leq \log(1 + \chi^2(P \parallel Q)) \leq \log e \cdot \chi^2(P \parallel Q). \quad (7.31)$$

The left-hand inequality states that no lower bound on KL in terms of χ^2 is possible.

- TV and Hellinger: see (7.20). Another bound [135]:

$$\text{TV}(P, Q) \leq \sqrt{-2 \ln \left(1 - \frac{H^2(P, Q)}{2} \right)}$$

7.7 Divergences between Gaussians 105

- Le Cam and Hellinger [189, p. 48]:

$$\frac{1}{2}H^2(P, Q) \leq LC(P, Q) \leq H^2(P, Q). \quad (7.32)$$

- Le Cam and Jensen-Shannon [302]:

$$LC(P, Q) \log e \leq JS(P, Q) \leq LC(P, Q) \cdot 2 \log 2 \quad (7.33)$$

- χ^2 and TV: The full joint range is given by (7.29). Two simple consequences are:

$$TV(P, Q) \leq \frac{1}{2} \sqrt{\chi^2(P||Q)} \quad (7.34)$$

$$TV(P, Q) \leq \max \left\{ \frac{1}{2}, \frac{\chi^2(P||Q)}{1 + \chi^2(P||Q)} \right\} \quad (7.35)$$

where the second is useful for bounding TV away from one.

- JS and TV: The full joint region is given by

$$2d \left(\frac{1 - TV(P, Q)}{2} \middle\| \frac{1}{2} \right) \leq JS(P, Q) \leq TV(P, Q) \cdot 2 \log 2. \quad (7.36)$$

The lower bound is a consequence of Fano's inequality. For the upper bound notice that for $p, q \in [0, 1]$ and $|p - q| = \tau$ the maximum of $d(p \parallel \frac{p+q}{2})$ is attained at $p = 0, q = \tau$ (from the convexity of $d(\cdot \parallel \cdot)$) and, thus, the binary joint-range is given by $\tau \mapsto d(\tau \parallel \tau/2) + d(1 - \tau \parallel 1 - \tau/2)$. Since the latter is convex, its concave envelope is a straightline connecting endpoints at $\tau = 0$ and $\tau = 1$.

7.7 Divergences between Gaussians

To get a better feel for the behavior of f -divergences, here we collect expressions (as well as asymptotic expansions near 0) of divergences between a pair of Gaussian distributions.

- 1 Total variation:

$$TV(\mathcal{N}(0, \sigma^2), \mathcal{N}(\mu, \sigma^2)) = 2\Phi\left(\frac{|\mu|}{2\sigma}\right) - 1 = \int_{-\frac{|\mu|}{2\sigma}}^{\frac{|\mu|}{2\sigma}} \varphi(x) dx = \frac{|\mu|}{\sqrt{2\pi}\sigma} + O(\mu^2), \quad \mu \rightarrow 0. \quad (7.37)$$

- 2 Hellinger distance:

$$H^2(\mathcal{N}(0, \sigma^2) \parallel \mathcal{N}(\mu, \sigma^2)) = 2 - 2e^{-\frac{\mu^2}{8\sigma^2}} = \frac{\mu^2}{4\sigma^2} + O(\mu^3), \quad \mu \rightarrow 0. \quad (7.38)$$

More generally,

$$H^2(\mathcal{N}(\mu_1, \Sigma_1) \parallel \mathcal{N}(\mu_2, \Sigma_2)) = 2 - 2 \frac{|\Sigma_1|^{\frac{1}{4}} |\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{8} (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} (\mu_1 - \mu_2) \right\},$$

where $\bar{\Sigma} = \frac{\Sigma_1 + \Sigma_2}{2}$.

3 KL divergence:

$$D(\mathcal{N}(\mu_1, \sigma_1^2) \parallel \mathcal{N}(\mu_2, \sigma_2^2)) = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{1}{2} \left(\frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right) \log e. \quad (7.39)$$

For a more general result see (2.8).

4 χ^2 -divergence:

$$\chi^2(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, \sigma^2)) = e^{\frac{\mu^2}{\sigma^2}} - 1 = \frac{\mu^2}{\sigma^2} + O(\mu^3), \quad \mu \rightarrow 0 \quad (7.40)$$

$$\chi^2(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, 1)) = \begin{cases} \frac{e^{\mu^2/(2-\sigma^2)}}{\sigma\sqrt{2-\sigma^2}} - 1 & \sigma^2 < 2 \\ \infty & \sigma^2 \geq 2 \end{cases} \quad (7.41)$$

5 χ^2 -divergence for Gaussian mixtures [161]:

$$\chi^2(P * \mathcal{N}(0, \Sigma) \parallel \mathcal{N}(0, \Sigma)) = \mathbb{E}[e^{\langle \Sigma^{-1}X, X' \rangle}] - 1, \quad X \perp\!\!\!\perp X' \sim P.$$

7.8 Mutual information based on f -divergence

Given an f -divergence D_f , we can define a version of mutual information

$$I_f(X; Y) \triangleq D_f(P_{X,Y} \parallel P_X P_Y). \quad (7.42)$$

Theorem 7.14 (Data processing). *For $U \rightarrow X \rightarrow Y$, we have $I_f(U; Y) \leq I_f(U; X)$.*

Proof. Note that $I_f(U; X) = D_f(P_{U,X} \parallel P_U P_X) \geq D_f(P_{U,Y} \parallel P_U P_Y) = I_f(U; Y)$, where we applied the data-processing Theorem 7.7 to the (possibly stochastic) map $(U, X) \mapsto (U, Y)$. See also Remark 3.10. \square

A useful property of mutual information is that $X \perp\!\!\!\perp Y$ iff $I(X; Y) = 0$. A generalization of it is the property that for $X \rightarrow Y \rightarrow Z$ we have $I(X; Y) = I(X; Z)$ iff $X \rightarrow Z \rightarrow Y$. Both of these may or may not hold for I_f depending on the strict convexity of f , see Ex. I.31.

Another often used property of the standard mutual information is the *subadditivity*: If $P_{A,B|X} = P_{A|X}P_{B|X}$ (i.e. A and B are conditionally independent given X), then

$$I(X; A, B) \leq I(X; A) + I(X; B). \quad (7.43)$$

However, other notions of f -information have complicated relationship with subadditivity:

1 The f -information corresponding to the χ^2 -divergence,

$$I_{\chi^2}(X; Y) \triangleq \chi^2(P_{X,Y} \parallel P_X P_Y) \quad (7.44)$$

is not subadditive.

7.8 Mutual information based on f -divergence 107

- 2 The f -information corresponding to total-variation $I_{\text{TV}}(X; Y) \triangleq \text{TV}(P_{X,Y}, P_X P_Y)$ is not subadditive. Even worse, it can get stuck. For example, take $X \sim \text{Ber}(1/2)$ and $A = \text{BSC}_\delta(X)$, $B = \text{BSC}_\delta(X)$ – two independent observations of X across the BSC. A simple computation shows:

$$I_{\text{TV}}(X; A, B) = I_{\text{TV}}(X; A) = I_{\text{TV}}(X; B).$$

In other words, an additional observation does not improve TV-information at all. This is the main reason for the famous herding effect in economics [20].

- 3 The symmetric KL-divergence⁴ $I_{\text{SKL}}(X; Y) \triangleq D(P_{X,Y} \| P_X P_Y) + D(P_X P_Y \| P_{X,Y})$ satisfies, quite amazingly [185], the *additivity property*:

$$I_{\text{SKL}}(X; A, B) = I_{\text{SKL}}(X; A) + I_{\text{SKL}}(X; B) \quad (7.45)$$

Let us prove this in the discrete case. First notice the following equivalent expression for I_{SKL} :

$$I_{\text{SKL}}(X; Y) = \sum_{x, x'} P_X(x) P_X(x') D(P_{Y|X=x} \| P_{Y|X=x'}) . \quad (7.46)$$

From (7.46) we get (7.45) by the additivity $D(P_{A,B|X=x} \| P_{A,B|X=x'}) = D(P_{A|X=x} \| P_{A|X=x'}) + D(P_{B|X=x} \| P_{B|X=x'})$. To prove (7.46) first consider the obvious identity:

$$\sum_{x, x'} P_X(x) P_X(x') [D(P_Y \| P_{Y|X=x'}) - D(P_Y \| P_{Y|X=x})] = 0$$

which is rewritten as

$$\sum_{x, x'} P_X(x) P_X(x') \sum_y P_Y(y) \log \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')} = 0. \quad (7.47)$$

Next, by definition,

$$I_{\text{SKL}}(X; Y) = \sum_{x, y} [P_{X,Y}(x, y) - P_X(x) P_Y(y)] \log \frac{P_{X,Y}(x, y)}{P_X(x) P_Y(y)}.$$

Since the marginals of $P_{X,Y}$ and $P_X P_Y$ coincide, we can replace $\log \frac{P_{X,Y}(x, y)}{P_X(x) P_Y(y)}$ by any $\log \frac{P_{Y|X}(y|x)}{f(y)}$ for any f . We choose $f(y) = P_{Y|X}(y|x')$ to get

$$I_{\text{SKL}}(X; Y) = \sum_{x, y} [P_{X,Y}(x, y) - P_X(x) P_Y(y)] \log \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')}.$$

Now averaging this over $P_X(x')$ and applying (7.47) to get rid of the second term in $[\dots]$, we obtain (7.46). For another interesting property of I_{SKL} , see Ex. I.43.

⁴ This is the f -information corresponding to the Jeffreys divergence $D(P \| Q) + D(Q \| P)$.

7.9 Empirical distribution and χ^2 -information

Consider an arbitrary channel $P_{Y|X}$ and some input distribution P_X . Suppose that we have $X_i \stackrel{\text{i.i.d.}}{\sim} P_X$ for $i = 1, \dots, n$. Let

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

denote the empirical distribution corresponding to this sample. Let $P_Y = P_{Y|X} \circ P_X$ be the output distribution corresponding to P_X and $P_{Y|X} \circ \hat{P}_n$ be the output distribution corresponding to \hat{P}_n (a random distribution). Note that when $P_{Y|X=x}(\cdot) = \phi(\cdot - x)$, where ϕ is a fixed density, we can think of $P_{Y|X} \circ \hat{P}_n$ as a *kernel density estimator (KDE)*, whose density is $\hat{p}_n(x) = (\phi * \hat{P}_n)(x) = \frac{1}{n} \sum_{i=1}^n \phi(X_i - x)$. Furthermore, using the fact that $\mathbb{E}[P_{Y|X} \circ \hat{P}_n] = P_Y$, we have

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_X)] = D(P_Y \| P_X) + \mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)],$$

where the first term represents the bias of the KDE due to convolution and increases with bandwidth of ϕ , while the second term represents the variability of the KDE and decreases with the bandwidth of ϕ . Surprisingly, the second term is sharply (within a factor of two) given by the I_{χ^2} information. More exactly, we prove the following result.

Proposition 7.15. *We have*

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \leq \log \left(1 + \frac{1}{n} I_{\chi^2}(X; Y) \right), \quad (7.48)$$

where $I_{\chi^2}(X; Y)$ is defined in (7.44). Furthermore,

$$\liminf_{n \rightarrow \infty} n \mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \geq \frac{\log e}{2} I_{\chi^2}(X; Y). \quad (7.49)$$

In particular, $\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = O(1/n)$ if $I_{\chi^2}(X; Y) < \infty$ and $\omega(1/n)$ otherwise.

In Section 25.4* we will discuss an extension of this simple bound, in particular showing that in many cases about $n = \exp\{I(X; Y) + K\}$ samples are sufficient to get $e^{-O(K)}$ bound on $D(P_{Y|X} \circ \hat{P}_n \| P_Y)$.

Proof. First, a simple calculation shows that

$$\mathbb{E}[\chi^2(P_{Y|X} \circ \hat{P}_n \| P_Y)] = \frac{1}{n} I_{\chi^2}(X; Y).$$

Then from (7.31) and Jensen's inequality we get (7.48).

To get the lower bound in (7.49), let \bar{X} be drawn uniformly at random from the sample $\{X_1, \dots, X_n\}$ and let \bar{Y} be the output of the $P_{Y|X}$ channel with input \bar{X} . With this definition we have:

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = I(X^n; \bar{Y}). \quad (7.50)$$

7.9 Empirical distribution and χ^2 -information 109

Next, apply (6.2) to get

$$I(X^n; \bar{Y}) \geq \sum_{i=1}^n I(X_i; \bar{Y}) = nI(X_1; \bar{Y}).$$

Finally, notice that

$$I(X_1; \bar{Y}) = D\left(\frac{n-1}{n}P_X P_Y + \frac{1}{n}P_{X,Y} \middle\| P_X P_Y\right)$$

and apply the local expansion of KL divergence (Proposition 2.19) to get (7.49). \square

In the discrete case, by taking $P_{Y|X}$ to be the identity ($Y = X$) we obtain the following guarantee on the closeness between the empirical and the population distribution. This fact can be used to test whether the sample was truly generated by the distribution P_X .

Corollary 7.16. *Suppose P_X is discrete with support \mathcal{X} . If \mathcal{X} is infinite, then*

$$\lim_{n \rightarrow \infty} n \mathbb{E}[D(\hat{P}_n \| P_X)] = \infty. \quad (7.51)$$

Otherwise, we have

$$\mathbb{E}[D(\hat{P}_n \| P_X)] \leq \frac{\log e}{n} (|\mathcal{X}| - 1). \quad (7.52)$$

Proof. Simply notice that $I_{\chi^2}(X; X) = |\mathcal{X}| - 1$. \square

Remark 7.9. For fixed P_X , the tight asymptotic result is

$$\lim_{n \rightarrow \infty} n \mathbb{E}[D(\hat{P}_n \| P_X)] = \frac{\log e}{2} (|\text{supp}(P_X)| - 1). \quad (7.53)$$

See Lemma 13.4 below.

Corollary 7.1 is also useful for the statistical application of *entropy estimation*. Given n iid samples, a natural estimator of the entropy of P_X is the empirical entropy $\hat{H}_{\text{emp}} = H(\hat{P}_n)$ (plug-in estimator). It is clear that empirical entropy is an *underestimate*, in the sense that the bias

$$\mathbb{E}[\hat{H}_{\text{emp}}] - H(P_X) = -\mathbb{E}[D(\hat{P}_n \| P_X)]$$

is always non-negative. For fixed P_X , \hat{H}_{emp} is known to be consistent even on countably infinite alphabets [14], although the convergence rate can be arbitrarily slow, which aligns with the conclusion of (7.51). However, for large alphabet of size $\Theta(n)$, the upper bound (7.52) does not vanish (this is tight for, e.g., uniform distribution). In this case, one need to de-bias the empirical entropy (e.g. on the basis of (7.53)) or employ different techniques in order to achieve consistent estimation.

7.10 Most f -divergences are locally χ^2 -like

In this section we prove analogs of Proposition 2.17 and Proposition 2.19 for the general f -divergences.

Theorem 7.17. Suppose that $D_f(P\|Q) < \infty$ and derivative of $f(x)$ at $x = 1$ exist. Then,

$$\lim_{\lambda \rightarrow 0} \frac{1}{\lambda} D_f(\lambda P + \bar{\lambda} Q \| Q) = (1 - P[\text{supp}(Q)]) f'(\infty),$$

where as usual we take $0 \cdot \infty = 0$ in the left-hand side.

Remark 7.10. Note that we do not need a separate theorem for $D_f(Q\|\lambda P + \bar{\lambda} Q)$ since the exchange of arguments leads to another f -divergence with $f(x)$ replaced by $xf(1/x)$.

Proof. Without loss of generality we may assume $f(1) = f'(1) = 0$ and $f \geq 0$. Then, decomposing $P = \mu P_1 + \bar{\mu} P_0$ with $P_0 \perp Q$ and $P_1 \ll Q$ we have

$$\frac{1}{\lambda} D_f(\lambda P + \bar{\lambda} Q \| Q) = \bar{\mu} f'(\infty) + \int dQ \frac{1}{\lambda} f\left(1 + \lambda(\mu \frac{dP_1}{dQ} - 1)\right).$$

Note that $g(\lambda) = f(1 + \lambda t)$ is positive and convex for every $t \in \mathbb{R}$ and hence $\frac{1}{\lambda} g(\lambda)$ is monotonically decreasing to $g'(0) = 0$ as $\lambda \searrow 0$. Since for $\lambda = 1$ the integrand is assumed to be Q -integrable, the dominated convergence theorem applies and we get the result. \square

Theorem 7.18. Let f be twice continuously differentiable on $(0, \infty)$ with

$$\limsup_{x \rightarrow +\infty} f''(x) < \infty.$$

If $\chi^2(P\|Q) < \infty$, then $D_f(\bar{\lambda} Q + \lambda P \| Q) < \infty$ for all $0 \leq \lambda < 1$ and

$$\lim_{\lambda \rightarrow 0} \frac{1}{\lambda^2} D_f(\bar{\lambda} Q + \lambda P \| Q) = \frac{f''(1)}{2} \chi^2(P\|Q). \quad (7.54)$$

If $\chi^2(P\|Q) = \infty$ and $f''(1) > 0$ then (7.54) also holds, i.e. $D_f(\bar{\lambda} Q + \lambda P \| Q) = \omega(\lambda^2)$.

Remark 7.11. Conditions of the theorem include D , D_{SKL} , H^2 , JS, LC and all Rényi-type divergences, with $f(x) = \frac{1}{p-1}(x^p - 1)$, of orders $p < 2$. A similar result holds also for the case when $f''(x) \rightarrow \infty$ with $x \rightarrow +\infty$ (e.g. Rényi-type divergences with $p > 2$), but then we need to make extra assumptions in order to guarantee applicability of the dominated convergence theorem (often just the finiteness of $D_f(P\|Q)$ is sufficient).

Proof. Assuming that $\chi^2(P\|Q) < \infty$ we must have $P \ll Q$ and hence we can use (7.1) as the definition of D_f . Note that under (7.1) without loss of generality we may assume $f'(1) = f(1) = 0$ (indeed, for that we can just add a multiple of $(x - 1)$ to $f(x)$, which does not change the value of $D_f(P\|Q)$). From the Taylor expansion we have then

$$f(1 + u) = u^2 \int_0^1 (1 - t) f''(1 + tu) dt.$$

7.11 f -divergences in parametric families: Fisher information 111

Applying this with $u = \lambda \frac{P-Q}{Q}$ we get

$$D_f(\bar{\lambda}Q + \lambda P\|Q) = \int dQ \int_0^1 dt (1-t) \lambda^2 \left(\frac{P-Q}{Q} \right)^2 f'' \left(1 + t\lambda \frac{P-Q}{Q} \right). \quad (7.55)$$

Note that for any $\epsilon > 0$ we have $\sup_{x \geq \epsilon} |f'(x)| \triangleq C_\epsilon < \infty$. Note that $\frac{P-Q}{Q} \geq -1$ and, thus, for every λ the integrand is non-negative and bounded by

$$\left(\frac{P-Q}{Q} \right)^2 C_{1-\lambda} \quad (7.56)$$

which is integrable over $dQ \times \text{Leb}[0, 1]$ (by finiteness of $\chi^2(P\|Q)$ and Fubini, which applies due to non-negativity). Thus, $D_f(\bar{\lambda}Q + \lambda P\|Q) < \infty$. Dividing (7.55) by λ^2 we see that the integrand is dominated by (7.56) and hence we can apply the dominated convergence theorem to conclude

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \frac{1}{\lambda^2} D_f(\bar{\lambda}Q + \lambda P\|Q) &\stackrel{(a)}{=} \int_0^1 dt (1-t) \int dQ \left(\frac{P-Q}{Q} \right)^2 \lim_{\lambda \rightarrow 0} f'' \left(1 + t\lambda \frac{P-Q}{Q} \right) \\ &= \int_0^1 dt (1-t) \int dQ \left(\frac{P-Q}{Q} \right)^2 f''(1) = \frac{f''(1)}{2} \chi^2(P\|Q), \end{aligned}$$

which proves (7.54).

We proceed to proving that $D_f(\lambda P + \bar{\lambda}Q\|Q) = \omega(\lambda^2)$ when $\chi^2(P\|Q) = \infty$. If $P \ll Q$ then this follows by replacing the equality in (a) with \geq due to Fatou lemma. If $P \not\ll Q$, we consider decomposition $P = \mu P_1 + \bar{\mu}P_0$ with $P_1 \ll Q$ and $P_0 \perp Q$. From definition (7.2) we have (for $\lambda_1 = \frac{\lambda\mu}{1-\lambda\bar{\mu}}$)

$$D_f(\lambda P + \bar{\lambda}Q\|Q) = (1 - \lambda\bar{\mu})D_f(\lambda_1 P_1 + \bar{\lambda}_1 Q\|Q) + \lambda\bar{\mu}D_f(P_0\|Q) \geq \lambda\bar{\mu}D_f(P_0\|Q).$$

Recall from Proposition 7.5 that $D_f(P_0\|Q) > 0$ unless $f(x) = c(x-1)$ for some constant c and the proof is complete. \square

7.11 f -divergences in parametric families: Fisher information

In Section 2.6.2* we have already previewed the fact that in parametric families of distributions, the Hessian of the KL divergence turns out to coincide with the Fisher information. Here we collect such facts and their proofs. These materials form the basis of sharp bounds on parameter estimation that we will study later in Chapter 29.

To start with an example, let us return to the Gaussian location family $P_t \triangleq \mathcal{N}(t, 1), t \in \mathbb{R}$. From the identities presented in Section 7.7 we obtain the following asymptotics:

$$\begin{aligned} \text{TV}(P_t, P_0) &= \frac{|t|}{\sqrt{2\pi}} + o(|t|), & H^2(P_t, P_0) &= \frac{t^2}{4} + o(t^2), \\ \chi^2(P_t\|P_0) &= t^2 + o(t^2), & D(P_t\|P_0) &= \frac{t^2}{2\log e} + o(t^2), \\ \text{LC}(P_t, P_0) &= \frac{1}{4}t^2 + o(t^2). \end{aligned}$$

We can see that with the exception of TV, other f -divergences behave quadratically under small displacement $t \rightarrow 0$. This turns out to be a general fact, and furthermore the coefficient in front of t^2 is given by the Fisher information (at $t = 0$). To proceed carefully, we need some technical assumptions on the family P_t .

Definition 7.19 (Regular single-parameter families). Fix $\tau > 0$, space \mathcal{X} and a family P_t of distributions on \mathcal{X} , $t \in [0, \tau]$. We define the following types of conditions that we call regularity at $t = 0$:

- (a) $P_t(dx) = p_t(x)\mu(dx)$, for some measurable $(t, x) \mapsto p_t(x) \in \mathbb{R}_+$ and a fixed measure μ on \mathcal{X} ;
- (b₀) There exists a measurable function $(s, x) \mapsto \dot{p}_s(x)$, $s \in [0, \tau)$, $x \in \mathcal{X}$, such that for μ -almost every x_0 we have $\int_0^\tau |\dot{p}_s(x_0)|ds < \infty$ and

$$p_t(x_0) = p_0(x_0) + \int_0^t \dot{p}_s(x_0)ds. \quad (7.57)$$

Furthermore, for μ -almost every x_0 we have $\lim_{t \searrow 0} \dot{p}_t(x_0) = \dot{p}_0(x_0)$.

- (b₁) We have $\dot{p}_t(x) = 0$ whenever $p_0(x) = 0$ and, furthermore,

$$\int_{\mathcal{X}} \mu(dx) \sup_{0 \leq t < \tau} \frac{(\dot{p}_t(x))^2}{p_0(x)} < \infty. \quad (7.58)$$

- (c₀) There exists a measurable function $(s, x) \mapsto \dot{h}_s(x)$, $s \in [0, \tau)$, $x \in \mathcal{X}$, such that for μ -almost every x_0 we have $\int_0^\tau |\dot{h}_s(x_0)|ds < \infty$ and

$$h_t(x_0) \triangleq \sqrt{p_t(x_0)} = \sqrt{p_0(x_0)} + \int_0^t \dot{h}_s(x_0)ds. \quad (7.59)$$

Furthermore, for μ -almost every x_0 we have $\lim_{t \searrow 0} \dot{h}_t(x_0) = \dot{h}_0(x_0)$.

- (c₁) The family of functions $\{(\dot{h}_t(x))^2 : t \in [0, \tau)\}$ is uniformly μ -integrable.

Remark 7.12. Recall that the uniform integrability condition (c₁) is implied by the following stronger (but easier to verify) condition:

$$\int_{\mathcal{X}} \mu(dx) \sup_{0 \leq t < \tau} (\dot{h}_t(x))^2 < \infty. \quad (7.60)$$

Impressively, if one also assumes the continuous differentiability of h_t then the uniform integrability condition becomes equivalent to the continuity of the Fisher information

$$t \mapsto J_F(t) \triangleq 4 \int \mu(dx) (\dot{h}_t(x))^2. \quad (7.61)$$

We refer to [43, Appendix V] for this finesse.

Theorem 7.20. *Let the family of distributions $\{P_t : t \in [0, \tau)\}$ satisfy the conditions (a), (b₀) and (b₁) in Definition 7.29. Then we have*

$$\chi^2(P_t \| P_0) = J_F(0)t^2 + o(t^2), \quad (7.62)$$

7.11 f -divergences in parametric families: Fisher information 113

$$D(P_t \| P_0) = \frac{\log e}{2} J_F(0) t^2 + o(t^2), \quad (7.63)$$

where $J_F(0) \triangleq \int_{\mathcal{X}} \mu(dx) \frac{(\dot{p}_0(x))^2}{p_0(x)} < \infty$ is the Fisher information at $t = 0$.

Proof. From assumption (b₁) we see that for any x_0 with $p_0(x_0) = 0$ we must have $\dot{p}_t(x_0) = 0$ and thus $p_t(x_0) = 0$ for all $t \in [0, \tau]$. Hence, we may restrict all integrals below to subset $\{x : p_0(x) > 0\}$, on which the ratio $\frac{(p_t(x) - p_0(x))^2}{p_0(x)}$ is well-defined. Consequently, we have by (7.57)

$$\begin{aligned} \frac{1}{t^2} \chi^2(P_t \| P_0) &= \frac{1}{t^2} \int \mu(dx) \frac{(p_t(x) - p_0(x))^2}{p_0(x)} \\ &= \frac{1}{t^2} \int \mu(dx) \frac{1}{p_0(x)} \left(t \int_0^1 du \dot{p}_{tu}(x) \right)^2 \\ &\stackrel{(a)}{=} \int \mu(dx) \int_0^1 du_1 \int_0^1 du_2 \frac{\dot{p}_{tu_1}(x) \dot{p}_{tu_2}(x)}{p_0(x)} \end{aligned}$$

Note that by the continuity assumption in (b₁) we have $\dot{p}_{tu_1}(x) \dot{p}_{tu_2}(x) \rightarrow \dot{p}_0^2(x)$ for every (u_1, u_2, x) as $t \rightarrow 0$. Furthermore, we also have $\left| \frac{\dot{p}_{tu_1}(x) \dot{p}_{tu_2}(x)}{p_0(x)} \right| \leq \sup_{0 \leq t < \tau} \frac{(\dot{p}_t(x))^2}{p_0(x)}$, which is integrable by (7.58). Consequently, application of the dominated convergence theorem to the integral in (a) concludes the proof of (7.62).

We next show that for any f -divergence with twice continuously differentiable f (and in fact, without assuming (7.58)) we have:

$$\liminf_{t \rightarrow 0} \frac{1}{t^2} D_f(P_t \| P_0) \geq \frac{f''(1)}{2} J_F(0). \quad (7.64)$$

Indeed, similar to (7.55) we get

$$D_f(P_t \| P_0) = \int_0^1 dz (1-z) \mathbb{E}_{X \sim P_0} \left[f' \left(1 + z \frac{p_t(X) - p_0(X)}{p_0(X)} \right) \left(\frac{p_t(X) - p_0(X)}{p_0(X)} \right)^2 \right]. \quad (7.65)$$

Dividing by t^2 notice that from (b₀) we have $\frac{p_t(X) - p_0(X)}{tp_0(X)} \xrightarrow{\text{a.s.}} \frac{\dot{p}_0(X)}{p_0(X)}$ and thus

$$f' \left(1 + z \frac{p_t(X) - p_0(X)}{p_0(X)} \right) \left(\frac{p_t(X) - p_0(X)}{tp_0(X)} \right)^2 \rightarrow f'(1) \left(\frac{\dot{p}_0(X)}{p_0(X)} \right)^2.$$

Thus, applying Fatou's lemma we recover (7.64).

Next, plugging $f(x) = x \log x$ in (7.65) we obtain for the KL divergence

$$\frac{1}{t^2} D(P_t \| P_0) = (\log e) \int_0^1 dz \mathbb{E}_{X \sim P_0} \left[\frac{1-z}{1+z \frac{p_t(X) - p_0(X)}{p_0(X)}} \left(\frac{p_t(X) - p_0(X)}{tp_0(X)} \right)^2 \right]. \quad (7.66)$$

The first fraction inside the bracket is between 0 and 1 and the second by $\sup_{0 < t < \tau} \left(\frac{\dot{p}_t(X)}{p_0(X)} \right)^2$, which is P_0 -integrable by (b₁). Thus, dominated convergence theorem applies to the double integral

in (7.65) and we obtain

$$\lim_{t \rightarrow 0} \frac{1}{t^2} D(P_t \| P_0) = (\log e) \int_0^1 dz \mathbb{E}_{X \sim P_0} \left[(1-z) \left(\frac{\dot{p}_0(X)}{p_0(X)} \right)^2 \right],$$

completing the proof of (7.63). \square

Remark 7.13. Theorem 7.31 extends to the case of multi-dimensional parameters as follows. Define the Fisher information matrix at $\theta \in \mathbb{R}^d$:

$$J_F(\theta) \triangleq \int \mu(dx) \nabla_\theta \sqrt{p_\theta(x)} \nabla_\theta \sqrt{p_\theta(x)}^\top \quad (7.67)$$

Then (7.62) becomes $\chi^2(P_t \| P_0) = t^\top J_F(0)t + o(\|t\|^2)$ as $t \rightarrow 0$ and similarly for (7.63), which has previously appeared in (2.33).

Theorem 7.31 applies to many cases (e.g. to smooth subfamilies of exponential families, for which one can take $\mu = P_0$ and $p_0(x) \equiv 1$), but it is not sufficiently general. To demonstrate the issue, consider the following example.

Example 7.1 (Location families with compact support). We say that family P_t is a (scalar) location family if $\mathcal{X} = \mathbb{R}$, $\mu = \text{Leb}$ and $p_t(x) = p_0(x - t)$. Consider the following example, for $\alpha > -1$:

$$p_0(x) = C_\alpha \times \begin{cases} x^\alpha, & x \in [0, 1], \\ (2-x)^\alpha, & x \in [1, 2], \\ 0, & \text{otherwise} \end{cases},$$

with C_α chosen from normalization. Clearly, here condition (7.58) is not satisfied and both $\chi^2(P_t \| P_0)$ and $D(P_t \| P_0)$ are infinite for $t > 0$, since $P_t \not\ll P_0$. But $J_F(0) < \infty$ whenever $\alpha > 1$ and thus one expects that a certain remedy should be possible. Indeed, one can compute those f -divergences that are finite for $P_t \not\ll P_0$ and find that for $\alpha > 1$ they are quadratic in t . As an illustration, we have

$$H^2(P_t, P_0) = \begin{cases} \Theta(t^{1+\alpha}), & 0 \leq \alpha < 1 \\ \Theta(t^2 \log \frac{1}{t}), & \alpha = 1 \\ \Theta(t^2), & \alpha > 1 \end{cases} \quad (7.68)$$

as $t \rightarrow 0$. This can be computed directly, or from a more general results of [158, Theorem VI.1.1].⁵

⁵ Statistical significance of this calculation is that if we were to estimate the location parameter t from n iid samples, then precision δ_n^* of the optimal estimator up to constant factors is given by solving $H^2(P_{\delta_n^*}, P_0) \asymp \frac{1}{n}$, cf. [158, Chapter VI]. For $\alpha < 1$ we have $\delta_n^* \asymp n^{-\frac{1}{1+\alpha}}$ which is notably better than the empirical mean estimator (attaining precision of only $n^{-\frac{1}{2}}$). For $\alpha = 1/2$ this fact was noted by D. Bernoulli in 1777 as a consequence of his (newly proposed) maximum likelihood estimation.

7.11 f -divergences in parametric families: Fisher information 115

The previous example suggests that quadratic behavior as $t \rightarrow 0$ can hold even when $P_t \not\ll P_0$, which is the case handled by the next (more technical) result, whose proof we placed in Section 7.14*). One can verify that condition (c_1) is indeed satisfied for all $\alpha > 1$ in Example 7.1, thus establishing the quadratic behavior. Also note that the stronger (7.60) only applies to $\alpha \geq 2$.

Theorem 7.21. *Given a family of distributions $\{P_t : t \in [0, \tau)\}$ satisfying the conditions (a), c_0 and (c_1) of Definition 7.29, we have*

$$\chi^2(P_t \parallel \bar{\epsilon}P_0 + \epsilon P_t) = t^2 \bar{\epsilon}^2 \left(J_F(0) + \frac{1 - 4\epsilon}{\epsilon} J^\#(0) \right) + o(t^2), \quad \forall \epsilon \in (0, 1) \quad (7.69)$$

$$H^2(P_t, P_0) = \frac{t^2}{4} J_F(0) + o(t^2), \quad (7.70)$$

where $J_F(0) = 4 \int \dot{h}_0^2 d\mu < \infty$ is the Fisher information and $J^\#(0) = \int \dot{h}_0^2 1_{\{h_0=0\}} d\mu$ can be called the Fisher defect at $t = 0$.

Example 7.2 (On Fisher defect). Note that in most cases of interest we will have the situation that $t \mapsto h_t(x)$ is actually differentiable for all t in some *two-sided* neighborhood $(-\tau, \tau)$ of 0. In such cases, $h_0(x) = 0$ implies that $t = 0$ is a local minima and thus $\dot{h}_0(x) = 0$, implying that the defect $J_F^\# = 0$. However, for other families this will not be so, sometimes even when $p_t(x)$ is smooth on $t \in (-\tau, \tau)$ (but not h_t). Here is such an example.

Consider $P_t = \text{Ber}(t^2)$. A straightforward calculation shows:

$$\chi^2(P_t \parallel \bar{\epsilon}P_0 + \epsilon P_t) = t^2 \frac{\bar{\epsilon}^2}{\epsilon} + O(t^4), \quad H^2(P_t, P_0) = 2(1 - \sqrt{1 - t^2}) = t^2 + O(t^4).$$

Taking $\mu(\{0\}) = \mu(\{1\}) = 1$ to be the counting measure, we get the following

$$h_t(x) = \begin{cases} \sqrt{1 - t^2}, & x = 0 \\ |t|, & x = 1 \end{cases}, \quad \dot{h}_t(x) = \begin{cases} \frac{-t}{\sqrt{1-t^2}}, & x = 0 \\ \text{sign}(t), & x = 1, t \neq 0 \\ 1, & x = 1, t = 0 \end{cases} \quad (\text{just as an agreement}).$$

Note that if we view P_t as a family on $t \in [0, \tau)$ for small τ , then all conditions (a), c_0 and (c_1) are clearly satisfied (\dot{h}_t is bounded on $t \in (-\tau, \tau)$). We have $J_F(0) = 4$ and $J^\#(0) = 1$ and thus (7.69) recovers the correct expansion for χ^2 and (7.70) for H^2 .

Notice that the non-smoothness of h_t only becomes visible if we extend the domain to $t \in (-\tau, \tau)$. In fact, this issue is not seen in terms of densities p_t . Indeed, let us compute the density p_t and its derivative \dot{p}_t explicitly too:

$$p_t(x) = \begin{cases} 1 - t^2, & x = 0 \\ t^2, & x = 1 \end{cases}, \quad \dot{p}_t(x) = \begin{cases} -2t, & x = 0 \\ 2t, & x = 1 \end{cases}.$$

Clearly, p_t is continuously differentiable on $t \in (-\tau, \tau)$. Furthermore, the following expectation (typically equal to $J_F(t)$ in (7.61))

$$\mathbb{E}_{X \sim P_t} \left[\left(\frac{\dot{p}_t(X)}{p_t(X)} \right)^2 \right] = \begin{cases} 0, & t = 0 \\ 4 + \frac{4t^2}{1-t^2}, & t \neq 0 \end{cases}$$

is discontinuous at $t = 0$. To make things worse, at $t = 0$ this expectation does not match our definition of the Fisher information $J_F(0)$ in Theorem 7.33, and thus does not yield the correct small- t behavior for either χ^2 or H^2 . In general, to avoid difficulties one should restrict to those families with $t \mapsto h_t(x)$ continuously differentiable in $t \in (-\tau, \tau)$.

7.12 Rényi divergences and tensorization

The following family of divergence measures introduced by Rényi is key in many applications involving product measures. Although these measures are not f -divergences, they are obtained as monotone transformation of an appropriate f -divergence and thus satisfy DPI and other properties of f -divergences. Later, Rényi divergence will feature prominently in characterizing the optimal error exponents in hypothesis testing (see Section 16.1 and especially Remark 16.2), in approximating of channel output statistic (see Section 25.4*), and in nonasymptotic bounds for composite hypothesis testing (see Section 32.2.1).

Definition 7.22. For any $\lambda \in \mathbb{R} \setminus 0, 1$ we define the Rényi divergence of order λ as

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^\lambda \right],$$

where $\mathbb{E}_Q[\cdot]$ is understood as an f -divergence $D_f(P\|Q)$ with $f(x) = x^\lambda$, see Definition 7.1. Conditional Rényi divergence is defined as

$$\begin{aligned} D_\lambda(P_{X|Y}\|Q_{X|Y}|P_Y) &\triangleq D_\lambda(P_Y \times P_{X|Y}\|P_Y \times Q_{X|Y}) \\ &= \frac{1}{\lambda - 1} \log \mathbb{E}_{Y \sim P_Y} \int_{\mathcal{X}} (dP_{X|Y}(x))^\lambda (dQ_{X|Y}(x))^{1-\lambda}. \end{aligned}$$

Numerous properties of Rényi divergences are known, see [310]. Here we only notice a few:

- Special cases of $\lambda = \frac{1}{2}, 1, 2$: Under mild regularity conditions $\lim_{\lambda \rightarrow 1} D_\lambda(P\|Q) = D(P\|Q)$. On the other hand, D_2 is a monotone transformation of χ^2 in (7.4), while $D_{\frac{1}{2}}$ is a monotone transformation of H^2 in (7.5).
- For all $\lambda \in \mathbb{R}$ the map $\lambda \rightarrow D_\lambda(P\|Q)$ is non-decreasing and the map $\lambda \rightarrow (1-\lambda)D_\lambda(P\|Q)$ is concave.
- For $\lambda \in [0, 1]$ the map $(P, Q) \mapsto D_\lambda(P\|Q)$ is convex.
- For $\lambda \geq 0$ the map $Q \mapsto D_\lambda(P\|Q)$ is convex.

7.12 Rényi divergences and tensorization 117

- There is a version of the chain rule:

$$D_\lambda(P_{A,B}||Q_{A,B}) = D_\lambda(P_B||Q_B) + D_\lambda(P_{A|B}||Q_{A|B}|P_B^{(\lambda)}), \quad (7.71)$$

where $P_B^{(\lambda)}$ is the λ -tilting of P_B towards Q_B given by

$$P_B^{(\lambda)}(b) \triangleq P_B^\lambda(b) Q_B^{1-\lambda}(b) \exp\{-(\lambda-1)D_\lambda(P_B||Q_B)\}. \quad (7.72)$$

- The key property is additivity under products, or *tensorization*:

$$D_\lambda\left(\prod_i P_{X_i} \middle\| \prod_i Q_{X_i}\right) = \sum_i D_\lambda(P_{X_i}||Q_{X_i}), \quad (7.73)$$

which is a simple consequence of (7.71). D_λ 's are the only divergences satisfying DPI and tensorization [217]. The most well-known special cases of (7.73) are for Hellinger distance, see (7.24) and for χ^2 :

$$1 + \chi^2\left(\prod_{i=1}^n P_i \middle\| \prod_{i=1}^n Q_i\right) = \prod_{i=1}^n (1 + \chi^2(P_i||Q_i)).$$

We can also obtain additive bounds for non-product distributions, see Ex. I.32 and I.33.

The following consequence of the chain rule will be crucial in statistical applications later (see Section 32.2, in particular, Theorem 32.8).

Proposition 7.23. Consider product channels $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$ and $Q_{Y^n|X^n} = \prod Q_{Y_i|X_i}$. We have (with all optimizations over all possible distributions)

$$\inf_{P_{X^n}, Q_{X^n}} D_\lambda(P_{Y^n}||Q_{Y^n}) = \sum_{i=1}^n \inf_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i}||Q_{Y_i}) \quad (7.74)$$

$$\sup_{P_{X^n}, Q_{X^n}} D_\lambda(P_{Y^n}||Q_{Y^n}) = \sum_{i=1}^n \sup_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i}||Q_{Y_i}) = \sum_{i=1}^n \sup_{x, x'} D_\lambda(P_{Y_i|X_i=x}||Q_{Y_i|X_i=x'}) \quad (7.75)$$

In particular, for any collections of distributions $\{P_\theta, \theta \in \Theta\}$ and $\{Q_\theta, \theta \in \Theta\}$:

$$\inf_{P \in \text{co}\{P_\theta^{\otimes n}\}, Q \in \text{co}\{Q_\theta^{\otimes n}\}} D_\lambda(P||Q) \geq n \inf_{P \in \text{co}\{P_\theta\}, Q \in \text{co}\{Q_\theta\}} D_\lambda(P||Q) \quad (7.76)$$

$$\sup_{P \in \text{co}\{P_\theta^{\otimes n}\}, Q \in \text{co}\{Q_\theta^{\otimes n}\}} D_\lambda(P||Q) \leq n \sup_{P \in \text{co}\{P_\theta\}, Q \in \text{co}\{Q_\theta\}} D_\lambda(P||Q) \quad (7.77)$$

Remark 7.14. The mnemonic for (7.76)-(7.77) is that “mixtures of products are less distinguishable than products of mixtures”. The former arise in statistical settings where iid observations are drawn from a single distribution whose parameter is drawn from a prior.

Proof. The second equality in (7.75) follows from the fact that D_λ is an increasing function of an f -divergence, and thus maximization should be attained at an extreme point of the space of probabilities, which are just the single-point masses. The main equalities (7.74)-(7.75) follow from a) restricting optimizations to product distributions and invoking (7.73); and b) the chain rule

for D_λ . For example for $n = 2$, we fix P_{X^2} and Q_{X^2} , which (via channels) induce joint distributions P_{X^2, Y^2} and Q_{X^2, Y^2} . Then we have

$$D_\lambda(P_{Y_1|Y_2=y} \| Q_{Y_1|Y_2=y'}) \geq \inf_{\tilde{P}_{X_1}, \tilde{Q}_{X_1}} D_\lambda(\tilde{P}_{Y_1} \| \tilde{Q}_{Y_1}),$$

since $P_{Y_1|Y_2=y}$ is a distribution induced by taking $\tilde{P}_{X_1} = P_{X_1|Y_2=y}$, and similarly for $Q_{Y_1|Y_2=y'}$. In all, we get

$$D_\lambda(P_{Y^2} \| Q_{Y^2}) = D_\lambda(P_{Y_2} \| Q_{Y_2}) + D_\lambda(P_{Y_1|Y_2} \| Q_{Y_1|Y_2} | P_{Y_2}^{(\lambda)}) \geq \sum_{i=1}^2 \inf_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i} \| Q_{Y_i}),$$

as claimed. The case of sup is handled similarly.

From (7.74)-(7.75), we get (7.76)-(7.77) by taking $\mathcal{X} = \Theta$ and specializing inf, sup to diagonal distributions P_{X^n} and Q_{X^n} , i.e. those with the property that $\mathbb{P}[X_1 = \dots = X_n] = 1$ and $\mathbb{Q}[X_1 = \dots = X_n] = 1$. \square

7.13 Variational representation of f -divergences

In Theorem 4.8 we had a very useful variational representation of KL-divergence due to Donsker and Varadhan. In this section we show how to derive such representations for other f -divergences in a principled way. The proofs are slightly technical and given in Section 7.14* at the end of this chapter.

Let $f : (0, +\infty) \rightarrow \mathbb{R}$ be a convex function. The convex conjugate $f^* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}_+} xy - f(x), \quad y \in \mathbb{R}. \quad (7.78)$$

Denote the domain of f^* by $\text{dom}(f^*) \triangleq \{y : f^*(y) < \infty\}$. Two important properties of the convex conjugates are

- 1 f^* is also convex (which holds regardless of f being convex or not);
- 2 Biconjugation: $(f^*)^* = f$, which means

$$f(x) = \sup_y xy - f^*(y)$$

and implies the following (for all $x > 0$ and y)

$$f(x) + f^*(g) \geq xy.$$

Similarly, we can define a convex conjugate for any convex functional $\Psi(P)$ defined on the space of measures, by setting

$$\Psi^*(g) = \sup_P \int g dP - \Psi(P). \quad (7.79)$$

7.13 Variational representation of f -divergences 119

Under appropriate conditions (e.g. finite \mathcal{X}), biconjugation then yields the sought-after variational representation

$$\Psi(P) = \sup_g \int g dP - \Psi^*(g). \quad (7.80)$$

Next we will now compute these conjugates for $\Psi(P) = D_f(P\|Q)$. It turns out to be convenient to first extend the definition of $D_f(P\|Q)$ to all finite signed measures P then compute the conjugate. To this end, let $f_{\text{ext}} : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ be an extension of f , such that $f_{\text{ext}}(x) = f(x)$ for $x \geq 0$ and f_{ext} is convex on \mathbb{R} . In general, we can always choose $f_{\text{ext}}(x) = \infty$ for all $x < 0$. In special cases e.g. $f(x) = |x - 1|/2$ or $f(x) = (x - 1)^2$ we can directly take $f_{\text{ext}}(x) = f(x)$ for all x . Now we can define $D_f(P\|Q)$ for all signed measure measures P in the same way as in Definition 7.1 using f_{ext} in place of f .

For each choice of f_{ext} we have a variational representation of f -divergence:

Theorem 7.24. *Let P and Q be probability measures on \mathcal{X} . Fix an extension f_{ext} of f and let f_{ext}^* is the conjugate of f_{ext} , i.e., $f_{\text{ext}}^*(y) = \sup_{x \in \mathbb{R}} xy - f_{\text{ext}}(x)$. Denote $\text{dom}(f_{\text{ext}}^*) \triangleq \{y : f_{\text{ext}}^*(y) < \infty\}$. Then*

$$D_f(P\|Q) = \sup_{g: \mathcal{X} \rightarrow \text{dom}(f_{\text{ext}}^*)} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f_{\text{ext}}^*(g(X))]. \quad (7.81)$$

where the supremum can be taken over either (a) all simple g or (b) over all g satisfying $\mathbb{E}_Q[f_{\text{ext}}^*(g(X))] < \infty$.

We remark that when $P \ll Q$ then both results (a) and (b) also hold for supremum over $g : \mathcal{X} \rightarrow \mathbb{R}$, i.e. without restricting $g(x) \in \text{dom}(f_{\text{ext}}^*)$.

As a consequence of the variational characterization, we get the following properties for f -divergences:

- 1 *Convexity:* First of all, note that $D_f(P\|Q)$ is expressed as a supremum of affine functions (since the expectation is a linear operation). As a result, we get that $(P, Q) \mapsto D_f(P\|Q)$ is convex, which was proved previously in Theorem 7.8 using different method.
- 2 *Weak lower semicontinuity:* Recall the example in Remark 4.11, where $\{X_i\}$ are i.i.d. Rademachers (± 1), and

$$\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

by the central limit theorem; however, by Proposition 7.5, for all n ,

$$D_f\left(\frac{P_{X_1+X_2+\dots+X_n}}{\sqrt{n}} \middle\| \mathcal{N}(0, 1)\right) = f(0) + f'(\infty) > 0,$$

since the former distribution is discrete and the latter is continuous. Therefore similar to the KL divergence, the best we can hope for f -divergence is semicontinuity. Indeed, if \mathcal{X} is a nice space (e.g., Euclidean space), in (7.81) we can restrict the function g to continuous bounded functions, in which case $D_f(P\|Q)$ is expressed as a supremum of weakly continuous functionals

(note that $f^* \circ g$ is also continuous and bounded since f^* is continuous) and is hence weakly lower semicontinuous, i.e., for any sequence of distributions P_n and Q_n such that $P_n \xrightarrow{w} P$ and $Q_n \xrightarrow{w} Q$, we have

$$\liminf_{n \rightarrow \infty} D_f(P_n \| Q_n) \geq D_f(P \| Q).$$

3 Relation to DPI: As discussed in (4.14) variational representations can be thought of as extensions of the DPI. As an exercise, one should try to derive the estimate

$$|P[A] - Q[A]| \leq \sqrt{Q[A] \cdot \chi^2(P \| Q)}$$

via both the DPI and (7.85).

Example 7.3 (Total variation and Hellinger). For total variation, we have $f(x) = \frac{1}{2}|x - 1|$. Consider the extension $f_{\text{ext}}(x) = \frac{1}{2}|x - 1|$ for $x \in \mathbb{R}$. Then

$$f_{\text{ext}}^*(y) = \sup_x \left\{ xy - \frac{1}{2}|x - 1| \right\} = \begin{cases} +\infty & \text{if } |y| > \frac{1}{2} \\ y & \text{if } |y| \leq \frac{1}{2} \end{cases}.$$

Thus (7.81) gives

$$\text{TV}(P, Q) = \sup_{g: |g| \leq \frac{1}{2}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)], \quad (7.82)$$

which previously appeared in (7.16). A similar calculation for Hellinger-squared yields (after changing from g to $h = 1 - g$ in (7.81)):

$$H^2(P, Q) = 2 - \inf_{h>0} \mathbb{E}_P[h] + \mathbb{E}_Q[\frac{1}{h}].$$

Example 7.4 (χ^2 -divergence). For χ^2 -divergence we have $f(x) = (x - 1)^2$. Take $f_{\text{ext}}(x) = (x - 1)^2$, whose conjugate is $f_{\text{ext}}^*(y) = y + \frac{y^2}{4}$. Applying (7.81) yields

$$\chi^2(P \| Q) = \sup_{g: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q \left[g(X) + \frac{g^2(X)}{4} \right] \quad (7.83)$$

$$= \sup_{g: \mathcal{X} \rightarrow \mathbb{R}} 2\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g^2(X)] - 1 \quad (7.84)$$

where the last step follows from a change of variable ($g \leftarrow \frac{1}{2}g - 1$).

To get another equivalent, but much more memorable representation, we notice that (7.84) it is not scale-invariant. To make it so, setting $g = \lambda h$ and optimizing over the $\lambda \in \mathbb{R}$ first we get

$$\chi^2(P \| Q) = \sup_{h: \mathcal{X} \rightarrow \mathbb{R}} \frac{(\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)])^2}{\text{Var}_Q(h(X))}. \quad (7.85)$$

The statistical interpretation of (7.85) is as follows: if a test statistic $h(X)$ is such that the separation between its expectation under P and Q far exceeds its standard deviation, then this suggests the two hypothesis can be distinguished reliably. The representation (7.85) will turn out useful in statistical applications in Chapter 29 for deriving the Hammersley-Chapman-Robbins (HCR) lower bound as well as its Bayesian version, see Section 29.1.2, and ultimately the Cramér-Rao and van Trees lower bounds.

7.13 Variational representation of f -divergences 121

Example 7.5 (KL-divergence). In this case we have $f(x) = x \log x$. Consider the extension $f_{\text{ext}}(x) = \infty$ for $x < 0$, whose convex conjugate is $f^*(y) = \frac{\log e}{e} \exp(y)$. Hence (7.81) yields

$$D(P\|Q) = \sup_{g:\mathcal{X}\rightarrow\mathbb{R}} \mathbb{E}_P[g(X)] - (\mathbb{E}_Q[\exp\{g(X)\}] - 1)\log e \quad (7.86)$$

Note that in the last example, the variational representation (7.86) we obtained for the KL divergence is not the same as the Donsker-Varadhan identity in Theorem 4.8, that is,

$$D(P\|Q) = \sup_{g:\mathcal{X}\rightarrow\mathbb{R}} \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[\exp\{g(X)\}]. \quad (7.87)$$

In fact, (7.86) is weaker than (7.87) in the sense that for each choice of g , the obtained lower bound on $D(P\|Q)$ in the RHS is smaller. Furthermore, regardless of the choice of f_{ext} , the Donsker-Varadhan representation can never be obtained from Theorem 7.37 because, unlike (7.87), the second term in (7.81) is always linear in Q . It turns out if we define $D_f(P\|Q) = \infty$ for all non-probability measure P , and compute its convex conjugate, we obtain in the next theorem a different type of variational representation, which, specialized to KL divergence in Example 7.5, recovers exactly the Donsker-Varadhan identity.

Theorem 7.25. *Consider the extension f_{ext} of f such that $f_{\text{ext}}(x) = \infty$ for $x < 0$. Let $S = \{x : q(x) > 0\}$ where q is as in (7.2). Then*

$$D_f(P\|Q) = f'(\infty)P[S^c] + \sup_g \mathbb{E}_P[g1_S] - \Psi_{Q,P}^*(g), \quad (7.88)$$

where

$$\Psi_{Q,P}^*(g) \triangleq \inf_{a \in \mathbb{R}} \mathbb{E}_Q[f_{\text{ext}}^*(g(X) - a)] + aP[S].$$

In the special case $f'(\infty) = \infty$, we have

$$D_f(P\|Q) = \sup_g \mathbb{E}_P[g] - \Psi_Q^*(g), \quad \Psi_Q^*(g) \triangleq \inf_{a \in \mathbb{R}} \mathbb{E}_Q[f_{\text{ext}}^*(g(X) - a)] + a. \quad (7.89)$$

Remark 7.15 (Marton's divergence). Recall that in Theorem 7.12 we have shown both the sup and inf characterizations for the TV. Do other f -divergences also possess inf characterizations? The only other known example (to us) is due to Marton. Let

$$D_m(P\|Q) = \int dQ \left(1 - \frac{dP}{dQ} \right)_+^2,$$

which is clearly an f -divergence with $f(x) = (1 - x)_+^2$. We have the following [44, Lemma 8.3]:

$$D_m(P\|Q) = \inf \{ \mathbb{E}[P[X \neq Y|Y]^2] : X \sim P, Y \sim Q \},$$

where the infimum is over all couplings of P and Q . See Ex. I.34.

Marton's D_m divergence plays a crucial role in the theory of concentration of measure [44, Chapter 8]. Note also that while Theorem 7.27 does not apply to D_m , due to the absence of twice continuous differentiability, it does apply to the symmetrized Marton divergence $D_{sm}(P\|Q) \triangleq D_m(P\|Q) + D_m(Q\|P)$.

7.14* Technical proofs: convexity, local expansions and variational representations

In this section we collect proofs of some technical theorems from this chapter.

Proof of Theorem 7.33. By definition we have

$$L(t) \triangleq \frac{1}{\bar{\epsilon}^2 t^2} \chi^2(P_t \| \bar{\epsilon}P_0 + \epsilon P_t) = \frac{1}{t^2} \int_{\mathcal{X}} \mu(dx) \frac{(p_t(x) - p_0(x))^2}{\bar{\epsilon}p_0(x) + \epsilon p_t(x)} = \frac{1}{t^2} \int \mu(dx) g(t, x)^2, \quad (7.90)$$

where

$$g(t, x) \triangleq \frac{p_t(x) - p_0(x)}{\sqrt{\bar{\epsilon}p_0(x) + \epsilon p_t(x)}} = \phi(h_t(x); x), \quad \phi(h; x) \triangleq \frac{h^2 - p_0(x)}{\sqrt{\bar{\epsilon}p_0(x) + \epsilon h^2}}.$$

By c_0) the function $t \mapsto h_t(x) \triangleq \sqrt{p_t(x)}$ is absolutely continuous (for μ -a.e. x). Below we will show that $\|\phi(\cdot; x)\|_{\text{Lip}} = \sup_{h \geq 0} |\phi'(h; x)| \leq \frac{2-\epsilon}{(1-\epsilon)\sqrt{\epsilon}}$. This implies that $t \mapsto g(t, x)$ is also absolutely continuous and hence differentiable almost everywhere. Consequently, we have

$$g(t, x) = t \int_0^1 du \dot{g}(tu, x), \quad \dot{g}(t, x) \triangleq \phi'(h_t(x); x) \dot{h}_t(x),$$

Since $\phi'(\cdot; x)$ is continuous with

$$\phi'(h_0(x); x) = \begin{cases} 2, & x : h_0(x) > 0, \\ \frac{1}{\sqrt{\epsilon}}, & x : h_0(x) = 0 \end{cases} \quad (7.91)$$

(we verify these facts below too), we conclude that

$$\lim_{s \rightarrow 0} \dot{g}(s, x) = \dot{g}(0, x) = \dot{h}_0(x) \left(2 \cdot 1\{h_0(x) > 0\} + \frac{1}{\sqrt{\epsilon}} 1\{h_0(x) = 0\} \right), \quad (7.92)$$

where we also used continuity $\dot{h}_t(x) \rightarrow \dot{h}_0(x)$ by assumption c_0).

Substituting the integral expression for $g(t, x)$ into (7.90) we obtain

$$L(t) = \int \mu(dx) \int_0^1 du_1 \int_0^1 du_2 \dot{g}(tu_1, x) \dot{g}(tu_2, x). \quad (7.93)$$

Since $|\dot{g}(s, x)| \leq C|h_s(x)|$ for some $C = C(\epsilon)$, we have from Cauchy-Schwarz

$$\int \mu(dx) |\dot{g}(s_1, x) \dot{g}(s_2, x)| \leq C^2 \sup_t \int_{\mathcal{X}} \mu(dx) \dot{h}_t(x)^2 < \infty. \quad (7.94)$$

where the last inequality follows from the uniform integrability assumption (c_1). This implies that Fubini's theorem applies in (7.93) and we obtain

$$L(t) = \int_0^1 du_1 \int_0^1 du_2 G(tu_1, tu_2), \quad G(s_1, s_2) \triangleq \int \mu(dx) \dot{g}(s_1, x) \dot{g}(s_2, x).$$

Notice that if a family of functions $\{f_{\alpha}(x) : \alpha \in I\}$ is uniformly square-integrable, then the family $\{f_{\alpha}(x)f_{\beta}(x) : \alpha \in I, \beta \in I\}$ is uniformly integrable simply because apply $|f_{\alpha}f_{\beta}| \leq \frac{1}{2}(f_{\alpha}^2 + f_{\beta}^2)$.

7.14* Technical proofs: convexity, local expansions and variational representations 123

Consequently, from the assumption (c_1) we see that the integral defining $G(s_1, s_2)$ allows passing the limit over s_1, s_2 inside the integral. From (7.92) we get as $t \rightarrow 0$

$$G(tu_1, tu_2) \rightarrow G(0, 0) = \int \mu(dx) \dot{h}_0(x)^2 \left(4 \cdot 1\{h_0 > 0\} + \frac{1}{\epsilon} 1\{h_0 = 0\} \right) = J_F(0) + \frac{1 - 4\epsilon}{\epsilon} J^\#(0).$$

From (7.94) we see that $G(s_1, s_2)$ is bounded and thus, the bounded convergence theorem applies and

$$\lim_{t \rightarrow 0} \int_0^1 du_1 \int_0^1 du_2 G(tu_1, tu_2) = G(0, 0),$$

which thus concludes the proof of $L(t) \rightarrow J_F(0)$ and of (7.69) assuming facts about ϕ . Let us verify those.

For simplicity, in the next paragraph we omit the argument x in $h_0(x)$ and $\phi(\cdot; x)$. A straightforward differentiation yields

$$\phi'(h) = 2h \frac{h_0^2(1 - \frac{\epsilon}{2}) + \frac{\epsilon}{2} h^2}{(\bar{\epsilon}h_0^2 + \epsilon h^2)^{3/2}}.$$

Since $\frac{h}{\sqrt{\bar{\epsilon}h_0^2 + \epsilon h^2}} \leq \frac{1}{\sqrt{\epsilon}}$ and $\frac{h_0^2(1 - \frac{\epsilon}{2}) + \frac{\epsilon}{2} h^2}{\bar{\epsilon}h_0^2 + \epsilon h^2} \leq \frac{1 - \epsilon/2}{1 - \epsilon}$ we obtain the finiteness of ϕ' . For the continuity of ϕ' notice that if $h_0 > 0$ then clearly the function is continuous, whereas for $h_0 = 0$ we have $\phi'(h) = \frac{1}{\sqrt{\epsilon}}$ for all h .

We next proceed to the Hellinger distance. Just like in the argument above, we define

$$M(t) \triangleq \frac{1}{t^2} H^2(P_t, P_0) = \int \mu(dx) \int_0^1 du_1 \int_0^1 du_2 \dot{h}_{tu_1}(x) \dot{h}_{tu_2}(x).$$

Exactly as above from Cauchy-Schwarz and $\sup_t \int \mu(dx) \dot{h}_t(x)^2 < \infty$ we conclude that Fubini applies and hence

$$M(t) = \int_0^1 du_1 \int_0^1 du_2 H(tu_1, tu_2), \quad H(s_1, s_2) \triangleq \int \mu(dx) \dot{h}_{s_1}(x) \dot{h}_{s_2}(x).$$

Again, the family $\{\dot{h}_{s_1} \dot{h}_{s_2} : s_1 \in [0, \tau], s_2 \in [0, \tau]\}$ is uniformly integrable and thus from c_0) we conclude $H(tu_1, tu_2) \rightarrow \frac{1}{4} J_F(0)$. Furthermore, similar to (7.94) we see that $H(s_1, s_2)$ is bounded and thus

$$\lim_{t \rightarrow 0} M(t) = \int_0^1 du_1 \int_0^1 du_2 \lim_{t \rightarrow 0} H(tu_1, tu_2) = \frac{1}{4} J_F(0),$$

concluding the proof of (7.70). \square

Proceeding to variational representations, we prove the counterpart of Gelfand-Yaglom-Perez Theorem 4.6, cf. [134].

Proof of Theorem 7.11. The lower bound $D_f(P\|Q) \geq D_f(P_\mathcal{E}\|Q_\mathcal{E})$ follows from the DPI. To prove an upper bound, first we reduce to the case of $f \geq 0$ by property 6 in Proposition 7.5.

Then define sets $S = \text{supp } Q$, $F_\infty = \{\frac{dP}{dQ} = 0\}$ and for a fixed $\epsilon > 0$ let

$$F_m = \left\{ \epsilon m \leq f\left(\frac{dP}{dQ}\right) < \epsilon(m+1) \right\}, m = 0, 1, \dots.$$

We have

$$\begin{aligned} \epsilon \sum_m mQ[F_m] &\leq \int_S dQ f\left(\frac{dP}{dQ}\right) \leq \epsilon \sum_m (m+1)Q[F_m] + f(0)Q[F_\infty] \\ &\leq \epsilon \sum_m mQ[F_m] + f(0)Q[F_\infty] + \epsilon. \end{aligned} \quad (7.95)$$

Notice that on the interval $I_m^+ = \{x > 1 : \epsilon m \leq f(x) < \epsilon(m+1)\}$ the function f is increasing and on $I_m^- = \{x \leq 1 : \epsilon m \leq f(x) < \epsilon(m+1)\}$ it is decreasing. Thus partition further every F_m into $F_m^+ = \{\frac{dP}{dQ} \in I_m^+\}$ and $F_m^- = \{\frac{dP}{dQ} \in I_m^-\}$. Then, we see that

$$f\left(\frac{P[F_m^\pm]}{Q[F_m^\pm]}\right) \geq \epsilon m.$$

Consequently, for a fixed n define the partition consisting of sets $\mathcal{E} = \{F_0^+, F_0^-, \dots, F_n^+, F_n^-, F_\infty, S^c, \cup_{m>n} F_m\}$. For this partition we have, by the previous display:

$$D(P_{\mathcal{E}} \| Q_{\mathcal{E}}) \geq \epsilon \sum_{m \leq n} mQ[F_m] + f(0)Q[F_\infty] + f'(\infty)P[S^c]. \quad (7.96)$$

We next show that with sufficiently large n and sufficiently small ϵ the RHS of (7.96) approaches $D_f(P \| Q)$. If $f(0)Q[F_\infty] = \infty$ (and hence $D_f(P \| Q) = \infty$) then clearly (7.96) is also infinite. Thus, assume that $f(0)Q[F_\infty] < \infty$.

If $\int_S dQ f\left(\frac{dP}{dQ}\right) = \infty$ then the sum over m on the RHS of (7.95) is also infinite, and hence for any $N > 0$ there exists some n such that $\sum_{m \leq n} mQ[F_m] \geq N$, thus showing that RHS for (7.96) can be made arbitrarily large. Thus assume $\int_S dQ f\left(\frac{dP}{dQ}\right) < \infty$. Considering LHS of (7.95) we conclude that for some large n we have $\sum_{m>n} mQ[F_m] \leq \frac{1}{2}$. Then, we must have again from (7.95)

$$\epsilon \sum_{m \leq n} mQ[F_m] + f(0)Q[F_\infty] \geq \int_S dQ f\left(\frac{dP}{dQ}\right) - \frac{3}{2}\epsilon.$$

Thus, we have shown that for arbitrary $\epsilon > 0$ the RHS of (7.96) can be made greater than $D_f(P \| Q) - \frac{3}{2}\epsilon$. \square

Proof of Theorem 7.37. First, we show that for any $g : \mathcal{X} \rightarrow \text{dom}(f_{\text{ext}}^*)$ we must have

$$\mathbb{E}_P[g(X)] \leq D_f(P \| Q) + \mathbb{E}_Q[f_{\text{ext}}^*(g(X))]. \quad (7.97)$$

Let $p(\cdot)$ and $q(\cdot)$ be the densities of P and Q . Then, from the definition of f_{ext}^* we have for every x s.t. $q(x) > 0$:

$$f_{\text{ext}}^*(g(x)) + f_{\text{ext}}\left(\frac{p(x)}{q(x)}\right) \geq g(x) \frac{p(x)}{q(x)}.$$

7.14* Technical proofs: convexity, local expansions and variational representations 125

Integrating this over $dQ = q d\mu$ restricted to the set $\{q > 0\}$ we get

$$\mathbb{E}_Q[f_{\text{ext}}^*(g(X))] + \int_{q>0} q(x) f_{\text{ext}}\left(\frac{P(x)}{Q(x)}\right) d\mu \geq \mathbb{E}_P[g(X) \mathbf{1}\{q(X) > 0\}]. \quad (7.98)$$

Now, notice that

$$\sup\{y : y \in \text{dom}(f_{\text{ext}}^*)\} = \lim_{x \rightarrow \infty} \frac{f_{\text{ext}}(x)}{x} = f'(\infty) \quad (7.99)$$

Therefore, $f'(\infty)P[q(X) = 0] \geq \mathbb{E}_P[g(X)\mathbf{1}\{q(X) = 0\}]$. Summing the latter inequality with (7.98) we obtain (7.97).

Next we prove that supremum in (7.81) over simple functions g does yield $D_f(P\|Q)$, so that inequality (7.97) is tight. Armed with Theorem 7.11, it suffices to show (7.81) for finite \mathcal{X} . Indeed, for general \mathcal{X} , given a finite partition $\mathcal{E} = \{E_1, \dots, E_n\}$ of \mathcal{X} , we say a function $g : \mathcal{X} \rightarrow \mathbb{R}$ is \mathcal{E} -compatible if g is constant on each $E_i \in \mathcal{E}$. Taking the supremum over all finite partitions \mathcal{E} we get

$$\begin{aligned} D_f(P\|Q) &= \sup_{\mathcal{E}} D_f(P_{\mathcal{E}}\|Q_{\mathcal{E}}) \\ &= \sup_{\mathcal{E}} \sup_{\substack{g: \mathcal{X} \rightarrow \text{dom}(f_{\text{ext}}^*) \\ g \text{ } \mathcal{E}\text{-compatible}}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f_{\text{ext}}^*(g(X))] \\ &= \sup_{\substack{g: \mathcal{X} \rightarrow \text{dom}(f_{\text{ext}}^*) \\ g \text{ simple}}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f_{\text{ext}}^*(g(X))], \end{aligned}$$

where the last step follows is because the two suprema combined is equivalent to the supremum over all simple (finitely-valued) functions g .

Next, consider finite \mathcal{X} . Let $S = \{x \in \mathcal{X} : Q(x) > 0\}$ denote the support of Q . We show the following statement

$$D_f(P\|Q) = \sup_{g: S \rightarrow \text{dom}(f_{\text{ext}}^*)} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f_{\text{ext}}^*(g(X))] + f'(\infty)P(S^c), \quad (7.100)$$

which is equivalent to (7.81) by (7.99). By definition,

$$D_f(P\|Q) = \underbrace{\sum_{x \in S} Q(x) f_{\text{ext}}\left(\frac{P(x)}{Q(x)}\right)}_{\triangleq \Psi(P)} + f'(\infty) \cdot P(S^c),$$

Consider the functional $\Psi(P)$ defined above where P takes values over all signed measures on S , which can be identified with \mathbb{R}^S . The convex conjugate of $\Psi(P)$ is as follows: for any $g : S \rightarrow \mathbb{R}$,

$$\begin{aligned} \Psi^*(g) &= \sup_P \sum_x P(x)g(x) - Q(x) \left\{ \sup_{h \in \text{dom}(f_{\text{ext}}^*)} \frac{P(x)}{Q(x)} h - f_{\text{ext}}^*(h) \right\} \\ &= \sup_P \inf_{h: S \rightarrow \text{dom}(f_{\text{ext}}^*)} \sum_x P(x)(g(x) - h(x)) + Q(x)f_{\text{ext}}^*(h(x)) \\ &\stackrel{(a)}{=} \inf_{h: S \rightarrow \text{dom}(f_{\text{ext}}^*)} \sup_P \sum_x P(x)(g(x) - h(x)) + \mathbb{E}_Q[f_{\text{ext}}^*(h)] \end{aligned}$$

$$= \begin{cases} \mathbb{E}_Q[f_{\text{ext}}^*(g(X))] & g : S \rightarrow \text{dom}(f_{\text{ext}}^*) \\ +\infty & \text{otherwise} \end{cases}.$$

where (a) follows from the minimax theorem (which applies due to finiteness of \mathcal{X}). Applying the convex duality in (7.80) yields the proof of the desired (7.100). \square

Proof of Theorem 7.38. First we argue that the supremum in the right-hand side of (7.88) can be taken over all simple functions g . Then thanks to Theorem 7.11, it will suffice to consider finite alphabet \mathcal{X} . To that end, fix any g . For any δ , there exists a such that $\mathbb{E}_Q[f_{\text{ext}}^*(g - a)] - aP[S] \leq \Psi_{Q,P}^*(g) + \delta$. Since $\mathbb{E}_Q[f_{\text{ext}}^*(g - a_n)]$ can be approximated arbitrarily well by simple functions we conclude that there exists a simple function \tilde{g} such that simultaneously $\mathbb{E}_P[\tilde{g}1_S] \geq \mathbb{E}_P[g1_S] - \delta$ and

$$\Psi_{Q,P}^*(\tilde{g}) \leq \mathbb{E}_Q[f_{\text{ext}}^*(\tilde{g} - a)] - aP[S] + \delta \leq \Psi_{Q,P}^*(g) + 2\delta.$$

This implies that restricting to simple functions in the supremization in (7.88) does not change the right-hand side.

Next consider finite \mathcal{X} . We proceed to compute the conjugate of Ψ , where $\Psi(P) \triangleq D_f(P\|Q)$ if P is a probability measure on \mathcal{X} and $+\infty$ otherwise. Then for any $g : \mathcal{X} \rightarrow \mathbb{R}$, maximizing over all probability measures P we have:

$$\begin{aligned} \Psi^*(g) &= \sup_P \sum_{x \in \mathcal{X}} P(x)g(x) - D_f(P\|Q) \\ &= \sup_P \sum_{x \in \mathcal{X}} P(x)g(x) - \sum_{x \in S^c} P(x)g(x) - \sum_{x \in S} Q(x)f\left(\frac{P(x)}{Q(x)}\right) \\ &= \sup_P \inf_{h: S \rightarrow \mathbb{R}} \sum_{x \in S} P(x)[g(x) - h(x)] + \sum_{x \in S^c} P(x)[g(x) - f'(\infty)] + \sum_{x \in S} Q(x)f_{\text{ext}}^*(h(x)) \\ &\stackrel{(a)}{=} \inf_{h: S \rightarrow \mathbb{R}} \left\{ \sup_P \left(\sum_{x \in S} P(x)[g(x) - h(x)] + \sum_{x \in S^c} P(x)[g(x) - f'(\infty)] \right) + \mathbb{E}_Q[f_{\text{ext}}^*(h(X))] \right\} \\ &\stackrel{(b)}{=} \inf_{h: S \rightarrow \mathbb{R}} \left\{ \max \left(\max_{x \in S} g(x) - h(x), \max_{x \in S^c} g(x) - f'(\infty) \right) + \mathbb{E}_Q[f_{\text{ext}}^*(h(X))] \right\} \\ &\stackrel{(c)}{=} \inf_{a \in \mathbb{R}} \left\{ \max \left(a, \max_{x \in S^c} g(x) - f'(\infty) \right) + \mathbb{E}_Q[f_{\text{ext}}^*(g(X) - a)] \right\} \end{aligned}$$

where (a) follows from the minimax theorem; (b) is due to P being a probability measure; (c) follows since we can restrict to $h(x) = g(x) - a$ for $x \in S$, thanks to the fact that f_{ext}^* is non-decreasing (since $\text{dom}(f_{\text{ext}}) = \mathbb{R}_+$).

From convex duality we have shown that $D_f(P\|Q) = \sup_g \mathbb{E}_P[g] - \Psi^*(g)$. Notice that without loss of generality we may take $g(x) = f'(\infty) + b$ for $x \in S^c$. Interchanging the optimization over b with that over a we find that

$$\sup_b bP[S^c] - \max(a, b) = -aP[S],$$

7.14* Technical proofs: convexity, local expansions and variational representations 127

which then recovers (7.88). To get (7.89) simply notice that if $P[S^c] > 0$, then both sides of (7.89) are infinite (since $\Psi_Q^*(g)$ does not depend on the values of g outside of S). Otherwise, (7.89) coincides with (7.88). \square

8

Entropy method in combinatorics and geometry

A commonly used method in combinatorics for bounding the number of certain objects from above involves a smart application of Shannon entropy. This method typically proceeds as follows: in order to count the cardinality of a given set \mathcal{C} , we draw an element uniformly at random from \mathcal{C} , whose entropy is given by $\log |\mathcal{C}|$. To bound $|\mathcal{C}|$ from above, we describe this random object by a random vector $X = (X_1, \dots, X_n)$, e.g., an indicator vector, then proceed to compute or upper-bound the joint entropy $H(X_1, \dots, X_n)$.

Notably, three methods of increasing precision are as follows:

- Marginal bound:

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$$

- Pairwise bound (Shearer's lemma and generalization cf. Theorem 1.9):

$$H(X_1, \dots, X_n) \leq \frac{1}{n-1} \sum_{i < j} H(X_i, X_j)$$

- Chain rule (exact calculation):

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1})$$

We give three applications using the above three methods, respectively, in the order of increasing difficulty:

- Enumerating binary vectors of a given average weights
- Counting triangles and other subgraphs
- Brégman's theorem

Finally, to demonstrate how entropy method can also be used for questions in Euclidean spaces, we prove the Loomis-Whitney and Bollobás-Thomason theorems based on analogous properties of *differential* entropy (Section 2.3).

8.1 Binary vectors of average weights

Lemma 8.1 (Massey [205]). *Let $\mathcal{C} \subset \{0, 1\}^n$ and let p be the average fraction of 1's in \mathcal{C} , i.e.*

$$p = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{w_H(x)}{n},$$

where $w_H(x)$ is the Hamming weight (number of 1's) of $x \in \{0, 1\}^n$. Then $|\mathcal{C}| \leq 2^{nh(p)}$.

Remark 8.1. This result holds even if $p > 1/2$.

Proof. Let $X = (X_1, \dots, X_n)$ be drawn uniformly at random from \mathcal{C} . Then

$$\log |\mathcal{C}| = H(X) = H(X_1, \dots, X_n) \leq \sum_i^n H(X_i) = \sum_{i=1}^n h(p_i),$$

where $p_i = \mathbb{P}[X_i = 1]$ is the fraction of vertices whose i -th bit is 1. Note that

$$p = \frac{1}{n} \sum_{i=1}^n p_i,$$

since we can either first average over vectors in \mathcal{C} or first average across different bits. By Jensen's inequality and the fact that $x \mapsto h(x)$ is concave,

$$\sum_{i=1}^n h(p_i) \leq nh \left(\frac{1}{n} \sum_{i=1}^n p_i \right) = nh(p).$$

Hence we have shown that $\log |\mathcal{C}| \leq nh(p)$. □

Theorem 8.2.

$$\sum_{j=0}^k \binom{n}{j} \leq 2^{nh(k/n)}, \quad k \leq n/2.$$

Proof. We take $\mathcal{C} = \{x \in \{0, 1\}^n : w_H(x) \leq k\}$ and invoke the previous lemma, which says that

$$\sum_{j=0}^k \binom{n}{j} = |\mathcal{C}| \leq 2^{nh(p)} \leq 2^{nh(k/n)},$$

where the last inequality follows from the fact that $x \mapsto h(x)$ is increasing for $x \leq 1/2$. □

Remark 8.2. Alternatively, we can prove Theorem 8.3 using the large-deviation bound in Part III. By the Chernoff bound on the binomial tail (see (15.19) in Example 15.1),

$$\frac{\text{LHS}}{2^n} = \mathbb{P}(\text{Bin}(n, 1/2) \leq k) \leq 2^{-nd(\frac{k}{n} \| \frac{1}{2})} = 2^{-n(1-h(k/n))} = \frac{\text{RHS}}{2^n}.$$

8.2 Shearer's lemma & counting subgraphs

Recall that a special case of Shearer's lemma Theorem 1.9 (or Han's inequality Theorem 1.8) says:

$$H(X_1, X_2, X_3) \leq \frac{1}{2}[H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3)].$$

A classical application of this result (see Remark 1.10) is to bound cardinality of a set in \mathbb{R}^3 given cardinalities of its projections.

For graphs H and G , define $N(H, G)$ to be the number of copies of H in G .¹ For example,

$$N(\text{---}, \text{---}) = 4, \quad N(\text{---}, \text{---}) = 8.$$

If we know G has m edges, what is the maximal number of H that are contained in G ? To study this quantity, let's define

$$N(H, m) = \max_{G: |E(G)| \leq m} N(H, G).$$

We will show that for the maximal number of triangles satisfies

$$N(K_3, m) \asymp m^{3/2}. \quad (8.1)$$

To show that $N(H, m) \gtrsim m^{3/2}$, consider $G = K_n$ which has $m = |E(G)| = \binom{n}{2} \asymp n^2$ and $N(K_3, K_n) = \binom{n}{3} \asymp n^3 \asymp m^{3/2}$.

To show the upper bound, fix a graph $G = (V, E)$ with m edges. Draw a labeled triangle uniformly at random and denote the vertices by (X_1, X_2, X_3) . Then by Shearer's Lemma,

$$\log(3!N(K_3, G)) = H(X_1, X_2, X_3) \leq \frac{1}{2}[H(X_1, X_2) + H(X_2, X_3) + H(X_1, X_3)] \leq \frac{3}{2}\log(2m).$$

Hence

$$N(K_3, G) \leq \frac{\sqrt{2}}{3}m^{3/2}. \quad (8.2)$$

Remark 8.3. Interestingly, linear algebra argument yields exactly the same upper bound as (8.2): Let A be the adjacency matrix of G with eigenvalues $\{\lambda_i\}$. Then

$$2|E(G)| = \text{tr}(A^2) = \sum \lambda_i^2$$

$$6N(K_3, G) = \text{tr}(A^3) = \sum \lambda_i^3$$

By Minkowski's inequality, $(6N(K_3, G))^{1/3} \leq (2|E(G)|)^{1/2}$ which yields $N(K_3, G) \leq \frac{\sqrt{2}}{3}m^{3/2}$.

Using Shearer's Theorem 1.9 Friedgut and Kahn [128] obtained the counterpart of (8.1) for arbitrary H ; this result was first proved by Alon [8]. We start by introducing the *fractional covering*

¹ To be precise, here $N(H, G)$ is the number of subgraphs of G (subsets of edges) isomorphic to H . If we denote by $\text{inj}(H, G)$ the number of injective maps $V(H) \rightarrow V(G)$ mapping edges of H to edges of G , then $N(H, G) = \frac{1}{|\text{Aut}(H)|}\text{inj}(H, G)$.

8.2 Shearer's lemma & counting subgraphs 131

number of a graph. For a graph $H = (V, E)$, define the fractional covering number as the value of the following linear program:²

$$\rho^*(H) = \min_w \left\{ \sum_{e \in E} w_H(e) : \sum_{e \in E, v \in e} w_H(e) \geq 1, \forall v \in V, w_H(e) \in [0, 1] \right\} \quad (8.3)$$

Theorem 8.3.

$$c_0(H)m^{\rho^*(H)} \leq N(H, m) \leq c_1(H)m^{\rho^*(H)}. \quad (8.4)$$

For example, for triangles we have $\rho^*(K_3) = 3/2$ and Theorem 8.6 is consistent with (8.1).

Proof. *Upper bound:* Let $V(H) = [n]$ and let $w^*(e)$ be the solution for $\rho^*(H)$. For any G with m edges, draw a subgraph of G , uniformly at random from all those that are isomorphic to H . Given such a random subgraph set $X_i \in V(G)$ to be the vertex corresponding to an i -th vertex of H , $i \in [n]$. Now define a random 2-subset S of $[n]$ by sampling an edge e from $E(H)$ with probability $\frac{w^*(e)}{\rho^*(H)}$. By the definition of $\rho^*(H)$ we have for any $i \in [n]$ that $\mathbb{P}[i \in S] \geq \frac{1}{\rho^*(H)}$. We are now ready to apply Shearer's Theorem 1.9:

$$\begin{aligned} \log N(H, G) &= H(X) \\ &\leq H(X_S | S) \rho^*(H) \leq \log(2m) \rho^*(H), \end{aligned}$$

where the last bound is as before: if $S = \{v, w\}$ then $X_S = (X_v, X_w)$ takes one of $2m$ values. Overall, we get³ $N(H, G) \leq (2m)^{\rho^*(H)}$.

Lower bound: It amounts to construct a graph G with m edges for which $N(H, G) \geq c(H)|e(G)|^{\rho^*(H)}$. Consider the dual LP of (8.3)

$$\alpha^*(H) = \max_{\psi} \left\{ \sum_{v \in V(H)} \psi(v) : \psi(v) + \psi(w) \leq 1, \forall (vw) \in E, \psi(v) \in [0, 1] \right\} \quad (8.5)$$

i.e., the *fractional packing number*. By the duality theorem of LP, we have $\alpha^*(H) = \rho^*(H)$. The graph G is constructed as follows: for each vertex v of H , replicate it for $m(v)$ times. For each edge $e = (vw)$ of H , replace it by a complete bipartite graph $K_{m(v), m(w)}$. Then the total number of edges of G is

$$|E(G)| = \sum_{(vw) \in E(H)} m(v)m(w).$$

² If the “ $\in [0, 1]$ ” constraints in (8.3) and (8.5) are replaced by “ $\in \{0, 1\}$ ”, we obtain the covering number $\rho(H)$ and the independence number $\alpha(H)$ of H , respectively.

³ Note that for $H = K_3$ this gives a bound weaker than (8.2). To recover (8.2) we need to take $X = (X_1, \dots, X_n)$ be uniform on all injective homomorphisms $H \rightarrow G$.

Furthermore, $N(G, H) \geq \prod_{v \in V(H)} m(v)$. To minimize the exponent $\frac{\log N(G, H)}{\log |E(G)|}$, fix a large number M and let $m(v) = \lceil M^{\psi(v)} \rceil$, where ψ is the maximizer in (8.5). Then

$$\begin{aligned} |E(G)| &\leq \sum_{(vw) \in E(H)} 4M^{\psi(v)+\psi(w)} \leq 4M|E(H)| \\ N(G, H) &\geq \prod_{v \in V(H)} M^{\psi(v)} = M^{\alpha^*(H)} \end{aligned}$$

and we are done. \square

8.3 Brégman's Theorem

In this section, we present an elegant entropy proof by Radhakrishnan [247] of Brégman's Theorem [50], which bounds the number of *perfect matchings* (1-regular spanning subgraphs) in a bipartite graphs.

We start with some definitions. The *permanent* of an $n \times n$ matrix A is defined as

$$\text{perm}(A) \triangleq \sum_{\pi \in S_n} \prod_{i=1}^n a_{i\pi(i)},$$

where S_n denotes the group of all permutations of $[n]$. For a bipartite graph G with n vertices on the left and right respectively, the number of perfect matchings in G is given by $\text{perm}(A)$, where A is the adjacency matrix. For example,

$$\text{perm} \left(\begin{array}{ccc} \circ & \circ & \\ \diagdown & \diagup & \\ \circ & \circ & \end{array} \right) = 1, \quad \text{perm} \left(\begin{array}{ccc} \circ & \circ & \\ \diagup & \diagdown & \\ \circ & \circ & \end{array} \right) = 2$$

Theorem 8.4 (Brégman's Theorem). *For any $n \times n$ bipartite graph with adjacency matrix A ,*

$$\text{perm}(A) \leq \prod_{i=1}^n (d_i!)^{\frac{1}{d_i}},$$

where d_i is the degree of left vertex i (i.e. sum of the i^{th} row of A).

As an example, consider $G = K_{n,n}$. Then $\text{perm}(G) = n!$, which coincides with the RHS $[(n!)^{1/n}]^n = n!$. More generally, if G consists of n/d copies of $K_{d,d}$, then Bregman's bound is tight and $\text{perm} = (d!)^{n/d}$.

As a first attempt of proving Theorem 8.7 using the entropy method, we select a perfect matching uniformly at random which matches the i^{th} left vertex to the X_i^{th} right one. Let $X =$

8.3 Brégman's Theorem 133

(X_1, \dots, X_n) . Then

$$\log \text{perm}(A) = H(X) = H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \leq \sum_{i=1}^n \log(d_i).$$

Hence $\text{perm}(A) \leq \prod_i d_i$. This is worse than Brégman's bound by an exponential factor, since by Stirling's formula

$$\prod_{i=1}^n (d_i!)^{\frac{1}{d_i}} \sim \left(\prod_{i=1}^n d_i \right) e^{-n}.$$

Here is our second attempt. The hope is to use the chain rule to expand the joint entropy and bound the conditional entropy more carefully. Let's write

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1}) \leq \sum_{i=1}^n \mathbb{E}[\log N_i].$$

where N_i , as a random variable, denotes the number of possible values X_i can take conditioned on X_1, \dots, X_{i-1} , i.e., how many possible matchings for left vertex i given the outcome of where $1, \dots, i-1$ are matched to. However, it is hard to proceed from this point as we only know the degree information, not the graph itself. In fact, since we do not know the relative positions of the vertices, there is no reason why we should order from 1 to n . The key idea is to *label the vertices randomly*, apply chain rule in this random order and average.

To this end, pick π uniformly at random from S_n and independent of X . Then

$$\begin{aligned} \log \text{perm}(A) &= H(X) = H(X|\pi) \\ &= H(X_{\pi(1)}, \dots, X_{\pi(n)}|\pi) \\ &= \sum_{k=1}^n H(X_{\pi(k)} | X_{\pi(1)}, \dots, X_{\pi(k-1)}, \pi) \\ &= \sum_{k=1}^n H(X_k | \{X_j : \pi^{-1}(j) < \pi^{-1}(k)\}, \pi) \\ &\leq \sum_{k=1}^n \mathbb{E} \log N_k, \end{aligned}$$

where N_k denotes the number of possible matchings for vertex k given the outcomes of $\{X_j : \pi^{-1}(j) < \pi^{-1}(k)\}$ and the expectation is with respect to (X, π) . The key observation is:

Lemma 8.5. N_k is uniformly distributed on $[d_k]$.

Example 8.1. As a concrete example for Lemma 8.8, consider the graph G on the right. For vertex $k = 1$, $d_k = 2$. Depending on the random ordering, if $\pi = 1 * *$, then $N_k = 2$ w.p. $1/3$; if $\pi = * * 1$, then $N_k = 1$ w.p. $1/3$; if $\pi = 213$, then $N_k = 2$ w.p. $1/3$; if $\pi = 312$, then $N_k = 1$ w.p. $1/3$. Combining everything, indeed N_k is equally likely to be 1 or 2.

Applying Lemma 8.8,

$$\mathbb{E}_{(X,\pi)} \log N_k = \frac{1}{d_k} \sum_{i=1}^{d_k} \log i = \log(d_i!)^{\frac{1}{d_i}}$$

and hence

$$\log \text{perm}(A) \leq \sum_{k=1}^n \log(d_i!)^{\frac{1}{d_i}} = \log \prod_{i=1}^n (d_i!)^{\frac{1}{d_i}}.$$

Finally, we prove Lemma 8.8:

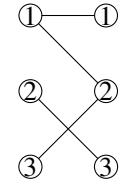
Proof. Note that $X_i = \sigma(i)$ for some random permutation σ . Let $T = \partial(k)$ be the neighbors of k . Then

$$N_k = |T \setminus \{\sigma(j) : \pi^{-1}(j) < \pi^{-1}(k)\}|$$

which is a function of (σ, π) . In fact, conditioned on any realization of σ , N_k is uniform over $[d_k]$. To see this, note that $\sigma^{-1}(T)$ is a fixed subset of $[n]$ of cardinality d_k , and $k \in \sigma^{-1}(T)$. On the other hand, $S \triangleq \{j : \pi^{-1}(j) < \pi^{-1}(k)\}$ is a uniformly random subset of $[n] \setminus \{k\}$. Then

$$N_k = |\sigma^{-1}(T) \setminus S| = 1 + \underbrace{|\sigma^{-1}(T) \setminus \{k\} \cap S|}_{\text{Unif}(\{0, \dots, d_k - 1\})},$$

which is uniform over $[d_k]$. □



8.4 Euclidean geometry: Bollobás-Thomason and Loomis-Whitney

The following famous result shows that n -dimensional rectangles simultaneously minimize the volumes of all coordinate projections:⁴

Theorem 8.6 (Bollobás-Thomason Box Theorem). *Let $K \subset \mathbb{R}^n$ be a compact set. For $S \subset [n]$, denote by $K_S \subset \mathbb{R}^S$ the projection of K onto those coordinates indexed by S . Then there exists a rectangle A s.t. $\text{Leb}(A) = \text{Leb}(K)$ and for all $S \subset [n]$:*

$$\text{Leb}(A_S) \leq \text{Leb}(K_S)$$

⁴ Note that since K is compact, its projection and slices are all compact and hence measurable.

8.4 Euclidean geometry: Bollobás-Thomason and Loomis-Whitney 135

Proof. Let X^n be uniformly distributed on K . Then $h(X^n) = \log \text{Leb}(K)$. Let A be a rectangle of size $a_1 \times \dots \times a_n$ where

$$\log a_i = h(X_i | X^{i-1}).$$

Then, we have by Theorem 2.7(a)

$$h(X_S) \leq \log \text{Leb}(K_S).$$

On the other hand, by the chain rule and the fact that conditioning reduces differential entropy (recall Theorem 2.7(a) and (c)),

$$\begin{aligned} h(X_S) &= \sum_{i=1}^n \mathbb{1}\{i \in S\} h(X_i | X_{[i-1] \cap S}) \\ &\geq \sum_{i \in S} h(X_i | X^{i-1}) \\ &= \log \prod_{i \in S} a_i \\ &= \log \text{Leb}(A_S) \end{aligned}$$

□

The following result is a continuous counterpart of Shearer's lemma (see Theorem 1.9 and Remark 1.10):

Corollary 8.7 (Loomis-Whitney). *Let K be a compact subset of \mathbb{R}^n and let K_{j^c} denote the projection of K onto coordinates in $[n] \setminus j$. Then*

$$\text{Leb}(K) \leq \prod_{j=1}^n \text{Leb}(K_{j^c})^{\frac{1}{n-1}}. \quad (8.6)$$

Proof. Let A be a rectangle having the same volume as K . Note that

$$\text{Leb}(K) = \text{Leb}(A) = \prod_{j=1}^n \text{Leb}(A_{j^c})^{\frac{1}{n-1}}$$

By the previous theorem, $\text{Leb}(A_{j^c}) \leq \text{Leb}(K_{j^c})$.

□

The meaning of the Loomis-Whitney inequality is best understood by introducing the average width of K in the j th direction: $w_j \triangleq \frac{\text{Leb}(K)}{\text{Leb}(K_{j^c})}$. Then (8.6) is equivalent to

$$\text{Leb}(K) \geq \prod_{j=1}^n w_j,$$

i.e. that volume of K is greater than that of the rectangle of average widths.

9

Random number generators

Consider the following problem: Given a stream of independent $\text{Ber}(p)$ bits, with *unknown* p , we want to turn them into pure random bits, i.e., independent $\text{Ber}(1/2)$ bits; Our goal is to find a universal way to extract the most number of bits. In other words, we want to extract as many fair coin flips as possible from possibly biased coin flips, without knowing the actual bias.

In 1951 von Neumann [316] proposed the following scheme: Divide the stream into pairs of bits, output 0 if 10, output 1 if 01, otherwise do nothing and move to the next pair. Since both 01 and 10 occur with probability pq (where $q \triangleq 1 - p$ throughout this chapter), regardless of the value of p , we obtain fair coin flips at the output. To measure the efficiency of von Neumann's scheme, note that, on average, we have $2n$ bits in and $2pqn$ bits out. So the efficiency (rate) is pq . The question is: Can we do better?

There are several choices to be made in the problem formulation. *Universal v.s. non-universal*: the source distribution can be unknown or partially known, respectively. *Exact v.s. approximately fair coin flips*: whether the generated coin flips are exactly fair or approximately, as measured by one of the f -divergences studied in Chapter 7 (e.g., the total variation or KL divergence). In this chapter, we only focus on the universal generation of exactly fair coins.

9.1 Setup

Let $\{0, 1\}^* = \cup_{k \geq 0} \{0, 1\}^k = \{\emptyset, 0, 1, 00, 01, \dots\}$ denote the set of all finite-length binary strings, where \emptyset denotes the empty string. For any $x \in \{0, 1\}^*$, $l(x)$ denotes the length of x .

Let us first introduce the definition of random number generator formally. If the input vector is X , denote the output (variable-length) vector by $Y \in \{0, 1\}^*$. Then the desired property of Y is the following: Conditioned on the length of Y being k , Y is uniformly distributed on $\{0, 1\}^k$.

Definition 9.1 (Extractor). We say $\Psi : \{0, 1\}^* \rightarrow \{0, 1\}^*$ is an *extractor* if

- 1 $\Psi(x)$ is a prefix of $\Psi(y)$ if x is a prefix of y .
- 2 For any n and any $p \in (0, 1)$, if $X^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$, then $\Psi(X^n) \sim \text{Ber}(1/2)^k$ conditioned on $l(\Psi(X^n)) = k$ for each $k \geq 1$.

The efficiency of an extractor Ψ is measured by its *rate*:

$$r_\Psi(p) = \limsup_{n \rightarrow \infty} \frac{\mathbb{E}[I(\Psi(X^n))]}{n}, \quad X^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p).$$

In other words, Ψ consumes a stream of n coins with bias p and outputs on average $nr_\Psi(p)$ fair coins.

Note that the von Neumann scheme above defines a valid extractor Ψ_{vN} (with $\Psi_{\text{vN}}(x^{2n+1}) = \Psi_{\text{vN}}(x^{2n})$), whose rate is $r_{\text{vN}}(p) = pq$. Clearly this is wasteful, because even if the input bits are already fair, we only get 25% in return.

9.2 Converse

We show that no extractor has a rate higher than the binary entropy function $h(p)$, even if the extractor is allowed to be non-universal (depending on p). The intuition is that the “information content” contained in each $\text{Ber}(p)$ variable is $h(p)$ bits; as such, it is impossible to extract more than that. This is easily made precise by the data processing inequality for entropy (since extractors are deterministic functions).

Theorem 9.2. *For any extractor Ψ and any $p \in (0, 1)$,*

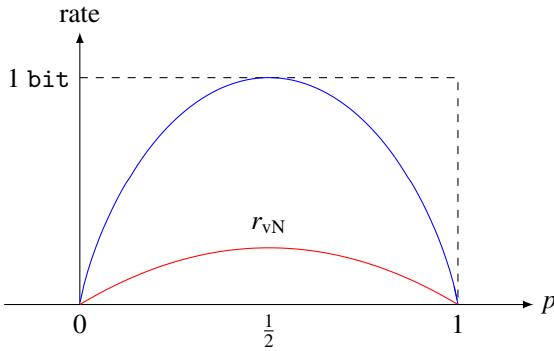
$$r_\Psi(p) \geq h(p) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}.$$

Proof. Let $L = \Psi(X^n)$. Then

$$nh(p) = H(X^n) \geq H(\Psi(X^n)) = H(\Psi(X^n)|L) + H(L) \geq H(\Psi(X^n)|L) = \mathbb{E}[L] \text{ bits,}$$

where the last step follows from the assumption on Ψ that $\Psi(X^n)$ is uniform over $\{0, 1\}^k$ conditioned on $L = k$. \square

The rate of von Neumann extractor and the entropy bound are plotted below. Next we present two extractors, due to Elias [111] and Peres [227] respectively, that attain the binary entropy function. (More precisely, both construct a sequence of extractors whose rate approaches the entropy bound).



9.3 Elias' construction from data compression

The intuition behind Elias' scheme is the following:

- 1 For iid X^n , the probability of each string only depends on its *type*, i.e., the number of 1's. (This is the main idea of the method of types for data compression.) Therefore conditioned on the number of 1's, X^n is uniformly distributed (over the type class). This observation holds universally for any p .
- 2 Given a uniformly distributed random variable on some finite set, we can easily turn it into *variable-length* string of fair coin flips. For example:
 - If U is uniform over $\{1, 2, 3\}$, we can map $1 \mapsto \emptyset, 2 \mapsto 0$ and $3 \mapsto 1$.
 - If U is uniform over $\{1, 2, \dots, 11\}$, we can map $1 \mapsto \emptyset, 2 \mapsto 0, 3 \mapsto 1$, and the remaining eight numbers $4, \dots, 11$ are assigned to 3-bit strings.

Lemma 9.3. *Given U uniformly distributed on $[M]$, there exists $f : [M] \rightarrow \{0, 1\}^*$ such that conditioned on $l(f(U)) = k$, $f(U)$ is uniformly over $\{0, 1\}^k$. Moreover,*

$$\log_2 M - 4 \leq \mathbb{E}[l(f(U))] \leq \log_2 M \text{ bits.}$$

Proof. We defined f by partitioning $[M]$ into subsets whose cardinalities are powers of two, and assign elements in each subset to binary strings of that length. Formally, denote the binary expansion of M by $M = \sum_{i=0}^n m_i 2^i$, where the most significant bit $m_n = 1$ and $n = \lfloor \log_2 M \rfloor + 1$. Those non-zero m_i 's defines a partition $[M] = \bigcup_{j=0}^t M_j$, where $|M_j| = 2^{i_j}$. Map the elements of M_j to $\{0, 1\}^{i_j}$. Finally, notice that uniform distribution conditioned on any subset is still uniform.

To prove the bound on the expected length, the upper bound follows from the same entropy argument $\log_2 M = H(U) \geq H(f(U)) \geq H(f(U)|l(f(U))) = \mathbb{E}[l(f(U))]$, and the lower bound follows from

$$\mathbb{E}[l(f(U))] = \frac{1}{M} \sum_{i=0}^n m_i 2^i \cdot i = n - \frac{1}{M} \sum_{i=0}^n m_i 2^i (n-i) \geq n - \frac{2^n}{M} \sum_{i=0}^n 2^{i-n} (n-i) \geq n - \frac{2^{n+1}}{M} \geq n - 4,$$

where the last step follows from $n \leq \log_2 M + 1$. □

9.4 Peres' iterated von Neumann's scheme 139

Elias' extractor Fix $n \geq 1$. Let $w_H(x^n)$ define the Hamming weight (number of ones) of a binary string x^n . Let $T_k = \{x^n \in \{0, 1\}^n : w_H(x^n) = k\}$ define the Hamming sphere of radius k . For each $0 \leq k \leq n$, we apply the function f from Lemma 9.3 to each T_k . This defines a mapping $\Psi_E : \{0, 1\}^n \rightarrow \{0, 1\}^*$ and then we extend it to $\Psi_E : \{0, 1\}^* \rightarrow \{0, 1\}^*$ by applying the mapping per n -bit block and discard the last incomplete block. Then it is clear that the rate is given by $\frac{1}{n}\mathbb{E}[l(\Psi_E(X^n))]$. By Lemma 9.3, we have

$$\mathbb{E} \log \binom{n}{w_H(X^n)} - 4 \leq \mathbb{E}[l(\Psi_E(X^n))] \leq \mathbb{E} \log \binom{n}{w_H(X^n)}$$

Using Stirling's approximation, we can show (see, e.g., [19, Lemma 4.7.1])

$$\frac{2^{nh(p)}}{\sqrt{8k(n-k)/n}} \leq \binom{n}{k} \leq \frac{2^{nh(p)}}{\sqrt{2\pi k(n-k)/n}} \quad (9.1)$$

whenever $1 \leq k \leq n-1$. Since $w_H(X^n) \sim \text{Bin}(n, p)$ and h is a continuous bounded function, applying the law of large numbers yields

$$\mathbb{E}[l(\Psi_E(X^n))] = nh(p) + o(n).$$

Therefore the extraction rate of Ψ_E approaches the optimum $h(p)$ as $n \rightarrow \infty$.

9.4 Peres' iterated von Neumann's scheme

The main idea is to recycle the bits thrown away in von Neumann's scheme and iterate. What von Neumann's extractor discarded are: (a) bits from equal pairs; (b) location of the distinct pairs. To achieve the entropy bound, we need to extract the randomness out of these two parts as well.

First, some notations: Given x^{2n} , let $k = l(\Psi_{vN}(x^{2n}))$ denote the number of consecutive distinct bit-pairs.

- Let $1 \leq m_1 < \dots < m_k \leq n$ denote the locations such that $x_{2m_j} \neq x_{2m_j-1}$.
- Let $1 \leq i_1 < \dots < i_{n-k} \leq n$ denote the locations such that $x_{2i_j} = x_{2i_j-1}$.
- $y_j = x_{2m_j}$, $v_j = x_{2i_j}$, $u_j = x_{2j} \oplus x_{2j+1}$.

Here y^k are the bits that von Neumann's scheme outputs and both v^{n-k} and u^n are discarded. Note that u^n is important because it encodes the location of the y^k and contains a lot of information. Therefore von Neumann's scheme can be improved if we can extract the randomness out of both v^{n-k} and u^n .

Peres' extractor For each $t \in \mathbb{N}$, recursively define an extractor Ψ_t as follows:

- Set Ψ_1 to be von Neumann's extractor Ψ_{vN} , i.e., $\Psi_1(x^{2n+1}) = \Psi_1(x^{2n}) = y^k$.
- Define Ψ_t by $\Psi_t(x^{2n}) = \Psi_t(x^{2n+1}) = (\Psi_1(x^{2n}), \Psi_{t-1}(u^n), \Psi_{t-1}(v^{n-k}))$.

Example: Input $x = 100111010011$ of length $2n = 12$. Output recursively:

$$\begin{aligned} & \overbrace{(011)}^y \overbrace{(110100)}^u \overbrace{(101)}^v \\ & (1)(010)(10)(0) \\ & (1)(0) \end{aligned}$$

Next we (a) verify Ψ_t is a valid extractor; (b) evaluate its efficiency (rate). Note that the bits that enter into the iteration are no longer i.i.d. To compute the rate of Ψ_t , it is convenient to introduce the notion of exchangeability. We say X^n are *exchangeable* if the joint distribution is invariant under permutation, that is, $P_{X_1, \dots, X_n} = P_{X_{\pi(1)}, \dots, X_{\pi(n)}}$ for any permutation π on $[n]$. In particular, if X_i 's are binary, then X^n are exchangeable if and only if the joint distribution only depends on the *Hamming weight*, i.e., $P_{X^n}(x^n) = f(w_H(x^n))$ for some function f . Examples: X^n is iid $\text{Ber}(p)$; X^n is uniform over the Hamming sphere T_k .

As an example, if X^{2n} are i.i.d. $\text{Ber}(p)$, then conditioned on $L = k$, V^{n-k} is iid $\text{Ber}(p^2/(p^2 + q^2))$, since $L \sim \text{Binom}(n, 2pq)$ and

$$\begin{aligned} \mathbb{P}[Y^k = y, U^n = u, V^{n-k} = v | L = k] &= \frac{p^{k+2m}q^{n-k-2m}}{\binom{n}{k}(p^2 + q^2)^{n-k}(2pq)^k} \\ &= 2^{-k} \cdot \binom{n}{k}^{-1} \cdot \left(\frac{p^2}{p^2 + q^2}\right)^m \left(\frac{q^2}{p^2 + q^2}\right)^{n-k-m} \\ &= \mathbb{P}[Y^k = y | L = k] \mathbb{P}[U^n = u | L = k] \mathbb{P}[V^{n-k} = v | L = k], \end{aligned}$$

where $m = w_H(v)$. In general, when X^{2n} are only exchangeable, we have the following:

Lemma 9.4 (Ψ_t preserves exchangeability). *Let X^{2n} be exchangeable and $L = \Psi_1(X^{2n})$. Then conditioned on $L = k$, Y^k , U^n and V^{n-k} are independent, each having an exchangeable distribution. Furthermore, $Y^k \stackrel{i.i.d.}{\sim} \text{Ber}(\frac{1}{2})$ and U^n is uniform over T_k .*

Proof. It suffices to show that $\forall y, y' \in \{0, 1\}^k, u, u' \in T_k$ and $v, v' \in \{0, 1\}^{n-k}$ such that $w_H(v) = w_H(v')$, we have

$$\mathbb{P}[Y^k = y, U^n = u, V^{n-k} = v | L = k] = \mathbb{P}[Y^k = y', U^n = u', V^{n-k} = v' | L = k],$$

which implies that $\mathbb{P}[Y^k = y, U^n = u, V^{n-k} = v | L = k] = f(w_H(v))$ for some function f . Note that the string X^{2n} and the triple (Y^k, U^n, V^{n-k}) are in one-to-one correspondence of each other. Indeed, to reconstruct X^{2n} , simply read the k distinct pairs from Y and fill them according to the locations of ones in U and fill the remaining equal pairs from V . [Examples: $(y, u, v) = (01, 1100, 01) \Rightarrow x = (10010011)$, $(y, u, v) = (11, 1010, 10) \Rightarrow x' = (01110100)$.] Finally, note that u, y, v and u', y', v' correspond to two input strings x and x' of identical Hamming weight ($w_H(x) = k + 2w_H(v)$) and hence of identical probability due to the exchangeability of X^{2n} . \square

Lemma 9.5 (Ψ_t is an extractor). *Let X^{2n} be exchangeable. Then $\Psi_t(X^{2n}) \stackrel{i.i.d.}{\sim} \text{Ber}(1/2)$ conditioned on $l(\Psi_t(X^{2n})) = m$.*

9.5 Bernoulli factory 141

Proof. Note that $\Psi_t(X^{2n}) \in \{0, 1\}^*$. It is equivalent to show that for all $s^m \in \{0, 1\}^m$,

$$\mathbb{P}[\Psi_t(X^{2n}) = s^m] = 2^{-m}\mathbb{P}[l(\Psi_t(X^{2n})) = m].$$

Proceed by induction on t . The base case of $t = 1$ follows from Lemma 9.4 (the distribution of the Y part). Assume Ψ_{t-1} is an extractor. Recall that $\Psi_t(X^{2n}) = (\Psi_1(X^{2n}), \Psi_{t-1}(U^n), \Psi_{t-1}(V^{n-k}))$ and write the length as $L = L_1 + L_2 + L_3$, where $L_2 \perp L_3 | L_1$ by Lemma 9.4. Then

$$\begin{aligned} & \mathbb{P}[\Psi_t(X^{2n}) = s^m] \\ &= \sum_{k=0}^m \mathbb{P}[\Psi_t(X^{2n}) = s^m | L_1 = k] \mathbb{P}[L_1 = k] \\ &\stackrel{\text{Lemma 9.4}}{=} \sum_{k=0}^m \sum_{r=0}^{m-k} \mathbb{P}[L_1 = k] \mathbb{P}[Y^k = s^k | L_1 = k] \mathbb{P}[\Psi_{t-1}(U^n) = s_{k+1}^{k+r} | L_1 = k] \mathbb{P}[\Psi_{t-1}(V^{n-k}) = s_{k+r+1}^m | L_1 = k] \\ &\stackrel{\text{induction}}{=} \sum_{k=0}^m \sum_{r=0}^{m-k} \mathbb{P}[L_1 = k] 2^{-k} 2^{-r} \mathbb{P}[L_2 = r | L_1 = k] 2^{-(m-k-r)} \mathbb{P}[L_3 = m - k - r | L_1 = k] \\ &= 2^{-m} \mathbb{P}[L = m]. \end{aligned} \quad \square$$

Next we compute the rate of Ψ_t . Let $X^{2n} \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Then by SLLN, $\frac{1}{2n}l(\Psi_1(X^{2n})) \triangleq \frac{L_n}{2n}$ converges a.s. to pq . Assume, again by induction, that $\frac{1}{2n}l(\Psi_{t-1}(X^{2n})) \xrightarrow{\text{a.s.}} r_{t-1}(p)$, with $r_1(p) = pq$. Then

$$\frac{1}{2n}l(\Psi_t(X^{2n})) = \frac{L_n}{2n} + \frac{1}{2n}l(\Psi_{t-1}(U^n)) + \frac{1}{2n}l(\Psi_{t-1}(V^{n-L_n})).$$

Note that $U^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(2pq)$, $V^{n-L_n} | L_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p^2/(p^2+q^2))$ and $L_n \xrightarrow{\text{a.s.}} \infty$. Then the induction hypothesis implies that $\frac{1}{n}l(\Psi_{t-1}(U^n)) \xrightarrow{\text{a.s.}} r_{t-1}(2pq)$ and $\frac{1}{2(n-L_n)}l(\Psi_{t-1}(V^{n-L_n})) \xrightarrow{\text{a.s.}} r_{t-1}(p^2/(p^2+q^2))$. We obtain the recursion:

$$r_t(p) = pq + \frac{1}{2}r_{t-1}(2pq) + \frac{p^2+q^2}{2}r_{t-1}\left(\frac{p^2}{p^2+q^2}\right) \triangleq (Tr_{t-1})(p), \quad (9.2)$$

where the operator T maps a continuous function on $[0, 1]$ to another. Furthermore, T is monotone in the sense that $f \leq g$ pointwise then $Tf \leq Tg$. Then it can be shown that r_t converges monotonically from below to the fixed point of T , which turns out to be exactly the binary entropy function h . Instead of directly verifying $Th = h$, here is a simple proof: Consider $X_1, X_2 \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Then $2h(p) = H(X_1, X_2) = H(X_1 \oplus X_2, X_1) = H(X_1 \oplus X_2) + H(X_1 | X_1 \oplus X_2) = h(2pq) + 2pqh(\frac{1}{2}) + (p^2+q^2)h(\frac{p^2}{p^2+q^2})$.

The convergence of r_t to h are shown in Fig. 9.1.

9.5 Bernoulli factory

Given a stream of $\text{Ber}(p)$ bits with unknown p , for what kind of function $f : [0, 1] \rightarrow [0, 1]$ can we simulate iid bits from $\text{Ber}(f(p))$. Our discussion above deals with $f(p) \equiv \frac{1}{2}$. The most famous

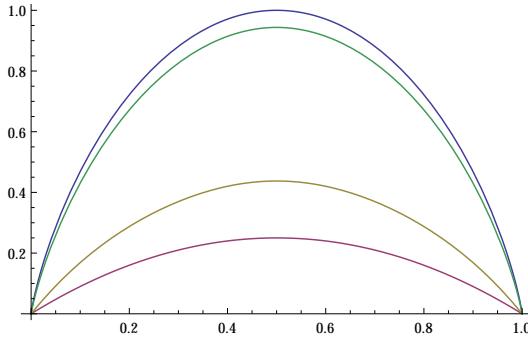


Figure 9.1 Rate function r_t for $t = 1, 4, 10$ versus the binary entropy function.

example is whether we can simulate $\text{Ber}(2p)$ from $\text{Ber}(p)$, i.e., $f(p) = 2p \wedge 1$. Keane and O'Brien [171] showed that all f that can be simulated are either constants or “polynomially bounded away from 0 or 1”: for all $0 < p < 1$, $\min\{f(p), 1-f(p)\} \geq \min\{p, 1-p\}^n$ for some $n \in \mathbb{N}$. In particular, doubling the bias is impossible.

The above result deals with what $f(p)$ can be simulated in principle. What type of computational devices are needed for such a task? Note that since $r_1(p)$ is quadratic in p , all rate functions r_t that arise from the iteration (9.2) are rational functions (ratios of polynomials), converging to the binary entropy function as Fig. 9.1 shows. It turns out that for any rational function f that satisfies $0 < f < 1$ on $(0, 1)$, we can generate independent $\text{Ber}(f(p))$ from $\text{Ber}(p)$ using either of the following schemes with finite memory [216]:

- 1 *Finite-state machine* (FSM): initial state (red), intermediate states (white) and final states (blue, output 0 or 1 then reset to initial state).
- 2 *Block simulation*: let A_0, A_1 be disjoint subsets of $\{0, 1\}^k$. For each k -bit segment, output 0 if falling in A_0 or 1 if falling in A_1 . If neither, discard and move to the next segment. The block size is at most the degree of the denominator polynomial of f .

The next table gives some examples of f that can be realized with these two architectures. (Exercise: How to generate $f(p) = 1/3$?)

It turns out that the only type of f that can be simulated using either FSM or block simulation is rational function. For $f(p) = \sqrt{p}$, which satisfies Keane-O'Brien's characterization, it cannot be simulated by FSM or block simulation, but it can be simulated by the so-called pushdown automata, which is a FSM operating with a stack (infinite memory) [216].

It is unknown how to find the optimal Bernoulli factory with the best rate. Clearly, a converse is the entropy bound $\frac{h(p)}{h(f(p))}$, which can be trivial (bigger than one).

9.5 Bernoulli factory 143

Goal	Block simulation	FSM
$f(p) = 1/2$	$A_0 = 10; A_1 = 01$	<pre> graph LR S(()) -- 0 --> I(()) S -- 1 --> G1(((1))) I -- 0 --> G1 I -- 1 --> S G1 -- 1 --> S </pre>
$f(p) = 2pq$	$A_0 = 00, 11; A_1 = 01, 10$	<pre> graph LR S(()) -- 0 --> I(()) S -- 1 --> G0(((0))) I -- 0 --> G0 I -- 1 --> G1(((1))) G0 -- 1 --> S G1 -- 1 --> S G1 -- 0 --> G0 </pre>
$f(p) = \frac{p^3}{p^3+q^3}$	$A_0 = 000; A_1 = 111$	<pre> graph LR S(()) -- 0 --> I1(()) S -- 0 --> I2(()) S -- 1 --> G00(((0))) I1 -- 0 --> G00 I1 -- 1 --> G01(((1))) I2 -- 0 --> G00 I2 -- 1 --> G01 G00 -- 1 --> S G01 -- 1 --> S G01 -- 0 --> G00 G01 -- 1 --> G00 </pre>

Table 9.1 Bernoulli factories realized by FSM or block simulation.

Exercises for Part I

I.1 (Combinatorial meaning of entropy)

1 Fix $n \geq 1$ and $0 \leq k \leq n$. Let $p = \frac{k}{n}$ and define $T_p \subset \{0, 1\}^n$ to be the set of all binary sequences with p fraction of ones. Show that if $k \in [1, n - 1]$ then

$$|T_p| = \sqrt{\frac{1}{np(1-p)}} \exp\{nh(p)\} C(n, k)$$

where $C(n, k)$ is bounded by two universal constants $C_0 \leq C(n, k) \leq C_1$, and $h(\cdot)$ is the binary entropy. Conclude that for all $0 \leq k \leq n$ we have

$$\log |T_p| = nh(p) + O(\log n).$$

Hint: Stirling's approximation:

$$e^{\frac{1}{12n+1}} \leq \frac{n!}{\sqrt{2\pi n}(n/e)^n} \leq e^{\frac{1}{12n}}, \quad n \geq 1 \quad (\text{I.1})$$

2 Let $Q^n = \text{Bern}(q)^n$ be iid Bernoulli distribution on $\{0, 1\}^n$. Show that

$$\log Q^n[T_p] = -nd(p\|q) + O(\log n)$$

3* More generally, let \mathcal{X} be a finite alphabet, \hat{P}, Q distributions on \mathcal{X} , and $T_{\hat{P}}$ a set of all strings in \mathcal{X}^n with composition \hat{P} . If $T_{\hat{P}}$ is non-empty (i.e. if $n\hat{P}(\cdot)$ is integral) then

$$\begin{aligned} \log |T_{\hat{P}}| &= nH(\hat{P}) + O(\log n) \\ \log Q^n[T_{\hat{P}}] &= -nD(\hat{P}\|Q) + O(\log n) \end{aligned}$$

and furthermore, both $O(\log n)$ terms can be bounded as $|O(\log n)| \leq |\mathcal{X}| \log(n+1)$. (Hint: show that number of non-empty $T_{\hat{P}}$ is $\leq (n+1)^{|\mathcal{X}|}$.)

1.2 (Refined method of types) The following refines Proposition 1.6. Let n_1, \dots be non-negative integers with $\sum_i n_i = n$ and let k_+ be the number of non-zero n_i 's. Then

$$\log \binom{n}{n_1, n_2, \dots} = nH(\hat{P}) - \frac{k_+ - 1}{2} \log(2\pi n) - \frac{1}{2} \sum_{i:n_i>0} \log \hat{P}_i - C_{k_+},$$

where $\hat{P}_i = \frac{n_i}{n}$ and $0 \leq C_{k_+} \leq \frac{\log e}{12}$. (Hint: use (I.1)).

1.3 (Conditional entropy and Markov types)

(a) Fix $n \geq 1$, a sequence $x^n \in \mathcal{X}^n$ and define

$$N_{x^n}(a, b) = |\{(x_i, x_{i+1}) : x_i = a, x_{i+1} = b, i = 1, \dots, n\}|,$$

Exercises for Part I 145

where we define $x_{n+1} = x_1$ (cyclic continuation). Show that $\frac{1}{n}N_{x^n}(\cdot, \cdot)$ defines a probability distribution $P_{A,B}$ on $\mathcal{X} \times \mathcal{X}$ with equal marginals $P_A = P_B$. Conclude that $H(A|B) = H(B|A)$.

Is $P_{A|B} = P_{B|A}$?

(b) Let $T_{x^n}^{(2)}$ (Markov type-class of x^n) be defined as

$$T_{x^n}^{(2)} = \{\tilde{x}^n \in \mathcal{X}^n : N_{\tilde{x}^n} = N_{x^n}\}.$$

Show that elements of $T_{x^n}^{(2)}$ can be identified with cycles in the complete directed graph G on \mathcal{X} , such that for each $(a, b) \in \mathcal{X} \times \mathcal{X}$ the cycle passes $N_{x^n}(a, b)$ times through edge (a, b) .

(c) Show that each such cycle can be uniquely specified by identifying the first node and by choosing at each vertex of the graph the order in which the outgoing edges are taken. From this and Stirling's approximation conclude that

$$\log |T_{x^n}^{(2)}| = nH(x_{T+1}|x_T) + O(\log n), \quad T \sim \text{Unif}([n]).$$

Check that $H(x_{T+1}|x_T) = H(A|B) = H(B|A)$.

(d) Show that for any time-homogeneous Markov chain X^n with $P_{X_1, X_2}(a_1, a_2) > 0 \forall a_1, a_2 \in \mathcal{X}$ we have

$$\log P_{X^n}(X^n \in T_{x^n}^{(2)}) = -nD(P_{B|A} \| P_{X_2|X_1}P_A) + O(\log n).$$

I.4 Find the entropy rate of a stationary ergodic Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix}$$

I.5 Let $\mathbb{X} = X_0^\infty$ be a stationary Markov chain. Let $P_{Y|X}$ be a Markov kernel. Define a new process $\mathbb{Y} = Y_0^\infty$ where $Y_i \sim P_{Y|X=X_i}$ conditionally independent of all other $X_j, j \neq i$. Prove that

$$H(Y_n|Y_2^{n-1}, X_1) \leq H(\mathbb{Y}) \leq H(Y_n|Y_1^{n-1}) \tag{I.2}$$

and

$$\lim_{n \rightarrow \infty} H(Y_n|Y_2^{n-1}, X_1) = H(\mathbb{Y}) = \lim_{n \rightarrow \infty} H(Y_n|Y_1^{n-1}). \tag{I.3}$$

I.6 (Robust version of the maximal entropy) Maximal differential entropy among all variables X supported on $[-b, b]$ is attained by a uniform distribution. Prove that as $\epsilon \rightarrow 0+$ we have

$$\sup\{h(M+Z) : M \in [-b, b], \mathbb{E}[Z] = 0, \text{Var}[Z] \leq \epsilon\} = \log(2b) + o(1).$$

where supremization is over all (not necessarily independent) random variables M, Z such that $M+Z$ possesses a density. (Hint: [118, Appendix C] proves $o(1) = O(\epsilon^{1/3} \log \frac{1}{\epsilon})$ bound.)

I.7 (Maximum entropy.) Prove that for any X taking values on $\mathbb{N} = \{1, 2, \dots\}$ such that $\mathbb{E}[X] < \infty$,

$$H(X) \leq \mathbb{E}[X]h\left(\frac{1}{\mathbb{E}[X]}\right),$$

146 Exercises for Part I

maximized uniquely by the geometric distribution. Here as usual $h(\cdot)$ denotes the binary entropy function. *Hint:* Find an appropriate Q such that RHS - LHS = $D(P_X||Q)$.

- I.8** (Finiteness of entropy) We have shown that any \mathbb{N} -valued random variable X , with $\mathbb{E}[X] < \infty$ has $H(X) \leq \mathbb{E}[X]h(1/\mathbb{E}[X]) < \infty$. Next let us improve this result.

- (a) Show that $\mathbb{E}[\log X] < \infty \Rightarrow H(X) < \infty$.

Moreover, show that the condition of X being integer-valued is not superfluous by giving a counterexample.

- (b) Show that if $k \mapsto P_X(k)$ is a decreasing sequence, then $H(X) < \infty \Rightarrow \mathbb{E}[\log X] < \infty$.

Moreover, show that the monotonicity assumption is not superfluous by giving a counterexample.

- I.9** (Maximum entropy under Hamming weight constraint.) For any $\alpha \leq 1/2$ and $d \in \mathbb{N}$,

$$\max\{H(Y) : Y \in \{0, 1\}^d, \mathbb{E}[w_H(Y)] \leq \alpha d\} = dh(\alpha),$$

achieved by the product distribution $Y \sim \text{Ber}(\alpha)^{\otimes d}$. *Hint:* Find an appropriate Q such that RHS - LHS = $D(P_Y||Q)$.

- I.10** Let $\mathcal{N}(\mathbf{m}, \Xi)$ be the Gaussian distribution on \mathbb{R}^n with mean $\mathbf{m} \in \mathbb{R}^n$ and covariance matrix Ξ .

- (a) Under what conditions on $\mathbf{m}_0, \Xi_0, \mathbf{m}_1, \Xi_1$ is

$$D(\mathcal{N}(\mathbf{m}_1, \Xi_1) \parallel \mathcal{N}(\mathbf{m}_0, \Xi_0)) < \infty$$

- (b) Compute $D(\mathcal{N}(\mathbf{m}, \Xi) \parallel \mathcal{N}(0, \mathbf{I}_n))$, where \mathbf{I}_n is the $n \times n$ identity matrix.

- (c) Compute $D(\mathcal{N}(\mathbf{m}_1, \Xi_1) \parallel \mathcal{N}(\mathbf{m}_0, \Xi_0))$ for non-singular Ξ_0 . (Hint: think how Gaussian distribution changes under shifts $\mathbf{x} \mapsto \mathbf{x} + \mathbf{a}$ and non-singular linear transformations $\mathbf{x} \mapsto \mathbf{Ax}$. Apply data-processing to reduce to previous case.)

- I.11** (Information lost in erasures) Let X, Y be a pair of random variables with $I(X; Y) < \infty$. Let Z be obtained from Y by passing the latter through an erasure channel, i.e., $X \rightarrow Y \rightarrow Z$ where

$$P_{Z|Y}(z|y) = \begin{cases} 1 - \delta, & z = y, \\ \delta, & z = ? \end{cases}$$

where $?$ is a symbol not in the alphabet of Y . Find $I(X; Z)$.

- I.12** (Information bottleneck) Let $X \rightarrow Y \rightarrow Z$ where Y is a discrete random variable taking values on a finite set \mathcal{Y} . Prove that

$$I(X; Z) \leq \log |\mathcal{Y}|.$$

- I.13** The Hewitt-Savage 0-1 law states that certain symmetric events have no randomness. Let $\{X_i\}_{i \geq 1}$ be a sequence of iid random variables. Let E be an event determined by this sequence. We say E is exchangeable if it is invariant under permutation of finitely many indices in the sequence of $\{X_i\}$'s, e.g., the occurrence of E is unchanged if we permute the values of (X_1, X_4, X_7) , etc.

Let's prove the Hewitt-Savage 0-1 law information-theoretically in the following steps:

- (a) (Warm-up) Verify that $E = \{\sum_{i \geq 1} X_i \text{ converges}\}$ and $E = \{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mathbb{E}[X_1]\}$ are exchangeable events.

Exercises for Part I 147

- (b) Let E be an exchangeable event and $W = 1_E$ is its indicator random variable. Show that for any k , $I(W; X_1, \dots, X_k) = 0$. (Hint: Use tensorization (6.2) to show that for arbitrary n , $nI(W; X_1, \dots, X_k) \leq 1$ bit.)
- (c) Since E is determined by the sequence $\{X_i\}_{i \geq 1}$, we have by continuity of mutual information:

$$H(W) = I(W; X_1, \dots) = \lim_{k \rightarrow \infty} I(W; X_1, \dots, X_k) = 0.$$

Conclude that E has no randomness, i.e., $P(E) = 0$ or $P(E) = 1$.

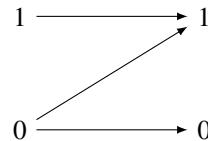
- (d) (Application to random walk) Often after the application of Hewitt-Savage, further efforts are needed to determine whether the probability is 0 or 1. As an example, consider X_i 's are iid ± 1 and $S_n = \sum_{i=1}^n X_i$ denotes the symmetric random walk. Verify that the event $E = \{S_n = 0 \text{ finitely often}\}$ is exchangeable. Now show that $P(E) = 0$. (Hint: consider $E^+ = \{S_n > 0 \text{ eventually}\}$ and E^- similarly. Apply Hewitt-Savage to them and invoke symmetry.)

- I.14** Conditioned on $X = x$, let Y be Poisson with mean x , i.e.,

$$P_{Y|X}[k|x] = e^{-x} \frac{x^k}{k!}, \quad k = 0, 1, 2, \dots$$

Let X be an exponential random variable with unit mean. Find $I(X; Y)$.

- I.15** Consider the following *Z-channel* given by $P_{Y|X}[1|1] = 1$ and $P_{Y|X}[1|0] = P_{Y|X}[0|0] = 1/2$.



- (a) Find the capacity

$$C = \max_X I(X; Y).$$

- (b) Find $D(P_{Y|X=0} \| P_Y^*)$ and $D(P_{Y|X=1} \| P_Y^*)$ where P_Y^* is the capacity-achieving output distribution, or *caod*, i.e., the distribution of Y induced by the maximizer of $I(X; Y)$.

- I.16** (a) For any X such that $\mathbb{E}[|X|] < \infty$, show that

$$D(P_X \| \mathcal{N}(0, 1)) \geq \frac{(\mathbb{E}[X])^2}{2} \quad \text{nats.}$$

- (b) For $a > 0$, find the minimum and minimizer of

$$\min_{P_X: \mathbb{E}[X] \geq a} D(P_X \| \mathcal{N}(0, 1)).$$

Is the minimizer unique? Why?

148 Exercises for Part I

I.17 (Entropy numbers and capacity.) Let $\{P_{Y|X=x} : x \in \mathcal{X}\}$ be a set of distributions and let $C = \sup_{P_X} I(X; Y)$ be its capacity. For every $\epsilon \geq 0$, define¹

$$N(\epsilon) = \min\{k : \exists Q_1 \dots Q_k : \forall x \in \mathcal{X}, \min_j D(P_{Y|X=x} \| Q_j) \leq \epsilon^2\}. \quad (\text{I.4})$$

(a) Prove that

$$C = \inf_{\epsilon \geq 0} (\epsilon^2 + \log N(\epsilon)). \quad (\text{I.5})$$

(Hint: when is $N(\epsilon) = 1$? See Theorem 32.4.)

(b) Similarly, show

$$I(X; Y) = \inf_{\epsilon \geq 0} (\epsilon + \log N(\epsilon; P_X)),$$

where the average-case covering number is

$$N(\epsilon; P_X) = \min\{k : \exists Q_1 \dots Q_k : \mathbb{E}_{x \sim P_X} [\min_j D(P_{Y|X=x} \| Q_j)] \leq \epsilon\} \quad (\text{I.6})$$

Comments: The reason these estimates are useful is because $N(\epsilon)$ for small ϵ roughly speaking depends on local (differential) properties of the map $x \mapsto P_{Y|X=x}$, unlike C which is global.

I.18 Consider the channel $P_{Y^m|X} : [0, 1] \mapsto \{0, 1\}^m$, where given $x \in [0, 1]$, Y^m is i.i.d. $\text{Ber}(x)$. Using the upper bound from Ex. I.17 prove

$$C(m) \triangleq \max_{P_X} I(X; Y^m) \leq \frac{1}{2} \log m + O(1), \quad m \rightarrow \infty.$$

Hint: Find a covering of the input space.

Show a lower bound to establish

$$C(m) \geq \frac{1}{2} \log m + o(\log m), \quad m \rightarrow \infty.$$

You may use without proof that $\forall \epsilon > 0$ there exists $K(\epsilon)$ such that for all $m \geq 1$ and all $p \in [\epsilon, 1 - \epsilon]$ we have $|H(\text{Binom}(m, p)) - \frac{1}{2} \log m| \leq K(\epsilon)$.

I.19 Show that

$$P_{Y_1 \dots Y_n | X_1 \dots X_n} = \prod_{i=1}^n P_{Y_i | X_i} \quad (\text{I.7})$$

if and only if for all $i = 1, \dots, n$,

$$Y_i \rightarrow X_i \rightarrow (X_{\setminus i}, Y_{\setminus i}) \quad (\text{I.8})$$

where $X_{\setminus i} = \{X_j, j \neq i\}$.

¹ $N(\epsilon)$ is the minimum number of points that cover the set $\{P_{Y|X=x} : x \in \mathcal{X}\}$ to within ϵ in divergence; $\log N(\epsilon)$ would be called (Kolmogorov) metric ϵ -entropy of the set $\{P_{Y|X=x} : x \in \mathcal{X}\}$ – see Chapter 27.

Exercises for Part I 149

- I.20** Suppose Z_1, \dots, Z_n are independent Poisson random variables with mean λ . Show that $\sum_{i=1}^n Z_i$ is a sufficient statistic of (Z_1, \dots, Z_n) for λ .
- I.21** Suppose Z_1, \dots, Z_n are independent uniformly distributed on the interval $[0, \lambda]$. Show that $\max_{1 \leq i \leq n} Z_i$ is a sufficient statistic of (Z_1, \dots, Z_n) for λ .
- I.22** Consider a binary symmetric random walk X_n on \mathbb{Z} that starts at zero. In other words, $X_n = \sum_{j=1}^n B_j$, where (B_1, B_2, \dots) are independent and equally likely to be ± 1 .
- (a) When $n \gg 1$ does knowing X_{2n} provide any information about X_n ? More exactly, prove

$$\liminf_{n \rightarrow \infty} I(X_n; X_{2n}) > 0.$$

(Hint: lower-semicontinuity and central-limit theorem)

(b) Bonus: Compute the exact value of the limit

$$\lim_{n \rightarrow \infty} I(X_n; X_{2n}).$$

- I.23** (Continuity of entropy on finite alphabet.) We have shown that entropy is continuous on on finite alphabet. Now let us study how continuous it is with respect to the total variation. Prove

$$|H(P) - H(Q)| \leq h(\text{TV}(P, Q)) + \text{TV}(P, Q) \log(|\mathcal{X}| - 1)$$

for any P and Q supported on \mathcal{X} .

Hint: Use Fano's inequality and the inf-representation (over coupling) of total variation in Theorem 7.12(a).

- I.24** Distributions and graphical models:

- (a) Draw all possible directed acyclic graphs (DAGs, or graphical models) compatible with the following distribution on $X, Y, Z \in \{0, 1\}$:

$$P_{X,Z}(x, z) = \begin{cases} 1/6, & x = 0, z \in \{0, 1\}, \\ 1/3, & x = 1, z \in \{0, 1\} \end{cases} \quad (\text{I.9})$$

$$Y = X + Z \pmod{2} \quad (\text{I.10})$$

You may include only the minimal DAGs (recall: the DAG is minimal for a given distribution if removal of any edge leads to a graphical model incompatible with the distribution).²

- (b) Draw the DAG describing the set of distributions $P_{X^n Y^n}$ satisfying:

$$P_{Y^n | X^n} = \prod_{i=1}^n P_{Y_i | X_i}$$

- (c) Recall that two DAGs G_1 and G_2 are called equivalent if they have the same vertex sets and each distribution factorizes w.r.t. G_1 if and only if it does so w.r.t. G_2 . For example, it is

² Note: $\{X \rightarrow Y\}$, $\{X \leftarrow Y\}$ and $\{X \perp Y\}$ are the three possible directed graphical modelss for two random variables. For example, the third graph describes the set of distributions for which X and Y are independent: $P_{XY} = P_X P_Y$. In fact, $P_X P_Y$ factorizes according to any of the three DAGs, but $\{X \perp Y\}$ is the unique minimal DAG.

150 Exercises for Part I

well known

$$X \rightarrow Y \rightarrow Z \iff X \leftarrow Y \leftarrow Z \iff X \leftarrow Y \rightarrow Z.$$

Consider the following two DAGs with countably many vertices:

$$X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n \rightarrow \dots$$

$$X_1 \leftarrow X_2 \leftarrow \dots \leftarrow X_n \leftarrow \dots$$

Are they equivalent?

I.25 Give a necessary and sufficient condition for

$$A \rightarrow B \rightarrow C$$

for jointly Gaussian (A, B, C) in terms of correlation coefficients.

For discrete (A, B, C) denote $x_{abc} = P_{ABC}(a, b, c)$ and write the Markov chain condition as a list of degree-2 polynomial equations in $\{x_{abc}, a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}\}$.

I.26 Let A, B, C be discrete with $P_{C|B}(c|b) > 0 \forall b, c$. Show

$$\begin{array}{l} A \rightarrow B \rightarrow C \\ A \rightarrow C \rightarrow B \end{array} \implies A \perp\!\!\!\perp (B, C)$$

Discuss implications for sufficient statistics.

Bonus: for binary (A, B, C) characterize all counter-examples.

Comment: Thus, a popular positivity condition $P_{ABC} > 0$ allows to infer conditional independence relations, which are not true in general. *Wisdom:* This example demonstrates that a set of distributions satisfying certain (conditional) independence relations does not equal to the closure of its intersection with $\{P_{ABC} > 0\}$.

I.27 Show that for jointly gaussian (A, B, C)

$$I(A; C) = I(B; C) = 0 \implies I(A, B; C) = 0. \quad (\text{I.11})$$

Find a counter-example for general (A, B, C) .

Prove or disprove: Implication (I.11) also holds for arbitrary discrete (A, B, C) under positivity condition $P_{ABC}(a, b, c) > 0 \forall abc$.

I.28 Let $I_{\text{TV}}(X; Y) = \text{TV}(P_{X,Y}, P_X P_Y)$. Let $X \sim \text{Ber}(1/2)$ and conditioned on X generate A and B independently setting them equal to X or $1 - X$ with probabilities $1 - \delta$ and δ , respectively (i.e. $A \leftarrow X \rightarrow B$). Show

$$I_{\text{TV}}(X; A, B) = I_{\text{TV}}(X; A) = \left| \frac{1}{2} - \delta \right|.$$

This means the second observation of X is “uninformative” (in the I_{TV} sense).

Similarly, show that when $X \sim \text{Ber}(\delta)$ for $\delta < 1/2$ there exists joint distribution $P_{X,Y}$ so that $\text{TV}(P_{Y|X=0}, P_{Y|X=1}) > 0$ (thus $I_{\text{TV}}(X; Y)$ and $I(X; Y)$ are strictly positive), but at the same time $\min_{\hat{X}(Y)} \mathbb{P}[X \neq \hat{X}] = \delta$. In other words, observation Y is informative about X , but does not improve the probability of error.

Exercises for Part I 151

I.29 (Rényi divergences and Blackwell order) Let $p_\epsilon = \frac{e^\epsilon}{1+e^\epsilon}$. Show that for all $\epsilon > 0$ and all $\alpha > 0$ we have

$$D_\alpha(\text{Ber}(p_\epsilon) \| \text{Ber}(1 - p_\epsilon)) < D_\alpha(\mathcal{N}(\epsilon, 1) \| \mathcal{N}(0, 1)).$$

Yet, for small enough ϵ we have

$$\text{TV}(\text{Ber}(p_\epsilon), \text{Ber}(1 - p_\epsilon)) > \text{TV}(\mathcal{N}(\epsilon, 1), \mathcal{N}(0, 1)).$$

Note: This shows that domination under all Rényi divergences does not imply a similar comparison in other f -divergences [?]. On the other hand, we have the equivalence [217]:

$$\begin{aligned} \forall \alpha > 0 : D_\alpha(P_1 \| P_0) \leq D_\alpha(Q_1 \| Q_0) \\ \iff \exists n_0 \forall n \geq n_0 \forall f : D_f(P_1^{\otimes n} \| P_0^{\otimes n}) \leq D_f(Q_1^{\otimes n} \| Q_0^{\otimes n}). \end{aligned}$$

(The latter is also equivalent to existence of a kernel K_n such that $K_n \circ P_i^{\otimes n} = Q_i^{\otimes n}$ – a so-called Blackwell order on pairs of measures).

I.30 (Rényi divergence as KL [271]) Show for all $\alpha \in \mathbb{R}$:

$$(1 - \alpha)D_\alpha(P \| Q) = \inf_R (\alpha D(R \| P) + (1 - \alpha)D(R \| Q)). \quad (\text{I.12})$$

Whenever the LHS is finite, derive the explicit form of a unique minimizer R .

I.31 For an f -divergence, consider the following statements:

- (i) If $I_f(X; Y) = 0$, then $X \perp\!\!\!\perp Y$.
- (ii) If $X - Y - Z$ and $I_f(X; Y) = I_f(X; Z) < \infty$, then $X - Z - Y$.

Recall that $f: (0, \infty) \rightarrow \mathbb{R}$ is a convex function with $f(1) = 0$.

- (a) Choose an f -divergence which is not a multiple of the KL divergence (i.e., f cannot be of form $c_1 x \log x + c_2(x - 1)$ for any $c_1, c_2 \in \mathbb{R}$). Prove both statements for I_f .
- (b) Choose an f -divergence which is non-linear (i.e., f cannot be of form $c(x - 1)$ for any $c \in \mathbb{R}$) and provide examples that violate (i) and (ii).
- (c) Choose an f -divergence. Prove that (i) holds, and provide an example that violates (ii).

I.32 (Chain rules I)

- (a) Show using (I.12) and the chain rule for KL that

$$(1 - \alpha)D_\alpha(P_{X^n} \| Q_{X^n}) \geq \sum_{i=1}^n \inf_a (1 - \alpha)D_\alpha(P_{X_i | X^{i-1}=a} \| Q_{X_i | X^{i-1}=a})$$

- (b) Derive two special cases:

$$\begin{aligned} 1 - \frac{1}{2}H^2(P_{X^n}, Q_{X^n}) &\leq \prod_{i=1}^n \sup_a (1 - \frac{1}{2}H^2(P_{X_i | X^{i-1}=a}, Q_{X_i | X^{i-1}=a})) \\ 1 + \chi^2(P_{X^n} \| Q_{X^n}) &\leq \prod_{i=1}^n \sup_a (1 + \chi^2(P_{X_i | X^{i-1}=a} \| Q_{X_i | X^{i-1}=a})) \end{aligned}$$

I.33 (Chain rules II)

152 Exercises for Part I

(a) Show that the chain rule for divergence can be restated as

$$D(P_{X^n} \| Q_{X^n}) = \sum_{i=1}^n D(P_i \| P_{i-1}),$$

where $P_i = P_{X^i} Q_{X_{i+1}^n | X^i}$, with $P_n = P_{X^n}$ and $P_0 = Q_{X^n}$. The identity above shows how KL-distance from P_{X^n} to Q_{X^n} can be traversed by summing distances between intermediate P_i 's.

(b) Using the same path and triangle inequality show that

$$\text{TV}(P_{X^n}, Q_{X^n}) \leq \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}} \text{TV}(P_{X_i | X^{i-1}}, Q_{X_i | X^{i-1}})$$

(c) Similarly, show for the Hellinger distance H :

$$H(P_{X^n}, Q_{X^n}) \leq \sum_{i=1}^n \sqrt{\mathbb{E}_{P_{X^{i-1}}} H^2(P_{X_i | X^{i-1}}, Q_{X_i | X^{i-1}})}$$

I.34 (a) Define *Marton's divergence*

$$D_m(P \| Q) = \int dQ \left(1 - \frac{dP}{dQ} \right)_+^2.$$

Prove that

$$D_m(P \| Q) = \inf_{P_{XY}} \{ \mathbb{E}[P[X \neq Y|Y]^2] : P_X = P, P_Y = Q \}$$

where the infimum is over all couplings. (Hint: For one direction use the same coupling achieving TV. For the other direction notice that $\mathbb{P}[X \neq Y|Y] \geq 1 - \frac{P(Y)}{Q(Y)}$.)

(b) Define *symmetrized Marton's divergence*

$$D_{sm}(P \| Q) = D_m(P \| Q) + D_m(Q \| P).$$

Prove that

$$D_{sm}(P \| Q) = \inf_{P_{XY}} \{ \mathbb{E}[\mathbb{P}^2[X \neq Y|Y]] + \mathbb{E}[\mathbb{P}^2[X \neq Y|X]] : P_X = P, P_Y = Q \}.$$

I.35 (Center of gravity under f -divergences.) Recall from Corollary 4.1 the fact that

$$\min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) = I(X; Y)$$

achieved at $Q_Y = P_Y$. Prove the following extensions to other f -divergences:

(a) Suppose that for P_X -a.e. x , $P_{Y|X=x} \ll \mu$ with density $p(y|x)$.³ Then

$$\min_{Q_Y} \chi^2(P_{Y|X} \| Q_Y | P_X) = \left(\int \mu(dy) \sqrt{\mathbb{E}[p_{Y|X}(y|X)^2]} \right)^2 - 1. \quad (\text{I.13})$$

If the right-hand side is finite, the minimum is achieved at $Q_Y(dy) \propto \sqrt{\mathbb{E}[p(y|X)^2]} \mu(dy)$.

³ Note that the results do not depend on the choice of μ , so we can take for example $\mu = P_Y$, in view of Lemma 3.3.

(b) Show that

$$\min_{Q_Y} D(Q_Y \| P_{Y|X} | P_X) = \log \int \mu(dy) \exp(\mathbb{E}[\log p(y|X)]). \quad (\text{I.14})$$

If the right-hand side is finite, the minimum is achieved at $Q_Y(dy) \propto \exp(\mathbb{E}[\log p(y|X)])\mu(dy)$.

Note: This exercise shows that the center of gravity with respect to other f -divergences need not be P_Y but its reweighted version. For statistical applications, see Exercise VI.6 and Exercise VI.9, where (I.13) is used to determine the form of the Bayes estimator.

- I.36** Let (X, Y) be uniformly distributed in the unit ℓ_p -ball $B_p \triangleq \{(x, y) : |x|^p + |y|^p \leq 1\}$, where $p \in (0, \infty)$. Also define the ℓ_∞ -ball $B_\infty \triangleq \{(x, y) : |x| \leq 1, |y| \leq 1\}$.

- (a) Compute $I(X; Y)$ for $p = 1/2$, $p = 1$ and $p = \infty$.
(b) (Bonus) What do you think $I(X; Y)$ converges to as $p \rightarrow 0$. Can you prove it?

- I.37** (Divergence of order statistics) Given $x^n = (x_1, \dots, x_n) \in \mathbb{R}^n$, let $x_{(1)} \leq \dots \leq x_{(n)}$ denote the ordered entries. Let P, Q be distributions on \mathbb{R} and $P_{X^n} = P^n, Q_{X^n} = Q^n$.

- (a) Prove that

$$D(P_{X_{(1)}, \dots, X_{(n)}} \| Q_{X_{(1)}, \dots, X_{(n)}}) = nD(P \| Q). \quad (\text{I.15})$$

- (b) Show that

$$D(\text{Bin}(n, p) \| \text{Bin}(n, q)) = nd(p \| q).$$

- I.38** (Sampling without replacement I, [285]) Consider two ways of generating a random vector $X^n = (X_1, \dots, X_n)$: Under P , X^n are sampled from the set $[n] = \{1, \dots, n\}$ without replacement; under Q , X^n are sampled from $[n]$ with replacement. Let's compare the joint distribution of the first k draws X_1, \dots, X_k for some $1 \leq k \leq n$.

- (a) Show that

$$\begin{aligned} \text{TV}(P_{X^k}, Q_{X^k}) &= 1 - \frac{k!}{n^k} \binom{n}{k} \\ D(P_{X^k} \| Q_{X^k}) &= -\log \frac{k!}{n^k} \binom{n}{k}. \end{aligned}$$

Conclude that D and TV are $o(1)$ iff $k = o(\sqrt{n})$. You may use the fact that TV between two discrete distributions is equal to half the ℓ_1 -distance between their PMFs.

- (b) Explain the specialness of \sqrt{n} by find an explicit test that distinguishes P and Q with high probability when $k \gg \sqrt{n}$. Hint: Birthday problem.
I.39 (Sampling without replacement II, [285]) Let X_1, \dots, X_k be a random sample of balls without replacement from an urn containing a_i balls of color $i \in [q]$, $\sum_{i=1}^q a_i = n$. Let $Q_X(i) = \frac{a_i}{n}$. Show that

$$D(P_{X^k} \| Q_X^k) \leq c \frac{k^2(q-1)}{(n-1)(n-k+1)}, \quad c = \frac{\log e}{2}.$$

Let R_{m,b_0,b_1} be the distribution of the number of 1's in the first $m \leq b_0 + b_1$ coordinates of a randomly permuted binary strings with b_0 zeros and b_1 ones.

154 Exercises for Part I

(a) Show that

$$D(P_{X_{m+1}|X^m} \| Q_X | P_{X^m}) = \sum_{i=1}^q \mathbb{E} \left[\frac{a_i - V_i}{N-m} \log \frac{a_i - V_i}{p_i(N-m)} \right],$$

where $V_i \sim R_{m,N-a_i,a_i}$.

(b) Show that the i -th term above also equals $p_i \mathbb{E}[\log \frac{a_i - \tilde{V}_i}{p_i(N-m)}]$, $\tilde{V}_i \sim R_{m,N-a_i,a_i-1}$.

(c) Use Jensen's inequality to show that the i -th term is upper bounded by $p_i \log \left(1 + \frac{m}{(n-1)(N-m)} \frac{1-p_i}{p_i} \right)$.

(d) Use the bound $\log(1+x) \leq x \log e$ to complete the proof.

I.40 (Effective de Finetti) We will show that for any distribution P_{X^n} invariant to permutation and $k < n$ there exists a mixture of iid distributions Q_{X^k} which approximates P_{X^k} :

$$\text{TV}(P_{X^k}, Q_{X^k}) \leq c \sqrt{\frac{k^2 H(X_1)}{n-k+1}}, \quad Q_{X_k} = \sum_{i=1}^m \lambda_i (Q_i)^n \quad (\text{I.16})$$

where $\sum_{i=1}^m \lambda_i = 1$, $\lambda_i \geq 0$ and Q_i are some distributions on \mathcal{X} and $c > 0$ is a universal constant.

Follow the steps:

(a) Show the identity (here P_{X^k} is arbitrary)

$$D\left(P_{X^k} \middle\| \prod_{j=1}^k P_{X_j}\right) = \sum_{j=1}^{k-1} I(X^j; X_{j+1}).$$

(b) Show that there must exist some $t \in \{k, k+1, \dots, n\}$ such that

$$I(X^{k-1}; X_k | X_{t+1}^n) \leq \frac{H(X^{k-1})}{n-k+1}.$$

(Hint: Expand $I(X^{k-1}; X_k^n)$ via chain rule.)

(c) Show from 1 and 2 that

$$D\left(P_{X^k|T} \middle\| \prod P_{X_j|T} | P_T\right) \leq \frac{kH(X^{k-1})}{n-k+1}$$

where $T = X_{t+1}^n$.

(d) By Pinsker's inequality

$$\mathbb{E}_T \left[\text{TV}\left(P_{X^k|T}, \prod P_{X_j|T}\right) \right] \leq c \sqrt{\frac{kH(X^{k-1})|\mathcal{X}|}{n-k+1}}, \quad c = \frac{1}{\sqrt{2 \log e}}.$$

Conclude the proof of (I.16) by convexity of total variation.

Note: Another estimate [285, 89] is easy to deduce from Exercise I.39 and Exercise I.38: there exists a mixture of iid Q_{X^k} such that

$$\text{TV}(Q_{X^k}, P_{X^k}) \leq \frac{k}{n} \min(2|\mathcal{X}|, k-1).$$

The bound (I.16) improves the above only when $H(X_1) \lesssim 1$.

Exercises for Part I 155

- I.41** (Wringing Lemma [104, 300]) Prove that for any $\delta > 0$ and any (U^n, V^n) there exists an index set $I \subset [n]$ of size $|I| \leq \frac{I(U^n; V^n)}{\delta}$ such that

$$I(U_t; V_t | U_I, V_I) \leq \delta \quad \forall t \in [n].$$

When $I(U^n; V^n) \ll n$, this shows that conditioning on a (relatively few) entries, one can make individual coordinates almost independent. (Hint: Show $I(A; B; C, D) \geq I(A; C) + I(B; D|A, C)$ first. Then start with $I = \emptyset$ and if there is any index t s.t. $I(U_t; V_t | U_I, V_I) > \delta$ then add it to I and repeat.)

- I.42** This exercise shows other ways of proving the Fano's inequality in its various forms.

- (a) Prove (6.5) as follows. Given any $P = (P_{\max}, P_2, \dots, P_M)$, apply a random permutation π to the last $M - 1$ atoms to obtain the distribution P_π . By comparing $H(P)$ and $H(Q)$, where Q is the average of P_π over all permutations, complete the proof.
- (b) Prove (6.5) by directly solving the convex optimization $\max\{H(P) : 0 \leq p_i \leq P_{\max}, i = 1, \dots, M, \sum_i p_i = 1\}$.
- (c) Prove (6.9) as follows. Let $P_e = \mathbb{P}[X \neq \hat{X}]$. First show that

$$I(X; Y) \geq I(X; \hat{X}) \geq \min_{P_{Z|X}} \{I(P_X, P_{Z|X}) : \mathbb{P}[X = Z] \geq 1 - P_e\}.$$

Notice that the minimum is non-zero unless $P_e = P_{\max}$. Second, solve the stated convex optimization problem. (Hint: look for invariants that the matrix $P_{Z|X}$ must satisfy under permutations $(X, Z) \mapsto (\pi(X), \pi(Z))$ then apply the convexity of $I(P_X, \cdot)$).

- I.43** (Generalization gap = I_{SKL} , [12]) A learning algorithm selects a parameter W based on observing (not necessarily independent) samples S_1, \dots, S_n , where all S_i have a common marginal law P_S , with the goal of minimizing the loss on a fresh sample = $\mathbb{E}[\ell(W, S)]$, where $S^n \perp\!\!\!\perp S \sim P_S$ and ℓ is an arbitrary loss function⁴. Consider a Gibbs algorithm (generalizing ERM and various regularizations) which chooses

$$W \sim P_{W|S^n}(w|s^n) = \frac{1}{Z(s^n)} \pi(w) \exp\left\{-\frac{\alpha}{n} \sum_{i=1}^n \ell(w, s_i)\right\},$$

where $\pi(\cdot)$ is a fixed prior on weights and $Z(\cdot)$ – normalization constant. Show that generalization gap of this algorithm is given by

$$\mathbb{E}[\ell(W, S)] - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \ell(W, s_i)\right] = \frac{1}{\alpha} I_{SKL}(W; S^n).$$

- I.44** (PAC-Bayes bounds [57]) Donsker-Varadhan characterizations of mutual information in (4.24) and (4.25) allows us to bound expectations of functions of weakly-dependent random variables, as shown in (4.26). Show the following extension: In the setting of (4.26) for any Q_X and any $\lambda > 0$ with probability $\geq 1 - \delta$ (over $y \sim P_Y$)

$$\mathbb{E}[h(X, Y) - h(X, \bar{Y}) | Y = y] \leq \frac{1}{\lambda} \left(\log \frac{1}{\delta} + D(P_{X|Y=y} \| Q_X) \right) + \frac{\epsilon^2 \lambda}{2}. \quad (\text{I.17})$$

⁴ For example, if $S = (X, Y)$ we may have $\ell(w, (x, y)) = 1\{f_w(x) \neq y\}$ where f_w denotes a neural network with weights w .

156 Exercises for Part I

(in other words, the typical deviation is of order $\sqrt{\epsilon^2 \log \frac{1}{\delta} + D(\cdot, \cdot)}$). Prove this inequality in two steps (assume that (X, Y, \bar{Y}) are all discrete):

- For convenience, let $\mathbb{E}_{X|Y}$ and $\mathbb{E}_{\bar{Y}}$ denote the respective (conditional) expectation operators. Show that the result follows from the following inequality (valid for all f and Q_X):

$$\mathbb{E}_Y \left[e^{\mathbb{E}_{X|Y}[f(X, Y)] - \ln \mathbb{E}_{\bar{Y}} e^{f(X, \bar{Y})}} \right] \leq 1 \quad (\text{I.18})$$

- Prove the previous inequality (Hint: a single application of Jensen's).
- Now consider a countable collection $\{f_i, i \in \mathbb{Z}_+\}$ of functions and an arbitrary distribution Q on \mathbb{Z}_+ . Show the following version of a union bound: with probability $1 - \delta$ over $Y \sim P_Y$ we have simultaneously for all i :

$$f_i(Y) \leq \ln \mathbb{E}_Y e^{f_i(Y)} + \log \frac{1}{\delta Q(i)}.$$

(Hint: choose $P_{X|Y}$ in (I.18)).

Note: In applications, P_Y is unknown (data) distribution and X is the chosen classifier. $h(X, Y)$ is the training loss and $\mathbb{E}[h(X, \bar{Y})|Y]$ is the (sample-dependent) test loss. Compared to (4.26), the bound (I.17) does not depend on (generally unknown) P_Y or P_X .

- I.45** Let $A = \{A_j : j \in J\}$ be a countable collection of random variables and T is a J -valued random index. Show that if each A_j is ϵ -subgaussian then

$$|\mathbb{E}[A_T]| \leq \sqrt{2\epsilon^2 I(A; T)}.$$

This generalizes (4.27), cf. [262].

- I.46** (Divergence for mixtures [153, 175]) Let $\bar{Q} = \sum_i \pi_i Q_i$ be a mixture distribution.

- (a) Prove

$$D(P\|\bar{Q}) \leq -\log \left(\sum_i \pi_i \exp(-D(P\|Q_i)) \right),$$

improving over the simple convexity estimate $D(P\|\bar{Q}) \leq \sum_i \pi_i D(P\|Q_i)$. (Hint: Prove that the function $Q \mapsto \exp\{-aD(P\|Q)\}$ is concave for every $a \leq 1$.)

- (b) Furthermore, for any distribution $\{\tilde{\pi}_j\}$, any $\lambda \in [0, 1]$ we have

$$\begin{aligned} \sum_j \tilde{\pi}_j D(Q_j\|\bar{Q}) + D(\pi\|\tilde{\pi}) &\geq -\sum_i \pi_i \log \sum_j \tilde{\pi}_j e^{-(1-\lambda)D_\lambda(P_i\|P_j)} \\ &\geq -\log \sum_{i,j} \pi_i \tilde{\pi}_j e^{-(1-\lambda)D_\lambda(P_i\|P_j)} \end{aligned}$$

(Hint: Prove $D(P_{A|B=b}\|Q_A) \geq -\mathbb{E}_{A|B=b}[\log \mathbb{E}_{A' \sim Q_A} \frac{g(A', b)}{g(A, b)}]$ via Donsker-Varadhan. Plug in $g(a, b) = P_{B|A}(b|a)^{1-\lambda}$, average over B and use Jensen to bring outer $\mathbb{E}_{B|A}$ inside the log.)

- I.47** (Mutual information and pairwise distances [153]) Suppose we have knowledge of pairwise distances $d_\lambda(x, x') \triangleq D_\lambda(P_{Y|X=x}\|P_{Y|X=x'})$, where D_λ is the Rényi divergence of order λ . What can be said about $I(X; Y)$? Let $X, X' \sim P_X$. Using Exercise I.46, prove that

$$I(X; Y) \leq -\mathbb{E}[\log \mathbb{E}[\exp(-d_1(X, X'))|X]]$$

and for every $\lambda \in [0, 1]$

$$I(X; Y) \geq -\mathbb{E}[\log \mathbb{E}[\exp(-(1-\lambda)d_\lambda(X, X'))|X]].$$

See Theorem 32.5 for an application.

I.48 ($D \lesssim H^2 \log \frac{1}{H^2}$ trick). Show that for any P, U, R and $0 < \epsilon < 2^{-5\frac{\lambda}{\lambda-1}}$ we have

$$D(P\|\epsilon U + \bar{\epsilon}R) \leq 8(H^2(P, R) + 2\epsilon) \left(\frac{\lambda}{\lambda-1} \log \frac{1}{\epsilon} + D_\lambda(P\|U) \right).$$

Thus, a Hellinger ϵ -net for a set of P 's can be converted into a KL $(\epsilon^2 \log \frac{1}{\epsilon})$ -net; see Section 32.2.4.)

(a) Start by proving the tail estimate for the divergence: For any $\lambda > 1$ and $b > e^{(\lambda-1)^{-1}}$

$$\mathbb{E}_P \left[\log \frac{dP}{dQ} \cdot 1\left\{ \frac{dP}{dQ} > b \right\} \right] \leq \frac{\log b}{b^{\lambda-1}} \exp\{(\lambda-1)D_\lambda(P\|Q)\}$$

(b) Show that for any $b > 1$ we have

$$D(P\|Q) \leq H^2(P, Q) \frac{b \log b}{(\sqrt{b}-1)^2} + \mathbb{E}_P \left[\log \frac{dP}{dQ} \cdot 1\left\{ \frac{dP}{dQ} > b \right\} \right]$$

(Hint: Write $D(P\|Q) = \mathbb{E}_P[h(\frac{dQ}{dP})]$ for $h(x) = -\log x + x - 1$ and notice that $\frac{h(x)}{(\sqrt{x}-1)^2}$ is monotonically decreasing on \mathbb{R}_+ .)

(c) Set $Q = \epsilon U + \bar{\epsilon}R$ and show that for every $\delta < e^{-\frac{1}{\lambda-1}} \wedge \frac{1}{4}$

$$D(P\|Q) \leq (4H^2(P, R) + 8\epsilon + c_\lambda \epsilon^{1-\lambda} \delta^{\lambda-1}) \log \frac{1}{\delta},$$

where $c_\lambda = \exp\{(\lambda-1)D_\lambda(P\|U)\}$. (Notice $H^2(P, Q) \leq H^2(P, R) + 2\epsilon$, $D_\lambda(P\|Q) \leq D_\lambda(P\|U) + \log \frac{1}{\epsilon}$ and set $b = 1/\delta$.)

(d) Complete the proof by setting $\delta^{\lambda-1} = \frac{4H^2(P, R) + 2\epsilon}{c_\lambda \epsilon^{\lambda-1}}$.

I.49 Let $G = (V, E)$ be a finite *directed* graph. Let

$$\begin{aligned} \Delta &= |\{(x, y, z) \in V^3 : (x, y), (y, z), (z, x) \in E\}|, \\ \wedge &= |\{(x, y, z) \in V^3 : (x, y), (x, z) \in E\}|. \end{aligned}$$

Prove that $\Delta \leq \wedge$.

Hint: Prove $H(X, Y, Z) \leq H(X) + 2H(Y|X)$ for random variables (X, Y, Z) distributed uniformly over the set of directed 3-cycles, i.e. subsets $X \rightarrow Y \rightarrow Z \rightarrow X$.



Part II

Lossless data compression



The principal engineering goal of data compression is to represent a given sequence a_1, a_2, \dots, a_n produced by a source as a sequence of bits of minimal possible length with possible algorithmic constraints. Of course, reducing the number of bits is generally impossible, unless the source satisfies certain statistical restrictions, that is, only a small subset of all sequences actually occur in practice. (Or, more precisely, only a small subset captures the majority of the overall probability distribution) Is this the case for real-world data?

As a simple demonstration, one may take two English novels and compute empirical frequencies of each letter. It will turn out to be the same for both novels (approximately). Thus, we can see that there is some underlying structure in English texts restricting possible output sequences. The structure goes beyond empirical frequencies of course, as further experimentation (involving digrams, word frequencies etc) may reveal. Thus, the main reason for the possibility of data compression is the *experimental (empirical) law: Real-world sources produce very restricted sets of sequences.*

How do we model these restrictions? Further experimentation (with language, music, images) reveals that frequently, the structure may be well described if we assume that sequences are generated probabilistically [268, Sec. III]. This is one of the main contributions of Shannon: *another empirical law states that real-world sources may be described probabilistically with increasing precision starting from i.i.d., first order Markov, second order Markov etc.* Note that sometimes one needs to find an appropriate basis in which this “law” holds – this is the case of images. (That is, rasterized sequence of pixels does not exhibit local probabilistic laws due to the 2-D constraints being ignored; instead, wavelets and local Fourier transform provide much better bases).⁵

Let us state upfront that the principal idea of strategically assigning shorter descriptions (bit strings) to more probable symbols is quite obvious, and was present already in the Morse code. The main innovation of Shannon was to show that compressing *groups* of symbols together can lead to dramatic savings, even when symbols are considered as i.i.d. This was a bold proposition at the time, as algorithmically it appears to be impossible to sort all possible 26^{10} realizations of the 10-letter English chunks, in the order of their decreasing frequency. Shannon’s idea became practical with the invention of Huffman, arithmetic and Lempel-Ziv compressors, decades after.

In the beginning of our investigation we will restrict attention to representing one random variable X in terms of (minimal number of) bits. Again, the understanding and the innovation comes when we replace a single X with an n -letter block $S^n = (S_1, \dots, S_n)$. The types of compression we will consider:

- Variable-length lossless compression. Here we require $\mathbb{P}[X \neq \hat{X}] = 0$, where \hat{X} is the decoded version. To make the question interesting, we compress X into a variable-length binary string. It will turn out that optimal compression length is $H(X) - O(\log(1 + H(X)))$. If we further restrict attention to so-called prefix-free or uniquely decodable codes, then the optimal compression length is $H(X) + O(1)$. Applying these results to n -letter variables $X = S^n$ we see that optimal

⁵ Of course, one should not take these “laws” too far. In regards to language modeling, (finite-state) Markov assumption is too simplistic to truly generate all proper sentences, cf. Chomsky [66].

compression length normalized by n converges to the entropy rate (Section 6.4) of the process $\{S_j\}$.

- Fixed-length, almost lossless compression. Here, we allow some very small (or vanishing with $n \rightarrow \infty$ when $X = S^n$) probability of error, i.e. $\mathbb{P}[X \neq \hat{X}] \leq \epsilon$. It turns out that under mild assumptions on the process $\{S_j\}$, here again we can compress to entropy rate but no more. This mode of compression permits various beautiful results in the presence of side-information (Slepian-Wolf, etc).
- Lossy compression. Here we require only $\mathbb{E}[d(X, \hat{X})] \leq \epsilon$ where $d(\cdot, \cdot)$ is some loss function. This type of compression problems is the central topic of Part V.

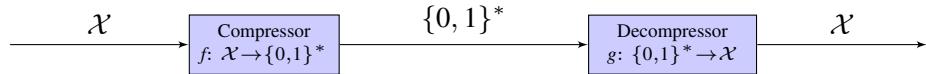
Note that more correctly we should have called all the examples above as “fixed-to-variable”, “fixed-to-fixed” and “fixed-to-lossy” codes, because they take fixed number of input letters. We omit discussion of the beautiful class of variable-to-fixed compressors, such as the famous Tunstall code [305], which consume an incoming stream of letters in variable-length chunks.

10

Variable-length lossless compression

10.1 Variable-length lossless compression

The coding paradigm of this compression is depicted in the following figure. Here a compressor



sor is a function f that maps each symbol $x \in \mathcal{X}$ into a variable-length string $f(x)$ in $\{0, 1\}^* \triangleq \bigcup_{k \geq 0} \{0, 1\}^k = \{\emptyset, 0, 1, 00, 01, \dots\}$. Each $f(x)$ is referred to as a *codeword* and the collection of codewords the *codebook*. We say f is a *lossless* compressor for a random variable X if there exists a decompressor $g : \{0, 1\}^* \rightarrow \mathcal{X}$ such that $\mathbb{P}[X = g(f(X))] = 1$, i.e., $g(f(x)) = x$ for all $x \in \mathcal{X}$ such that $P_X(x) > 0$. (As such, f is injective on the support of P_X). We are interested in the most economical way to compress the data. So let us introduce the length function $l : \{0, 1\}^* \rightarrow \mathbb{Z}_+$, e.g., $l(\emptyset) = 0, l(01001) = 5$.

Notice that since $\{0, 1\}^*$ is countable, lossless compression is only possible for discrete X . Also, without loss of generality, we can relabel \mathcal{X} such that $\mathcal{X} = \mathbb{N} = \{1, 2, \dots\}$ and sort the PMF decreasingly: $P_X(1) \geq P_X(2) \geq \dots$. At this point we do not impose any other constraints on the map f ; later in Section 10.3 we will introduce conditions such as prefix-freeness and unique-decodability. The unconstrained setting is sometimes called a *single-shot compression* setting, cf. [177].

We could consider different objectives for selecting the best compressor f , for example, minimizing any of $\mathbb{E}[l(f(X))]$, esssup $l(f(X))$, median $[l(f(X))]$ would be reasonable. It turns out that there is a compressor f^* that minimizes all objectives simultaneously. As mentioned in the preface of this chapter, the main idea is to assign longer codewords to less likely symbols, and reserve the shorter codewords for more probable symbols. To make precise of the optimality of f^* , let us recall the concept of stochastic dominance.

Definition 10.1 (Stochastic dominance). For real-valued random variables X and Y , we say Y stochastically dominates (or, is stochastically larger than) X , denoted by $X \leq^{\text{st.}} Y$, if $\mathbb{P}[Y \leq t] \leq \mathbb{P}[X \leq t]$ for all $t \in \mathbb{R}$.

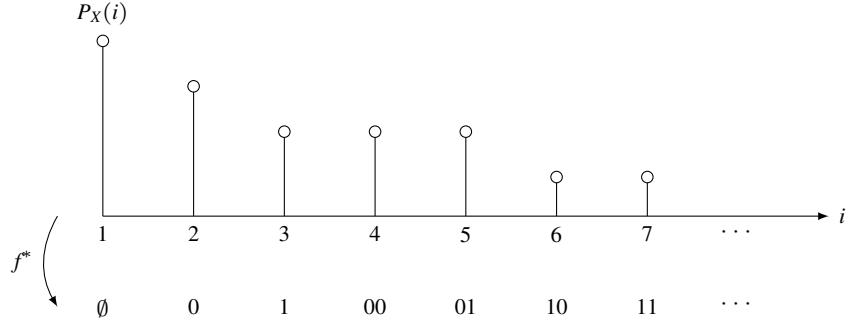


Figure 10.1 Illustration of the optimal variable-length lossless compressor f^* .

By definition, $X \leq^{\text{st.}} Y$ if and only if the CDF of X is larger than that of Y pointwise; in other words, the distribution of X assigns more probability to lower values than that of Y does. In particular, if X is dominated by Y stochastically, so are their means, medians, supremum, etc.

Theorem 10.2 (Optimal f^*). *Consider the compressor f^* defined (for a down-sorted PMF P_X) by $f^*(1) = \emptyset, f^*(2) = 0, f^*(3) = 1, f^*(4) = 00, \text{etc}$, assigning strings with increasing lengths to symbols $i \in \mathcal{X}$. (See Fig. 10.1 for an illustration.) Then*

1 *Length of codeword:*

$$l(f^*(i)) = \lfloor \log_2 i \rfloor.$$

2 *$l(f^*(X))$ is stochastically the smallest: For any lossless compressor $f: \mathcal{X} \rightarrow \{0, 1\}^*$,*

$$l(f^*(X)) \stackrel{\text{st.}}{\leq} l(f(X))$$

i.e., for any k , $\mathbb{P}[l(f(X)) \leq k] \leq \mathbb{P}[l(f^(X)) \leq k]$. As a result, $\mathbb{E}[l(f^*(X))] \leq \mathbb{E}[l(f(X))]$.*

Proof. Note that

$$|A_k| \triangleq |\{x : l(f(x)) \leq k\}| \leq \sum_{i=0}^k 2^i = 2^{k+1} - 1 = |\{x : l(f^*(x)) \leq k\}| \triangleq |A_k^*|.$$

Here the inequality is because f is lossless so that $|A_k|$ can at most be the total number of binary strings of length up to k . Then

$$\mathbb{P}[l(f(X)) \leq k] = \sum_{x \in A_k} P_X(x) \leq \sum_{x \in A_k^*} P_X(x) = \mathbb{P}[l(f^*(X)) \leq k], \quad (10.1)$$

since $|A_k| \leq |A_k^*|$ and A_k^* contains all $2^{k+1} - 1$ most likely symbols. \square

The following lemma (see Ex. I.7) is useful in bounding the expected code length of f^* . It says if the random variable is integer-valued, then its entropy can be controlled using its mean.

10.1 Variable-length lossless compression 165

Lemma 10.3. For any $Z \in \mathbb{N}$ s.t. $\mathbb{E}[Z] < \infty$, $H(Z) \leq \mathbb{E}[Z]h(\frac{1}{\mathbb{E}[Z]})$, where $h(\cdot)$ is the binary entropy function.

Theorem 10.4 (Optimal average code length: exact expression). Suppose $X \in \mathbb{N}$ and $P_X(1) \geq P_X(2) \geq \dots$. Then

$$\mathbb{E}[l(f^*(X))] = \sum_{k=1}^{\infty} \mathbb{P}[X \geq 2^k].$$

Proof. Recall that expectation of $U \in \mathbb{Z}_+$ can be written as $\mathbb{E}[U] = \sum_{k \geq 1} \mathbb{P}[U \geq k]$. Then by Theorem 10.2, $\mathbb{E}[l(f^*(X))] = \mathbb{E}[\lfloor \log_2 X \rfloor] = \sum_{k \geq 1} \mathbb{P}[\lfloor \log_2 X \rfloor \geq k] = \sum_{k \geq 1} \mathbb{P}[\log_2 X \geq k]$. \square

Theorem 10.5 (Optimal average code length vs. entropy [9]).

$$H(X) \text{ bits} - \log_2[e(H(X) + 1)] \leq \mathbb{E}[l(f^*(X))] \leq H(X) \text{ bits}$$

Remark 10.1. Theorem 10.5 is the first example of a *coding theorem* in this book, which relates the fundamental limit $\mathbb{E}[l(f^*(X))]$ (an operational quantity) to the entropy $H(X)$ (an information measure).

Proof. Define $L(X) = l(f^*(X))$. For the upper bound, observe that since the PMF are ordered decreasingly by assumption, $P_X(m) \leq 1/m$, so $L(m) \leq \log_2 m \leq \log_2(1/P_X(m))$. Taking expectation yields $\mathbb{E}[L(X)] \leq H(X)$.

For the lower bound,

$$\begin{aligned} H(X) &= H(X, L) = H(X|L) + H(L) \stackrel{(a)}{\leq} \mathbb{E}[L] + H(L) \\ &\stackrel{(b)}{\leq} \mathbb{E}[L] + h\left(\frac{1}{1 + \mathbb{E}[L]}\right)(1 + \mathbb{E}[L]) \\ &= \mathbb{E}[L] + \log_2(1 + \mathbb{E}[L]) + \mathbb{E}[L] \log_2\left(1 + \frac{1}{\mathbb{E}[L]}\right) \\ &\stackrel{(c)}{\leq} \mathbb{E}[L] + \log_2(1 + \mathbb{E}[L]) + \log_2 e \\ &\stackrel{(d)}{\leq} \mathbb{E}[L] + \log_2(e(1 + H(X))) \end{aligned} \tag{10.2}$$

where in (a) we have used the fact that $H(X|L = k) \leq k \text{ bits}$, because f^* is lossless, so that given $f^*(X) \in \{0, 1\}^k$, X can take at most 2^k values; (b) follows by Lemma 10.3; (c) is via $x \log(1+1/x) \leq \log e, \forall x > 0$; and (d) is by the previously shown upper bound $H(X) \leq \mathbb{E}[L]$. \square

To give an illustration, we need to introduce an important method of going from a *single-letter* source to a *multi-letter* one. Suppose that $S_j \stackrel{\text{i.i.d.}}{\sim} P_S$ (this is called a *memoryless source*). We can group n letters of S_j together and consider $X = S^n$ as one super-letter. Applying our results to

random variable X we obtain:

$$nH(S) \geq \mathbb{E}[l(f^*(S^n))] \geq nH(S) - \log_2 n + O(1).$$

In fact for memoryless sources, the exact asymptotic behavior is found in [292, Theorem 4]:

$$\mathbb{E}[l(f^*(S^n))] = \begin{cases} nH(S) + O(1), & P_S = \text{Unif} \\ nH(S) - \frac{1}{2} \log_2 n + O(1), & P_S \neq \text{Unif} \end{cases}.$$

For the case of sources for which $\log_2 \frac{1}{P_S}$ has non-lattice distribution, it is further shown in [292, Theorem 3]:

$$\mathbb{E}[l(f^*(S^n))] = nH(S) - \frac{1}{2} \log_2(8\pi e V(S)n) + o(1), \quad (10.3)$$

where $V(S)$ is the *varentropy* of the source S :

$$V(S) \triangleq \text{Var} \left[\log_2 \frac{1}{P_S(S)} \right]. \quad (10.4)$$

Theorem 10.5 relates the *mean* of $l(f^*(X))$ to that of $\log_2 \frac{1}{P_X(X)}$ (entropy). It turns out that distributions of these random variables are also closely related.

Theorem 10.6 (Code length distribution of f^*). $\forall \tau > 0, k \geq 0$,

$$\mathbb{P} \left[\log_2 \frac{1}{P_X(X)} \leq k \right] \leq \mathbb{P} [l(f^*(X)) \leq k] \leq \mathbb{P} \left[\log_2 \frac{1}{P_X(X)} \leq k + \tau \right] + 2^{-\tau+1}.$$

Proof. Lower bound (achievability): Use $P_X(m) \leq 1/m$. Then similarly as in Theorem 10.5, $L(m) = \lfloor \log_2 m \rfloor \leq \log_2 m \leq \log_2 \frac{1}{P_X(m)}$. Hence $L(X) \leq \log_2 \frac{1}{P_X(X)}$ a.s.

Upper bound (converse): By truncation,

$$\begin{aligned} \mathbb{P}[L \leq k] &= \mathbb{P} \left[L \leq k, \log_2 \frac{1}{P_X(X)} \leq k + \tau \right] + \mathbb{P} \left[L \leq k, \log_2 \frac{1}{P_X(X)} > k + \tau \right] \\ &\leq \mathbb{P} \left[\log_2 \frac{1}{P_X(X)} \leq k + \tau \right] + \sum_{x \in \mathcal{X}} P_X(x) \mathbf{1}_{\{l(f^*(x)) \leq k\}} \mathbf{1}_{\{P_X(x) \leq 2^{-k-\tau}\}} \\ &\leq \mathbb{P} \left[\log_2 \frac{1}{P_X(X)} \leq k + \tau \right] + (2^{k+1} - 1) \cdot 2^{-k-\tau} \end{aligned} \quad \square$$

So far our discussion applies to an arbitrary random variable X . Next we consider the source as a random process (S_1, S_2, \dots) and introduce *blocklength* n . We apply our results to $X = S^n$, that is, by treating the first n symbols as a supersymbol. The following corollary states that the limiting behavior of $l(f^*(S^n))$ and $\log \frac{1}{P_{S^n}(S^n)}$ always coincide.

Corollary 10.7. Let (S_1, S_2, \dots) be a random process and U, V real-valued random variable. Then

$$\frac{1}{n} \log_2 \frac{1}{P_{S^n}(S^n)} \xrightarrow{\text{d}} U \iff \frac{1}{n} l(f^*(S^n)) \xrightarrow{\text{d}} U \quad (10.5)$$

10.1 Variable-length lossless compression 167

and

$$\frac{1}{\sqrt{n}} \left(\log_2 \frac{1}{P_{S^n}(S^n)} - H(S^n) \right) \xrightarrow{d} V \Leftrightarrow \frac{1}{\sqrt{n}} (l(f^*(S^n)) - H(S^n)) \xrightarrow{d} V \quad (10.6)$$

Proof. First recall that convergence in distribution is equivalent to convergence of CDF at all continuity point, i.e., $U_n \xrightarrow{d} U \Leftrightarrow \mathbb{P}[U_n \leq u] \rightarrow \mathbb{P}[U \leq u]$ for all u at which point the CDF of U is continuous (i.e., not an atom of U).

To get (10.5), apply Theorem 10.7 with $k = un$ and $\tau = \sqrt{n}$:

$$\mathbb{P} \left[\frac{1}{n} \log_2 \frac{1}{P_X(X)} \leq u \right] \leq \mathbb{P} \left[\frac{1}{n} l(f^*(X)) \leq u \right] \leq \mathbb{P} \left[\frac{1}{n} \log_2 \frac{1}{P_X(X)} \leq u + \frac{1}{\sqrt{n}} \right] + 2^{-\sqrt{n}+1}.$$

To get (10.6), apply Theorem 10.7 with $k = H(S^n) + \sqrt{n}u$ and $\tau = n^{1/4}$:

$$\begin{aligned} \mathbb{P} \left[\frac{1}{\sqrt{n}} \left(\log \frac{1}{P_{S^n}(S^n)} - H(S^n) \right) \leq u \right] &\leq \mathbb{P} \left[\frac{l(f^*(S^n)) - H(S^n)}{\sqrt{n}} \leq u \right] \\ &\leq \mathbb{P} \left[\frac{1}{\sqrt{n}} \left(\log \frac{1}{P_{S^n}(S^n)} - H(S^n) \right) \leq u + n^{-1/4} \right] + 2^{-n^{1/4}} \end{aligned} \quad \blacksquare$$

Now let us particularize the preceding theorem to memoryless sources of i.i.d. S_j 's. The important observation is that the log likelihood becomes an i.i.d. sum:

$$\log \frac{1}{P_{S^n}(S^n)} = \sum_{i=1}^n \underbrace{\log \frac{1}{P_S(S_i)}}_{i.i.d.}.$$

- 1 By the Law of Large Numbers (LLN), we know that $\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} \mathbb{E} \log \frac{1}{P_S(S)} = H(S)$. Therefore in (10.5) the limiting distribution U is degenerate, i.e., $U = H(S)$, and we have $\frac{1}{n} l(f^*(S^n)) \xrightarrow{\mathbb{P}} \mathbb{E} \log \frac{1}{P_S(S)} = H(S)$. [Note: convergence in distribution to a constant \Leftrightarrow convergence in probability to a constant]
- 2 By the Central Limit Theorem (CLT), if variance $V(S) < \infty$, then we know that V in (10.6) is Gaussian, i.e.,

$$\frac{1}{\sqrt{nV(S)}} \left(\log \frac{1}{P_{S^n}(S^n)} - nH(S) \right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Consequently, we have the following Gaussian approximation for the probability law of the optimal code length

$$\frac{1}{\sqrt{nV(S)}} (l(f^*(S^n)) - nH(S)) \xrightarrow{d} \mathcal{N}(0, 1),$$

or, in shorthand,

$$l(f^*(S^n)) \sim nH(S) + \sqrt{nV(S)} \mathcal{N}(0, 1) \text{ in distribution.}$$

Gaussian approximation tells us the speed of $\frac{1}{n} l(f^*(S^n))$ to entropy and give us a good approximation at finite n .

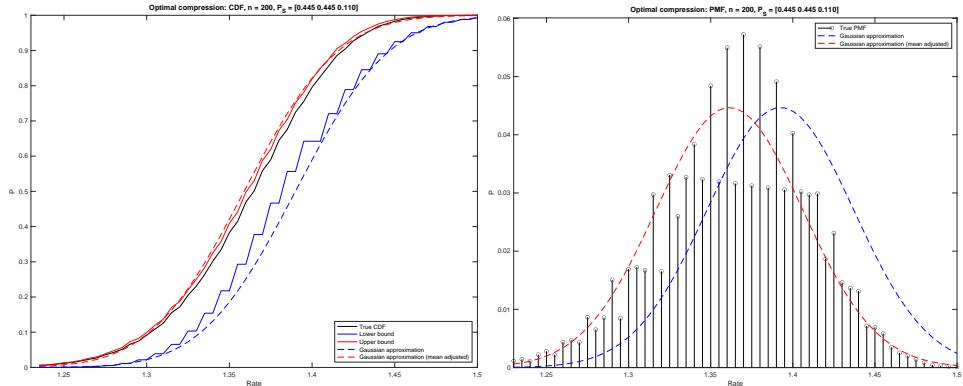


Figure 10.2 Left plot: Comparison of the true CDF of $l(f^*(S^n))$, bounds of Theorem 10.7 (optimized over τ), and the Gaussian approximations in (10.7) and (10.8). Right plot: PMF of the optimal compression length $l(f^*(S^n))$ and the two Gaussian approximations.

Example 10.1 (Ternary source). Next we apply our bounds to approximate the distribution of $l(f^*(S^n))$ in a concrete example. Consider a memoryless ternary source outputting i.i.d. n symbols from the distribution $P_S = [0.445, 0.445, 0.11]$. We first compare different results on the minimal expected length $\mathbb{E}[l(f^*(S^n))]$ in the following table:

Blocklength	Lower bound (10.5)	$\mathbb{E}[l(f^*(S^n))]$	$H(S^n)$ (upper bound)	asymptotics (10.3)
$n = 20$	21.5	24.3	27.8	$23.3 + o(1)$
$n = 100$	130.4	134.4	139.0	$133.3 + o(1)$
$n = 500$	684.1	689.2	695.0	$688.1 + o(1)$

In all cases above $\mathbb{E}[l(f^*(S))]$ is close to a midpoint between the bounds.

Next we consider the distribution of $l(f^*(S^n))$. Its Gaussian approximation is defined as

$$nH(S) + \sqrt{nV(S)}Z, \quad Z \sim \mathcal{N}(0, 1). \quad (10.7)$$

However, in view of (10.3) we also define the *mean-adjusted* Gaussian approximation as

$$nH(S) - \frac{1}{2} \log_2(8\pi e V(S)n) + \sqrt{nV(S)}Z, \quad Z \sim \mathcal{N}(0, 1). \quad (10.8)$$

Fig. 10.2 compares the true distribution of $l(f^*(S^n))$ with bounds and two Gaussian approximations.

10.2 Mandelbrot's argument for universality of Zipf's (power) law

Given a corpus of text it is natural to plot its *rank-frequency* table by sorting the word frequencies according to their rank $p_1 \geq p_2 \geq \dots$. The resulting tables, as noticed by Zipf [338], satisfy

10.2 Mandelbrot's argument for universality of Zipf's (power) law 169

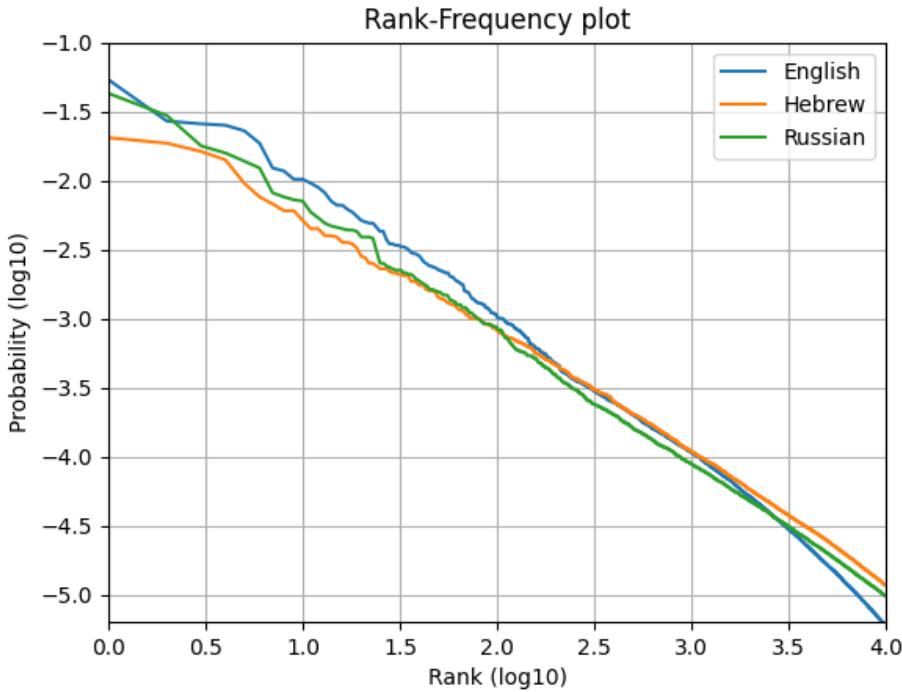


Figure 10.3 The log-log frequency-rank plots of the most used words in various languages exhibit a power law tail with exponent close to 1, as popularized by Zipf [338]. Data from [282].

$p_r \asymp r^{-\alpha}$ for some value of α . Remarkably, this holds across various corpora of text in multiple different languages (and with $\alpha \approx 1$) – see Fig. 10.3 for an illustration. Even more surprisingly, a lot of other similar tables possess the power-law distribution: “city populations, the sizes of earthquakes, moon craters, solar flares, computer files, wars, personal names in most cultures, number of papers scientists write, number of citations a paper receives, number of hits on web pages, sales of books and music recordings, number of species in biological taxa, people’s incomes” (quoting from [220], which gives references for each study). This spectacular universality of the power law continues to provoke scientists from many disciplines to suggest explanations for its occurrence; see [215] for a survey of such. One of the earliest (in the context of natural language of Zipf) is due to Mandelbrot [204] and is in fact intimately related to the topic of this Chapter.

Let us go back to the question of minimal expected length of the representation of source X . We have shown bounds on this quantity in terms of the entropy of X in Theorem 10.5. Let us introduce the following function

$$\mathcal{H}(\Lambda) = \sup_{f, P_X} \{H(X) : \mathbb{E}[l(f(X))] \leq \Lambda\},$$

where optimization is over lossless encoders and probability distributions $P_X = \{p_j : j = 1, \dots\}$. Theorem 10.5 (or more precisely, the intermediate result (10.2)) shows that

$$\Lambda \log 2 \leq \mathcal{H}(\Lambda) \leq \Lambda \log 2 + (1 + \Lambda) \log(1 + \Lambda) - \Lambda \log \Lambda.$$

It turns out that the upper bound is in fact tight. Furthermore, among all distributions the optimal tradeoff between entropy and minimal compression length is attained at power law distributions.

To show that, notice that in computing $\mathcal{H}(\Lambda)$, we can restrict attention to sorted PMFs $p_1 \geq p_2 \geq \dots$ (call this class \mathcal{P}^\downarrow), for which the optimal encoder is such that $l(f(j)) = \lfloor \log_2 j \rfloor$ (Theorem 10.2). Thus, we have shown

$$\mathcal{H}(\Lambda) = \sup_{P \in \mathcal{P}^\downarrow} \{H(P) : \sum_j p_j \lfloor \log_2 j \rfloor \leq \Lambda\}.$$

Next, let us fix the base of the logarithm of H to be 2, for convenience. (We will convert to arbitrary base at the end). Applying Example 5.2 we obtain:

$$\mathcal{H}(\Lambda) \leq \inf_{\lambda > 0} \lambda \Lambda + \log_2 Z(\lambda), \quad (10.9)$$

where $Z(\lambda) = \sum_{n=1}^{\infty} 2^{-\lambda \lfloor \log_2 n \rfloor} = \sum_{m=0}^{\infty} 2^{(1-\lambda)m} = \frac{1}{1-2^{1-\lambda}}$ if $\lambda > 1$ and $Z(\lambda) = \infty$ otherwise. Clearly, the infimum over $\lambda > 0$ is a minimum attained at a value $\lambda^* > 1$ satisfying

$$\Lambda = - \left. \frac{d}{d\lambda} \right|_{\lambda=\lambda^*} \log_2 Z(\lambda).$$

Define the distribution

$$P_\lambda(n) \triangleq \frac{1}{Z(\lambda)} 2^{-\lambda \lfloor \log_2 n \rfloor}, \quad n \geq 1$$

and notice that

$$\begin{aligned} \mathbb{E}_{P_\lambda}[\lfloor \log_2 X \rfloor] &= - \frac{d}{d\lambda} \log_2 Z(\lambda) = \frac{2^{1-\lambda}}{1-2^{1-\lambda}} \\ H(P_\lambda) &= \log_2 Z(\lambda) + \lambda \mathbb{E}_{P_\lambda}[\lfloor \log_2 X \rfloor]. \end{aligned}$$

Comparing with (10.9) we find that the upper bound in (10.9) is tight and attained by P_{λ^*} . From the first equation above, we also find $\lambda^* = \log_2 \frac{2+2\Lambda}{\Lambda}$. Altogether this yields

$$\mathcal{H}(\Lambda) = \Lambda \log 2 + (\Lambda + 1) \log(\Lambda + 1) - \Lambda \log \Lambda,$$

and the extremal distribution $P_{\lambda^*}(n) \asymp n^{-\lambda^*}$ is power-law distribution with the exponent $\lambda^* \rightarrow 1$ as $\Lambda \rightarrow \infty$.

The argument of Mandelbrot [204] The above derivation shows a special (extremality) property of the power law, but falls short of explaining its empirical ubiquity. Here is a way to connect the optimization problem $\mathcal{H}(\Lambda)$ to the evolution of the natural language. Suppose that there is a countable set S of elementary concepts that are used by the brain as building blocks of perception and communication with the outside world. As an approximation we can think that concepts are in one-to-one correspondence with language words. Now every concept x is represented internally

10.3 Uniquely decodable codes, prefix codes and Huffman codes 171

by the brain as a certain pattern, in the simplest case – a sequence of zeros and ones of length $l(f(x))$ ([204] considers more general representations). Now we have seen that the number of sequences of concepts with a composition P grows exponentially (in length) with the exponent given by $H(P)$, see Proposition 1.6. Thus in the long run the probability distribution P over the concepts results in the rate of information transfer equal to $\frac{H(P)}{\mathbb{E}_P[l(f(X))]}$. Mandelbrot concludes that in order to transfer maximal information per unit, *language and brain representation co-evolve in such a way as to maximize this ratio*. Note that

$$\sup_{P,f} \frac{H(P)}{\mathbb{E}_P[l(f(X))]} = \sup_{\Lambda} \frac{\mathcal{H}(\Lambda)}{\Lambda}.$$

It is not hard to show that $\mathcal{H}(\Lambda)$ is concave and thus the supremum is achieved at $\Lambda = 0+$ and equals infinity. This appears to have not been observed by Mandelbrot. To fix this issue, we can postulate that for some unknown reason there is a requirement of also having a certain minimal entropy $H(P) \geq h_0$. In this case

$$\sup_{P,f:H(P)\geq h_0} \frac{H(P)}{\mathbb{E}_P[l(f(X))]} = \frac{h_0}{\mathcal{H}^{-1}(h_0)}$$

and the supremum is achieved at a power law distribution P . Thus, the implication is that *the frequency of word usage in human languages evolves until a power law is attained*, at which point it maximizes information transfer within the brain. That's the gist of the argument of [204]. It is clear that this does not explain appearance of the power law in other domains, for which other explanations such as preferential attachment models are more plausible, see [215]. Finally, we mention that the P_λ distributions take discrete values $2^{-\lambda m - \log_2 Z(\lambda)}$, $m = 0, 1, 2, \dots$ with multiplicities 2^m . Thus P_λ appears as a rather unsightly staircase on frequency-rank plots such as Fig. 10.3. This artifact can be alleviated by considering non-binary brain representations with *unequal lengths* of signals.

10.3 Uniquely decodable codes, prefix codes and Huffman codes

In the previous sections we have studied f^* , which achieves the stochastically (in particular, in expectation) shortest code length among all variable-length lossless compressors. Note that f^* is obtained by ordering the PMF and assigning shorter codewords to more likely symbols. In this section we focus on a specific class of compressors with good algorithmic properties which lead to low complexity decoding and short delay when decoding from a stream of compressed bits. This part is more combinatorial in nature.

We start with a few definitions. Let $\mathcal{A}^+ = \bigcup_{n \geq 1} \mathcal{A}^n$ denotes all non-empty finite-length strings consisting of symbols from the alphabet \mathcal{A} . Throughout this chapter \mathcal{A} is a countable set.

Definition 10.8 (Extension of a code). The (symbol-by-symbol) extension of $f: \mathcal{A} \rightarrow \{0, 1\}^*$ is $f: \mathcal{A}^+ \rightarrow \{0, 1\}^*$ where $f(a_1, \dots, a_n) = (f(a_1), \dots, f(a_n))$ is defined by concatenating the bits.

Definition 10.9 (Uniquely decodable codes). $f : \mathcal{A} \rightarrow \{0, 1\}^*$ is *uniquely decodable* if its extension $f : \mathcal{A}^+ \rightarrow \{0, 1\}^*$ is injective.

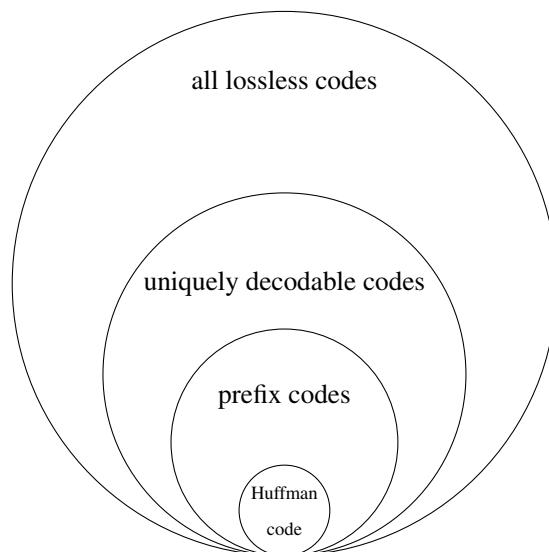
Definition 10.10 (Prefix codes). $f : \mathcal{A} \rightarrow \{0, 1\}^*$ is a *prefix code*¹ if no codeword is a prefix of another (e.g., 010 is a prefix of 0101).

Example 10.2. $\mathcal{A} = \{a, b, c\}$.

- $f(a) = 0, f(b) = 1, f(c) = 10$. Not uniquely decodable, since $f(ba) = f(c) = 10$.
- $f(a) = 0, f(b) = 10, f(c) = 11$. Uniquely decodable and a prefix code.
- $f(a) = 0, f(b) = 01, f(c) = 011, f(d) = 0111$ Uniquely decodable but not a prefix code, since as long as 0 appears, we know that the previous codeword has terminated.²

Remark 10.2.

- 1 Prefix codes are uniquely decodable and hence lossless, as illustrated in the following picture:



- 2 Similar to prefix-free codes, one can define suffix-free codes. Those are also uniquely decodable (one should start decoding in reverse direction).
- 3 By definition, any uniquely decodable code does not have the empty string as a codeword. Hence $f : \mathcal{X} \rightarrow \{0, 1\}^+$ in both Definition 10.9 and Definition 10.10.
- 4 Unique decodability means that one can decode from a stream of bits without ambiguity, but one might need to look ahead in order to decide the termination of a codeword. (Think of the

¹ Also known as prefix-free/comma-free/self-punctuating/instantaneous code.

² In this example, if 0 is placed at the very end of each codeword, the code is uniquely decodable, known as the *unary code*.

10.3 Uniquely decodable codes, prefix codes and Huffman codes 173

last example). In contrast, prefix codes allow the decoder to decode instantaneously without looking ahead.

- 5 Prefix codes are in one-to-one correspondence with binary trees (with codewords at leaves). It is also equivalent to strategies to ask “yes/no” questions previously mentioned at the end of Section 1.1.

Theorem 10.11 (Kraft-McMillan).

1 Let $f: \mathcal{A} \rightarrow \{0, 1\}^*$ be uniquely decodable. Set $l_a = l(f(a))$. Then f satisfies the Kraft inequality

$$\sum_{a \in \mathcal{A}} 2^{-l_a} \leq 1. \quad (10.10)$$

2 Conversely, for any set of code length $\{l_a : a \in \mathcal{A}\}$ satisfying (10.10), there exists a prefix code f , such that $l_a = l(f(a))$. Moreover, such an f can be computed efficiently.

Remark 10.3. The consequence of Theorem 10.12 is that as far as compression efficiency is concerned, we can ignore those uniquely decodable codes that are not prefix codes.

Proof. We prove the Kraft inequality for prefix codes and uniquely decodable codes separately. The proof for the former is probabilistic, following ideas in [10, Exercise 1.8, p. 12]. Let f be a prefix code. Let us construct a probability space such that the LHS of (10.10) is the probability of some event, which cannot exceed one. To this end, consider the following scenario: Generate independent $\text{Ber}(\frac{1}{2})$ bits. Stop if a codeword has been written, otherwise continue. This process terminates with probability $\sum_{a \in \mathcal{A}} 2^{-l_a}$. The summation makes sense because the events that a given codeword is written are mutually exclusive, thanks to the prefix condition.

Now let f be a uniquely decodable code. The proof uses *generating function* as a device for counting. (The analogy in coding theory is the weight enumerator function.) First assume \mathcal{A} is finite. Then $L = \max_{a \in \mathcal{A}} l_a$ is finite. Let $G_f(z) = \sum_{a \in \mathcal{A}} z^{l_a} = \sum_{l=0}^L A_l(f)z^l$, where $A_l(f)$ denotes the number of codewords of length l in f . For $k \geq 1$, define $f^k : \mathcal{A}^k \rightarrow \{0, 1\}^+$ as the symbol-by-symbol extension of f . Then $G_{f^k}(z) = \sum_{a^k \in \mathcal{A}^k} z^{l(f^k(a^k))} = \sum_{a_1} \cdots \sum_{a_k} z^{l_{a_1} + \cdots + l_{a_k}} = [G_f(z)]^k = \sum_{l=0}^{KL} A_l(f^k)z^l$. By the unique decodability of f , f^k is lossless. Hence $A_l(f^k) \leq 2^l$. Therefore we have $G_f(1/2)^k = G_{f^k}(1/2) \leq kL$ for all k . Then $\sum_{a \in \mathcal{A}} 2^{-l_a} = G_f(1/2) \leq \lim_{k \rightarrow \infty} (kL)^{1/k} = 1$. If \mathcal{A} is countably infinite, for any finite subset $\mathcal{A}' \subset \mathcal{A}$, repeating the same argument gives $\sum_{a \in \mathcal{A}'} 2^{-l_a} \leq 1$. The proof is complete by the arbitrariness of \mathcal{A}' .

Conversely, given a set of code lengths $\{l_a : a \in \mathcal{A}\}$ s.t. $\sum_{a \in \mathcal{A}} 2^{-l_a} \leq 1$, construct a prefix code f as follows: First relabel \mathcal{A} to \mathbb{N} and assume that $1 \leq l_1 \leq l_2 \leq \dots$. For each i , define

$$a_i \triangleq \sum_{k=1}^{i-1} 2^{-l_k}$$

with $a_1 = 0$. Then $a_i < 1$ by Kraft inequality. Thus we define the codeword $f(i) \in \{0, 1\}^+$ as the first l_i bits in the binary expansion of a_i . Finally, we prove that f is a prefix code by contradiction:

Suppose for some $j > i$, $f(i)$ is the prefix of $f(j)$, since $l_j \geq l_i$. Then $a_j - a_i \leq 2^{-l_i}$, since they agree on the most significant l_i bits. But $a_j - a_i = 2^{-l_i} + 2^{-l_{i+1}} + \dots > 2^{-l_i}$, which is a contradiction. \square

Remark 10.4. A conjecture of Ahlswede et al [4] states that for any set of lengths for which $\sum 2^{-l_a} \leq \frac{3}{4}$ there exists a fix-free code (i.e. one which is simultaneously prefix-free and suffix-free). So far, existence has only been shown when the Kraft sum is $\leq \frac{5}{8}$, cf. [331].

In view of Theorem 10.12, the optimal average code length among all prefix (or uniquely decodable) codes is given by the following optimization problem

$$\begin{aligned} L^*(X) &\triangleq \min \sum_{a \in \mathcal{A}} P_X(a) l_a && (10.11) \\ \text{s.t. } &\sum_{a \in \mathcal{A}} 2^{-l_a} \leq 1 \\ &l_a \in \mathbb{N} \end{aligned}$$

This is an *integer programming* (IP) problem, which, in general, is computationally hard to solve. It is remarkable that this particular IP can be solved in *near-linear* time, thanks to the Huffman algorithm. Before describing the construction of Huffman codes, let us give bounds to $L^*(X)$ in terms of entropy:

Theorem 10.12.

$$H(X) \leq L^*(X) \leq H(X) + 1 \text{ bit.} \quad (10.12)$$

Proof. Right inequality: Consider the following length assignment $l_a = \lceil \log_2 \frac{1}{P_X(a)} \rceil$,³ which satisfies Kraft since $\sum_{a \in \mathcal{A}} 2^{-l_a} \leq \sum_{a \in \mathcal{A}} P_X(a) = 1$. By Theorem 10.12, there exists a prefix code f such that $l(f(a)) = \lceil \log_2 \frac{1}{P_X(a)} \rceil$ and $\mathbb{E}l(f(X)) \leq H(X) + 1$.

Left inequality: We give two proofs for this converse. One of the commonly used ideas to deal with combinatorial optimization is *relaxation*. Our first idea is to drop the integer constraints in (10.11) and relax it into the following optimization problem, which obviously provides a lower bound

$$L^*(X) \triangleq \min \sum_{a \in \mathcal{A}} P_X(a) l_a \quad (10.13)$$

$$\text{s.t. } \sum_{a \in \mathcal{A}} 2^{-l_a} \leq 1 \quad (10.14)$$

This is a nice *convex optimization* problem, with affine objective function and a convex feasible set. Solving (10.13) by Lagrange multipliers (Exercise!) yields that the minimum is equal to $H(X)$ (achieved at $l_a = \log_2 \frac{1}{P_X(a)}$).

³ Such a code is called a Shannon code.

10.3 Uniquely decodable codes, prefix codes and Huffman codes 175

Another proof is the following: For any f whose codelengths $\{l_a\}$ satisfying the Kraft inequality, define a probability measure $Q(a) = \frac{2^{-l_a}}{\sum_{a \in \mathcal{A}} 2^{-l_a}}$. Then

$$\mathbb{E}l(f(X)) - H(X) = D(P\|Q) - \log \sum_{a \in \mathcal{A}} 2^{-l_a} \geq 0.$$

□

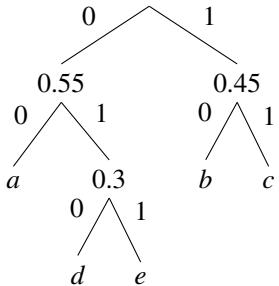
Next we describe the Huffman code, which achieves the optimum in (10.11). In view of the fact that prefix codes and binary trees are one-to-one, the main idea of the Huffman code is to build the binary tree from the bottom up: Given a PMF $\{P_X(a) : a \in \mathcal{A}\}$,

- 1 Choose the two least-probable symbols in the alphabet.
- 2 Delete the two symbols and add a new symbol (with combined probabilities). Add the new symbol as the parent node of the previous two symbols in the binary tree.

The algorithm terminates in $|\mathcal{A}| - 1$ steps. Given the binary tree, the code assignment can be obtained by assigning 0/1 to the branches. Therefore the time complexity is $O(|\mathcal{A}|)$ (sorted PMF) or $O(|\mathcal{A}| \log |\mathcal{A}|)$ (unsorted PMF).

Example 10.3. $\mathcal{A} = \{a, b, c, d, e\}, P_X = \{0.25, 0.25, 0.2, 0.15, 0.15\}$.

Huffman tree:



Codebook:

$f(a) = 00$
$f(b) = 10$
$f(c) = 11$
$f(d) = 010$
$f(e) = 011$

Theorem 10.13 (Optimality of Huffman codes). *The Huffman code achieves the minimal average code length (10.11) among all prefix (or uniquely decodable) codes.*

Proof. See [75, Sec. 5.8]. □

Remark 10.5 (Drawbacks of Huffman codes).

- 1 As Shannon pointed out in his 1948 paper, in compressing English texts, in addition to exploiting the nonequiprobability of English letters, working with pairs (or more generally, n -grams) of letters achieves even more compression. To compress a block of symbols (S_1, \dots, S_n) , while a natural idea is to apply the Huffman codes on a symbol-by-symbol basis (i.e., applying the corresponding Huffman code for each P_{S_i}). By Theorem 10.15, this is only guaranteed to achieve an average length at most $\sum_{i=1}^n H(S_i) + n$ bits, which also fails to exploit the memory in the source

when $\sum_{i=1}^n H(S_i)$ is significantly larger than $H(S_1, \dots, S_n)$. The solution is to apply block Huffman coding. Indeed, compressing the block (S_1, \dots, S_n) using its Huffman code (designed for P_{S_1, \dots, S_n}) achieves $H(S_1, \dots, S_n)$ within one bit, but the complexity is $|\mathcal{A}|^n!$

- 2 Constructing the Huffman code requires knowing the source distribution. This brings us the question: Is it possible to design universal compressor which achieves entropy for a class of source distributions? And what is the price to pay? These questions are addressed in Chapter 13.

There are much more elegant solutions, e.g.,

- 1 Arithmetic coding: sequential encoding, linear complexity in compressing (S_1, \dots, S_n) – Section 13.1.
- 2 Lempel-Ziv algorithm: low-complexity, universal, provably optimal in a very strong sense – Section 13.7.

As a summary of this chapter, we state the following comparison of average code length (in bits) for lossless codes:

$$H(X) - \log_2[e(H(X) + 1)] \leq \mathbb{E}[l(f^*(X))] \leq H(X) \leq \mathbb{E}[l(f_{\text{Huffman}}(X))] \leq H(X) + 1.$$

11

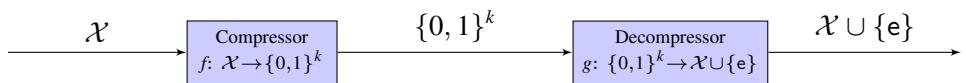
Fixed-length (almost lossless) compression. Slepian-Wolf.

In the previous chapter we introduced the concept of variable-length compression and studied its fundamental limits (with and without prefix-free condition). In some situations, however, one may desire that the output of the compressor always has a fixed length, say, k bits. Unless k is unreasonably large, then, this will require relaxing the losslessness condition. This is the focus of this chapter: compression in the presence of (typically vanishingly small) probability of error. It turns out allowing even very small error enables several beautiful effects:

- The possibility to compress data via matrix multiplication over finite fields (Linear Compression).
- The possibility to reduce compression length from $H(X)$ to $H(X|Y)$ if side information Y is available at the decompressor (Slepian-Wolf).
- The possibility to reduce compression length below $H(X)$ if access to a compressed representation of side-information Y is available at the decompressor (Ahlswede-Körner-Wyner).

11.1 Fixed-length almost lossless code. Asymptotic Equipartition Property (AEP).

The coding paradigm in this section is illustrated as follows: Note that if we insist like in Chapter 10



that $g(f(X)) = X$ with probability one, then $k \geq \log_2 |\text{supp}(P_X)|$ and no meaningful compression can be achieved. It turns out that by tolerating a small error probability, we can gain a lot in terms of code length! So, instead of requiring $g(f(x)) = x$ for all $x \in \mathcal{X}$, consider only lossless decompression for a subset $\mathcal{S} \subset \mathcal{X}$:

$$g(f(x)) = \begin{cases} x & x \in \mathcal{S} \\ e & x \notin \mathcal{S} \end{cases}$$

and the probability of error is:

$$\mathbb{P}[g(f(X)) \neq X] = \mathbb{P}[g(f(X)) = e] = \mathbb{P}[X \notin \mathcal{S}].$$

Definition 11.1. A compressor-decompressor pair (f, g) is called a (k, ϵ) -code if:

$$\begin{aligned} f: \mathcal{X} &\rightarrow \{0, 1\}^k \\ g: \{0, 1\}^k &\rightarrow \mathcal{X} \cup \{\text{e}\} \end{aligned}$$

such that $g(f(x)) \in \{x, \text{e}\}$ for all $x \in \mathcal{X}$ and $\mathbb{P}[g(f(X)) = \text{e}] \leq \epsilon$.

The minimum probability of error is defined as

$$\epsilon^*(X, k) \triangleq \inf\{\epsilon : \exists (k, \epsilon)\text{-code for } X\}$$

The following result connects the respective fundamental limits of fixed-length almost lossless compression and variable-length lossless compression (Chapter 10):

Theorem 11.2 (Fundamental limit of fixed-length compression). *Recall the optimal variable-length compressor f^* defined in Theorem 10.2. Then*

$$\epsilon^*(X, k) = \mathbb{P}[I(f^*(X)) \geq k] = 1 - \text{total probability of the } 2^k - 1 \text{ most likely symbols of } X.$$

Proof. The proof is essentially tautological. Note $1 + 2 + \dots + 2^{k-1} = 2^k - 1$. Let $\mathcal{S} = \{2^k - 1 \text{ most likely realizations of } X\}$. Then

$$\epsilon^*(X, k) = \mathbb{P}[X \notin \mathcal{S}] = \mathbb{P}[I(f^*(X)) \geq k].$$

The last equality follows from (10.1). \square

Comparing Theorems 10.2 and 11.2, we see that the optimal codes in these two settings work as follows:

- Variable-length: f^* encodes the $2^k - 1$ symbols with the highest probabilities to $\{\phi, 0, 1, 00, \dots, 1^{k-1}\}$.
- Fixed-length: The optimal compressor f maps the elements of \mathcal{S} into $(00\dots 00), \dots, (11\dots 10)$ and the rest in $\mathcal{X} \setminus \mathcal{S}$ to $(11\dots 11)$. The decompressor g decodes perfectly except for outputting e upon receipt of $(11\dots 11)$.

Remark 11.1. In Definition 11.1 we require that the errors are always *detectable*, i.e., $g(f(x)) = x$ or e . Alternatively, we can drop this requirement and allow *undetectable* errors, in which case we can of course do better since we have more freedom in designing codes. It turns out that we do not gain much by this relaxation. Indeed, if we define

$$\tilde{\epsilon}^*(X, k) = \inf\{\mathbb{P}[g(f(X)) \neq X] : f: \mathcal{X} \rightarrow \{0, 1\}^k, g: \{0, 1\}^k \rightarrow \mathcal{X} \cup \{\text{e}\}\},$$

then $\tilde{\epsilon}^*(X, k) = 1 - \text{sum of } 2^k \text{ largest masses of } X$. This follows immediately from $\mathbb{P}[g(f(X)) = X] = \sum_{x \in \mathcal{S}} P_X(x)$ where $\mathcal{S} \triangleq \{x : g(f(x)) = x\}$ satisfies $|\mathcal{S}| \leq 2^k$, because f takes no more than 2^k values. Compared to Theorem 11.2, we see that $\tilde{\epsilon}^*(X, k)$ and $\epsilon^*(X, k)$ do not differ much. In particular, $\epsilon^*(X, k+1) \leq \tilde{\epsilon}^*(X, k) \leq \epsilon^*(X, k)$.

11.1 Fixed-length almost lossless code. Asymptotic Equipartition Property (AEP). 179

Corollary 11.3 (Shannon). *Let S^n be i.i.d. Then*

$$\lim_{n \rightarrow \infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases}$$

$$\lim_{n \rightarrow \infty} \epsilon^*(S^n, nH(S) + \sqrt{nV(S)}\gamma) = 1 - \Phi(\gamma).$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0, 1)$, $H(S) = \mathbb{E}[\log \frac{1}{P_S(S)}]$ is the entropy, $V(S) = \text{Var}[\log \frac{1}{P_S(S)}]$ is the variance which is assumed to be finite.

Proof. Combine Theorem 11.2 with Corollary 10.1. \square

Next we give separate achievability and converse bounds complementing the exact result in Theorem 11.2.

Theorem 11.4 (Converse).

$$\epsilon^*(X, k) \geq \tilde{\epsilon}^*(X, k) \geq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} > k + \tau\right] - 2^{-\tau}, \quad \forall \tau > 0.$$

Proof. The argument identical to the converse of Theorem 10.7. Let $\mathcal{S} = \{x : g(f(x)) = x\}$. Then $|\mathcal{S}| \leq 2^k$ and $\mathbb{P}[X \in \mathcal{S}] \leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \leq k + \tau\right] + \underbrace{\mathbb{P}\left[X \in \mathcal{S}, \log_2 \frac{1}{P_X(X)} > k + \tau\right]}_{\leq 2^{-\tau}}$. \square

We state two achievability bounds.

Theorem 11.5.

$$\epsilon^*(X, k) \leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \geq k\right]. \quad (11.1)$$

Theorem 11.6.

$$\epsilon^*(X, k) \leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} > k - \tau\right] + 2^{-\tau}, \quad \forall \tau > 0. \quad (11.2)$$

Note that Theorem 11.5 is in fact always stronger than Theorem 11.6. Still, we present the proof of Theorem 11.6 and the technology behind it – **random coding** – a powerful technique introduced by Shannon for proving existence of good codes (achievability). This technique is used throughout in this book and Theorem 11.6 is its first appearance. To see that Theorem 11.5 gives a better bound, note that even the first term in (11.2) exceeds (11.1). Nevertheless, the random coding argument for proving this weaker bound is much more important and generalizable. We will apply it again for linear compression in Section 11.2 and the Slepian-Wolf problem in Section 11.4 in this chapter;

later for data transmission and lossy data compression in Parts IV and V it will take the central stage as the method of choice for most achievability proofs.

Proof of Theorem 11.5. Construction: use those $2^k - 1$ symbols with the highest probabilities.

The analysis is essentially the same as the lower bound in Theorem 10.7 from Chapter 10. Note that the m^{th} largest mass $P_X(m) \leq \frac{1}{m}$. Therefore

$$\epsilon^*(X, k) = \sum_{m \geq 2^k} P_X(m) = \sum_{m \geq 2^k} 1_{\{m \geq 2^k\}} P_X(m) \leq \sum_{\left\{ \frac{1}{P_X(m)} \geq 2^k \right\}} P_X(m) = \mathbb{E} 1_{\left\{ \log_2 \frac{1}{P_X(X)} \geq k \right\}}.$$

□

Proof of Theorem 11.6. (Random coding.) For a given compressor f , the optimal decompressor which minimizes the error probability is the maximum a posteriori (MAP) decoder, i.e.,

$$g^*(w) = \operatorname{argmax}_x P_{X|f(X)}(x|w) = \operatorname{argmax}_{x:f(x)=w} P_X(x),$$

which can be hard to analyze. Instead, let us consider the following (suboptimal) decompressor g :

$$g(w) = \begin{cases} x, & \exists! x \in \mathcal{X} \text{ s.t. } f(x) = w \text{ and } \log_2 \frac{1}{P_X(x)} \leq k - \tau, \\ & (\text{exists unique high-probability } x \text{ that is mapped to } w) \\ e, & \text{o.w.} \end{cases}$$

Note that $\log_2 \frac{1}{P_X(x)} \leq k - \tau \iff P_X(x) \geq 2^{-(k-\tau)}$. We call those x “high-probability”.

Denote $f(x) = c_x$ and the codebook $\mathcal{C} = \{c_x : x \in \mathcal{X}\} \subset \{0, 1\}^k$. It is instructive to think of \mathcal{C} as a hashing table.

Error probability analysis: There are two ways to make an error \Rightarrow apply union bound. Before proceeding, define

$$J(x, \mathcal{C}) \triangleq \left\{ x' \in \mathcal{X} : c_{x'} = c_x, x' \neq x, \log_2 \frac{1}{P_X(x')} \leq k - \tau \right\}$$

to be the set of high-probability inputs whose hashes collide with that of x . Then we have the following estimate for probability of error:

$$\begin{aligned} \mathbb{P}[g(f(X)) = e] &= \mathbb{P} \left[\left\{ \log_2 \frac{1}{P_X(X)} > k - \tau \right\} \cup \{J(X, \mathcal{C}) \neq \emptyset\} \right] \\ &\leq \mathbb{P} \left[\log_2 \frac{1}{P_X(X)} > k - \tau \right] + \mathbb{P}[J(X, \mathcal{C}) \neq \emptyset] \end{aligned}$$

The first term does not depend on the codebook \mathcal{C} , while the second term does. The idea now is to randomize over \mathcal{C} and show that when we average over all possible choices of codebook, the second term is smaller than $2^{-\tau}$. Therefore there exists at least one codebook that achieves the desired bound. Specifically, let us consider \mathcal{C} which is uniformly distributed over all codebooks and independently of X . Equivalently, since \mathcal{C} can be represented by an $|\mathcal{X}| \times k$ binary matrix, whose rows correspond to codewords, we choose each entry to be independent fair coin flips.

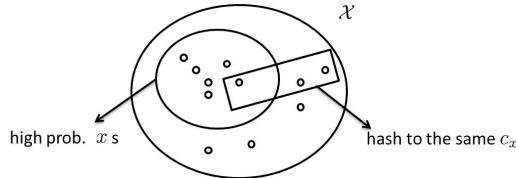
11.1 Fixed-length almost lossless code. Asymptotic Equipartition Property (AEP). 181

Averaging the error probability (over \mathcal{C} and over X), we have

$$\begin{aligned}
 \mathbb{E}_{\mathcal{C}}[\mathbb{P}[J(X, \mathcal{C}) \neq \phi]] &= \mathbb{E}_{\mathcal{C}, X} \left[\mathbb{1}_{\left\{ \exists x' \neq X : \log_2 \frac{1}{P_X(x')} \leq k - \tau, c_{x'} = c_X \right\}} \right] \\
 &\leq \mathbb{E}_{\mathcal{C}, X} \left[\sum_{x' \neq X} \mathbb{1}_{\left\{ \log_2 \frac{1}{P_X(x')} \leq k - \tau \right\}} \mathbb{1}_{\{c_{x'} = c_X\}} \right] \quad (\text{union bound}) \\
 &= 2^{-k} \mathbb{E}_X \left[\sum_{x' \neq X} \mathbb{1}_{\{P_X(x') \geq 2^{-k+\tau}\}} \right] \\
 &\leq 2^{-k} \sum_{x' \in \mathcal{X}} \mathbb{1}_{\{P_X(x') \geq 2^{-k+\tau}\}} \\
 &\leq 2^{-k} 2^{k-\tau} = 2^{-\tau}.
 \end{aligned}$$

□

Remark 11.2 (Why random coding works). The compressor $f(x) = c_x$ can be thought as hashing $x \in \mathcal{X}$ to a random k -bit string $c_x \in \{0, 1\}^k$, as illustrated below:



Here, x has high probability $\Leftrightarrow \log_2 \frac{1}{P_X(x)} \leq k - \tau \Leftrightarrow P_X(x) \geq 2^{-k+\tau}$. Therefore the number of those high-probability x 's is at most $2^{k-\tau}$, which is far smaller than 2^k , the total number of k -bit codewords. Hence the chance of collision is small.

Remark 11.3. The random coding argument is a canonical example of *probabilistic method*: To prove the existence of an object with certain property, we construct a probability distribution (randomize) and show that on average the property is satisfied. Hence there exists at least one realization with the desired property. The downside of this argument is that it is not constructive, i.e., does not give us an algorithm to find the object.

Remark 11.4. This is a subtle point: Notice that in the proof we choose the random codebook to be uniform over all possible codebooks. In other words, $\mathcal{C} = \{c_x : x \in \mathcal{X}\}$ consists of iid k -bit strings. In fact, in the proof we only need pairwise independence, i.e., $c_x \perp\!\!\!\perp c_{x'}$ for any $x \neq x'$ (Why?). Now, why should we care about this? In fact, having access to external randomness is also a lot of resources. It is more desirable to use less randomness in the random coding argument. Indeed, if we use zero randomness, then it is a deterministic construction, which is the best situation! Using pairwise independent codebook requires significantly less randomness than complete random coding which needs $|\mathcal{X}|k$ bits. To see this intuitively, note that one can use 2 independent random bits to generate 3 random bits that is pairwise independent but not mutually independent, e.g., $\{b_1, b_2, b_1 \oplus b_2\}$. This observation is related to linear compression studied in the next section, where the codeword we generated are not iid, but elements of a random linear subspace.

Remark 11.5 (AEP for memoryless sources). Consider iid S^n . By WLLN,

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} H(S). \quad (11.3)$$

For any $\delta > 0$, define the set

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n} \log \frac{1}{P_{S^n}(s^n)} - H(S) \right| \leq \delta \right\}.$$

For example: $S^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(p)$, since $P_{S^n}(s^n) = p^{w(s^n)} q^{n-w(s^n)}$, the typical set corresponds to those sequences whose Hamming weight is close to the expectation: $T_n^\delta = \{s^n \in \{0, 1\}^n : w(s^n) \in [p \pm \delta']n\}$, where δ' is a constant depending on δ .

As a consequence of (11.3),

- 1 $\mathbb{P}[S^n \in T_n^\delta] \rightarrow 1$ as $n \rightarrow \infty$.
- 2 $|T_n^\delta| \leq 2^{(H(S)+\delta)n} \ll |\mathcal{S}|^n$.

In other words, S^n is concentrated on the set T_n^δ which is exponentially smaller than the whole space. In almost lossless compression we can simply encode this set losslessly. Although this is different than the optimal encoding, Corollary 11.1 indicates that in the large- n limit the optimal compressor is no better.

The property (11.3) is often referred as the *Asymptotic Equipartition Property* (AEP), in the sense that the random vector is concentrated on a set wherein each realization is roughly equally likely up to the exponent. Indeed, Note that for any $s^n \in T_n^\delta$, its likelihood is concentrated around $P_{S^n}(s^n) \in 2^{-(H(S)\pm\delta)n}$, called δ -typical sequences.

11.2 Linear Compression

Recall from Shannon's theorem (Corollary 11.1) that as the blocklength $n \rightarrow \infty$, the optimal probability of error $\epsilon^*(X, nR)$ tends to zero (resp. one) if the compression rate is strictly above (resp. below) the entropy. Complementing the achievability result in Section 11.1, for example, Theorem 11.6 obtained by randomizing over all compressors, the goal of this section is to find compressor with *structures*. The simplest conceivable case is probably linear functions, which is also highly desirable for its simplicity (low complexity). Of course, we have to be on a vector space where we can define linear operations. In this part, we assume that the source takes the form $X = S^n$, where each coordinate is an element of a finite field (Galois field), i.e., $S_i \in \mathbb{F}_q$, where q is the cardinality of \mathbb{F}_q . (This is only possible if $q = p^k$ for some prime number p and $k \in \mathbb{N}$.)

Definition 11.7 (Galois field). F is a finite set with operations $(+, \cdot)$ where

- The addition operation $+$ is associative and commutative.
- The multiplication operation \cdot is associative and commutative.
- There exist elements $0, 1 \in F$ s.t. $0 + a = 1 \cdot a = a$.

11.2 Linear Compression 183

- $\forall a, \exists -a$, s.t. $a + (-a) = 0$
- $\forall a \neq 0, \exists a^{-1}$, s.t. $a^{-1}a = 1$
- Distributive: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$

Simple examples of finite fields:

- $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, where p is prime.
- $\mathbb{F}_4 = \{0, 1, x, x+1\}$ with addition and multiplication as polynomials in $\mathbb{F}_2[x]$ modulo $x^2 + x + 1$.

A linear compressor is a linear function $H : \mathbb{F}_q^n \rightarrow \mathbb{F}_q^k$ (represented by a matrix $H \in \mathbb{F}_q^{k \times n}$) that maps each $x \in \mathbb{F}_q^n$ to its codeword $w = Hx$, namely

$$\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & & \vdots \\ h_{k1} & \dots & h_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Compression is achieved if $k < n$, i.e., H is a fat matrix, which, again, is only possible in the almost lossless sense.

Theorem 11.8 (Achievability). *Let $X \in \mathbb{F}_q^n$ be a random vector. $\forall \tau > 0, \exists$ linear compressor $H : \mathbb{F}_q^n \rightarrow \mathbb{F}_q^k$ and decompressor $g : \mathbb{F}_q^k \rightarrow \mathbb{F}_q^n \cup \{\epsilon\}$, s.t.*

$$\mathbb{P}[g(HX) \neq X] \leq \mathbb{P}\left[\log_q \frac{1}{P_X(X)} > k - \tau\right] + q^{-\tau}$$

Remark 11.6. Consider the Hamming space $q = 2$. In comparison with Shannon's random coding achievability, which uses $k2^n$ bits to construct a completely random codebook, here for linear codes we need kn bits to randomly generate the matrix H , and the codebook is a k -dimensional linear subspace of the Hamming space.

Proof. Fix τ . As pointed in the proof of Shannon's random coding theorem (Theorem 11.6), given the compressor H , the optimal decompressor is the MAP decoder, i.e., $g(w) = \operatorname{argmax}_{x: Hx=w} P_X(x)$, which outputs the most likely symbol that is compatible with the codeword received. Instead, let us consider the following (suboptimal) decoder for its ease of analysis:

$$g(w) = \begin{cases} x & \exists! x \in \mathbb{F}_q^n : w = Hx, x - h.p. \\ \epsilon & \text{otherwise} \end{cases}$$

where we used the short-hand:

$$x - h.p. \text{ (high probability)} \Leftrightarrow \log_q \frac{1}{P_X(x)} < k - \tau \Leftrightarrow P_X(x) \geq q^{-k+\tau}.$$

Note that this decoder is the same as in the proof of Theorem 11.6. The proof is also mostly the same, except now hash collisions occur under the linear map H . By union bound,

$$\mathbb{P}[g(HX) = \epsilon] \leq \mathbb{P}\left[\log_q \frac{1}{P_X(x)} > k - \tau\right] + \mathbb{P}[\exists x' - h.p. : x' \neq X, Hx' = HX]$$

184

$$(\text{union bound}) \leq \mathbb{P} \left[\log_q \frac{1}{P_X(x)} > k - \tau \right] + \sum_x P_X(x) \sum_{x' - h.p., x' \neq x} 1\{Hx' = Hx\}$$

Now we use random coding to average the second term over all possible choices of H . Specifically, choose H as a matrix independent of X where each entry is iid and uniform on \mathbb{F}_q . For distinct x_0 and x_1 , the collision probability is

$$\begin{aligned} \mathbb{P}_H[Hx_1 = Hx_0] &= \mathbb{P}_H[Hx_2 = 0] && (x_2 \triangleq x_1 - x_0 \neq 0) \\ &= \mathbb{P}_H[H_1 \cdot x_2 = 0]^k && (\text{iid rows}) \end{aligned}$$

where H_1 is the first row of the matrix H , and each row of H is independent. This is the probability that H_i is in the orthogonal complement of x_2 . On \mathbb{F}_q^n , the orthogonal complement of a given non-zero vector has cardinality q^{n-1} . So the probability for the first row to lie in this subspace is $q^{n-1}/q^n = 1/q$, hence the collision probability $1/q^k$. Averaging over H gives

$$\mathbb{E}_H \sum_{x' - h.p., x' \neq x} 1\{Hx' = Hx\} = \sum_{x' - h.p., x' \neq x} \mathbb{P}_H[Hx' = Hx] = |\{x' : x' - h.p., x' \neq x\}| q^{-k} \leq q^{k-\tau} q^{-k} = q^{-\tau}$$

Thus the bound holds. \square

Remark 11.7. 1 Compared to Theorem 11.6, which is obtained by randomizing over all possible compressors, Theorem 11.12 is obtained by randomizing over only linear compressors, and the bound we obtained is identical. Therefore restricting on linear compression almost does not lose anything.

- 2 Note that in this case it is not possible to make all errors detectable.
- 3 Can we loosen the requirement on \mathbb{F}_q to instead be a commutative ring? In general, no, since zero divisors in the commutative ring ruin the key proof ingredient of low collision probability in the random hashing. E.g. in $\mathbb{Z}/6\mathbb{Z}$

$$\mathbb{P} \left[H \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0 \right] = 6^{-k} \quad \text{but} \quad \mathbb{P} \left[H \begin{bmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0 \right] = 3^{-k},$$

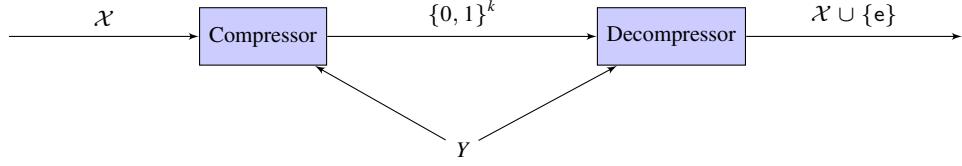
since $0 \cdot 2 = 3 \cdot 2 = 0$ in $\mathbb{Z}/6\mathbb{Z}$.

11.3 Compression with side information at both compressor and decompressor

Definition 11.9 (Compression with Side Information). Given $P_{X,Y}$,

- $f: \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}^k$
- $g: \{0, 1\}^k \times \mathcal{Y} \rightarrow \mathcal{X} \cup \{\epsilon\}$
- $\mathbb{P}[g(f(X, Y), Y) \neq X] < \epsilon$
- Fundamental Limit: $\epsilon^*(X|Y, k) = \inf\{\epsilon : \exists(k, \epsilon) - S.I. code\}$

11.4 Slepian-Wolf (Compression with side information at decompressor only) 185



Note that here unlike the source X , the side information Y need not be discrete. Conditioned on $Y = y$, the problem reduces to compression without side information studied in Section 11.1, where the source X is distributed according to $P_{X|Y=y}$. Since Y is known to both the compressor and decompressor, they can use the best code tailored for this distribution. Recall $\epsilon^*(X, k)$ defined in Definition 11.1, the optimal probability of error for compressing X using k bits, which can also be denoted by $\epsilon^*(P_X, k)$. Then we have the following relationship

$$\epsilon^*(X|Y, k) = \mathbb{E}_{y \sim P_Y}[\epsilon^*(P_{X|Y=y}, k)],$$

which allows us to apply various bounds developed before.

Theorem 11.10.

$$\mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} > k + \tau\right] - 2^{-\tau} \leq \epsilon^*(X|Y, k) \leq \mathbb{P}\left[\log_2 \frac{1}{P_{X|Y}(X|Y)} > k - \tau\right] + 2^{-\tau}, \quad \forall \tau > 0$$

Corollary 11.11. Let $(X, Y) = (S^n, T^n)$ where the pairs $(S_i, T_i) \stackrel{i.i.d.}{\sim} P_{ST}$. Then

$$\lim_{n \rightarrow \infty} \epsilon^*(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Proof. Using the converse Theorem 11.4 and achievability Theorem 11.6 (or Theorem 11.5) for compression without side information, we have

$$\mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|y)} > k + \tau \mid Y = y\right] - 2^{-\tau} \leq \epsilon^*(P_{X|Y=y}, k) \leq \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|y)} > k \mid Y = y\right]$$

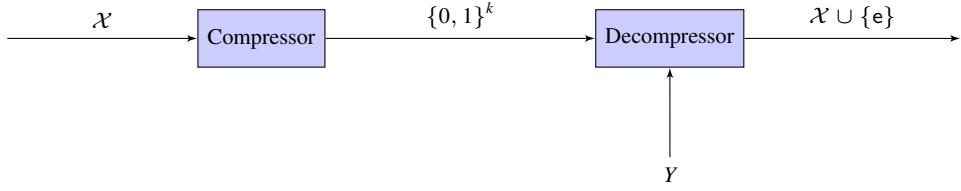
By taking the average over all $y \sim P_Y$, we get the theorem. For the corollary

$$\frac{1}{n} \log \frac{1}{P_{S^n|T^n}(S^n|T^n)} = \frac{1}{n} \sum_{i=1}^n \log \frac{1}{P_{S|T}(S_i|T_i)} \xrightarrow{\mathbb{P}} H(S|T)$$

as $n \rightarrow \infty$, using the WLLN. \square

11.4 Slepian-Wolf (Compression with side information at decompressor only)

Consider the compression with side information problem, except now the compressor has no access to the side information.



Definition 11.12 (S.W. code). Given $P_{X,Y}$,

- $f: \mathcal{X} \rightarrow \{0, 1\}^k$
- $g: \{0, 1\}^k \times \mathcal{Y} \rightarrow \mathcal{X} \cup \{\text{e}\}$
- $\mathbb{P}[g(f(X), Y) \neq X] \leq \epsilon$
- Fundamental Limit: $\epsilon_{\text{SW}}^*(X|Y, k) = \inf\{\epsilon : \exists (k, \epsilon)\text{-S.W. code}\}$

Now the very surprising result: Even without side information at the compressor, we can still compress down to the conditional entropy!

Theorem 11.13 (Slepian-Wolf, '73).

$$\epsilon^*(X|Y, k) \leq \epsilon_{\text{SW}}^*(X|Y, k) \leq \mathbb{P} \left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \right] + 2^{-\tau}$$

Corollary 11.14.

$$\lim_{n \rightarrow \infty} \epsilon_{\text{SW}}^*(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Remark 11.8. Definition 11.17 does not include the zero-undected-error condition (that is $g(f(x), y) = x$ or e). In other words, we allow for the possibility of undetected errors. Indeed, if we require this condition, the side-information savings will be mostly gone. Indeed, assuming $P_{X,Y}(x,y) > 0$ for all (x,y) it is clear that under zero-undected-error condition, if $f(x_1) = f(x_2) = c$ then $g(c) = \text{e}$. Thus except for c all other elements in $\{0, 1\}^k$ must have unique preimages. Similarly, one can show that Slepian-Wolf theorem does not hold in the setting of variable-length lossless compression (i.e. average length is $H(X)$ not $H(X|Y)$.)

Proof. LHS is obvious, since side information at the compressor and decompressor is better than only at the decompressor.

For the RHS, first generate a random codebook with iid uniform codewords: $C = \{c_x \in \{0, 1\}^k : x \in \mathcal{X}\}$ independently of (X, Y) , then define the compressor and decoder as

$$f(x) = c_x$$

11.5 Multi-terminal Slepian Wolf 187

$$g(w, y) = \begin{cases} x & \exists!x : c_x = w, x - h.p.|y \\ 0 & \text{o.w.} \end{cases}$$

where we used the shorthand $x - h.p.|y \Leftrightarrow \log_2 \frac{1}{P_{X|Y}(x|y)} < k - \tau$. The error probability of this scheme, as a function of the code book C , is

$$\begin{aligned} \mathcal{E}(C) &= \mathbb{P} \left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \text{ or } J(X, C|Y) \neq \emptyset \right] \\ &\leq \mathbb{P} \left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \right] + \mathbb{P} [J(X, C|Y) \neq \emptyset] \\ &= \mathbb{P} \left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \right] + \sum_{x,y} P_{X,Y}(x,y) \mathbf{1}_{\{J(x,C|y) \neq \emptyset\}}. \end{aligned}$$

where $J(x, C|y) \triangleq \{x' \neq x : x' - h.p.|y, c_{x'} = c_x\}$.

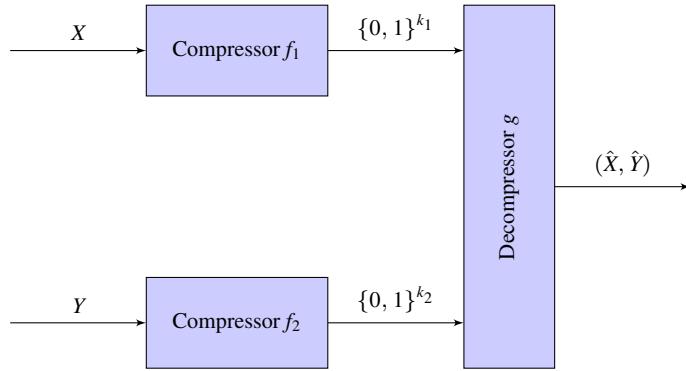
Now averaging over C and applying the union bound: use $|\{x' : x' - h.p.|y\}| \leq 2^{k-\tau}$ and $\mathbb{P}[c_{x'} = c_x] = 2^{-k}$ for any $x \neq x'$,

$$\begin{aligned} \mathbb{P}_C[J(x, C|y) \neq \emptyset] &\leq \mathbb{E}_C \left[\sum_{x' \neq x} \mathbf{1}_{\{x' - h.p.|y\}} \mathbf{1}_{\{c_{x'} = c_x\}} \right] \\ &= 2^{k-\tau} \mathbb{P}[c_{x'} = c_x] \\ &= 2^{-\tau} \end{aligned}$$

Hence the theorem follows as usual from two terms in the union bound. \square

11.5 Multi-terminal Slepian Wolf

Distributed compression: Two sources are correlated. Compress individually, decompress jointly. What are those rate pairs that guarantee successful reconstruction?



Definition 11.15. Given $P_{X,Y}$,

- (f_1, f_2, g) is (k_1, k_2, ϵ) -code if $f_1 : \mathcal{X} \rightarrow \{0, 1\}^{k_1}, f_2 : \mathcal{Y} \rightarrow \{0, 1\}^{k_2}, g : \{0, 1\}^{k_1} \times \{0, 1\}^{k_2} \rightarrow \mathcal{X} \times \mathcal{Y}$, s.t. $\mathbb{P}[(\hat{X}, \hat{Y}) \neq (X, Y)] \leq \epsilon$, where $(\hat{X}, \hat{Y}) = g(f_1(X), f_2(Y))$.
- Fundamental limit: $\epsilon_{\text{SW}}^*(X, Y, k_1, k_2) = \inf\{\epsilon : \exists (k_1, k_2, \epsilon)\text{-code}\}$.

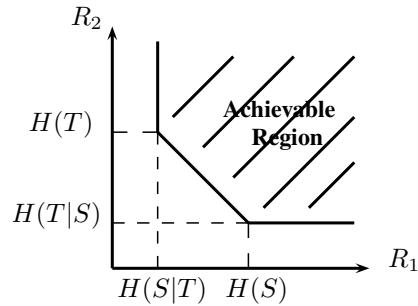
Theorem 11.16. $(X, Y) = (S^n, T^n)$ - iid pairs

$$\lim_{n \rightarrow \infty} \epsilon_{\text{SW}}^*(S^n, T^n, nR_1, nR_2) = \begin{cases} 0 & (R_1, R_2) \in \text{int}(\mathcal{R}_{\text{SW}}) \\ 1 & (R_1, R_2) \notin \mathcal{R}_{\text{SW}} \end{cases}$$

where \mathcal{R}_{SW} denotes the Slepian-Wolf rate region

$$\mathcal{R}_{\text{SW}} = \left\{ (a, b) : \begin{array}{l} a \geq H(S|T) \\ b \geq H(T|S) \\ a + b \geq H(S, T) \end{array} \right\}$$

The rate region \mathcal{R}_{SW} typically looks like:



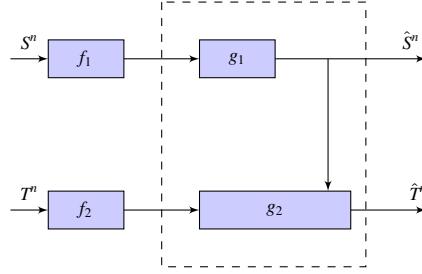
Since $H(T) - H(T|S) = H(S) - H(S|T) = I(S; T)$, the slope is -1 .

Proof. Converse: Take $(R_1, R_2) \notin \mathcal{R}_{\text{SW}}$. Then one of three cases must occur:

- 1 $R_1 < H(S|T)$. Then even if encoder and decoder had full T^n , still can't achieve this (from compression with side info result – Corollary 11.2).
- 2 $R_2 < H(T|S)$ (same).
- 3 $R_1 + R_2 < H(S, T)$. Can't compress below the joint entropy of the pair (S, T) .

Achievability: First note that we can achieve the two corner points. The point $(H(S), H(T|S))$ can be approached by almost lossless compressing S at entropy and compressing T with side information S at the decoder. To make this rigorous, let $k_1 = n(H(S) + \delta)$ and $k_2 = n(H(T|S) + \delta)$. By Corollary 11.1, there exist $f_1 : \mathcal{S}^n \rightarrow \{0, 1\}^{k_1}$ and $g_1 : \{0, 1\}^{k_1} \rightarrow \mathcal{S}^n$ s.t. $\mathbb{P}[g_1(f_1(S^n)) \neq S^n] \leq \epsilon_n \rightarrow 0$. By Theorem 11.18, there exist $f_2 : \mathcal{T}^n \rightarrow \{0, 1\}^{k_2}$ and $g_2 : \{0, 1\}^{k_2} \times \mathcal{S}^n \rightarrow \mathcal{T}^n$ s.t. $\mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \leq \epsilon_n \rightarrow 0$. Now that S^n is not available, feed the S.W. decompressor with $g(f(S^n))$ and define the joint decompressor by $g(w_1, w_2) = (g_1(w_1), g_2(w_2, g_1(w_1)))$ (see below):

11.6* Source-coding with a helper (Ahlswede-Körner-Wyner) 189



Apply union bound:

$$\begin{aligned}
 & \mathbb{P}[g(f_1(S^n), f_2(T^n)) \neq (S^n, T^n)] \\
 &= \mathbb{P}[g_1(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n)), g(f_1(S^n))) \neq T^n, g_1(f_1(S^n)) = S^n] \\
 &\leq \mathbb{P}[g_1(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \\
 &\leq 2\epsilon_n \rightarrow 0.
 \end{aligned}$$

Similarly, the point $(H(S), H(T|S))$ can be approached.

To achieve other points in the region, use the idea of **time sharing**: If you can achieve with vanishing error probability any two points (R_1, R_2) and (R'_1, R'_2) , then you can achieve for $\lambda \in [0, 1]$, $(\lambda R_1 + \bar{\lambda} R'_1, \lambda R_2 + \bar{\lambda} R'_2)$ by dividing the block of length n into two blocks of length λn and $\bar{\lambda} n$ and apply the two codes respectively

$$\begin{aligned}
 (S_1^{\lambda n}, T_1^{\lambda n}) &\rightarrow \begin{bmatrix} \lambda n R_1 \\ \lambda n R_2 \end{bmatrix} \quad \text{using } (R_1, R_2) \text{ code} \\
 (S_{\lambda n+1}^n, T_{\lambda n+1}^n) &\rightarrow \begin{bmatrix} \bar{\lambda} n R'_1 \\ \bar{\lambda} n R'_2 \end{bmatrix} \quad \text{using } (R'_1, R'_2) \text{ code}
 \end{aligned}$$

(Exercise: Write down the details rigorously.) Therefore, all convex combinations of points in the achievable regions are also achievable, so the achievable region must be convex. \square

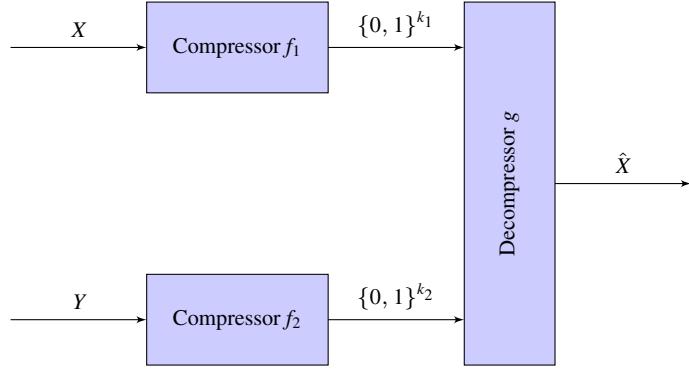
11.6* Source-coding with a helper (Ahlswede-Körner-Wyner)

Yet another variation of distributed compression problem is compressing X with a helper, see figure below. Note that the main difference from the previous section is that decompressor is only required to produce the estimate of X , using rate-limited help from an observer who has access to Y . Characterization of rate pairs R_1, R_2 is harder than in the previous section.

Theorem 11.17 (Ahlswede-Körner-Wyner). *Consider i.i.d. source $(X^n, Y^n) \sim P_{X,Y}$ with X discrete. If rate pair (R_1, R_2) is achievable with vanishing probability of error $\mathbb{P}[\hat{X}^n \neq X^n] \rightarrow 0$, then there exists an auxiliary random variable U taking values on alphabet of cardinality $|\mathcal{Y}| + 1$ such that $P_{X,Y,U} = P_{X,Y}P_{U|X,Y}$ and*

$$R_1 \geq H(X|U), R_2 \geq I(Y; U). \quad (11.4)$$

190



Furthermore, for every such random variable U the rate pair $(H(X|U), I(Y; U))$ is achievable with vanishing error.

Proof. We only sketch some crucial details.

First, note that iterating over all possible random variables U (without cardinality constraint) the set of pairs (R_1, R_2) satisfying (11.4) is convex. Next, consider a compressor $W_1 = f_1(X^n)$ and $W_2 = f_2(Y^n)$. Then from Fano's inequality (6.9) assuming $\mathbb{P}[X^n \neq \hat{X}^n] = o(1)$ we have

$$H(X^n|W_1, W_2) = o(n).$$

Thus, from chain rule and conditioning-decreases-entropy, we get

$$nR_1 \geq I(X^n; W_1|W_2) \geq H(X^n|W_2) - o(n) \quad (11.5)$$

$$= \sum_{k=1}^n H(X_k|W_2, X^{k-1}) - o(n) \quad (11.6)$$

$$\geq \sum_{k=1}^n H(X_k| \underbrace{W_2, X^{k-1}, Y^{k-1}}_{\triangleq U_k}) - o(n) \quad (11.7)$$

On the other hand, from (6.2) we have

$$nR_2 \geq I(W_2; Y^n) = \sum_{k=1}^n I(W_2; Y_k|Y^{k-1}) \quad (11.8)$$

$$= \sum_{k=1}^n I(W_2, X^{k-1}; Y_k|Y^{k-1}) \quad (11.9)$$

$$= \sum_{k=1}^n I(W_2, X^{k-1}, Y^{k-1}; Y_k) \quad (11.10)$$

where (11.9) follows from $I(W_2, X^{k-1}; Y_k|Y^{k-1}) = I(W_2; Y_k|Y^{k-1}) + I(X^{k-1}; Y_k|W_2, Y^{k-1})$ and the fact that $(W_2, Y_k) \perp\!\!\!\perp X^{k-1}|Y^{k-1}$; and (11.10) from $Y^{k-1} \perp\!\!\!\perp Y_k$. Comparing (11.7) and (11.10) we

11.6* Source-coding with a helper (Ahlswede-Körner-Wyner) 191

notice that denoting $U_k = (W_2, X^{k-1}, Y^{k-1})$ we have

$$(R_1, R_2) \geq \frac{1}{n} \sum_{k=1}^n (H(X_k|U_k), I(U_k; Y_k))$$

and thus (from convexity) the rate pair must belong to the region spanned by all pairs $(H(X|U), I(U; Y))$.

To show that without loss of generality the auxiliary random variable U can be chosen to take at most $|\mathcal{Y}| + 1$ values, one can invoke Carathéodory's theorem (see Lemma 7.20). We omit the details.

Finally, showing that for each U the mentioned rate-pair is achievable, we first notice that if there were side information at the decompressor in the form of the i.i.d. sequence U^n correlated to X^n , then Slepian-Wolf theorem implies that only rate $R_1 = H(X|U)$ would be sufficient to reconstruct X^n . Thus, the question boils down to creating a correlated sequence U^n at the decompressor by using the minimal rate R_2 . This is the content of the so called covering lemma – see Theorem 25.7: It is sufficient to use rate $I(U; Y)$ to do so. We omit further details. \square

12

Compressing stationary ergodic sources

We have studyig the compression of i.i.d. sequence $\{S_i\}$, for which

$$\frac{1}{n}l(f^*(S^n)) \xrightarrow{\mathbb{P}} H(S) \quad (12.1)$$

$$\lim_{n \rightarrow \infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases} \quad (12.2)$$

In this chapter, we shall examine similar results for ergodic processes and we first state the main theory as follows:

Theorem 12.1 (Shannon-McMillan). *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process, then*

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} \mathcal{H}, \quad \text{also a.s. and in } L_1 \quad (12.3)$$

where $\mathcal{H} = \lim_{n \rightarrow \infty} \frac{1}{n} H(S^n)$ is the entropy rate.

Corollary 12.2. *For any stationary and ergodic discrete process $\{S_1, S_2, \dots\}$, (12.1)–(12.2) hold with $H(S)$ replaced by \mathcal{H} .*

Proof. Shannon-McMillan (we only need convergence in probability) + Theorem 10.7 + Theorem 11.2 which tie together the respective CDF of the random variable $l(f^*(S^n))$ and $\log \frac{1}{P_{S^n}(S^n)}$. \square

In Chapter 11 we learned the asymptotic equipartition property (AEP) for iid sources. Here we generalize it to stationary ergodic sources thanks to Shannon-McMillan.

Corollary 12.3 (AEP for stationary ergodic sources). *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process. For any $\delta > 0$, define the set*

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \leq \delta \right\}.$$

Then

- 1 $\mathbb{P}[S^n \in T_n^\delta] \rightarrow 1$ as $n \rightarrow \infty$.
- 2 $2^{n(\mathcal{H}-\delta)}(1+o(1)) \leq |T_n^\delta| \leq 2^{n(\mathcal{H}+\delta)}(1+o(1))$.

Some historical notes are in order. Convergence in probability for stationary ergodic Markov chains was already shown in [268]. The extension to convergence in L_1 for all stationary ergodic processes is due to McMillan in [211], and to almost sure convergence to Breiman [51]. A modern proof is in [6]. Note also that for a Markov chain, existence of typical sequences and the AEP can be anticipated by thinking of a Markov process as a sequence of independent decisions regarding which transitions to take at each state. It is then clear that Markov process's trajectory is simply a transformation of trajectories of an iid process, hence must concentrate similarly.

12.1 Bits of ergodic theory

Let's start with a dynamic system view and introduce a few definitions:

Definition 12.4 (Measure preserving transformation). $\tau : \Omega \rightarrow \Omega$ is measure preserving (more precisely, probability preserving) if

$$\forall E \in \mathcal{F}, P(E) = P(\tau^{-1}E).$$

The set E is called τ -invariant if $E = \tau^{-1}E$. The set of all τ -invariant sets forms a σ -algebra (exercise) denoted \mathcal{F}_{inv} .

Definition 12.5 (Stationary process). A process $\{S_n, n = 0, \dots\}$ is stationary if there exists a measure preserving transformation $\tau : \Omega \rightarrow \Omega$ such that:

$$S_j = S_{j-1} \circ \tau = S_0 \circ \tau^j$$

Therefore a stationary process can be described by the tuple $(\Omega, \mathcal{F}, \mathbb{P}, \tau, S_0)$ and $S_k = S_0 \circ \tau^k$.

Remark 12.1.

- 1 Alternatively, a random process (S_0, S_1, S_2, \dots) is stationary if its joint distribution is invariant with respect to shifts in time, i.e., $P_{S_n^m} = P_{S_{n+t}^{m+t}}$, $\forall n, m, t$. Indeed, given such a process we can define a m.p.t. as follows:

$$(s_0, s_1, \dots) \xrightarrow{\tau} (s_1, s_2, \dots) \tag{12.4}$$

So τ is a shift to the right.

- 2 An event $E \in \mathcal{F}$ is shift-invariant if

$$(s_1, s_2, \dots) \in E \Rightarrow (s_0, s_1, s_2, \dots) \in E, \quad s_0$$

or equivalently $E = \tau^{-1}E$ (exercise). Thus τ -invariant events are also called shift-invariant, when τ is interpreted as (12.4).

- 3 Some examples of shift-invariant events are $\{\exists n : x_i = 0 \forall i \geq n\}$, $\{\limsup x_i < 1\}$ etc. A non shift-invariant event is $A = \{x_0 = x_1 = \dots = 0\}$, since $\tau(1, 0, 0, \dots) \in A$ but $(1, 0, \dots) \notin A$.

4 Also recall that the tail σ -algebra is defined as

$$\mathcal{F}_{tail} \triangleq \bigcap_{n \geq 1} \sigma\{S_n, S_{n+1}, \dots\}.$$

It is easy to check that all shift-invariant events belong to \mathcal{F}_{tail} . The inclusion is strict, as for example the event

$$\{\exists n : x_i = 0, \forall \text{odd } i \geq n\}$$

is in \mathcal{F}_{tail} but not shift-invariant.

Proposition 12.6 (Poincare recurrence). *Let τ be measure-preserving for $(\Omega, \mathcal{F}, \mathbb{P})$. Then for any measurable A with $\mathbb{P}[A] > 0$ we have*

$$\mathbb{P}\left[\bigcup_{k \geq 1} \tau^{-k} A | A\right] = \mathbb{P}[\tau^k(\omega) \in A \text{ occurs infinitely often} | A] = 1.$$

Proof. Let $B = \bigcup_{k \geq 1} \tau^{-k} A$. It is sufficient to show that $\mathbb{P}[A \cap B] = \mathbb{P}[A]$ or equivalently

$$\mathbb{P}[A \cup B] = \mathbb{P}[B]. \quad (12.5)$$

To that end notice that $\tau^{-1}A \cup \tau^{-1}B = B$ and thus

$$\mathbb{P}[\tau^{-1}(A \cup B)] = \mathbb{P}[B],$$

but the left-hand side equals $\mathbb{P}[A \cup B]$ by the measure-preservation of τ , proving (12.5). \square

Consider τ mapping initial state of the conservative (Hamiltonian) mechanical system to its state after passage of a given unit of time. It is known that τ preserves Lebesgue measure in phase space (Liouville's theorem). Thus Poincare recurrence leads to a rather counter-intuitive conclusions. For example, opening the barrier separating two gases in a cylinder allows them to mix. Poincare recurrence says that eventually they will return back to the original separated state (with each gas occupying roughly its half of the cylinder). Of course, the “paradox” is resolved by observing that it will take unphysically long for this to happen.

Definition 12.7 (Ergodicity). A transformation τ is ergodic if $\forall E \in \mathcal{F}_{inv}$ we have $\mathbb{P}[E] = 0$ or 1. A process $\{S_i\}$ is ergodic if all shift invariant events are deterministic, i.e., for any shift invariant event E , $\mathbb{P}[S_1^\infty \in E] = 0$ or 1.

Here are some examples:

- $\{S_k = k^2\}$: ergodic but not stationary
- $\{S_k = S_0\}$: stationary but not ergodic (unless S_0 is a constant). Note that the singleton set $E = \{(s, s, \dots)\}$ is shift invariant and $\mathbb{P}[S_1^\infty \in E] = \mathbb{P}[S_0 = s] \in (0, 1)$ – not deterministic.
- $\{S_k\}$ i.i.d. is stationary and ergodic (by Kolmogorov's 0-1 law, tail events have no randomness)

12.1 Bits of ergodic theory 195

- (Sliding-window construction of ergodic processes)

If $\{S_i\}$ is ergodic, then $\{X_i = f(S_i, S_{i+1}, \dots)\}$ is also ergodic. It is called a **B-process** if S_i is i.i.d.

Example, $S_i \sim \text{Ber}(\frac{1}{2})$ i.i.d., $X_k = \sum_{n=0}^{\infty} 2^{-n-1} S_{k+n} = 2X_{k-1} \pmod{1}$. The marginal distribution of X_i is uniform on $[0, 1]$. Note that X_k 's behavior is completely deterministic: given X_0 , all the future X_k 's are determined exactly. This example shows that certain deterministic maps exhibit ergodic/chaotic behavior under iterative application: although the trajectory is completely deterministic, its time-averages converge to expectations and in general “look random”.

- There are also stronger conditions than ergodicity. Namely, we say that τ is mixing (or strong mixing) if

$$\mathbb{P}[A \cap \tau^{-n}B] \rightarrow \mathbb{P}[A]\mathbb{P}[B].$$

We say that τ is weakly mixing if

$$\sum_{k=1}^n \frac{1}{n} |\mathbb{P}[A \cap \tau^{-n}B] - \mathbb{P}[A]\mathbb{P}[B]| \rightarrow 0.$$

Strong mixing implies weak mixing, which implies ergodicity (exercise).

- $\{S_i\}$: finite irreducible Markov chain with recurrent states is ergodic (in fact strong mixing), regardless of initial distribution.

Toy example: kernel $P(0|1) = P(1|0) = 1$ with initial dist. $P(S_0 = 0) = 0.5$. This process only has two sample paths: $\mathbb{P}[S_1^\infty = (010101\dots)] = \mathbb{P}[S_1^\infty = (101010\dots)] = \frac{1}{2}$. It is easy to verify this process is ergodic (in the sense defined above!). Note however, that in Markov-chain literature a chain is called ergodic if it is irreducible, aperiodic and recurrent. This example does not satisfy this definition (this clash of terminology is a frequent source of confusion).

- (optional) $\{S_i\}$: stationary zero-mean Gaussian process with autocovariance function $R(n) = \mathbb{E}[S_0 S_n^*]$.

$$\lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{t=0}^n R[t] = 0 \Leftrightarrow \{S_i\} \text{ ergodic} \Leftrightarrow \{S_i\} \text{ weakly mixing}$$

$$\lim_{n \rightarrow \infty} R[n] = 0 \Leftrightarrow \{S_i\} \text{ mixing}$$

Intuitively speaking, an ergodic process can have infinite memory in general, but the memory is weak. Indeed, we see that for a stationary Gaussian process ergodicity means the correlation dies (in the Cesaro-mean sense).

The *spectral measure* is defined as the (discrete time) Fourier transform of the autocovariance sequence $\{R(n)\}$, in the sense that there exists a unique probability measure μ on $[-\frac{1}{2}, \frac{1}{2}]$ such that $R(n) = \mathbb{E} \exp(i2n\pi X)$ where $X \sim \mu$. The spectral criteria can be formulated as follows:

$\{S_i\}$ ergodic \Leftrightarrow spectral measure has no atoms (CDF is continuous)

$\{S_i\}$ B-process \Leftrightarrow spectral measure has density

Detailed exposition on stationary Gaussian processes can be found in [100, Theorem 9.3.2, pp. 474, Theorem 9.7.1, pp. 493–494].

12.2 Proof of the Shannon-McMillan Theorem

We shall show the L_1 -convergence, which implies convergence in probability automatically. To this end, let us first introduce Birkhoff-Khintchine's convergence theorem for ergodic processes, the proof of which is presented in the next subsection. The interpretation of this result is that time averages converge to the ensemble average.

Theorem 12.8 (Birkhoff-Khintchine's Ergodic Theorem). *Let $\{S_i\}$ be a stationary and ergodic process. For any integral function f , i.e., $\mathbb{E}|f(S_1, \dots)| < \infty$,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(S_k, \dots) = \mathbb{E}f(S_1, \dots). \quad \text{a.s. and in } L_1.$$

In the special case where f depends on finitely many coordinates, say, $f = f(S_1, \dots, S_m)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(S_k, \dots, S_{k+m-1}) = \mathbb{E}f(S_1, \dots, S_m). \quad \text{a.s. and in } L_1.$$

Example 12.1. Consider $f = f(S_1)$.

- For iid $\{S_i\}$, Theorem 12.7 is SLLN (strong LLN).
- For $\{S_i\}$ such that $S_i = S_1$ for all i , which is non-ergodic, Theorem 12.7 fails unless S_1 is a constant.

Definition 12.9. $\{S_i : i \in \mathbb{N}\}$ is an m^{th} order Markov chain if $P_{S_{t+1}|S'_1} = P_{S_{t+1}|S'_{t-m+1}}$ for all $t \geq m$. It is called time homogeneous if $P_{S_{t+1}|S'_{t-m+1}} = P_{S_{m+1}|S'_1}$.

Remark 12.2. Showing (12.3) for an m^{th} order time homogeneous Markov chain $\{S_i\}$ is a direct application of Birkhoff-Khintchine.

$$\begin{aligned} \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} &= \frac{1}{n} \sum_{t=1}^n \log \frac{1}{P_{S_t|S^{t-1}}(S_t|S^{t-1})} \\ &= \frac{1}{n} \log \frac{1}{P_{S^m}(S^m)} + \frac{1}{n} \sum_{t=m+1}^n \log \frac{1}{P_{S_t|S^{t-1}_{t-m}}(S_t|S^{t-1}_{t-m})} \\ &= \underbrace{\frac{1}{n} \log \frac{1}{P_{S_1}(S_1^m)}}_{\rightarrow 0} + \underbrace{\frac{1}{n} \sum_{t=m+1}^n \log \frac{1}{P_{S_{m+1}|S_1^m}(S_t|S^{t-1}_{t-m})}}_{\rightarrow H(S_{m+1}|S_1^m) \text{ by Birkhoff-Khintchine}}, \end{aligned} \tag{12.6}$$

where we applied Theorem 12.7 with $f(s_1, s_2, \dots) = \log \frac{1}{P_{S_{m+1}|S_1^m}(s_{m+1}|s_1^m)}$.

12.2 Proof of the Shannon-McMillan Theorem 197

Now let's prove (12.3) for a general stationary ergodic process $\{S_i\}$ which might have infinite memory. The idea is to approximate the distribution of that ergodic process by an m -th order MC (finite memory) and make use of (12.6); then let $m \rightarrow \infty$ to make the approximation accurate (*Markov approximation*).

Proof of Theorem 12.1 in L_1 . To show that (12.3) converges in L_1 , we want to show that

$$\mathbb{E} \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \rightarrow 0, \quad n \rightarrow \infty.$$

To this end, fix an $m \in \mathbb{N}$. Define the following auxiliary distribution for the process:

$$\begin{aligned} Q^{(m)}(S_1^\infty) &= P_{S_1^m}(S_1^m) \prod_{t=m+1}^{\infty} P_{S_t|S_{t-m}^{t-1}}(S_t|S_{t-m}^{t-1}) \\ &\stackrel{\text{stat.}}{=} P_{S_1^m}(S_1^m) \prod_{t=m+1}^{\infty} P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1}) \end{aligned}$$

Note that under $Q^{(m)}$, $\{S_i\}$ is an m th-order time-homogeneous Markov chain.

By triangle inequality,

$$\begin{aligned} \mathbb{E} \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| &\leq \underbrace{\mathbb{E} \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \frac{1}{n} \log \frac{1}{Q_{S^n}^{(m)}(S^n)} \right|}_{\triangleq A} \\ &\quad + \underbrace{\mathbb{E} \left| \frac{1}{n} \log \frac{1}{Q_{S^n}^{(m)}(S^n)} - H_m \right|}_{\triangleq B} + \underbrace{|H_m - \mathcal{H}|}_{\triangleq C} \end{aligned}$$

where $H_m \triangleq H(S_{m+1}|S_1^m)$.

Now

- $C = |H_m - \mathcal{H}| \rightarrow 0$ as $m \rightarrow \infty$ by Theorem 5.4 (Recall that for stationary processes: $H(S_{m+1}|S_1^m) \rightarrow H$ from above).
- As shown in Remark 12.9, for any fixed m , $B \rightarrow 0$ in L_1 as $n \rightarrow \infty$, as a consequence of Birkhoff-Khintchine. Hence for any fixed m , $\mathbb{E}B \rightarrow 0$ as $n \rightarrow \infty$.
- For term A ,

$$\mathbb{E}[A] = \frac{1}{n} \mathbb{E}_P \left| \log \frac{dP_{S^n}}{dQ_{S^n}^{(m)}} \right| \leq \frac{1}{n} D(P_{S^n} \| Q_{S^n}^{(m)}) + \frac{2 \log e}{en}$$

where

$$\begin{aligned} \frac{1}{n} D(P_{S^n} \| Q_{S^n}^{(m)}) &= \frac{1}{n} \mathbb{E} \left[\log \frac{P_{S^n}(S^n)}{P_{S^n}(S^n) \prod_{t=m+1}^n P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1})} \right] \\ &\stackrel{\text{stat.}}{=} \frac{1}{n} (-H(S^n) + H(S^m) + (n-m)H_m) \\ &\rightarrow H_m - \mathcal{H} \text{ as } n \rightarrow \infty \end{aligned}$$

and the next Lemma 12.10.

Combining all three terms and sending $n \rightarrow \infty$, we obtain for any m ,

$$\limsup_{n \rightarrow \infty} \mathbb{E} \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \leq 2(H_m - \mathcal{H}).$$

Sending $m \rightarrow \infty$ completes the proof of L_1 -convergence. \square

Lemma 12.10.

$$\mathbb{E}_P \left[\left| \log \frac{dP}{dQ} \right| \right] \leq D(P||Q) + \frac{2 \log e}{e}.$$

Proof. $|x \log x| - x \log x \leq \frac{2 \log e}{e}, \forall x > 0$, since LHS is zero if $x \geq 1$, and otherwise upper bounded by $2 \sup_{0 \leq x \leq 1} x \log \frac{1}{x} = \frac{2 \log e}{e}$. \square

12.3* Proof of the Birkhoff-Khintchine Theorem

Proof of Theorem 12.7. \forall function $\tilde{f} \in L_1, \forall \epsilon$, there exists a decomposition $\tilde{f} = f + h$ such that f is bounded, and $h \in L_1, \|h\|_1 \leq \epsilon$.

Let us first focus on the bounded function f . Note that in the bounded domain $\mathcal{L}_1 \subset \mathcal{L}_2$, thus $f \in \mathcal{L}_2$. Furthermore, \mathcal{L}_2 is a Hilbert space with inner product $(f, g) = \mathbb{E}[f(S_1^\infty) \overline{g(S_1^\infty)}]$.

For the measure preserving transformation τ that generates the stationary process $\{S_i\}$, define the operator $T(f) = f \circ \tau$. Since τ is measure preserving, we know that $\|Tf\|_2^2 = \|f\|_2^2$, thus T is a unitary and bounded operator.

Define the operator

$$A_n(f) = \frac{1}{n} \sum_{k=1}^n f \circ \tau^k$$

Intuitively:

$$A_n = \frac{1}{n} \sum_{k=1}^n T^k = \frac{1}{n} (I - T^n)(I - T)^{-1}$$

Then, if $f \perp \ker(I - T)$ we should have $A_n f \rightarrow 0$, since only components in the kernel can blow up. This intuition is formalized in the proof below.

Let's further decompose f into two parts $f = f_1 + f_2$, where $f_1 \in \ker(I - T)$ and $f_2 \in \ker(I - T)^\perp$.

Observations:

- if $g \in \ker(I - T)$, g must be a constant function. This is due to the ergodicity. Consider indicator function 1_A , if $1_A = 1_A \circ \tau = 1_{\tau^{-1}A}$, then $\mathbb{P}[A] = 0$ or 1 . For a general case, suppose $g = Tg$ and g is not constant, then at least some set $\{g \in (a, b)\}$ will be shift-invariant and have non-trivial measure, violating ergodicity.

12.3* Proof of the Birkhoff-Khintchine Theorem 199

- $\ker(I - T) = \ker(I - T^*)$. This is due to the fact that T is unitary:

$$g = Tg \Rightarrow \|g\|^2 = (Tg, g) = (g, T^*g) \Rightarrow (T^*g, g) = \|g\| \|T^*g\| \Rightarrow T^*g = g$$

where in the last step we used the fact that Cauchy-Schwarz $(f, g) \leq \|f\| \cdot \|g\|$ only holds with equality for $g = cf$ for some constant c .

- $\ker(I - T)^\perp = \ker(I - T^*)^\perp = [\text{Im}(I - T)]$, where $[\text{Im}(I - T)]$ is an L_2 closure.
- $g \in \ker(I - T)^\perp \iff \mathbb{E}[g] = 0$. Indeed, only zero-mean functions are orthogonal to constants.

With these observations, we know that $f_1 = m$ is a const. Also, $f_2 \in [\text{Im}(I - T)]$ so we further approximate it by $f_2 = f_0 + h_1$, where $f_0 \in \text{Im}(I - T)$, namely $f_0 = g - g \circ \tau$ for some function $g \in L_2$, and $\|h_1\|_1 \leq \|h_1\|_2 < \epsilon$. Therefore we have

$$\begin{aligned} A_n f_1 &= f_1 = \mathbb{E}[f] \\ A_n f_0 &= \frac{1}{n}(g - g \circ \tau^n) \rightarrow 0 \text{ a.s. and } L_1 \end{aligned}$$

since $\mathbb{E}[\sum_{n \geq 1} (\frac{g \circ \tau^n}{n})^2] = \mathbb{E}[g^2] \sum \frac{1}{n^2} < \infty \implies \frac{1}{n} g \circ \tau^n \xrightarrow{\text{a.s.}} 0$.

The proof is completed by showing

$$\mathbb{P} \left[\limsup_n A_n(h + h_1) \geq \delta \right] \leq \frac{2\epsilon}{\delta}. \quad (12.7)$$

Indeed, then by taking $\epsilon \rightarrow 0$ we will have shown

$$\mathbb{P} \left[\limsup_n A_n(f) \geq \mathbb{E}[f] + \delta \right] = 0$$

as required. \square

Proof of (12.7) makes use of the Maximal Ergodic Lemma stated as follows:

Theorem 12.11 (Maximal Ergodic Lemma). *Let (\mathbb{P}, τ) be a probability measure and a measure-preserving transformation. Then for any $f \in L_1(\mathbb{P})$ we have*

$$\mathbb{P} \left[\sup_{n \geq 1} A_n f > a \right] \leq \frac{\mathbb{E}[f \mathbf{1}_{\{\sup_{n \geq 1} A_n f > a\}}]}{a} \leq \frac{\|f\|_1}{a}$$

where $A_n f = \frac{1}{n} \sum_{k=0}^{n-1} f \circ \tau^k$.

This is a so-called “weak L_1 ” estimate for a sublinear operator $\sup_n A_n(\cdot)$. In fact, this theorem is exactly equivalent to the following result:

Lemma 12.12 (Estimate for the maximum of averages). *Let $\{Z_n, n = 1, \dots\}$ be a stationary process with $\mathbb{E}[|Z|] < \infty$ then*

$$\mathbb{P} \left[\sup_{n \geq 1} \frac{|Z_1 + \dots + Z_n|}{n} > a \right] \leq \frac{\mathbb{E}[|Z|]}{a} \quad \forall a > 0$$

Proof. The argument for this Lemma has originally been quite involved, until a dramatically simple proof (below) was found by A. Garcia.

Define

$$S_n = \sum_{k=1}^n Z_k \quad (12.8)$$

$$L_n = \max\{0, Z_1, \dots, Z_1 + \dots + Z_n\} \quad (12.9)$$

$$M_n = \max\{0, Z_2, Z_2 + Z_3, \dots, Z_2 + \dots + Z_n\} \quad (12.10)$$

$$Z^* = \sup_{n \geq 1} \frac{S_n}{n} \quad (12.11)$$

It is sufficient to show that

$$\mathbb{E}[Z_1 1_{\{Z^* > 0\}}] \geq 0. \quad (12.12)$$

Indeed, applying (12.12) to $\tilde{Z}_1 = Z_1 - a$ and noticing that $\tilde{Z}^* = Z^* - a$ we obtain

$$\mathbb{E}[Z_1 1_{\{Z^* > a\}}] \geq a \mathbb{P}[Z^* > a],$$

from which Lemma follows by upper-bounding the left-hand side with $\mathbb{E}[|Z_1|]$.

In order to show (12.12) we first notice that $\{L_n > 0\} \nearrow \{Z^* > 0\}$. Next we notice that

$$Z_1 + M_n = \max\{S_1, \dots, S_n\}$$

and furthermore

$$Z_1 + M_n = L_n \quad \text{on } \{L_n > 0\}$$

Thus, we have

$$Z_1 1_{\{L_n > 0\}} = L_n - M_n 1_{\{L_n > 0\}}$$

where we do not need indicator in the first term since $L_n = 0$ on $\{L_n > 0\}^c$. Taking expectation we get

$$\mathbb{E}[Z_1 1_{\{L_n > 0\}}] = \mathbb{E}[L_n] - \mathbb{E}[M_n 1_{\{L_n > 0\}}] \quad (12.13)$$

$$\geq \mathbb{E}[L_n] - \mathbb{E}[M_n] \quad (12.14)$$

$$= \mathbb{E}[L_n] - \mathbb{E}[L_{n-1}] = \mathbb{E}[L_n - L_{n-1}] \geq 0, \quad (12.15)$$

where we used $M_n \geq 0$, the fact that M_n has the same distribution as L_{n-1} , and $L_n \geq L_{n-1}$, respectively. Taking limit as $n \rightarrow \infty$ in (12.15) we obtain (12.12). \square

12.4* Sinai's generator theorem

It turns out there is a way to associate to every probability-preserving transformation (p.p.t.) τ a number, called Kolmogorov-Sinai entropy. This number is invariant to isomorphisms of p.p.t.'s (appropriately defined).

12.4* Sinai's generator theorem 201

Definition 12.13. Fix a probability-preserving transformation τ acting on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Kolmogorov-Sinai entropy of τ is defined as

$$\mathcal{H}(\tau) \triangleq \sup_{X_0} \lim_{n \rightarrow \infty} \frac{1}{n} H(X_0, X_0 \circ \tau, \dots, X_0 \circ \tau^{n-1}),$$

where supremum is taken over all finitely-valued random variables $X_0 : \Omega \rightarrow \mathcal{X}$ and measurable with respect to \mathcal{F} .

Note that every random variable X_0 generates a stationary process adapted to τ , that is

$$X_k \triangleq X_0 \circ \tau^k.$$

In this way, Kolmogorov-Sinai entropy of τ equals the maximal entropy rate among all stationary processes adapted to τ . This quantity may be extremely hard to evaluate, however. One help comes in the form of the famous criterion of Y. Sinai. We need to elaborate on some more concepts before:

- σ -algebra $\mathcal{G} \subset \mathcal{F}$ is \mathbb{P} -dense in \mathcal{F} , or sometimes we also say $\mathcal{G} = \mathcal{F} \pmod{\mathbb{P}}$ or even $\mathcal{G} = \mathcal{F} \pmod{0}$, if for every $E \in \mathcal{F}$ there exists $E' \in \mathcal{G}$ s.t.

$$\mathbb{P}[E \Delta E'] = 0.$$

- Partition $\mathcal{A} = \{A_i, i = 1, 2, \dots\}$ measurable with respect to \mathcal{F} is called generating if

$$\bigvee_{n=0}^{\infty} \sigma\{\tau^{-n}\mathcal{A}\} = \mathcal{F} \pmod{\mathbb{P}}.$$

- Random variable $Y : \Omega \rightarrow \mathcal{Y}$ with a *countable* alphabet \mathcal{Y} is called a generator of $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$ if

$$\sigma\{Y, Y \circ \tau, \dots, Y \circ \tau^n, \dots\} = \mathcal{F} \pmod{\mathbb{P}}$$

Theorem 12.14 (Sinai's generator theorem). *Let Y be the generator of a p.p.t. $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$. Let $H(\mathbb{Y})$ be the entropy rate of the process $\mathbb{Y} = \{Y_k = Y \circ \tau^k, k = 0, \dots\}$. If $H(\mathbb{Y})$ is finite, then $\mathcal{H}(\tau) = H(\mathbb{Y})$.*

Proof. Notice that since $H(\mathbb{Y})$ is finite, we must have $H(Y_0^n) < \infty$ and thus $H(Y) < \infty$. First, we argue that $\mathcal{H}(\tau) \geq H(\mathbb{Y})$. If Y has finite alphabet, then it is simply from the definition. Otherwise let Y be \mathbb{Z}_+ -valued. Define a truncated version $\tilde{Y}_m = \min(Y, m)$, then since $\tilde{Y}_m \rightarrow Y$ as $m \rightarrow \infty$ we have from lower semicontinuity of mutual information, cf. (4.28), that

$$\lim_{m \rightarrow \infty} I(Y; \tilde{Y}_m) \geq H(Y),$$

and consequently for arbitrarily small ϵ and sufficiently large m

$$H(Y|\tilde{Y}) \leq \epsilon,$$

Then, consider the chain

$$H(Y_0^n) = H(\tilde{Y}_0^n, Y_0^n) = H(\tilde{Y}_0^n) + H(Y_0^n|\tilde{Y}_0^n)$$

$$\begin{aligned}
&= H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i | \tilde{Y}_0^n, Y_0^{i-1}) \\
&\leq H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i | \tilde{Y}_i) \\
&= H(\tilde{Y}_0^n) + nH(Y | \tilde{Y}) \leq H(\tilde{Y}_0^n) + n\epsilon
\end{aligned}$$

Thus, entropy rate of $\tilde{\mathbb{Y}}$ (which has finite-alphabet) can be made arbitrarily close to the entropy rate of \mathbb{Y} , concluding that $\mathcal{H}(\tau) \geq \mathcal{H}(\mathbb{Y})$.

The main part is showing that for any stationary process \mathbb{X} adapted to τ the entropy rate is upper bounded by $H(\mathbb{Y})$. To that end, consider $X : \Omega \rightarrow \mathcal{X}$ with finite \mathcal{X} and define as usual the process $\mathbb{X} = \{X \circ \tau^k, k = 0, 1, \dots\}$. By generating property of \mathbb{Y} we have that X (perhaps after modification on a set of measure zero) is a function of Y_0^∞ . So are all X_k . Thus

$$H(X_0) = I(X_0; Y_0^\infty) = \lim_{n \rightarrow \infty} I(X_0; Y_0^n),$$

where we used the continuity-in- σ -algebra property of mutual information, cf. (4.30). Rewriting the latter limit differently, we have

$$\lim_{n \rightarrow \infty} H(X_0 | Y_0^n) = 0.$$

Fix $\epsilon > 0$ and choose m so that $H(X_0 | Y_0^m) \leq \epsilon$. Then consider the following chain:

$$\begin{aligned}
H(X_0^n) &\leq H(X_0^n, Y_0^n) = H(Y_0^n) + H(X_0^n | Y_0^n) \\
&\leq H(Y_0^n) + \sum_{i=0}^n H(X_i | Y_i^n) \\
&= H(Y_0^n) + \sum_{i=0}^n H(X_0 | Y_0^{n-i}) \\
&\leq H(Y_0^n) + m \log |\mathcal{X}| + (n - m)\epsilon,
\end{aligned}$$

where we used stationarity of (X_k, Y_k) and the fact that $H(X_0 | Y_0^{n-i}) < \epsilon$ for $i \leq n - m$. After dividing by n and passing to the limit our argument implies

$$H(\mathbb{X}) \leq H(\mathbb{Y}) + \epsilon.$$

Taking here $\epsilon \rightarrow 0$ completes the proof.

Alternative proof: Suppose X_0 is taking values on a finite alphabet \mathcal{X} and $X_0 = f(Y_0^\infty)$. Then (this is a measure-theoretic fact) for every $\epsilon > 0$ there exists $m = m(\epsilon)$ and a function $f_\epsilon : \mathcal{Y}^{m+1} \rightarrow \mathcal{X}$ s.t.

$$\mathbb{P}[f(Y_0^\infty) \neq f_\epsilon(Y_0^m)] \leq \epsilon.$$

(This is just another way to say that $\bigcup_n \sigma(Y_0^n)$ is \mathbb{P} -dense in $\sigma(Y_0^\infty)$.) Define a stationary process $\tilde{\mathbb{X}}$ as

$$\tilde{X}_j \triangleq f_\epsilon(Y_j^{m+j}).$$

12.4* Sinai's generator theorem 203

Notice that since \tilde{X}_0^n is a function of Y_0^{n+m} we have

$$H(\tilde{X}_0^n) \leq H(Y_0^{n+m}).$$

Dividing by m and passing to the limit we obtain that for entropy rates

$$H(\tilde{\mathbb{X}}) \leq H(\mathbb{Y}).$$

Finally, to relate $\tilde{\mathbb{X}}$ to \mathbb{X} notice that by construction for every j

$$\mathbb{P}[\tilde{X}_j \neq X_j] \leq \epsilon.$$

Since both processes take values on a fixed finite alphabet, from Corollary 6.2 we infer that

$$|H(\mathbb{X}) - H(\tilde{\mathbb{X}})| \leq \epsilon \log |\mathcal{X}| + h(\epsilon).$$

Altogether, we have shown that

$$H(\mathbb{X}) \leq H(\mathbb{Y}) + \epsilon \log |\mathcal{X}| + h(\epsilon).$$

Taking $\epsilon \rightarrow 0$ we conclude the proof. \square

Examples:

- Let $\Omega = [0, 1]$, \mathcal{F} the Borel σ -algebra, $\mathbb{P} = \text{Leb}$ and

$$\tau(\omega) = 2\omega \pmod{1} = \begin{cases} 2\omega, & \omega < 1/2 \\ 2\omega - 1, & \omega \geq 1/2 \end{cases}$$

It is easy to show that $Y(\omega) = 1\{\omega < 1/2\}$ is a generator and that \mathbb{Y} is an i.i.d. Bernoulli(1/2) process. Thus, we get that Kolmogorov-Sinai entropy is $\mathcal{H}(\tau) = \log 2$.

- Let Ω be the unit circle \mathbb{S}^1 , \mathcal{F} the Borel σ -algebra, and \mathbb{P} the normalized length and

$$\tau(\omega) = \omega + \gamma$$

i.e. τ is a rotation by the angle γ . (When $\frac{\gamma}{2\pi}$ is irrational, this is known to be an ergodic p.p.t.). Here $Y = 1\{|\omega| < 2\pi\epsilon\}$ is a generator for arbitrarily small ϵ and hence

$$\mathcal{H}(\tau) \leq H(\mathbb{X}) \leq H(Y_0) = h(\epsilon) \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0.$$

This is an example of a zero-entropy p.p.t.

Remark 12.3. Two p.p.t.'s $(\Omega_1, \tau_1, \mathbb{P}_1)$ and $(\Omega_0, \tau_0, \mathbb{P}_0)$ are called isomorphic if there exists $f_i : \Omega_i \rightarrow \Omega_{1-i}$ defined \mathbb{P}_i -almost everywhere and such that 1) $\tau_{1-i} \circ f_i = f_{1-i} \circ \tau_i$; 2) $f_i \circ f_{1-i}$ is identity on Ω_i (a.e.); 3) $\mathbb{P}_i[f_{1-i}^{-1}E] = \mathbb{P}_{1-i}[E]$. It is easy to see that Kolmogorov-Sinai entropies of isomorphic p.p.t.s are equal. This observation was made by Kolmogorov in 1958. It was revolutionary, since it allowed to show that p.p.t.s corresponding shifts of iid $\text{Ber}(1/2)$ and iid $\text{Ber}(1/3)$ processes are not isomorphic. Before, the only invariants known were those obtained from studying the spectrum of a unitary operator

$$U_\tau : L_2(\Omega, \mathbb{P}) \rightarrow L_2(\Omega, \mathbb{P}) \tag{12.16}$$

$$\phi(x) \mapsto \phi(\tau(x)). \quad (12.17)$$

However, the spectrum of τ corresponding to any non-constant i.i.d. process consists of the entire unit circle, and thus is unable to distinguish $\text{Ber}(1/2)$ from $\text{Ber}(1/3)$.¹

¹ To see the statement about the spectrum, let X_i be iid with zero mean and unit variance. Then consider $\phi(x_1^\infty)$ defined as $\frac{1}{\sqrt{m}} \sum_{k=1}^m e^{i\omega k} x_k$. This ϕ has unit energy and as $m \rightarrow \infty$ we have $\|U_\tau \phi - e^{i\omega} \phi\|_{L_2} \rightarrow 0$. Hence every $e^{i\omega}$ belongs to the spectrum of U_τ .

13 Universal compression

In this chapter we will discuss how to produce compression schemes that do not require apriori knowledge of the distribution. Here, compressor is a map $\mathcal{X}^n \rightarrow \{0, 1\}^*$. Now, however, there is no one fixed probability distribution P_{X^n} on \mathcal{X}^n . The plan for this chapter is as follows:

- 1 We will start by discussing the earliest example of a universal compression algorithm (of Fitingof). It does not talk about probability distributions at all. However, it turns out to be asymptotically optimal simultaneously for all i.i.d. distributions and with small modifications for all finite-order Markov chains.
- 2 Next class of universal compressors is based on assuming that the true distribution P_{X^n} belongs to a given class. These methods proceed by choosing a good model distribution Q_{X^n} serving as the minimax approximation to each distribution in the class. The compression algorithm for a single distribution Q_{X^n} is then designed as in previous chapters.
- 3 Finally, an entirely different idea are algorithms of Lempel-Ziv type. These automatically adapt to the distribution of the source, without any prior assumptions required.

Throughout this chapter, all logarithms are binary. Instead of describing each compression algorithm, we will merely specify some distribution Q_{X^n} and apply one of the following constructions:

- Sort all x^n in the order of decreasing $Q_{X^n}(x^n)$ and assign values from $\{0, 1\}^*$ as in Theorem 10.2, this compressor has lengths satisfying

$$\ell(f(x^n)) \leq \log \frac{1}{Q_{X^n}(x^n)}.$$

- Set lengths to be

$$\ell(f(x^n)) \triangleq \left\lceil \log \frac{1}{Q_{X^n}(x^n)} \right\rceil$$

and apply Kraft's inequality Theorem 10.12 to construct a prefix code.

- Use arithmetic coding (see next section).

The important conclusion is that in all these cases we have

$$\ell(f(x^n)) \leq \log \frac{1}{Q_{X^n}(x^n)} + \text{universal constant},$$

and in this way we may and will always replace lengths with $\log \frac{1}{Q_{X^n}(x^n)}$. In this architecture, the only task of a universal compression algorithm is to specify the *probability assignment* Q_{X^n} .

Remark 13.1. Furthermore, if we only restrict attention to prefix codes, then any code $f: \mathcal{X}^n \rightarrow \{0, 1\}^*$ defines a distribution $Q_{X^n}(x^n) = 2^{-\ell(f(x^n))}$. (We assume the code's binary tree is full such that the Kraft sum equals one). In this way, for prefix-free codes results on redundancy, stated in terms of optimizing the choice of Q_{X^n} , imply tight converses too. For one-shot codes without prefix constraints the optimal answers are slightly different, however. (For example, the optimal universal code for all i.i.d. sources satisfies $\mathbb{E}[\ell(f(X^n))] \approx H(X^n) + \frac{|\mathcal{X}| - 3}{2} \log n$ in contrast with $\frac{|\mathcal{X}| - 1}{2} \log n$ for prefix-free codes, cf. [26, 180].)

If one factorizes $Q_{X^n} = \prod_{t=1}^n Q_{X_t|X_1^{t-1}}$ then we arrive at a crucial conclusion: (*universal*) compression is equivalent to *sequential (online) prediction* under the log-loss. As of 2022 the best performing text compression algorithms (cf. the leaderboard at [202]) use a deep neural network (transformer model) that starts from a fixed initialization. As the input text is processed, parameters of the network are continuously updated via stochastic gradient descent causing progressively better prediction (and hence compression) performance.

13.1 Arithmetic coding

Constructing an encoder table from Q_{X^n} may require a lot of resources if n is large. Arithmetic coding provides a convenient workaround by allowing the encoder to output bits sequentially. Notice that to do so, it requires that not only Q_{X^n} but also its marginalizations Q_{X^1}, Q_{X^2}, \dots be easily computable. (This is not the case, for example, for Shtarkov distributions (13.12)-(13.13), which are not compatible for different n .)

Let us agree upon some ordering on the alphabet of \mathcal{X} (e.g. $a < b < \dots < z$) and extend this order lexicographically to \mathcal{X}^n (that is for $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, we say $x < y$ if $x_i < y_i$ for the first i such that $x_i \neq y_i$, e.g., $baba < babb$). Then let

$$F_n(x^n) = \sum_{y^n < x^n} Q_{X^n}(y^n).$$

Associate to each x^n an interval $I_{x^n} = [F_n(x^n), F_n(x^n) + Q_{X^n}(x^n)]$. These intervals are disjoint subintervals of $[0, 1]$. As such, each x^n can be represented uniquely by any point in the interval I_{x^n} . A specific choice is as follows. Encode

$$x^n \mapsto \text{largest dyadic interval } D_{x^n} \text{ contained in } I_{x^n} \quad (13.1)$$

and we agree to select the left-most dyadic interval when there are two possibilities. Recall that dyadic intervals are intervals of the type $[m2^{-k}, (m+1)2^{-k}]$ where m is an integer. We encode such interval by the k -bit (zero-padded) binary expansion of the fractional number $m2^{-k} = 0.b_1b_2\dots b_k = \sum_{i=1}^k b_i 2^{-i}$. For example, $[3/4, 7/8] \mapsto 110$, $[3/4, 13/16] \mapsto 1100$. We set the

13.2 Combinatorial construction of Fitingof 207

codeword $f(x^n)$ to be that string. The resulting code is a prefix code satisfying

$$\log_2 \frac{1}{Q_{X^n}(x^n)} \leq \ell(f(x^n)) \leq \left\lceil \log_2 \frac{1}{Q_{X^n}(x^n)} \right\rceil + 1. \quad (13.2)$$

(This is an exercise, see Ex. II.11.)

Observe that

$$F_n(x^n) = F_{n-1}(x^{n-1}) + Q_{X^{n-1}}(x^{n-1}) \sum_{y < x_n} Q_{X_n|X^{n-1}}(y|x^{n-1})$$

and thus $F_n(x^n)$ can be computed sequentially if $Q_{X^{n-1}}$ and $Q_{X_n|X^{n-1}}$ are easy to compute. This method is the method of choice in many modern compression algorithms because it allows to dynamically incorporate the learned information about the stream, in the form of updating $Q_{X_n|X^{n-1}}$ (e.g. if the algorithm detects that an executable file contains a long chunk of English text, it may temporarily switch to $Q_{X_n|X^{n-1}}$ modeling the English language).

We note that efficient implementation of arithmetic encoder and decoder is a continuing research area. Indeed, performance depends on number-theoretic properties of denominators of distributions $Q_{X_i|X^{i-1}}$, because as encoder/decoder progress along the string, they need to periodically renormalize the current interval I_{x^i} to be $[0, 1]$ but this requires carefully realigning the dyadic boundaries. A recent idea, known as *asymmetric numeral system (ANS)* [102], lead to such impressive computational gains that in less than a decade it was adopted by most compression libraries handling diverse data streams (e.g., the Linux kernel images, Dropbox and Facebook traffic, etc).

13.2 Combinatorial construction of Fitingof

Fitingof [126] suggested that a sequence $x^n \in \mathcal{X}^n$ should be prescribed information $\Phi_0(x^n)$ equal to the logarithm of the number of all possible permutations obtainable from x^n (i.e. log-size of the type-class containing x^n). As we have shown in Proposition 1.6:

$$\Phi_0(x^n) = nH(x_T) + O(\log n) \quad T \sim \text{Unif}([n]) \quad (13.3)$$

$$= nH(\hat{P}_{x^n}) + O(\log n), \quad (13.4)$$

where \hat{P}_{x^n} is the empirical distribution of the sequence x^n :

$$\hat{P}_{x^n}(a) \triangleq \frac{1}{n} \sum_{i=1}^n 1\{x_i = a\}. \quad (13.5)$$

Then Fitingof argues that it should be possible to produce a prefix code with

$$\ell(f(x^n)) = \Phi_0(x^n) + O(\log n). \quad (13.6)$$

This can be done in many ways. In the spirit of what comes next, let us define

$$Q_{X^n}(x^n) \triangleq \exp\{-\Phi_0(x^n)\} c_n, \quad (13.7)$$

where the normalization constant c_n is determined by the number of types, namely, $c_n = 1/\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$. Counting the number of different possible empirical distributions (types), we get

$$c_n = O(n^{-(|\mathcal{X}|-1)}),$$

and thus, by Kraft inequality, there must exist a prefix code with lengths satisfying (13.6).¹ Now taking expectation over $X^n \stackrel{\text{i.i.d.}}{\sim} P_X$ we get

$$\mathbb{E}[\ell(f(X^n))] = nH(P_X) + (|\mathcal{X}| - 1) \log n + O(1),$$

for every i.i.d. source on \mathcal{X} .

Universal compressor for all finite-order Markov chains. Fitingof's idea can be extended as follows. Define now the first order information content $\Phi_1(x^n)$ to be the log of the number of all sequences, obtainable by permuting x^n with extra restriction that the new sequence should have the same statistics on digrams. Asymptotically, Φ_1 is just the conditional entropy

$$\Phi_1(x^n) = nH(x_T|x_{T-1}) + O(\log n), \quad T \sim \text{Unif}([n]),$$

where $T - 1$ is understood in the sense of modulo n . Again, it can be shown that there exists a code such that lengths

$$\ell(f(x^n)) = \Phi_1(x^n) + O(\log n).$$

This implies that for every first order stationary Markov chain $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ we have

$$\mathbb{E}[\ell(f(X^n))] = nH(X_2|X_1) + O(\log n).$$

This can be further continued to define $\Phi_2(x^n)$ and build a universal code, asymptotically optimal for all second order Markov chains etc.

13.3 Optimal compressors for a class of sources. Redundancy.

So we have seen that we can construct compressor $f: \mathcal{X}^n \rightarrow \{0, 1\}^*$ that achieves

$$\mathbb{E}[\ell(f(X^n))] \leq H(X^n) + o(n),$$

simultaneously for all i.i.d. sources (or even all r -th order Markov chains). What should we do next? Krichevsky suggested that the next barrier should be to minimize the regret, or *redundancy*:

$$\mathbb{E}[\ell(f(X^n))] - H(X^n)$$

simultaneously for all sources in a given class. We proceed to rigorous definitions.

¹ Explicitly, we can do a two-part encoding: first describe the type class of x^n (takes $(|\mathcal{X}| - 1) \log n$ bits) and then describe the element of the class (takes $\Phi_0(x^n)$ bits).

13.3 Optimal compressors for a class of sources. Redundancy. 209

Given a collection $\{P_{X^n|\theta} : \theta \in \Theta\}$ of sources, and a compressor $f : \mathcal{X}^n \rightarrow \{0, 1\}^*$ we define its redundancy as

$$\sup_{\theta_0} \mathbb{E}[\ell(f(X^n)) | \theta = \theta_0] - H(X^n | \theta = \theta_0).$$

Replacing code lengths with $\log \frac{1}{Q_{X^n}}$, we define redundancy of the distribution Q_{X^n} as

$$\sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n}).$$

Thus, the question of designing the best universal compressor (in the sense of optimizing worst-case deviation of the average length from the entropy) becomes the question of finding solution of:

$$Q_{X^n}^* = \operatorname{argmin}_{Q_{X^n}} \sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n}).$$

We therefore get to the following definition

Definition 13.1 (Redundancy in universal compression). Given a class of sources $\{P_{X^n|\theta=\theta_0} : \theta_0 \in \Theta, n = 1, \dots\}$ we define its minimax redundancy as

$$\mathcal{R}_n^* \equiv \mathcal{R}_n^*(\Theta) \triangleq \min_{Q_{X^n}} \sup_{\theta_0 \in \Theta} D(P_{X^n|\theta=\theta_0} \| Q_{X^n}). \quad (13.8)$$

Assuming the finiteness of \mathcal{R}_n^* , Theorem 5.8 gives the maximin and capacity representation

$$\mathcal{R}_n^* = \sup_{\pi} \min_{Q_{X^n}} D(P_{X^n|\theta} \| Q_{X^n} | \pi) \quad (13.9)$$

$$= \sup_{\pi} I(\theta; X^n), \quad (13.10)$$

where optimization is over priors $\pi \in \mathcal{P}(\Theta)$ on θ . Thus redundancy is simply the capacity of the channel $\theta \rightarrow X^n$. This result, obvious in hindsight, was rather surprising in the early days of universal compression. It is known as *capacity-redundancy theorem*.

Finding exact Q_{X^n} -minimizer in (13.8) is a daunting task even for the simple class of all i.i.d. Bernoulli sources (i.e. $\Theta = [0, 1]$, $P_{X^n|\theta} = \text{Ber}^n(\theta)$). In fact, for smooth parametric families the capacity-achieving input distribution is rather ugly: it is a discrete distribution with a k_n atoms, k_n slowly growing as $n \rightarrow \infty$. A provocative conjecture was put forward by physicists [208, 1] that there is a certain universality relation:

$$\mathcal{R}_n^* = \frac{3}{4} \log k_n + o(\log k_n)$$

satisfied for all parametric families simultaneously. For the Bernoulli example this implies $k_n \asymp n^{2/3}$, but even this is open. However, as we will see below it turns out that these unwieldy capacity-achieving input distributions converge as $n \rightarrow \infty$ to a beautiful limiting law, known as the Jeffreys prior.

Remark 13.2. (Shtarkov, Fitingof and individual sequence approach) There is a connection between the combinatorial method of Fitingof and the method of optimality for a class. Indeed,

following Shtarkov we may want to choose distribution $Q_{X^n}^{(S)}$ so as to minimize the worst-case redundancy *for each realization x^n* (not average!):

$$\min_{Q_{X^n}} \max_{x^n} \sup_{\theta_0} \log \frac{P_{X^n|\theta}(x^n|\theta_0)}{Q_{X^n}(x^n)} \quad (13.11)$$

This leads to Shtarkov's distribution (also known as the *normalized maximal likelihood* (NML) code):

$$Q_{X^n}^{(S)}(x^n) = c \sup_{\theta_0} P_{X^n|\theta}(x^n|\theta_0), \quad (13.12)$$

where c is the normalization constant. If class $\{P_{X^n|\theta}, \theta \in \Theta\}$ is chosen to be all i.i.d. distributions on \mathcal{X} then

$$\text{i.i.d. } Q_{X^n}^{(S)}(x^n) = c \exp\{-nH(\hat{P}_{x^n})\}, \quad (13.13)$$

and thus compressing w.r.t. $Q_{X^n}^{(S)}$ recovers Fitingof's construction Φ_0 up to $O(\log n)$ differences between $nH(\hat{P}_{x^n})$ and $\Phi_0(x^n)$. If we take $P_{X^n|\theta}$ to be all first order Markov chains, then we get construction Φ_1 etc. Note also, that the problem (13.11) can also be written as minimization of the regret for each *individual sequence* (under log-loss, with respect to a parameter class $P_{X^n|\theta}$):

$$\min_{Q_{X^n}} \max_{x^n} \left\{ \log \frac{1}{Q_{X^n}(x^n)} - \inf_{\theta_0} \log \frac{1}{P_{X^n|\theta}(x^n|\theta_0)} \right\}. \quad (13.14)$$

The gospel is that if there is a reason to believe that real-world data x^n are likely to be generated by one of the models $P_{X^n|\theta}$, then using minimizer of (13.14) will result in the compressor that both learns the right model (in the sense of $Q_{X_n|X^{n-1}} \approx \text{true } P_{X_n|X^{n-1}}$) and compresses with respect to it. See more in Section 13.6.

13.4* Approximate minimax solution: Jeffreys prior

In this section we will only consider the simple setting of a class of sources consisting of all i.i.d. distributions on a given finite alphabet $|\mathcal{X}| = d + 1$, which defines a d -parameter family of distributions. We will show that the prior, asymptotically achieving the capacity (13.10), is given by the Dirichlet distribution with parameters set to 1/2. Recall that the Dirichlet distribution $\text{Dirichlet}(\alpha_0, \dots, \alpha_d)$ with parameters $\alpha_j > 0$ is a distribution for a probability vector $(\theta_0, \dots, \theta_d)$ such that $(\theta_1, \dots, \theta_d)$ has a joint density

$$c(\alpha_0, \dots, \alpha_d) \prod_{j=0}^d \theta_j^{\alpha_j - 1} \quad (13.15)$$

where $c(\alpha_0, \dots, \alpha_d) = \frac{\Gamma(\alpha_0 + \dots + \alpha_d)}{\prod_{j=0}^d \Gamma(\alpha_j)}$ is the normalizing constant.

First, we give the formal setting as follows:

- Fix a finite alphabet \mathcal{X} of size $|\mathcal{X}| = d + 1$, which we will enumerate as $\mathcal{X} = \{0, \dots, d\}$.

13.4* Approximate minimax solution: Jeffreys prior 211

- As in Example 2.6, let $\Theta = \{(\theta_1, \dots, \theta_d) : \sum_{j=1}^d \theta_j \leq 1, \theta_j \geq 0\}$ parametrizes the collection of all probability distributions on \mathcal{X} . Note that Θ is a d -dimensional simplex. We will also define

$$\theta_0 \triangleq 1 - \sum_{j=1}^d \theta_j.$$

- The source class is

$$P_{X^n|\theta}(x^n|\theta) \triangleq \prod_{j=1}^n \theta_{x_j} = \exp \left\{ -n \sum_{a \in \mathcal{X}} \theta_a \log \frac{1}{\hat{P}_{x^n}(a)} \right\},$$

where as before \hat{P}_{x^n} is the empirical distribution of x^n , cf. (13.5).

In order to find the (near) optimal Q_{X^n} , we need to guess an (almost) optimal prior $\pi \in \mathcal{P}(\Theta)$ in (13.10) and take Q_{X^n} to be the mixture of $P_{X^n|\theta}$'s. We will search for π in the class of smooth densities on Θ and set

$$Q_{X^n}(x^n) \triangleq \int_{\Theta} P_{X^n|\theta}(x^n|\theta') \pi(\theta') d\theta'. \quad (13.16)$$

Before proceeding further, we recall the *Laplace method* of approximating exponential integrals. Suppose that $f(\theta)$ has a unique minimum at the interior point $\hat{\theta}$ of Θ and that Hessian $\text{Hess}f$ is uniformly lower-bounded by a multiple of identity (in particular, $f(\theta)$ is strongly convex). Then taking Taylor expansion of π and f we get

$$\int_{\Theta} \pi(\theta) e^{-nf(\theta)} d\theta = \int (\pi(\hat{\theta}) + O(\|t\|)) e^{-n(f(\hat{\theta}) - \frac{1}{2} t^T \text{Hess}f(\hat{\theta}) t + o(\|t\|^2))} dt \quad (13.17)$$

$$= \pi(\hat{\theta}) e^{-nf(\hat{\theta})} \int_{\mathbb{R}^d} e^{-x^T \text{Hess}f(\hat{\theta}) x} \frac{dx}{\sqrt{n^d}} (1 + O(n^{-1/2})) \quad (13.18)$$

$$= \pi(\hat{\theta}) e^{-nf(\hat{\theta})} \left(\frac{2\pi}{n} \right)^{\frac{d}{2}} \frac{1}{\sqrt{\det \text{Hess}f(\hat{\theta})}} (1 + O(n^{-1/2})) \quad (13.19)$$

where in the last step we computed Gaussian integral.

Next, we notice that

$$P_{X^n|\theta}(x^n|\theta') = \exp\{-n(D(\hat{P}_{x^n}\|P_{X|\theta=\theta'}) + H(\hat{P}_{x^n}))\},$$

and therefore, denoting

$$\hat{\theta}(x^n) \triangleq \hat{P}_{x^n}$$

we get from applying (13.19) to (13.16)

$$\log Q_{X^n}(x^n) = -nH(\hat{\theta}) + \frac{d}{2} \log \frac{2\pi}{n \log e} + \log \frac{P_{\theta}(\hat{\theta})}{\sqrt{\det J_F(\hat{\theta})}} + O(n^{-\frac{1}{2}}),$$

where we used the fact that $\text{Hess}_{\theta'} D(\hat{P} \| P_{X|\theta=\theta'})|_{\theta'=\hat{\theta}} = \frac{1}{\log e} J_F(\hat{\theta})$ with J_F being the Fisher information matrix introduced previously in (2.33). From here, using the fact that under $X^n \sim P_{X^n|\theta=\theta'}$ the random variable $\hat{\theta} = \theta' + O(n^{-1/2})$ we get by approximating $J_F(\hat{\theta})$ and $P_\theta(\hat{\theta})$

$$D(P_{X^n|\theta=\theta'} \| Q_{X^n}) = n(\mathbb{E}[H(\hat{\theta})] - H(X|\theta=\theta')) + \frac{d}{2} \log n - \log \frac{P_\theta(\theta')}{\sqrt{\det J_F(\theta')}} + C + O(n^{-\frac{1}{2}}), \quad (13.20)$$

where C is some constant (independent of the prior P_θ or θ'). The first term is handled by the next result, refining Corollary 7.1.

Lemma 13.2. *Let $X^n \stackrel{i.i.d.}{\sim} P$ on a finite alphabet \mathcal{X} such that $P(x) > 0$ for all $x \in \mathcal{X}$. Let $\hat{P} = \hat{P}_{X^n}$ be the empirical distribution of X^n , then*

$$\mathbb{E}[D(\hat{P} \| P)] = \frac{|\mathcal{X}| - 1}{2n} \log e + o\left(\frac{1}{n}\right).$$

In fact, $nD(\hat{P} \| P) \rightarrow \frac{\log e}{2} \chi^2(|\mathcal{X}| - 1)$ in distribution.

Proof. By Central Limit Theorem, $\sqrt{n}(\hat{P} - P)$ converges in distribution to $\mathcal{N}(0, \Sigma)$, where $\Sigma = \text{diag}(P) - PP^T$, where P is an $|\mathcal{X}|$ -by-1 column vector. Thus, computing second-order Taylor expansion of $D(\cdot \| P)$, cf. (2.33) and (2.36), we get the result. \square

Continuing (13.20) we get in the end

$$D(P_{X^n|\theta=\theta'} \| Q_{X^n}) = \frac{d}{2} \log n - \log \frac{\pi(\theta')}{\sqrt{\det J_F(\theta')}} + \text{const} + O(n^{-\frac{1}{2}}) \quad (13.21)$$

under the assumption of smoothness of prior π and that θ' is not on the boundary of Θ . Consequently, we can see that in order for the prior π be the saddle point solution, we should have

$$\pi(\theta') \propto \sqrt{\det J_F(\theta')},$$

provided that the right side is integrable. Prior proportional to square-root of the determinant of Fisher information matrix is known as the *Jeffreys prior*. In our case, using the explicit expression for Fisher information (2.38), we conclude that π^* is the $\text{Dirichlet}(1/2, 1/2, \dots, 1/2)$ prior, with density:

$$\pi^*(\theta) = c_d \frac{1}{\sqrt{\prod_{j=0}^d \theta_j}}, \quad (13.22)$$

where $c_d = \frac{\Gamma(\frac{d+1}{2})}{\Gamma(1/2)^{d+1}}$ is the normalization constant. The corresponding redundancy is then

$$\mathcal{R}_n^* = \frac{d}{2} \log \frac{n}{2\pi e} - \log \frac{\Gamma(\frac{d+1}{2})}{\Gamma(1/2)^{d+1}} + o(1). \quad (13.23)$$

Making the above derivation rigorous is far from trivial, and was completed in [325]. Surprisingly, while the Jeffreys prior π^* that we derived does attain the claimed value (13.23) of the mutual

13.5 Sequential probability assignment: Krichevsky-Trofimov 213

information $I(\theta; X^n)$, the corresponding mixture Q_{X^n} does not yield (13.23). In other words, this Q_{X^n} when plugged into (13.8) results in the value of \sup_{Θ_0} that is much larger than the optimal value (13.23). The way (13.23) was proved is by patching the Jeffreys prior near the boundary of the simplex.

Extension to general smooth parametric families. The fact that Jeffreys prior $\theta \sim \pi$ maximizes the value of mutual information $I(\theta; X^n)$ for general parametric families was conjectured in [29] in the context of selecting priors in Bayesian inference. This result was proved rigorously in [67, 68]. We briefly summarize the results of the latter.

Let $\{P_\theta : \theta \in \Theta_0\}$ be a smooth parametric family admitting a continuous and bounded Fisher information matrix $J_F(\theta)$ everywhere on the interior of $\Theta_0 \subset \mathbb{R}^d$. Then for every compact Θ contained in the interior of Θ_0 we have

$$\mathcal{R}_n^*(\Theta) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \int_{\Theta} \sqrt{\det J_F(\theta)} d\theta + o(1). \quad (13.24)$$

Although Jeffreys prior on Θ achieves (up to $o(1)$) the optimal value of $\sup_{\pi} I(\theta; X^n)$, to produce an approximate capacity-achieving output distribution Q_{X^n} , however, one needs to take a mixture with respect to a Jeffreys prior on a slightly larger set $\Theta_\epsilon = \{\theta : d(\theta, \Theta) \leq \epsilon\}$ and take $\epsilon \rightarrow 0$ slowly with $n \rightarrow \infty$. This sequence of Q_{X^n} 's does achieve the optimal redundancy up to $o(1)$.

Remark 13.3. In statistics Jeffreys prior is justified as being invariant to smooth reparametrization, as evidenced by (2.34). For example, in answering “will the sun rise tomorrow”, Laplace proposed to estimate the probability by modeling sunrise as i.i.d. Bernoulli process with a uniform prior on $\theta \in [0, 1]$. However, this is clearly not very logical, as one may equally well postulate uniformity of $\alpha = \theta^{10}$ or $\beta = \sqrt{\theta}$. Jeffreys prior $\theta \sim \frac{1}{\sqrt{\theta(1-\theta)}}$ is invariant to reparametrization in the sense that if one computed $\sqrt{\det J_F(\alpha)}$ under α -parametrization the result would be exactly the pushforward of the $\frac{1}{\sqrt{\theta(1-\theta)}}$ along the map $\theta \mapsto \theta^{10}$.

13.5 Sequential probability assignment: Krichevsky-Trofimov

From (13.22) it is not hard to derive the (asymptotically) optimal universal probability assignment Q_{X^n} . For simplicity we consider Bernoulli case, i.e. $d = 1$ and $\theta \in [0, 1]$ is the 1-dimensional parameter. Then,²

$$P_\theta^* = \frac{1}{\pi \sqrt{\theta(1-\theta)}} \quad (13.25)$$

$$Q_{X^n}^{(KT)}(x^n) = \frac{(2t_0 - 1)!! \cdot (2t_1 - 1)!!}{2^n n!}, \quad t_a = \#\{j \leq n : x_j = a\} \quad (13.26)$$

² This is obtained from the identity $\int_0^1 \frac{\theta^a (1-\theta)^b}{\sqrt{\theta(1-\theta)}} d\theta = \pi^{\frac{1-3\cdots(2a-1)\cdot 1\cdot 3\cdots(2b-1)}{2a+b}(a+b)!}$ for integer $a, b \geq 0$. This identity can be derived by change of variable $z = \frac{\theta}{1-\theta}$ and using the standard keyhole contour on the complex plane.

This assignment can now be used to create a universal compressor via one of the methods outlined in the beginning of this chapter. However, what is remarkable is that it has a very nice sequential interpretation (as does any assignment obtained via $Q_{X^n} = \int P_\theta P_{X^n|\theta}$ with P_θ not depending on n).

$$Q_{X_n|X^{n-1}}^{(KT)}(1|x^{n-1}) = \frac{t_1 + \frac{1}{2}}{n}, \quad t_1 = \#\{j \leq n-1 : x_j = 1\} \quad (13.27)$$

$$Q_{X_n|X^{n-1}}^{(KT)}(0|x^{n-1}) = \frac{t_0 + \frac{1}{2}}{n}, \quad t_0 = \#\{j \leq n-1 : x_j = 0\} \quad (13.28)$$

This is the famous ‘‘add 1/2’’ rule of Krichevsky and Trofimov. As mentioned in Section 13.1, this sequential assignment is very convenient for use in prediction as well as in implementing an arithmetic coder. The version for a general (non-binary) alphabet is equally simple:

$$Q_{X_n|X^{n-1}}^{(KT)}(a|x^{n-1}) = \frac{t_a + \frac{1}{2}}{n + \frac{|\mathcal{X}| - 2}{2}}, \quad t_a = \#\{j \leq n-1 : x_j = a\}$$

Remark 13.4 (Laplace ‘‘add 1’’ rule). A slightly less optimal choice of Q_{X^n} results from Laplace prior: just take P_θ to be uniform on $[0, 1]$. Then, in the Bernoulli ($d = 1$) case we get

$$Q_{X^n}^{(Lap)} = \frac{1}{\binom{n}{w}(n+1)}, \quad w = \#\{j : x_j = 1\}. \quad (13.29)$$

The corresponding successive probability is given by

$$Q_{X_n|X^{n-1}}^{(Lap)}(1|x^{n-1}) = \frac{t_1 + 1}{n+1}, \quad t_1 = \#\{j \leq n-1 : x_j = 1\}.$$

We notice two things. First, the distribution (13.29) is *exactly* the same as Fitingof’s (13.7). Second, this distribution ‘‘almost’’ attains the optimal first-order term in (13.23). Indeed, when X^n is iid $\text{Ber}(\theta)$ we have for the redundancy:

$$\mathbb{E} \left[\log \frac{1}{Q_{X^n}^{(Lap)}(X^n)} \right] - H(X^n) = \log(n+1) + \mathbb{E} \left[\log \binom{n}{W} \right] - nh(\theta), \quad W \sim \text{Bin}(n, \theta). \quad (13.30)$$

From Stirling’s expansion we know that as $n \rightarrow \infty$ this redundancy evaluates to $\frac{1}{2} \log n + O(1)$, uniformly in θ over compact subsets of $(0, 1)$. However, for $\theta = 0$ or $\theta = 1$ the Laplace redundancy (13.30) clearly equals $\log(n+1)$. Thus, supremum over $\theta \in [0, 1]$ is achieved close to endpoints and results in suboptimal redundancy $\log n + O(1)$. The Jeffreys prior (13.25) fixes the problem at the endpoints.

13.6 Individual sequence and universal prediction

The problem of selecting one Q_{X^n} serving as good prior for a whole class of distributions can also be interpreted in terms of so-called ‘‘universal prediction’’. An excellent textbook on the topic is [59]. We discuss this connection briefly.

13.6 Individual sequence and universal prediction 215

Consider the following problem: a sequence x^n is observed sequentially and our goal is to predict (by making a soft decision) the next symbol given the past observations. The experiment proceeds as follows:

- 1 A string $x^n \in \mathcal{X}^n$ is selected by the nature.
- 2 Having observed samples x_1, \dots, x_{t-1} we are requested to output a probability distribution $Q_t(\cdot | x^{t-1})$ on \mathcal{X}^n .
- 3 After that nature reveals the next sample x_t and our loss for t -th prediction is evaluated as

$$\log \frac{1}{Q_t(x_t | x^{t-1})}.$$

Goal (informal): Find a sequence of predictors $\{Q_t\}$ that minimizes the *cumulative* loss:

$$\ell(\{Q_t\}, x^n) \triangleq \sum_{t=1}^n \log \frac{1}{Q_t(x_t | x^{t-1})}.$$

Note that to make this goal formal, we need to explain how x^n is generated. Consider first a naive requirement that the worst-case loss is minimized:

$$\min_{\{Q_t\}_{t=1}^n} \max_{x^n} \ell(\{Q_t\}, x^n).$$

This is clearly hopeless. Indeed, at any step t the distribution Q_t must have at least one atom with weight $\leq \frac{1}{|\mathcal{X}|}$, and hence for any predictor

$$\max_{x^n} \ell(\{Q_t\}, x^n) \geq n \log |\mathcal{X}|,$$

which is clearly achieved iff $Q_t(\cdot) \equiv \frac{1}{|\mathcal{X}|}$, i.e. if predictor simply makes uniform random guesses. This triviality is not surprising: In the absence of whatsoever prior information on x^n it is impossible to predict anything.

The exciting idea, originated by Feder, Merhav and Gutman, cf. [117, 213], is to replace loss with *regret*, i.e. the gap to the best possible *static oracle*. More precisely, suppose a non-causal oracle can examine the entire string x^n and output a constant $Q_t \equiv Q$. From non-negativity of divergence this non-causal oracle achieves:

$$\ell_{\text{oracle}}(x^n) = \min_Q \sum_{t=1}^n \log \frac{1}{Q(x_t)} = nH(\hat{P}_{x^n}).$$

Can causal (but time-varying) predictor come close to this performance? In other words, we define *regret* of a sequential predictor as the excess risk over the static oracle

$$\text{reg}(\{Q_t\}, x^n) \triangleq \ell(\{Q_t\}, x^n) - nH(\hat{P}_{x^n})$$

and ask to minimize the worst-case regret:

$$\text{Reg}_n^* \triangleq \min_{\{Q_t\}} \max_{x^n} \text{reg}(\{Q_t\}, x^n). \quad (13.31)$$

Excitingly, non-trivial predictors emerge as solutions to the above problem, which furthermore do not rely on any assumptions on the prior distribution of x^n .

We next consider the case of $\mathcal{X} = \{0, 1\}$ for simplicity. To solve (13.31), first notice that designing a sequence $\{Q_t(\cdot|x^{t-1})\}$ is equivalent to defining one joint distribution Q_{X^n} and then factorizing the latter as $Q_{X^n}(x^n) = \prod_t Q_t(x_t|x^{t-1})$. Then the problem (13.31) becomes simply

$$\text{Reg}_n^* = \min_{Q_{X^n}} \max_{x^n} \log \frac{1}{Q_{X^n}(x^n)} - nH(\hat{P}_{x^n}).$$

First, we notice that generally we have that optimal Q_{X^n} is Shtarkov distribution (13.12), which implies that that regret is just the log of normalization constant in Shtarkov distribution. In the iid case we are considering, we get

$$\text{Reg}_n^* = \log \sum_{x^n} \max_Q \prod_{i=1}^n Q(x_i) = \log \sum_{x^n} \exp\{-nH(\hat{P}_{x^n})\}.$$

This is, however, frequently a not very convenient expression to analyze, so instead we consider upper and lower bounds. We may lower-bound the max over x^n with the average over the $X^n \sim \text{Ber}(\theta)^n$ and obtain (also applying Lemma 13.4):

$$\text{Reg}_n^* \geq \mathcal{R}_n^* + \frac{|\mathcal{X}| - 1}{2} \log e + o(1),$$

where \mathcal{R}_n^* is the universal compression redundancy defined in (13.8), whose asymptotics we derived in (13.23).

On the other hand, taking $Q_{X^n}^{(KT)}$ from Krichevsky-Trofimov (13.26) we find after some algebra and Stirling's expansion:

$$\max_{x^n} \log \frac{1}{Q_{X^n}^{(KT)}(x^n)} - nH(\hat{P}_{x^n}) = \frac{1}{2} \log n + O(1).$$

In all, we conclude that,

$$\text{Reg}_n^* = \mathcal{R}_n^* + O(1) = \frac{|\mathcal{X}| - 1}{2} \log n + O(1),$$

and remarkably, the per-letter regret $\frac{1}{n} \text{Reg}_n^*$ converges to zero. That is, *there exists a causal predictor that can predict (under log-loss) almost as well as any constant one, even if it is adapted to a particular sequence x^n non-causally*.

Explicit (asymptotically optimal) sequential prediction rules are given by Krichevsky-Trofimov's "add 1/2" rules (13.28). We note that the resulting rules are also independent of n ("horizon-free"). This is a very desirable property not shared by the optimal sequential predictors derived from factorizing the Shtarkov's distribution (13.12).

General parametric families. The general definition of (cumulative) individual-sequence (or worst-case) regret for a model class $\{P_{X^n|\theta=\theta_0}, \theta_0 \in \Theta\}$ is given by

$$\text{Reg}_n^*(\Theta) = \min_{Q_{X^n}} \sup_{x^n} \log \frac{1}{Q_{X^n}(x^n)} - \inf_{\theta_0 \in \Theta} \log \frac{1}{P_{X^n|\theta=\theta_0}(x^n)},$$

13.7 Lempel-Ziv compressor 217

This regret can be *interpreted* as worst-case loss of a given estimator compared to the best possible one from a class $P_{X^n|\theta}$, when the latter is selected optimally for each sequence. In this sense, regret gives a uniform (in x^n) bound on the performance of an algorithm against a class.

It turns out that similarly to (13.24) the individual sequence redundancy for general d -parametric families (under smoothness conditions) can be shown to satisfy [258]:

$$\text{Reg}_n^*(\Theta) = \mathcal{R}_n^*(\Theta) + \frac{d}{2} \log e + o(1) = \frac{d}{2} \log \frac{n}{2\pi} + \log \int_{\Theta} \sqrt{\det J_F(\theta)} d\theta + o(1).$$

In machine learning terms, we say that $\mathcal{R}_n^*(\Theta)$ in (13.8) is a cumulative sequential prediction regret under the well-specified setting (i.e. data X^n is generated by a distribution inside the model class Θ), while here $\text{Reg}_n^*(\Theta)$ corresponds to a fully mis-specified setting (i.e. data is completely arbitrary). There are also interesting settings in between these extremes, e.g. when data is iid but not from a model class Θ , cf. [118].

13.7 Lempel-Ziv compressor

So given a class of sources $\{P_{X^n|\theta} : \theta \in \Theta\}$ we have shown how to produce an asymptotically optimal compressors by using Jeffreys' prior. In the case of a class of i.i.d. processes, the resulting sequential probability of Krichevsky-Trofimov, see (13.5), had a very simple algorithmic description. When extended to more general classes (such as r -th order Markov chains), however, the sequential probability rules become rather complex. The Lempel-Ziv approach was to forego the path “design Q_{X^n} , convert to $Q_{X_r|X^{r-1}}$, extract compressor” and attempt to directly construct a reasonable sequential compressor or, equivalently, derive an algorithmically simple sequential estimator $Q_{X_r|X^{r-1}}$. The corresponding joint distribution Q_{X^n} is hard to imagine, and the achieved redundancy is not easy to derive, but the the algorithm becomes very transparent.

In principle, the problem is rather straightforward: as we observe a stationary process, we may estimate with better and better precision the conditional probability $\hat{P}_{X_n|X_{n-r}^{n-1}}$ and then use it as the basis for arithmetic coding. As long as \hat{P} converges to the actual conditional probability, we will attain the entropy rate of $H(X_n|X_{n-r}^{n-1})$. Note that Krichevsky-Trofimov assignment (13.28) is clearly learning the distribution too: as n grows, the estimator $Q_{X_n|X^{n-1}}$ converges to the true P_X (provided that the sequence is i.i.d.). So in some sense the converse is also true: *any good universal compression scheme is inherently learning the true distribution*.

The main drawback of the learn-then-compress approach is the following. Once we extend the class of sources to include those with memory, we invariably are lead to the problem of learning the joint distribution $P_{X_0^{r-1}}$ of r -blocks. However, the number of samples required to obtain a good estimate of $P_{X_0^{r-1}}$ is exponential in r . Thus learning may proceed rather slowly. Lempel-Ziv family of algorithms works around this in an ingeniously elegant way:

- First, estimating probabilities of rare substrings takes longest, but it is also the least useful, as these substrings almost never appear at the input.

- Second, *and the most crucial*, point is that an unbiased estimate of $P_{X^r}(x^r)$ is given by the reciprocal of the time since the last observation of x^r in the data stream.
- Third, there is a prefix code³ mapping any integer n to binary string of length roughly $\log_2 n$:

$$f_{int} : \mathbb{Z}_+ \rightarrow \{0, 1\}^+, \quad \ell(f_{int}(n)) = \log_2 n + O(\log \log n). \quad (13.32)$$

Thus, by encoding the pointer to the last observation of x^r via such a code we get a string of length roughly $\log P_{X^r}(x^r)$ automatically.

There are a number of variations of these basic ideas, so we will only attempt to give a rough explanation of why it works, without analyzing any particular algorithm.

We proceed to formal details. First, we need to establish Kac's lemma.

Lemma 13.3 (Kac). *Consider a finite-alphabet stationary ergodic process $\dots, X_{-1}, X_0, X_1 \dots$. Let $L = \inf\{t > 0 : X_{-t} = X_0\}$ be the last appearance of symbol X_0 in the sequence $X_{-\infty}^{-1}$. Then for any u such that $\mathbb{P}[X_0 = u] > 0$ we have*

$$\mathbb{E}[L|X_0 = u] = \frac{1}{\mathbb{P}[X_0 = u]}.$$

In particular, mean recurrence time $\mathbb{E}[L] = |\text{supp}(P_X)|$.

Proof. Note that from stationarity the following probability

$$\mathbb{P}[\exists t \geq k : X_t = u]$$

does not depend on $k \in \mathbb{Z}$. Thus by continuity of probability we can take $k = -\infty$ to get

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = \mathbb{P}[\exists t \in \mathbb{Z} : X_t = u].$$

However, the last event is shift-invariant and thus must have probability zero or one by ergodic assumption. But since $\mathbb{P}[X_0 = u] > 0$ it cannot be zero. So we conclude

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = 1. \quad (13.33)$$

Next, we have

$$\mathbb{E}[L|X_0 = u] = \sum_{t \geq 1} \mathbb{P}[L \geq t | X_0 = u] \quad (13.34)$$

$$= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t \geq 1} \mathbb{P}[L \geq t, X_0 = u] \quad (13.35)$$

$$= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t \geq 1} \mathbb{P}[X_{-t+1} \neq u, \dots, X_{-1} \neq u, X_0 = u] \quad (13.36)$$

$$= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t \geq 1} \mathbb{P}[X_0 \neq u, \dots, X_{t-2} \neq u, X_{t-1} = u] \quad (13.37)$$

³ For this just notice that $\sum_{k \geq 1} 2^{-\log_2 k - 2 \log_2 \log(k+1)} < \infty$ and use Kraft's inequality. See also Ex. II.14.

13.7 Lempel-Ziv compressor 219

$$= \frac{1}{\mathbb{P}[X_0 = u]} \mathbb{P}[\exists t \geq 0 : X_t = u] \quad (13.38)$$

$$= \frac{1}{\mathbb{P}[X_0 = u]}, \quad (13.39)$$

where (13.34) is the standard expression for the expectation of a \mathbb{Z}_+ -valued random variable, (13.37) is from stationarity, (13.38) is because the events corresponding to different t are disjoint, and (13.39) is from (13.33). \square

The following proposition serves to explain the basic principle behind operation of Lempel-Ziv:

Theorem 13.4. *Consider a finite-alphabet stationary ergodic process $\dots, X_{-1}, X_0, X_1 \dots$ with entropy rate H . Suppose that $X_{-\infty}^{-1}$ is known to the decoder. Then there exists a sequence of prefix-codes $f_n(x_0^{n-1}, x_{-\infty}^{-1})$ with expected length*

$$\frac{1}{n} \mathbb{E}[\ell(f_n(X_0^{n-1}, X_{-\infty}^{-1}))] \rightarrow H,$$

Proof. Let L_n be the last occurrence of the block x_0^{n-1} in the string $x_{-\infty}^{-1}$ (recall that the latter is known to decoder), namely

$$L_n = \inf\{t > 0 : x_{-t}^{-t+n-1} = x_0^{n-1}\}.$$

Then, by Kac's lemma applied to the process $Y_t^{(n)} = X_t^{t+n-1}$ we have

$$\mathbb{E}[L_n | X_0^{n-1} = x_0^{n-1}] = \frac{1}{\mathbb{P}[X_0^{n-1} = x_0^{n-1}]}.$$

We know encode L_n using the code (13.32). Note that there is crucial subtlety: even if $L_n < n$ and thus $[-t, -t+n-1]$ and $[0, n-1]$ overlap, the substring x_0^{n-1} can be decoded from the knowledge of L_n .

We have, by applying Jensen's inequality twice and noticing that $\frac{1}{n}H(X_0^{n-1}) \searrow H$ and $\frac{1}{n}\log H(X_0^{n-1}) \rightarrow 0$ that

$$\frac{1}{n} \mathbb{E}[\ell(f_{int}(L_n))] \leq \frac{1}{n} \mathbb{E}[\log \frac{1}{P_{X_0^{n-1}}(X_0^{n-1})}] + o(1) \rightarrow H.$$

From Kraft's inequality we know that for any prefix code we must have

$$\frac{1}{n} \mathbb{E}[\ell(f_{int}(L_n))] \geq \frac{1}{n} H(X_0^{n-1} | X_{-\infty}^{-1}) = H.$$

\square

The result shown above demonstrates that LZ algorithm has asymptotically optimal compression rate for every stationary ergodic process. Recall, however, that previously discussed compressors also enjoyed non-stochastic (individual sequence) guarantees. For example, we have seen in Section 13.6 that Krichevsky-Trofimov's compressor achieves on *every input sequence* a compression ratio that is at most $O(\frac{\log n}{n})$ worse than the arithmetic encoder built with the best possible (for this sequence!) static probability assignment. It turns out that LZ algorithm is also

220

special from this point of view. In [?] (see also [?, Theorem 4]) it was shown that the LZ compression rate on *every input sequence* is better than that achieved by any finite state machine (FSM) up to correction terms $O(\frac{\log \log n}{\log n})$. Consequently, investing via LZ achieves capital growth that is competitive against any possible FSM investor [?].

Altogether we can see that LZ compression enjoys certain optimality guarantees in both the stochastic and individual sequence senses.

Exercises for Part II

II.1 Let $S_j \in \{\pm 1\}$ be a stationary two-state Markov process with

$$P_{S_j|S_{j-1}}(s|s') = \begin{cases} \tau, & s \neq s' \\ 1 - \tau, & s = s' \end{cases}.$$

Let $E_j \stackrel{iid}{\sim} \text{Ber}(\delta)$, with $E_j \in \{0, 1\}$ and let Y_j be the observation of S_j through the binary erasure channel with erasure probability δ , i.e.

$$Y_j = S_j E_j.$$

Find entropy rate of Y_j (you can give answer in the form of a convergent series). Evaluate at $\tau = 0.11$, $\delta = 1/2$ and compare with $H(Y_1)$.

II.2 Recall that an entropy rate of a process $\{X_j : j = 1, \dots\}$ is defined as follows provided the limit exists:

$$\mathcal{H} = \lim_{n \rightarrow \infty} \frac{1}{n} H(X^n).$$

Consider a 4-state Markov chain with transition probability matrix

$$\begin{bmatrix} 0.89 & 0.11 & 0 & 0 \\ 0.11 & 0.89 & 0 & 0 \\ 0 & 0 & 0.11 & 0.89 \\ 0 & 0 & 0.89 & 0.11 \end{bmatrix}$$

The distribution of the initial state is $[p, 0, 0, 1 - p]$.

- (a) Does the entropy rate of such a Markov chain exist? If it does, find it.
- (b) Describe the asymptotic behavior of the optimum variable-length rate $\frac{1}{n} \ell(f^*(X_1, \dots, X_n))$. Consider convergence in probability and in distribution.
- (c) Repeat with transition matrix:

$$\begin{bmatrix} 0.89 & 0.11 & 0 & 0 \\ 0.11 & 0.89 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

II.3 Consider a three-state Markov chain S_1, S_2, \dots with the following transition probability matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix}.$$

222 Exercises for Part II

Compute the limit of $\frac{1}{n}\mathbb{E}[l(f^*(S^n))]$ when $n \rightarrow \infty$. Does your answer depend on the distribution of the initial state S_1 ?

- II.4** (a) Let X take values on a finite alphabet \mathcal{X} . Prove that

$$\epsilon^*(X, k) \geq \frac{H(X) - k - 1}{\log(|\mathcal{X}| - 1)}.$$

- (b) Deduce the following converse result: For a stationary process $\{S_k : k \geq 1\}$ on a finite alphabet \mathcal{S} ,

$$\liminf_{n \rightarrow \infty} \epsilon^*(S^n, nR) \geq \frac{\mathcal{H} - R}{\log |\mathcal{S}|}.$$

where $\mathcal{H} = \lim_{n \rightarrow \infty} \frac{H(S^n)}{n}$ is the entropy rate of the process.

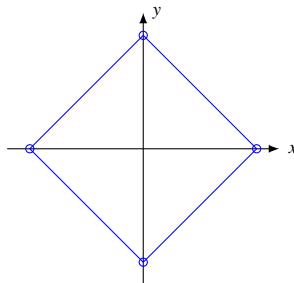
- II.5** *Run-length encoding* is a popular variable-length lossless compressor used in fax machines, image compression, etc. Consider compression of S^n – an i.i.d. $\text{Ber}(\delta)$ source with very small $\delta = \frac{1}{128}$ using run-length encoding: A chunk of consecutive $r \leq 255$ zeros (resp. ones) is encoded into a zero (resp. one) followed by an 8-bit binary encoding of r (If there are > 255 consecutive zeros then two or more 9-bit blocks will be output). Compute the average achieved compression rate

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\ell(f(S^n))]$$

How does it compare with the optimal lossless compressor?

Hint: Compute the expected number of 9-bit blocks output per chunk of consecutive zeros/ones; normalize by the expected length of the chunk.

- II.6** Draw n random points independently and uniformly from the vertices of the following square.



Denote the coordinates by $(X_1, Y_1), \dots, (X_n, Y_n)$. Suppose Alice only observes X^n and Bob only observes Y^n . They want to encode their observation using R_X and R_Y bits per symbol respectively and send the codewords to Charlie who will be able to reconstruct the sequence of pairs.

- (a) Find the optimal rate region for (R_X, R_Y) .
(b) What if the square is rotated by 45° ?

- II.7** Recall a bound on the probability of error for the Slepian-Wolf compression to k bits:

$$\epsilon_{SW}^*(k) \leq \min_{\tau > 0} \mathbb{P} \left[\log_{|\mathcal{A}|} \frac{1}{P_{X^n|Y}(X^n|Y)} > k - \tau \right] + |\mathcal{A}|^{-\tau} \quad (\text{II.1})$$

Consider the following case: $X^n = (X_1, \dots, X_n)$ – uniform on $\{0, 1\}^n$ and

$$Y = (X_1, \dots, X_n) + (N_1, \dots, N_n),$$

where N_i are iid Gaussian with zero mean and variance 0.1

Let $n = 10$. Propose a method to numerically compute or approximate the bound (II.1) as a function of $k = 1, \dots, 10$. Plot the results.

- II.8** Consider a probability measure \mathbb{P} and a measure-preserving transformation $\tau : \Omega \rightarrow \Omega$. Prove: τ -ergodic iff for any measurable A, B we have

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{P}[A \cap \tau^{-k}B] \rightarrow \mathbb{P}[A]\mathbb{P}[B].$$

Comment: Thus ergodicity is a weaker condition than *mixing*: $\mathbb{P}[A \cap \tau^{-n}B] \rightarrow \mathbb{P}[A]\mathbb{P}[B]$.

- II.9** Consider a ternary fixed length (almost lossless) compression $\mathcal{X} \rightarrow \{0, 1, 2\}^k$ with an additional requirement that the string in $w^k \in \{0, 1, 2\}^k$ should satisfy

$$\sum_{j=1}^k w_j \leq \frac{k}{2} \tag{II.2}$$

For example, $(0, 0, 0, 0)$, $(0, 0, 0, 2)$ and $(1, 1, 0, 0)$ satisfy the constraint but $(0, 0, 1, 2)$ does not. Let $\epsilon^*(S^n, k)$ denote the minimum probability of error among all possible compressors of $S^n = \{S_j, j = 1, \dots, n\}$ with i.i.d. entries of finite entropy $H(S) < \infty$. Compute

$$\lim_{n \rightarrow \infty} \epsilon^*(S^n, nR)$$

as a function of $R \geq 0$.

Hint: Relate to $\mathbb{P}[\ell(f^*(S^n)) \geq \gamma n]$ and use Stirling's formula (or Theorems 11.1.1, 11.1.3 in [75]) to find γ .

- II.10 Mismatched compression.** Let P, Q be distributions on some discrete alphabet \mathcal{A} .

- (a) Let $f_P^* : \mathcal{A} \rightarrow \{0, 1\}$ denote the optimal variable-length lossless compressor for $X \sim P$. Show that under Q ,

$$\mathbb{E}_Q[l(f_P^*(X))] \leq H(Q) + D(Q||P).$$

- (b) The Shannon code for $X \sim P$ is a prefix code f_P with the code length $l(f_P(a)) = \lceil \log_2 \frac{1}{P(a)} \rceil$, $a \in \mathcal{A}$. Show that if X is distributed according to Q instead, then

$$H(Q) + D(Q||P) \leq \mathbb{E}_Q[l(f_P(X))] \leq H(Q) + D(Q||P) + 1 \text{ bit.}$$

Comments: This can be interpreted as a robustness result for compression with model misspecification: When a compressor designed for P is applied to a source whose distribution is in fact Q , the suboptimality incurred by this mismatch can be related to divergence $D(Q||P)$.

- II.11 Arithmetic Coding.** We analyze the encoder defined by (13.1) for iid source. Let P be a distribution on some ordered finite alphabet, say, $a < b < \dots < z$. For each n , define $p(x^n) = \prod_{i=1}^n P(x_i)$ and $q(x^n) = \sum_{y^n < x^n} p(y^n)$ according to the lexicographic ordering, so that $F_n(x^n) = q(x^n)$ and $|I_{x^n}| = p(x^n)$.

224 Exercises for Part II

(a) Show that if $x^{n-1} = (x_1, \dots, x_{n-1})$, then

$$q(x^n) = q(x^{n-1}) + p(x^{n-1}) \sum_{\alpha < x_n} P(\alpha).$$

Conclude that $q(x^n)$ can be computed in $O(n)$ steps sequentially.

- (b) Show that intervals I_{x^n} are disjoint subintervals of $[0, 1)$.
- (c) *Encoding.* Show that the codelength $l(f(x^n))$ defined in (13.1) satisfies the constraint (13.2), namely, $\log_2 \frac{1}{p(x^n)} \leq l(f(x^n)) \leq \left\lceil \log_2 \frac{1}{p(x^n)} \right\rceil + 1$. Furthermore, verify that the map $x^n \mapsto f(x^n)$ defines a prefix code. (*Warning:* This is not about checking Kraft's inequality.)
- (d) *Decoding.* Upon receipt of the codeword, we can reconstruct the interval D_{x^n} . Divide the unit interval according to the distribution P , i.e., partition $[0, 1)$ into disjoint subintervals I_a, \dots, I_z . Output the index that contains D_{x^n} . Show that this gives the first symbol x_1 . Continue in this fashion by dividing I_{x_1} into $I_{x_1,a}, \dots, I_{x_1,z}$ and etc. Argue that x^n can be decoded losslessly. How many steps are needed?
- (e) Suppose $P_X(e) = 0.5, P_X(o) = 0.3, P_X(t) = 0.2$. Encode etoo (write the binary codewords) and describe how to decode.
- (f) Show that the average length of this code satisfies

$$nH(P) \leq \mathbb{E}[l(f(X^n))] \leq nH(P) + 2 \text{ bits.}$$

- (g) Assume that $X = (X_1, \dots, X_n)$ is not iid but $P_{X_1}, P_{X_2|X_1}, \dots, P_{X_n|X^{n-1}}$ are known. How would you modify the scheme so that we have

$$H(X^n) \leq \mathbb{E}[l(f(X^n))] \leq H(X^n) + 2 \text{ bits.}$$

II.12 Enumerative Codes. Consider the following simple universal compressor for binary sequences:

Given $x^n \in \{0, 1\}^n$, denote by $n_1 = \sum_{i=1}^n x_i$ and $n_0 = n - n_1$ the number of ones and zeros in x^n . First encode $n_1 \in \{0, 1, \dots, n\}$ using $\lceil \log_2(n+1) \rceil$ bits, then encode the index of x^n in the set of all strings with n_1 number of ones using $\lceil \log_2 \binom{n}{n_1} \rceil$ bits. Concatenating two binary strings, we obtain the codeword of x^n . This defines a lossless compressor $f: \{0, 1\}^n \rightarrow \{0, 1\}^*$.

- (a) Verify that f is a prefix code.
- (b) Let $S_\theta^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$. Show that for any $\theta \in [0, 1]$,

$$\mathbb{E}[l(f(S_\theta^n))] \leq nh(\theta) + \log n + O(1),$$

where $h(\cdot)$ is the binary entropy function. Conclude that

$$\sup_{0 \leq \theta \leq 1} \{\mathbb{E}[l(f(S_\theta^n))] - nh(\theta)\} \geq \log n + O(1).$$

[Optional: Explain why enumerative coding fails to achieve the optimal redundancy.]

Hint: The following non-asymptotic version of Stirling approximation *might* be useful

$$1 \leq \frac{n!}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} \leq \frac{e}{\sqrt{2\pi}}, \quad \forall n \in \mathbb{N}.$$

Exercises for Part II 225

II.13 Krichevsky-Trofimov codes. From Kraft's inequality we know that any probability distribution Q_{X^n} on $\{0, 1\}^n$ gives rise to a prefix code f such that $l(f(x^n)) = \left\lceil \log_2 \frac{1}{Q_{X^n}(x^n)} \right\rceil$ for all x^n . Consider the following Q_{X^n} defined by the factorization $Q_{X^n} = Q_{X_1} Q_{X_2|X_1} \cdots Q_{X_n|X^{n-1}}$,

$$Q_{X_1}(1) = \frac{1}{2}, \quad Q_{X_{t+1}|X^t}(1|x^t) = \frac{n_1(x^t) + \frac{1}{2}}{t+1}, \quad (\text{II.3})$$

where $n_1(x^t)$ denotes the number of ones in x^t . Denote the prefix code corresponding to this Q_{X^n} by $f_{\text{KT}} : \{0, 1\}^n \rightarrow \{0, 1\}^*$.

(a) Prove that for any n and any $x^n \in \{0, 1\}^n$,

$$Q_{X^n}(x^n) \geq \frac{1}{2} \frac{1}{\sqrt{n_0 + n_1}} \left(\frac{n_0}{n_0 + n_1} \right)^{n_0} \left(\frac{n_1}{n_0 + n_1} \right)^{n_1}.$$

where $n_0 = n_0(x^n)$ and $n_1 = n_1(x^n)$ denote the number of zeros and ones in x^n .

Hint: Use induction on n .

(b) Conclude that the K-T code length satisfies:

$$l(f_{\text{KT}}(x^n)) \leq nh\left(\frac{n_1}{n}\right) + \frac{1}{2} \log n + 2, \quad \forall x^n \in \{0, 1\}^n.$$

(c) Conclude that for K-T codes :

$$\sup_{0 \leq \theta \leq 1} \{\mathbb{E}[l(f_{\text{KT}}(S_\theta^n))] - nh(\theta)\} \leq \frac{1}{2} \log n + O(1).$$

This value is known as the *redundancy* of a universal code. It turns out that $\frac{1}{2} \log n + O(1)$ is optimal for the class of all Bernoulli sources (see (13.23)).

Comments:

- (a) The probability assignment (II.3) is known as the “add- $\frac{1}{2}$ ” estimator: Upon observing x^t which contains n_1 number of ones, a natural probability assignment to $x_{t+1} = 1$ is the empirical average $\frac{n_1}{t}$. Instead, K-T codes assign probability $\frac{n_1 + \frac{1}{2}}{t+1}$, or equivalently, adding $\frac{1}{2}$ to both n_0 and n_1 . This is a crucial modification to Laplace’s “add-one estimator”.⁴
- (b) By construction, the probability assignment Q_{X^n} can be sequentially computed, which allows us implement sequential encoding and encode a stream of bits on the fly. This is a highly desirable feature of the K-T codes. Of course, we need to resort to construction other than the one in Kraft's inequality construction, e.g., arithmetic coding.

II.14 (Elias coding) In this problem all logarithms and entropy units are binary.

- (a) Consider the following universal compressor for natural numbers: For $x \in \mathbb{N} = \{1, 2, \dots\}$, let $k(x)$ denote the length of its binary representation. Define its codeword $c(x)$ to be $k(x)$ zeros followed by the binary representation of x . Compute $c(10)$. Show that c is a prefix code and describe how to decode a stream of codewords.

⁴ Interested readers should check *Laplace's rule of succession* and the sunrise problem https://en.wikipedia.org/wiki/Rule_of_succession.

226 Exercises for Part II

- (b) Next we construct another code using the one above: Define the codeword $c'(x)$ to be $c(k(x))$ followed by the binary representation of x . Compute $c'(10)$. Show that c' is a prefix code and describe how to decode a stream of codewords.
- (c) Let X be a random variable on \mathbb{N} whose probability mass function is decreasing. Show that $\mathbb{E}[\log(X)] \leq H(X)$.
- (d) Show that the average code length of c satisfies $\mathbb{E}[\ell(c(X))] \leq 2H(X) + 2$ bit.
- (e) Show that the average code length of c' satisfies $\mathbb{E}[\ell(c'(X))] \leq H(X) + 2 \log(H(X) + 1) + 3$ bit.

Comments: The two coding schemes are known as Elias γ -codes and δ -codes.

Part III

Binary hypothesis testing



In this part we study the topic of binary hypothesis testing (BHT). This is an important area of statistics, with a definitive treatment given in [193]. Historically, there has been two schools of thought on how to approach this question. One is the so-called *significance testing* of Karl Pearson and Ronald Fisher. This is perhaps the most widely used approach in modern biomedical and social sciences. The concepts of null hypothesis, p -value, χ^2 -test, goodness-of-fit belong to this world. We will not be discussing these.

The other school was pioneered by Jerzy Neyman and Egon Pearson, and is our topic in this part. The concepts of type-I and type-II errors, likelihood-ratio tests, Chernoff exponent are from this domain. This is, arguably, a more popular way of looking at the problem among the engineering disciplines (perhaps explained by its foundational role in radar and electronic signal detection.)

The conceptual difference between the two is that in the first approach the full probabilistic model is specified *only* under the null hypothesis. (It still could be very specific like $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, contain unknown parameters, like $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ with $\theta \in \mathbb{R}$ arbitrary, or be nonparametric, like $(X_i, Y_i) \stackrel{\text{i.i.d.}}{\sim} P_{X,Y} = P_X P_Y$ denoting that observables X and Y are statistically independent). The main goal of the statistician in this setting is inventing a testing process that is able to find statistically significant deviations from the postulated null behavior. If such deviation is found then the null is rejected and (in scientific fields) a discovery is announced. The role of the alternative hypothesis (if one is specified at all) is to roughly suggest what feature of the null are most likely to be violated and motivates the choice of test procedures. For example, if under the null $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then both of the following are reasonable tests:

$$\frac{1}{n} \sum_{i=1}^n X_i \stackrel{?}{\approx} 0 \quad \frac{1}{n} \sum_{i=1}^n X_i^2 \stackrel{?}{\approx} 1.$$

However, the first one would be preferred if, under the alternative, “data has non-zero mean”, and the second if “data has zero mean but variance not equal to one”. Whichever of the alternatives is selected does not imply in any way the validity of the alternative. In addition, theoretical properties of the test are mostly studied under the null rather than the alternative. For this approach the null hypothesis (out of the two) plays a very special role.

The second approach treats hypotheses in complete symmetry. Exact specifications of probability distributions are required for both hypotheses and the precision of a proposed test is to be analyzed under both. This is the setting that is most useful for our treatment of forthcoming topics of channel coding (Part IV) and statistical estimation (Part VI).

The outline of this part is the following. First, we define the performance metric $\mathcal{R}(P, Q)$ giving a full description of the BHT problem. A key result in this theory, the Neyman-Pearson lemma determines the form of the optimal test and, at the same time, characterizes $\mathcal{R}(P, Q)$. We then specialize to the setting of iid observations and consider two types of asymptotics (as the sample size n goes to infinity): Stein’s regime (where type-I error is held constant) and Chernoff’s regime (where errors of both types are required to decay exponentially). The fundamental limit in the former regime is simply a scalar (given by $D(P||Q)$), while in the latter it is a region. To describe this region (Chapter 16) we will need to understand the problem of large deviations and the information projection (Chapter 15).

14 Neyman-Pearson lemma

14.1 Neyman-Pearson formulation

Consider the situation where we have two possible distributions on a space \mathcal{X} and

$$\begin{aligned} H_0 &: X \sim P \\ H_1 &: X \sim Q. \end{aligned}$$

What this means is that under hypothesis H_0 (the null hypothesis) X is distributed according to P , and under H_1 (the alternative hypothesis) X is distributed according to Q . A *test* (or decision rule) between two distributions chooses either H_0 or H_1 based on an observation of X . We will consider

- Deterministic tests: $f : \mathcal{X} \rightarrow \{0, 1\}$, or equivalently, $f(x) = 1_{\{x \in E\}}$ where E is known as a *decision region*; and more generally,
- Randomized tests: $P_{Z|X} : \mathcal{X} \rightarrow \{0, 1\}$, so that $P_{Z|X}(1|x) \in [0, 1]$ is the probability of rejecting the null upon observing $X = x$.

Let $Z = 0$ denote that the test chooses P (accepting the null) and $Z = 1$ that the test chooses Q (rejecting the null).

This setting is called “testing simple hypothesis against simple hypothesis”. Here “simple” refers to the fact that under each hypothesis there is only one distribution that could generate the data. In comparison, composite hypothesis postulates that $X \sim P$ for *some* P is a given class of distributions; see Section 32.2.1.

In order to quantify the “effectiveness” of a test, we focus on two metrics. Let π_{ij} denote the probability of the test choosing i when the correct hypothesis is j , with $i, j \in \{0, 1\}$. For every test $P_{Z|X}$ we associate a pair of numbers:

$$\begin{aligned} \alpha &= \pi_{0|0} = P[Z = 0] \quad (\text{Probability of success given } H_0 \text{ is true}) \\ \beta &= \pi_{0|1} = Q[Z = 0] \quad (\text{Probability of error given } H_1 \text{ is true}), \end{aligned}$$

where $P[Z = 0] = \int P_{Z|X}(0|x)P(dx)$ and $Q[Z = 0] = \int P_{Z|X}(0|x)Q(dx)$. There are many alternative names for these quantities: $1 - \alpha$ is called significance level, size, type-I error, false positive, false alarm rate of a test; β is called type-II error, false negative, missed detection rate; $1 - \beta$ or $\pi_{1|1}$ is known as true positive or the *power* of a test.

There are a few ways to determine the “best test”:

14.1 Neyman-Pearson formulation 231

- Bayesian: Assuming the prior distribution $\mathbb{P}[H_0] = \pi_0$ and $\mathbb{P}[H_1] = \pi_1$, we minimize the average probability of error:

$$P_b^* = \min_{P_{Z|X}: \mathcal{X} \rightarrow \{0,1\}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}. \quad (14.1)$$

- Minimax: Assuming there is an unknown prior distribution, we choose the test that performs the best for the worst-case prior

$$P_m^* = \min_{P_{Z|X}: \mathcal{X} \rightarrow \{0,1\}} \max\{\pi_{1|0}, \pi_{0|1}\}.$$

- Neyman-Pearson: Minimize the type-II error β subject to that the success probability under the null is at least α .

In this book the Neyman-Pearson formulation and the following quantities play important roles:

Definition 14.1. Given (P, Q) , the Neyman-Pearson region consists of achievable points for all randomized tests

$$\mathcal{R}(P, Q) = \{(P[Z = 0], Q[Z = 0]) : P_{Z|X} : \mathcal{X} \rightarrow \{0, 1\}\} \subset [0, 1]^2. \quad (14.2)$$

In particular, its lower boundary is defined as (see Fig. 14.1 for an illustration)

$$\beta_\alpha(P, Q) \triangleq \inf_{P[Z=0] \geq \alpha} Q[Z = 0] \quad (14.3)$$

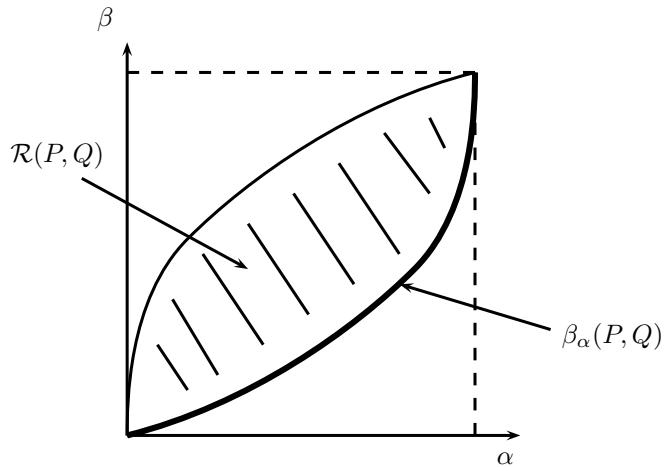
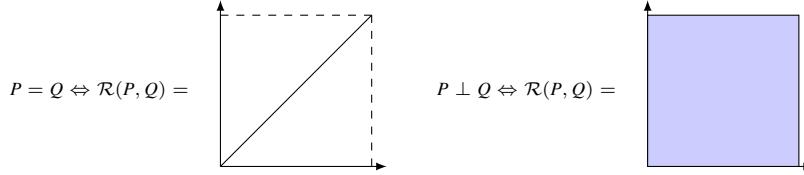


Figure 14.1 Illustration of the Neyman-Pearson region.

Remark 14.1. The Neyman-Pearson region encodes much useful information about the relationship between P and Q . For example, we have the following extreme cases¹

¹ Recall that P is mutually singular w.r.t. Q , denoted by $P \perp Q$, if $P[E] = 0$ and $Q[E] = 1$ for some E .

232



Moreover, $\text{TV}(P, Q)$ coincides with half the length of the longest vertical segment contained in $\mathcal{R}(P, Q)$ (Exercise III.2).

Theorem 14.2 (Properties of $\mathcal{R}(P, Q)$).

- (a) $\mathcal{R}(P, Q)$ is a closed, convex subset of $[0, 1]^2$.
- (b) $\mathcal{R}(P, Q)$ contains the diagonal.
- (c) Symmetry: $(\alpha, \beta) \in \mathcal{R}(P, Q) \Leftrightarrow (1 - \alpha, 1 - \beta) \in \mathcal{R}(P, Q)$.

Proof. (a) For convexity, suppose that $(\alpha_0, \beta_0), (\alpha_1, \beta_1) \in \mathcal{R}(P, Q)$, corresponding to tests $P_{Z_0|X}, P_{Z_1|X}$, respectively. Randomizing between these two tests, we obtain the test $\lambda P_{Z_0|X} + \bar{\lambda} P_{Z_1|X}$ for $\lambda \in [0, 1]$, which achieves the point $(\lambda\alpha_0 + \bar{\lambda}\alpha_1, \lambda\beta_0 + \bar{\lambda}\beta_1) \in \mathcal{R}(P, Q)$.

The closedness of $\mathcal{R}(P, Q)$ will follow from the explicit determination of all boundary points via the Neyman-Pearson lemma – see Remark 14.12. In more complicated situations (e.g. in testing against composite hypothesis) simple explicit solutions similar to Neyman-Pearson Lemma are not available but closedness of the region can frequently be argued still. The basic reason is that the collection of bounded functions $\{g : \mathcal{X} \rightarrow [0, 1]\}$ (with $g(x) = P_{Z|X}(0|x)$) forms a weakly compact set and hence its image under the linear functional $g \mapsto (\int g dP, \int g dQ)$ is closed.

- (b) Testing by random guessing, i.e., $Z \sim \text{Ber}(1 - \alpha) \perp\!\!\!\perp X$, achieves the point (α, α) .
- (c) If $(\alpha, \beta) \in \mathcal{R}(P, Q)$ is achieved by $P_{Z|X}, P_{1-Z|X}$ achieves $(1 - \alpha, 1 - \beta)$.

□

The region $\mathcal{R}(P, Q)$ consists of the operating points of all randomized tests, which include as special cases those of deterministic tests, namely

$$\mathcal{R}_{\text{det}}(P, Q) = \{(P(E), Q(E)) : E \text{ measurable}\}. \quad (14.4)$$

As the next result shows, the former is in fact the closed convex hull of the latter. Recall that $\text{cl}(E)$ (resp. $\text{co}(E)$) denote the closure and convex hull of a set E , namely, the smallest closed (resp. convex) set containing E . A useful example: For a subset E of an Euclidean space, and measurable functions $f, g : \mathbb{R} \rightarrow E$, we have $(\mathbb{E}[f(X)], \mathbb{E}[g(X)]) \in \text{cl}(\text{co}(E))$ for any real-valued random variable X .

Theorem 14.3 (Randomized test v.s. deterministic tests).

$$\mathcal{R}(P, Q) = \text{cl}(\text{co}(\mathcal{R}_{\text{det}}(P, Q))).$$

Consequently, if P and Q are on a finite alphabet \mathcal{X} , then $\mathcal{R}(P, Q)$ is a polygon of at most $2^{|\mathcal{X}|}$ vertices.

14.2 Likelihood ratio tests 233

Proof. “ \supset ”: Comparing (14.2) and (14.4), by definition, $\mathcal{R}(P, Q) \supset \mathcal{R}_{\text{det}}(P, Q)$, the former of which is closed and convex, by Theorem 14.3.

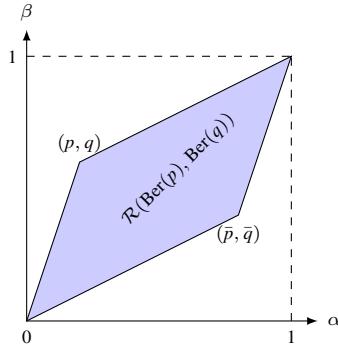
“ \subset ”: Given any randomized test $P_{Z|X}$, define a measurable function $g : \mathcal{X} \rightarrow [0, 1]$ by $g(x) = P_{Z|X}(0|x)$. Then

$$\begin{aligned} P[Z = 0] &= \sum_x g(x)P(x) = \mathbb{E}_P[g(X)] = \int_0^1 P[g(X) \geq t]dt \\ Q[Z = 0] &= \sum_x g(x)Q(x) = \mathbb{E}_Q[g(X)] = \int_0^1 Q[g(X) \geq t]dt \end{aligned}$$

where we applied the “area rule” that $E[U] = \int_{\mathbb{R}_+} \mathbb{P}[U \geq t] dt$ for any non-negative random variable U . Therefore the point $(P[Z = 0], Q[Z = 0]) \in \mathcal{R}$ is a mixture of points $(P[g(X) \geq t], Q[g(X) \geq t]) \in \mathcal{R}_{\text{det}}$, averaged according to t uniformly distributed on the unit interval. Hence $\mathcal{R} \subset \text{cl}(\text{co}(\mathcal{R}_{\text{det}}))$.

The last claim follows because there are at most $2^{|\mathcal{X}|}$ subsets in (14.4). \square

Example 14.1 (Testing $\text{Ber}(p)$ versus $\text{Ber}(q)$). Assume that $p < \frac{1}{2} < q$. Using Theorem 14.4, note that there are $2^2 = 4$ events $E = \emptyset, \{0\}, \{1\}, \{0, 1\}$. Then $\mathcal{R}(\text{Ber}(p), \text{Ber}(q))$ is given by



14.2 Likelihood ratio tests

To define optimal hypothesis tests, we need to define the concept of the log-likelihood ratio (LLR). In the simple case when $P \ll Q$ we can define the LLR $T(x) = \log \frac{dP}{dQ}(x)$ as a function $T : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ by thinking of $\log 0 = -\infty$. In order to handle also the case of $P \not\ll Q$, we can leverage our concept of the Log function, cf. (2.10).

Definition 14.4 (Extended log likelihood ratio). Assume that $dP = p(x)d\mu$ and $dQ = q(x)d\mu$ for some dominating measure μ (e.g. $\mu = P + Q$). Recalling the definition of Log from (2.10) we

define the extended LLR as

$$T(x) \triangleq \text{Log} \frac{p(x)}{q(x)} = \begin{cases} \log \frac{p(x)}{q(x)}, & p(x) > 0, q(x) > 0 \\ +\infty, & p(x) > 0, q(x) = 0 \\ -\infty, & p(x) = 0, q(x) > 0 \\ 0, & p(x) = 0, q(x) = 0, \end{cases}$$

The likelihood ratio test (LRT) with threshold $\tau \in \mathbb{R} \cup \{\pm\infty\}$ is $1\{x : T(x) \leq \tau\}$.

When $P \ll Q$ it is clear that $T(x) = \log \frac{dP}{dQ}(x)$ for P - and Q -almost every x . For this reason, everywhere in this Part we abuse notation and write simply $\log \frac{dP}{dQ}$ to denote the *extended* (!) LLR as defined above. Notice that LRT is a deterministic test, and that it does make intuitive sense: upon observing x , if $\frac{Q(x)}{P(x)}$ is large then Q is more likely and one should reject the null hypothesis P .

Note that for a discrete alphabet \mathcal{X} and assuming $Q \ll P$ we can see

$$Q[T = t] = \exp(-t)P[T = t] \quad \forall t \in \mathbb{R} \cup \{+\infty\}.$$

Indeed, this is shown by the following chain:

$$\begin{aligned} Q_T(t) &= \sum_x Q(x)1\{\log \frac{P(x)}{Q(x)} = t\} = \sum_x Q(x)1\{e^t Q(x) = P(x)\} \\ &= e^{-t} \sum_x P(x)1\{\log \frac{P(x)}{Q(x)} = t\} = e^{-t}P_T(t) \end{aligned}$$

We see that taking expectation over P and over Q are equivalent upon multiplying the expectant by $\exp(\pm T)$. The next result gives precise details in the general case.

Theorem 14.5 (Change of measure $P \leftrightarrow Q$). *The following hold:*

1 For any $h : \mathcal{X} \rightarrow \mathbb{R}$ we have

$$\mathbb{E}_Q[h(X)1\{T > -\infty\}] = \mathbb{E}_P[h(X)\exp(-T)] \quad (14.5)$$

$$\mathbb{E}_P[h(X)1\{T < +\infty\}] = \mathbb{E}_Q[h(X)\exp(T)] \quad (14.6)$$

2 For any $f \geq 0$ and any $-\infty < \tau < \infty$ we have

$$\begin{aligned} \mathbb{E}_Q[f(X)1\{T \geq \tau\}] &\leq \mathbb{E}_P[f(X)1\{T \geq \tau\}] \cdot \exp(-\tau) \\ \mathbb{E}_Q[f(X)1\{T \leq \tau\}] &\geq \mathbb{E}_P[f(X)1\{T \geq \tau\}] \cdot \exp(-\tau) \end{aligned} \quad (14.7)$$

Proof. We first observe that

$$Q[T = +\infty] = P[T = -\infty] = 0. \quad (14.8)$$

Then consider the chain

$$\mathbb{E}_Q[h(X)1\{T > -\infty\}] \stackrel{(a)}{=} \int_{\{-\infty < T(x) < \infty\}} d\mu q(x)h(x) \stackrel{(b)}{=} \int_{\{-\infty < T(x) < \infty\}} d\mu p(x)\exp(-T(x))h(x)$$

14.3 Converse bounds on $\mathcal{R}(P, Q)$ 235

$$\stackrel{(c)}{=} \int_{\{-\infty < T(x) \leq \infty\}} d\mu p(x) \exp(-T(x)) h(x) = \mathbb{E}_P[\exp(-T)g(T)],$$

where in (a) we used (14.8) to justify restriction to finite values of T ; in (b) we used $\exp(-T(x)) = \frac{q(x)}{p(x)}$ for $p, q > 0$; and (c) follows from the fact that $\exp(-T(x)) = 0$ whenever $T = \infty$. Exchanging the roles of P and Q proves (14.6).

The last part follows upon taking $h(x) = f(x)1\{T(x) \geq \tau\}$ and $h(x) = f(x)1\{T(x) \leq \tau\}$ in (14.5) and (14.6), respectively. \square

The importance of the LLR is that it is a sufficient statistic for testing the two hypotheses (recall Section 3.5 and in particular Example 3.8), as the following result shows.

Corollary 14.6. *$T = T(X)$ is a sufficient statistic for testing P versus Q .*

Proof. For part 2, sufficiency of T would be implied by $P_{X|T} = Q_{X|T}$. For the case of X being discrete we have:

$$\begin{aligned} P_{X|T}(x|t) &= \frac{P_X(x)P_{T|X}(t|x)}{P_T(t)} = \frac{P(x)1\{\frac{P(x)}{Q(x)} = e^t\}}{P_T(t)} = \frac{e^t Q(x)1\{\frac{P(x)}{Q(x)} = e^t\}}{P_T(t)} \\ &= \frac{Q_{XT}(xt)}{e^{-t}P_T(t)} = \frac{Q_{XT}}{Q_T} = Q_{X|T}(x|t). \end{aligned}$$

\square

We leave the general case as an exercise.

From Theorem 14.4 we know that to obtain the achievable region $\mathcal{R}(P, Q)$, one can iterate over all decision regions and compute the region $\mathcal{R}_{\text{det}}(P, Q)$ first, then take its closed convex hull. But this is a formidable task if the alphabet is large or infinite. On the other hand, we know that the LLR is a sufficient statistic. Next we give bounds to the region $\mathcal{R}(P, Q)$ in terms of the statistics of the LLR. As usual, there are two types of statements:

- Converse (outer bounds): any point in $\mathcal{R}(P, Q)$ must satisfy certain constraints;
- Achievability (inner bounds): points satisfying certain constraints belong to $\mathcal{R}(P, Q)$.

14.3 Converse bounds on $\mathcal{R}(P, Q)$

Theorem 14.7 (Weak converse). $\forall(\alpha, \beta) \in \mathcal{R}(P, Q)$,

$$\begin{aligned} d(\alpha\|\beta) &\leq D(P\|Q) \\ d(\beta\|\alpha) &\leq D(Q\|P) \end{aligned}$$

where $d(\cdot\|\cdot)$ is the binary divergence function in (2.6).

Proof. Use the data processing inequality for KL divergence with $P_{Z|X}$; cf. Corollary 2.2. \square

We will strengthen this bound with the aid of the following result.

Lemma 14.8. *For any test Z and any $\gamma > 0$ we have*

$$P[Z = 0] - \gamma Q[Z = 0] \leq P[T > \log \gamma],$$

where $T = \log \frac{dP}{dQ}$ is understood in the extended sense of Definition 14.5.

Note that we do not need to assume $P \ll Q$ precisely because $\pm\infty$ are admissible values for the (extended) LLR.

Proof. Defining $\tau = \log \gamma$ and $g(x) = P_{Z|X}(0|x)$ we get from (14.7):

$$P[Z = 0, T \leq \tau] - \gamma Q[Z = 0, T \leq \tau] \leq 0.$$

Decomposing $P[Z = 0] = P[Z = 0, T \leq \tau] + P[Z = 0, T > \tau]$ and similarly for Q we obtain then

$$P[Z = 0] - \gamma Q[Z = 0] \leq P[T > \log \gamma, Z = 0] - \gamma Q[T > \log \gamma, Z = 0] \leq P[T > \log \gamma]$$

□

Theorem 14.9 (Strong converse). $\forall(\alpha, \beta) \in \mathcal{R}(P, Q), \forall \gamma > 0,$

$$\alpha - \gamma\beta \leq P\left[\log \frac{dP}{dQ} > \log \gamma\right] \quad (14.9)$$

$$\beta - \frac{1}{\gamma}\alpha \leq Q\left[\log \frac{dP}{dQ} < \log \gamma\right] \quad (14.10)$$

Proof. Apply Lemma 14.8 to (P, Q, γ) and $(Q, P, 1/\gamma)$. □

Remark 14.2.

- Theorem 14.9 provides an outer bound for the region $\mathcal{R}(P, Q)$ in terms of half-spaces. To see this, fix $\gamma > 0$ and consider the line $\alpha - \gamma\beta = c$ by gradually increasing c from zero. There exists a maximal c , say c^* , at which point the line touches the lower boundary of the region. Then (14.9) says that c^* cannot exceed $P[\log \frac{dP}{dQ} > \log \gamma]$. Hence \mathcal{R} must lie to the left of the line. Similarly, (14.10) provides bounds for the upper boundary. Altogether Theorem 14.9 states that $\mathcal{R}(P, Q)$ is contained in the intersection of an infinite collection of half-spaces indexed by γ .
- To apply the strong converse Theorem 14.9, we need to know the CDF of the LLR, whereas to apply the weak converse Theorem 14.7 we need only to know the expectation of the LLR, i.e., the divergence.

14.4 Achievability bounds on $\mathcal{R}(P, Q)$

Given the convexity of the set $\mathcal{R}(P, Q)$, it is natural to try to find all of its supporting lines (hyperplanes), as it is well-known that closed convex set equals the intersection of all halfspaces that are supporting hyperplanes. We are thus lead to the following problem: for $t > 0$,

$$\max\{\alpha - t\beta : (\alpha, \beta) \in \mathcal{R}(P, Q)\},$$

14.4 Achievability bounds on $\mathcal{R}(P, Q)$ 237

which is equivalent to minimizing the average probability of error in (14.1), with $t = \frac{\pi_1}{\pi_0}$. This can be solved without much effort. For simplicity, consider the discrete case. Then

$$\alpha^* - t\beta^* = \max_{(\alpha, \beta) \in \mathcal{R}} (\alpha - t\beta) = \max_{P_{Z|X}} \sum_{x \in \mathcal{X}} (P(x) - tQ(x)) P_{Z|X}(0|x) = \sum_{x \in \mathcal{X}} |P(x) - tQ(x)|^+$$

where the last equality follows from the fact that we are free to choose $P_{Z|X}(0|x)$, and the best choice is obvious:

$$P_{Z|X}(0|x) = 1 \left\{ \log \frac{P(x)}{Q(x)} \geq \log t \right\}.$$

Thus, we have shown that all supporting hyperplanes are parameterized by LRT. This completely recovers the region $\mathcal{R}(P, Q)$ except for the points corresponding to the faces (flat pieces) of the region. The precise result is stated as follows:

Theorem 14.10 (Neyman-Pearson Lemma: “LRT is optimal”). *For each α, β_α in (14.3) is attained by the following test:*

$$P_{Z|X}(0|x) = \begin{cases} 1 & \log \frac{dP}{dQ} > \tau \\ \lambda & \log \frac{dP}{dQ} = \tau \\ 0 & \log \frac{dP}{dQ} < \tau \end{cases} \quad (14.11)$$

where $\tau \in \mathbb{R}$ and $\lambda \in [0, 1]$ are the unique solutions to $\alpha = P[\log \frac{dP}{dQ} > \tau] + \lambda P[\log \frac{dP}{dQ} = \tau]$.

Proof of Theorem 14.11. Let $t = \exp(\tau)$. Given any test $P_{Z|X}$, let $g(x) = P_{Z|X}(0|x) \in [0, 1]$. We want to show that

$$\alpha = P[Z = 0] = \mathbb{E}_P[g(X)] = P\left[\frac{dP}{dQ} > t\right] + \lambda P\left[\frac{dP}{dQ} = t\right] \quad (14.12)$$

$$\Rightarrow \beta = Q[Z = 0] = \mathbb{E}_Q[g(X)] \stackrel{\text{goal}}{\geq} Q\left[\frac{dP}{dQ} > t\right] + \lambda Q\left[\frac{dP}{dQ} = t\right] \quad (14.13)$$

Using the simple fact that $\mathbb{E}_Q[f(X)1_{\{\frac{dP}{dQ} \leq t\}}] \geq t^{-1}\mathbb{E}_P[f(X)1_{\{\frac{dP}{dQ} \leq t\}}]$ for any $f \geq 0$ twice, we have

$$\begin{aligned} \beta &= \mathbb{E}_Q[g(X)1_{\{\frac{dP}{dQ} \leq t\}}] + \mathbb{E}_Q[g(X)1_{\{\frac{dP}{dQ} > t\}}] \\ &\geq \frac{1}{t} \underbrace{\mathbb{E}_P[g(X)1_{\{\frac{dP}{dQ} \leq t\}}]}_{\geq \mathbb{E}_P[(1-g(X))1_{\{\frac{dP}{dQ} > t\}}]} + \mathbb{E}_Q[g(X)1_{\{\frac{dP}{dQ} > t\}}] \\ &\stackrel{(14.12)}{=} \frac{1}{t} \left(\underbrace{\mathbb{E}_P[(1-g(X))1_{\{\frac{dP}{dQ} > t\}}]}_{\geq \mathbb{E}_Q[(1-g(X))1_{\{\frac{dP}{dQ} > t\}}]} + \lambda P\left[\frac{dP}{dQ} = t\right] \right) + \mathbb{E}_Q[g(X)1_{\{\frac{dP}{dQ} > t\}}] \\ &\geq \mathbb{E}_Q[(1-g(X))1_{\{\frac{dP}{dQ} > t\}}] + \lambda Q\left[\frac{dP}{dQ} = t\right] + \mathbb{E}_Q[g(X)1_{\{\frac{dP}{dQ} > t\}}] \\ &= Q\left[\frac{dP}{dQ} > t\right] + \lambda Q\left[\frac{dP}{dQ} = t\right]. \end{aligned} \quad \square$$

238

Remark 14.3. As a consequence of the Neyman-Pearson lemma, all the points on the boundary of the region $\mathcal{R}(P, Q)$ are attainable. Therefore

$$\mathcal{R}(P, Q) = \{(\alpha, \beta) : \beta_\alpha \leq \beta \leq 1 - \beta_{1-\alpha}\}.$$

Since $\alpha \mapsto \beta_\alpha$ is convex on $[0, 1]$, hence continuous, the region $\mathcal{R}(P, Q)$ is a closed convex set, as previously stated in Theorem 14.3. Consequently, the infimum in the definition of β_α is in fact a minimum.

Furthermore, the lower half of the region $\mathcal{R}(P, Q)$ is the convex hull of the union of the following two sets:

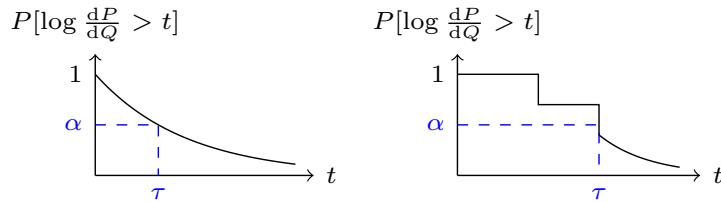
$$\begin{cases} \alpha = P[\log \frac{dP}{dQ} > \tau] \\ \beta = Q[\log \frac{dP}{dQ} > \tau] \end{cases} \quad \tau \in \mathbb{R} \cup \{\pm\infty\}.$$

and

$$\begin{cases} \alpha = P[\log \frac{dP}{dQ} \geq \tau] \\ \beta = Q[\log \frac{dP}{dQ} \geq \tau] \end{cases} \quad \tau \in \mathbb{R} \cup \{\pm\infty\}.$$

Therefore it does not lose optimality to restrict our attention on tests of the form $1\{\log \frac{dP}{dQ} \geq \tau\}$ or $1\{\log \frac{dP}{dQ} > \tau\}$. The convex combination (randomization) of the above two styles of tests lead to the achievability of the Neyman-Pearson lemma (Theorem 14.11).

Remark 14.4. The Neyman-Pearson test (14.11) is related to the LRT² as follows:



- Left figure: If $\alpha = P[\log \frac{dP}{dQ} > \tau]$ for some τ , then $\lambda = 0$, and (14.11) becomes the LRT $Z = 1\{\log \frac{dP}{dQ} \leq \tau\}$.
- Right figure: If $\alpha \neq P[\log \frac{dP}{dQ} > \tau]$ for any τ , then we have $\lambda \in (0, 1)$, and (14.11) is equivalent to randomize over tests: $Z = 1\{\log \frac{dP}{dQ} \leq \tau\}$ with probability $1 - \lambda$ or $1\{\log \frac{dP}{dQ} < \tau\}$ with probability λ .

Corollary 14.11. $\forall \tau \in \mathbb{R}$, there exists $(\alpha, \beta) \in \mathcal{R}(P, Q)$ s.t.

$$\alpha = P\left[\log \frac{dP}{dQ} > \tau\right]$$

² Note that it so happens that in Definition 14.5 the LRT is defined with an \leq instead of $<$.

14.5 Stein's regime 239

$$\beta \leq \exp(-\tau)P\left[\log \frac{dP}{dQ} > \tau\right] \leq \exp(-\tau)$$

Proof. For the case of discrete \mathcal{X} it is easy to give an explicit proof

$$\begin{aligned} Q\left[\log \frac{dP}{dQ} > \tau\right] &= \sum Q(x) \mathbb{1}\left\{\frac{P(x)}{Q(x)} > \exp(\tau)\right\} \\ &\leq \sum P(x) \exp(-\tau) \mathbb{1}\left\{\frac{P(x)}{Q(x)} > \exp(\tau)\right\} = \exp(-\tau)P\left[\log \frac{dP}{dQ} > \tau\right]. \end{aligned}$$

The general case is just an application of (14.7). \square

In the remainder of the chapter, we focus on the special case of iid observations in the large-sample asymptotics. Consider

$$\begin{aligned} H_0 : X_1, \dots, X_n &\stackrel{\text{i.i.d.}}{\sim} P \\ H_1 : X_1, \dots, X_n &\stackrel{\text{i.i.d.}}{\sim} Q, \end{aligned} \tag{14.14}$$

where P and Q do not depend on n ; this is a particular case of our general setting with P and Q replaced by their n -fold product distributions. We are interested in the asymptotics of the error probabilities $\pi_{0|1}$ and $\pi_{1|0}$ as $n \rightarrow \infty$ in the following two regimes:

- Stein regime: When $\pi_{1|0}$ is constrained to be at most ϵ , what is the best exponential rate of convergence for $\pi_{0|1}$?
- Chernoff regime: When both $\pi_{1|0}$ and $\pi_{0|1}$ are required to vanish exponentially, what is the optimal tradeoff between their exponents?

14.5 Stein's regime

Recall that we are in the iid setting (14.14) and are interested in tests satisfying $1 - \alpha = \pi_{1|0} \leq \epsilon$ and $\beta = \pi_{0|1} \leq \exp(-nE)$ for some exponent $E > 0$. Motivation of this asymmetric objective is that often a “missed detection” ($\pi_{0|1}$) is far more disastrous than a “false alarm” ($\pi_{1|0}$). For example, a false alarm could simply result in extra computations (attempting to decode a packet when there is in fact only noise has been received). The formal definition of the best exponent is as follows.

Definition 14.12. The ϵ -optimal exponent in Stein's regime is

$$V_\epsilon \triangleq \sup\{E : \exists n_0, \forall n \geq n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1 - \epsilon, \beta < \exp(-nE)\}.$$

and *Stein's exponent* is defined as $V \triangleq \lim_{\epsilon \rightarrow 0} V_\epsilon$.

It is an exercise to check the following equivalent definition

$$V_\epsilon = \liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\beta_{1-\epsilon}(P_{X^n}, Q_{X^n})}$$

where β_α is defined in (14.3).

Here is the main result of this section.

Theorem 14.13 (Stein's lemma). *Consider the iid setting (14.14) where $P_{X^n} = P^n$ and $Q_{X^n} = Q^n$. Then*

$$V_\epsilon = D(P\|Q), \quad \forall \epsilon \in (0, 1).$$

Consequently, $V = D(P\|Q)$.

The way to use this result in practice is the following. Suppose it is required that $\alpha \geq 0.999$, and $\beta \leq 10^{-40}$, what is the required sample size? Stein's lemma provides a rule of thumb: $n \geq -\frac{\log 10^{-40}}{D(P\|Q)}$.

Proof. We first assume that $P \ll Q$ so that $\frac{dP}{dQ}$ is well defined. Define the LLR

$$F_n = \log \frac{dP_{X^n}}{dQ_{X^n}} = \sum_{i=1}^n \log \frac{dP}{dQ}(X_i), \quad (14.15)$$

which is an iid sum under both hypotheses. As such, by WLLN, under P , as $n \rightarrow \infty$,

$$\frac{1}{n} F_n = \frac{1}{n} \sum_{i=1}^n \log \frac{dP}{dQ}(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = D(P\|Q). \quad (14.16)$$

Alternatively, under Q , we have

$$\frac{1}{n} F_n \xrightarrow{\mathbb{P}} \mathbb{E}_Q \left[\log \frac{dP}{dQ} \right] = -D(Q\|P). \quad (14.17)$$

Note that both convergence results hold even if the divergence is infinite.

(Achievability) We show that $V_\epsilon \geq D(P\|Q) \equiv D$ for any $\epsilon > 0$. First assume that $D < \infty$. Pick $\tau = n(D - \delta)$ for some small $\delta > 0$. Then Corollary 14.2 yields

$$\begin{aligned} \alpha &= P(F_n > n(D - \delta)) \rightarrow 1, \text{ by (14.16)} \\ \beta &\leq e^{-n(D-\delta)} \end{aligned}$$

then pick n large enough (depends on ϵ, δ) such that $\alpha \geq 1 - \epsilon$, we have the exponent $E = D - \delta$ achievable, $V_\epsilon \geq E$. Sending $\delta \rightarrow 0$ yields $V_\epsilon \geq D$. Finally, if $D = \infty$, the above argument holds for arbitrary $\tau > 0$, proving that $V_\epsilon = \infty$.

(Converse) We show that $V_\epsilon \leq D$ for any $\epsilon < 1$, to which end it suffices to consider $D < \infty$. As a warm-up, we first show a weak converse by applying Theorem 14.7 based on data processing inequality. For any $(\alpha, \beta) \in \mathcal{R}(P_{X^n}, Q_{X^n})$, we have

$$-h(\alpha) + \alpha \log \frac{1}{\beta} \leq d(\alpha\|\beta) \leq D(P_{X^n}\|Q_{X^n}) \quad (14.18)$$

14.5 Stein's regime 241

For any achievable exponent $E < V_\epsilon$, by definition, there exists a sequence of tests such that $\alpha_n \geq 1 - \epsilon$ and $\beta_n \leq \exp(-nE)$. Plugging this into (14.18) and using $h \leq \log 2$, we have $E \leq \frac{D(P\|Q)}{1-\epsilon} + \frac{\log 2}{n(1-\epsilon)}$. Sending $n \rightarrow \infty$ yields

$$V_\epsilon \leq \frac{D(P\|Q)}{1-\epsilon},$$

which is weaker than what we set out to prove; nevertheless, this weak converse is tight for $\epsilon \rightarrow 0$, so that for Stein's exponent we have succeeded in proving the desired result of $V = \lim_{\epsilon \rightarrow 0} V_\epsilon \geq D(P\|Q)$. So the question remains: if we allow the type-I error to be $\epsilon = 0.999$, is it possible for the type-II error to decay faster? This is shown impossible by the strong converse next.

To this end, note that, in proving the weak converse, we only made use of the *expectation* of F_n in (14.18), we need to make use of the *entire distribution* (CDF) in order to obtain better results. Applying the strong converse Theorem 14.9 to testing P_{X^n} versus Q_{X^n} and $\alpha = 1 - \epsilon$ and $\beta = \exp(-nE)$, we have

$$1 - \epsilon - \gamma \exp(-nE) \leq \alpha_n - \gamma \beta_n \leq P_{X^n}[F_n > \log \gamma].$$

Pick $\gamma = \exp(n(D + \delta))$ for $\delta > 0$, by WLLN (14.16) the probability on the right side goes to 0, which implies that for any fixed $\epsilon < 1$, we have $E \leq D + \delta$ and hence $V_\epsilon \leq D + \delta$. Sending $\delta \rightarrow 0$ complete the proof.

Finally, let us address the case of $P \not\ll Q$, in which case $D(P\|Q) = \infty$. By definition, there exists a subset A such that $Q(A) = 0$ but $P(A) > 0$. Consider the test that selects P if $X_i \in A$ for some $i \in [n]$. It is clear that this test achieves $\beta = 0$ and $1 - \alpha = (1 - P(A))^n$, which can be made less than any ϵ for large n . This shows $V_\epsilon = \infty$, as desired. \square

Remark 14.5 (Non-iid data). Just like in Chapter 12 on data compression, Theorem 14.15 can be extended to stationary ergodic processes:

$$V_\epsilon = \lim_{n \rightarrow \infty} \frac{1}{n} D(P_{X^n}\|Q_{X^n})$$

where $\{X_i\}$ is stationary and ergodic under both P and Q . Indeed, the counterpart of (14.16) based on WLLN, which is the key for choosing the appropriate threshold τ , for ergodic processes is the Birkhoff-Khintchine convergence theorem (cf. Theorem 12.7).

Remark 14.6. The theoretical importance of Stein's exponent is that:

$$\forall E \subset \mathcal{X}^n, \quad P_{X^n}[E] \geq 1 - \epsilon \Rightarrow Q_{X^n}[E] \geq \exp(-nV_\epsilon + o(n))$$

Thus knowledge of Stein's exponent V_ϵ allows one to prove exponential bounds on probabilities of arbitrary sets; this technique is known as “change of measure”, which will be applied in large deviations analysis in Chapter 15.

14.6 Chernoff regime: preview

We are still considering iid setting (14.14), namely, testing

$$H_0 : X^n \sim P^n \quad \text{versus} \quad H_1 : X^n \sim Q^n,$$

but the objective in the Chernoff regime is to achieve exponentially small error probability of both types simultaneously. We say a pair of exponents (E_0, E_1) is achievable if there exists a sequence of tests such that

$$\begin{aligned} 1 - \alpha &= \pi_{1|0} \leq \exp(-nE_0) \\ \beta &= \pi_{0|1} \leq \exp(-nE_1). \end{aligned}$$

Intuitively, one exponent can be made large at the expense of making the other small. So the interesting question is to find their optimal tradeoff by characterizing the achievable region of (E_0, E_1) . This problem was solved by [155, 38] and is the topic of Chapter 16. (See Fig. 16.2 for an illustration of the optimal (E_0, E_1) -tradeoff.)

Let us explain what we already know about the region of achievable pairs of exponents (E_0, E_1) .

First, Stein's regime corresponds to corner points of this achievable region. Indeed, Theorem 14.15 tells us that when fixing $\alpha_n = 1 - \epsilon$, namely $E_0 = 0$, picking $\tau = D(P\|Q) - \delta$ ($\delta \rightarrow 0$) gives the exponential convergence rate of β_n as $E_1 = D(P\|Q)$. Similarly, exchanging the role of P and Q , we can achieve the point $(E_0, E_1) = (D(Q\|P), 0)$.

Second, we have shown in Section 7.3 that the minimum total error probabilities over all tests satisfies

$$\min_{(\alpha, \beta) \in \mathcal{R}(P^n, Q^n)} 1 - \alpha + \beta = 1 - \text{TV}(P^n, Q^n).$$

As $n \rightarrow \infty$, P^n and Q^n becomes increasingly distinguishable and their total variation converges to 1 exponentially, with exponent E given by $\max \min(E_0, E_1)$ over all achievable pairs. From the bounds (7.20) and tensorization of the Hellinger distance (7.23), we obtain

$$1 - \sqrt{1 - \exp(-2nE_H)} \leq 1 - \text{TV}(P^n, Q^n) \leq \exp(-nE_H), \quad (14.19)$$

where we denoted

$$E_H \triangleq \log \left(1 - \frac{1}{2} H^2(P, Q) \right).$$

Thus, we can see that

$$E_H \leq E \leq 2E_H.$$

This characterization is valid even if P and Q depends on the sample size n which will prove useful later when we study *composite hypothesis testing* in Section 32.2.1. However, for fixed P and Q this is not precise enough. In order to determine the full set of achievable pairs, we need to make

14.6 Chernoff regime: preview 243

a detour into the topic of large deviations next. To see how this connection arises, notice that the (optimal) likelihood ratio tests give us explicit expressions for both error probabilities:

$$1 - \alpha_n = P \left[\frac{1}{n} F_n \leq \tau \right], \quad \beta_n = Q \left[\frac{1}{n} F_n > \tau \right]$$

where F_n is the LLR in (14.15). When τ falls in the range of $(-D(Q||P), D(P||Q))$, both probabilities are vanishing thanks to WLLN – see (14.16) and (14.17), and we are interested in their exponential convergence rate. This falls under the purview of large deviations theory.

15 Information projection and large deviations

In this chapter we discuss the following set of topics:

- 1 Basics of large deviation: log moment generating function (MGF) ψ_X and its conjugate (rate function) ψ_X^* , tilting.
- 2 Information projection problem:

$$\min_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q||P) = \psi^*(\gamma).$$

- 3 Use information projection to prove tight Chernoff bound: for iid copies X_1, \dots, X_n of X ,

$$\mathbb{P}\left[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma\right] = \exp(-n\psi^*(\gamma) + o(n)).$$

In the next chapter, we apply these results to characterize the achievable (E_0, E_1) -region (as defined in Section 14.6) to get

$$(E_0(\theta) = \psi_P^*(\theta), \quad E_1(\theta) = \psi_P^*(\theta) - \theta),$$

with ψ_P^* being the rate function of $\log \frac{dP}{dQ}$ (under P). This gives us a complete (parametric) description of the sought-after tradeoff between the two exponents.

15.1 Basics of large deviations theory

Let X_1, \dots, X_n be an iid sequence drawn from P and $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ their empirical distribution. The large deviation theory focuses on establishing sharp exponential estimates of the kind

$$\mathbb{P}[\hat{P}_n \in \mathcal{E}] = \exp\{-nE + o(n)\}.$$

The full account of such theory requires delicate consideration of topological properties of \mathcal{E} , and is the subject of classical treatments e.g. [86]. We focus here on a simple special case which, however, suffices for the purpose of establishing the Chernoff exponents in hypothesis testing, and also showcases all the relevant information-theoretic ideas. Our ultimate goal is to show the following result:

15.1 Basics of large deviations theory 245

Theorem 15.1. Consider a random variable X whose log MGF $\psi_X(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$ is finite for all $\lambda \in \mathbb{R}$. Let $B = \text{esssup } X$ and let $\mathbb{E}[X] < \gamma < B$. Then

$$P\left[\sum_{i=1}^n X_i \geq n\gamma\right] = \exp\{-nE(\gamma) + o(n)\},$$

where $E(\gamma) = \sup_{\lambda \geq 0} \lambda\gamma - \psi_X(\lambda) = \psi_X^*(\gamma)$, known as the rate function.

The concepts of log MGF and the rate function will be elaborated in subsequent sections. We provide the proof below that should be revisited after reading the rest of the chapter.

Proof. Let us recall the usual Chernoff bound: For iid X^n , for any $\lambda \geq 0$,

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right) \geq \exp(n\lambda\gamma)\right] \\ &\stackrel{\text{(Markov ineq.)}}{\leq} \exp(-n\lambda\gamma) \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] \\ &= \exp(-n\lambda\gamma + n\underbrace{\log \mathbb{E}[\exp(\lambda X)]}_{\psi_X(\lambda)}). \end{aligned}$$

Optimizing over $\lambda \geq 0$ gives the *non-asymptotic* upper bound (concentration inequality) which holds for any n :

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] \leq \exp\left\{-n \sup_{\lambda \geq 0} (\lambda\gamma - \psi_X(\lambda))\right\}. \quad (15.1)$$

This proves the upper bound part of Theorem 15.1. Our goal, thus, is to show the lower bound. This will be accomplished by first expressing $E(\gamma)$ as a certain KL-minimization problem (see Theorem 15.10), known as information projection, and then solving this problem (see (15.26)) to obtain the desired value of $E(\gamma)$. In the process of this proof we will also understand why the apparently naive Chernoff bound is in fact sharp. The explanation is that, essentially, inequality (15.1) performs a change of measure to a tilted distribution P_λ , which is the closest to P (in KL divergence) among all distributions Q with $\mathbb{E}_Q[X] \geq \gamma$. \square

15.1.1 Log MGF and rate function

Definition 15.2. The log moment-generating function (also known as the *cumulant generating function*) of a real-valued random variable X is

$$\psi_X(\lambda) = \log \mathbb{E}[\exp(\lambda X)], \quad \lambda \in \mathbb{R}.$$

Per convention in information theory, we will denote $\psi_P(\lambda) = \psi_X(\lambda)$ if $X \sim P$.

As an example, for a standard Gaussian $Z \sim \mathcal{N}(0, 1)$, we have $\psi_Z(\lambda) = \frac{\lambda^2}{2}$. Taking $X = Z^3$ yields a random variable such that $\psi_X(\lambda)$ is infinite for all non-zero λ .

In the remaining of the chapter, we shall assume that the MGF of random variable X is finite, namely $\psi_X(\lambda) < \infty$ for all $\lambda \in \mathbb{R}$. This, in particular, implies that all moments of X is finite.

Theorem 15.3 (Properties of ψ_X). (a) ψ_X is convex;
 (b) ψ_X is continuous;
 (c) ψ_X is infinitely differentiable and

$$\psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = e^{-\psi_X(\lambda)} \mathbb{E}[Xe^{\lambda X}].$$

In particular, $\psi_X(0) = 0, \psi'_X(0) = \mathbb{E}[X]$.

- (d) If $a \leq X \leq b$ a.s., then $a \leq \psi'_X \leq b$;
 (e) Conversely, if

$$A = \inf_{\lambda \in \mathbb{R}} \psi'_X(\lambda), \quad B = \sup_{\lambda \in \mathbb{R}} \psi'_X(\lambda),$$

then $A \leq X \leq B$ a.s.;

- (f) If X is not a constant, then ψ_X is strictly convex, and consequently, ψ'_X is strictly increasing.
 (g) Chernoff bound:

$$P(X \geq \gamma) \leq \exp(-\lambda\gamma + \psi_X(\lambda)), \quad \lambda \geq 0. \quad (15.2)$$

Remark 15.1. The slope of log MGF encodes the range of X . Indeed, Theorem 15.3(d) and (e) together show that the smallest closed interval containing the support of P_X equals (closure of) the range of ψ'_X . In other words, A and B coincide with the essential infimum and supremum (min and max of RV in the probabilistic sense) of X respectively,

$$A = \text{essinf } X \triangleq \sup\{a : X \geq a \text{ a.s.}\}$$

$$B = \text{esssup } X \triangleq \inf\{b : X \leq b \text{ a.s.}\}$$

See Fig. ?? for an illustration.

Proof. Note that (g) is already proved in (15.1). The proof of (e)–(f) relies on Theorem 15.9 and can be revisited later.

- (a) Fix $\theta \in (0, 1)$. Recall Hölder's inequality:

$$\mathbb{E}[|UV|] \leq \|U\|_p \|V\|_q, \quad \text{for } p, q \geq 1, \frac{1}{p} + \frac{1}{q} = 1$$

where the L_p -norm of a random variable U is defined by $\|U\|_p = (\mathbb{E}|U|^p)^{1/p}$. Applying to $\mathbb{E}[e^{(\theta\lambda_1 + \bar{\theta}\lambda_2)X}]$ with $p = 1/\theta, q = 1/\bar{\theta}$, we get

$$\mathbb{E}[\exp((\lambda_1/p + \lambda_2/q)X)] \leq \|\exp(\lambda_1 X/p)\|_p \|\exp(\lambda_2 X/q)\|_q = \mathbb{E}[\exp(\lambda_1 X)]^\theta \mathbb{E}[\exp(\lambda_2 X)]^{\bar{\theta}},$$

$$\text{i.e., } e^{\psi_X(\theta\lambda_1 + \bar{\theta}\lambda_2)} \leq e^{\psi_X(\lambda_1)\theta} e^{\psi_X(\lambda_2)\bar{\theta}}.$$

- (b) By our assumptions on X , the domain of ψ_X is \mathbb{R} . By the fact that a convex function must be continuous on the interior of its domain, we conclude that ψ_X is continuous on \mathbb{R} .

15.1 Basics of large deviations theory 247

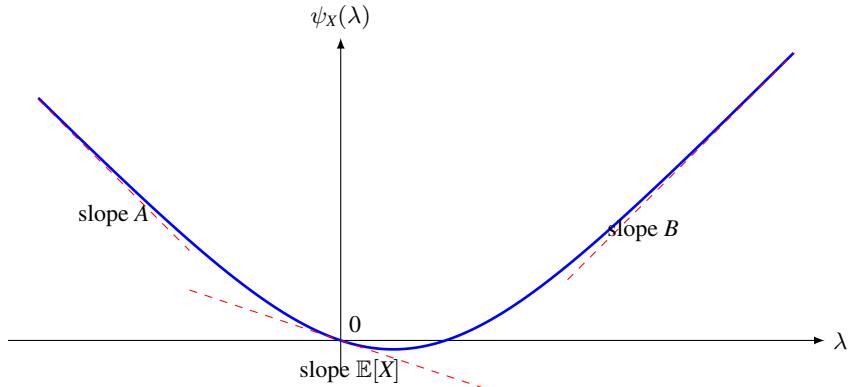


Figure 15.1 Example of a log MGF $\psi_X(\gamma)$ with P_X supported on $[A, B]$. The limiting maximal and minimal slope is A and B respectively. The slope at $\gamma = 0$ is $\psi'_X(0) = \mathbb{E}[X]$. Here we plot for $X = \pm 1$ with $\mathbb{P}[X = 1] = 1/3$.

- (c) The subtlety here is that we need to be careful when exchanging the order of differentiation and expectation.

Assume without loss of generality that $\lambda \geq 0$. First, we show that $\mathbb{E}[|Xe^{\lambda X}|]$ exists. Since

$$\begin{aligned} e^{|X|} &\leq e^X + e^{-X} \\ |Xe^{\lambda X}| &\leq e^{|\lambda+1|X} \leq e^{(\lambda+1)X} + e^{-(\lambda+1)X} \end{aligned}$$

by assumption on X , both of the summands are absolutely integrable in X . Therefore by the dominated convergence theorem, $\mathbb{E}[|Xe^{\lambda X}|]$ exists and is continuous in λ .

Second, by the existence and continuity of $\mathbb{E}[|Xe^{\lambda X}|]$, $u \mapsto \mathbb{E}[|Xe^{uX}|]$ is integrable on $[0, \lambda]$, we can switch order of integration and differentiation as follows:

$$\begin{aligned} e^{\psi_X(\lambda)} &= \mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[1 + \int_0^\lambda Xe^{uX} du\right] \stackrel{\text{Fubini}}{=} 1 + \int_0^\lambda \mathbb{E}[Xe^{uX}] du \\ \Rightarrow \psi'_X(\lambda)e^{\psi_X(\lambda)} &= \mathbb{E}[Xe^{\lambda X}] \end{aligned}$$

thus $\psi'_X(\lambda) = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}]$ exists and is continuous in λ on \mathbb{R} .

Furthermore, using similar application of the dominated convergence theorem we can extend to $\lambda \in \mathbb{C}$ and show that $\lambda \mapsto \mathbb{E}[e^{\lambda X}]$ is a holomorphic function. Thus it is infinitely differentiable.

- (d) $A \leq X \leq B \Rightarrow \psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \in [A, B]$.
- (e) Suppose (for contradiction) that $P_X[X > B] > 0$. Then $P_X[X > B + 2\epsilon] > 0$ for some small $\epsilon > 0$. But then $P_\lambda[X \leq B + \epsilon] \rightarrow 0$ for $\lambda \rightarrow \infty$ (see Theorem 15.9.3 below). On the other hand, we know from Theorem 15.9.2 that $\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) \leq B$. This is not yet a contradiction, since P_λ might still have some very small mass at a very negative value. To show that this cannot happen, we first assume that $B - \epsilon > 0$ (otherwise just replace X with $X - 2B$). Next note that

$$B \geq \mathbb{E}_{P_\lambda}[X] = \mathbb{E}_{P_\lambda}[X1_{\{X < B-\epsilon\}}] + \mathbb{E}_{P_\lambda}[X1_{\{B-\epsilon \leq X \leq B+\epsilon\}}] + \mathbb{E}_{P_\lambda}[X1_{\{X > B+\epsilon\}}]$$

$$\begin{aligned} &\geq \mathbb{E}_{P_\lambda}[X \mathbf{1}_{\{X < B - \epsilon\}}] + \mathbb{E}_{P_\lambda}[X \mathbf{1}_{\{X > B + \epsilon\}}] \\ &\geq -\mathbb{E}_{P_\lambda}[|X| \mathbf{1}_{\{X < B - \epsilon\}}] + (B + \epsilon) \underbrace{P_\lambda[X > B + \epsilon]}_{\rightarrow 1} \end{aligned} \quad (15.3)$$

therefore we will obtain a contradiction if we can show that $\mathbb{E}_{P_\lambda}[|X| \mathbf{1}_{\{X < B - \epsilon\}}] \rightarrow 0$ as $\lambda \rightarrow \infty$. To that end, notice that convexity of ψ_X implies that $\psi'_X \nearrow B$. Thus, for all $\lambda \geq \lambda_0$ we have $\psi'_X(\lambda) \geq B - \frac{\epsilon}{2}$. Thus, we have for all $\lambda \geq \lambda_0$

$$\psi_X(\lambda) \geq \psi_X(\lambda_0) + (\lambda - \lambda_0)(B - \frac{\epsilon}{2}) = c + \lambda(B - \frac{\epsilon}{2}), \quad (15.4)$$

for some constant c . Then,

$$\begin{aligned} \mathbb{E}_{P_\lambda}[|X| \mathbf{1}_{\{X < B - \epsilon\}}] &= \mathbb{E}[|X| e^{\lambda X - \psi_X(\lambda)} \mathbf{1}_{\{X < B - \epsilon\}}] \\ &\leq \mathbb{E}[|X| e^{\lambda X - c - \lambda(B - \frac{\epsilon}{2})} \mathbf{1}_{\{X < B - \epsilon\}}] \\ &\leq \mathbb{E}[|X| e^{\lambda(B - \epsilon) - c - \lambda(B - \frac{\epsilon}{2})}] \\ &= \mathbb{E}[|X|] e^{-\lambda \frac{\epsilon}{2} - c} \rightarrow 0 \quad \lambda \rightarrow \infty \end{aligned}$$

where the first inequality is from (15.4) and the second from $X < B - \epsilon$. Thus, the first term in (15.3) goes to 0 implying the desired contradiction.

- (f) Suppose ψ_X is not strictly convex. Since ψ_X is convex from part (f), ψ_X must be “flat” (affine) near some point. That is, there exists a small neighborhood of some λ_0 such that $\psi_X(\lambda_0 + u) = \psi_X(\lambda_0) + ur$ for some $r \in \mathbb{R}$. Then $\psi_{P_\lambda}(u) = ur$ for all u in small neighborhood of zero, or equivalently $\mathbb{E}_{P_\lambda}[e^{u(X-r)}] = 1$ for u small. The following Lemma 15.5 implies $P_\lambda[X = r] = 1$, but then $P[X = r] = 1$, contradicting the assumption $X \neq \text{const}$.

□

Lemma 15.4. $\mathbb{E}[e^{uS}] = 1$ for all $u \in (-\epsilon, \epsilon)$ then $S = 0$.

Proof. Expand in Taylor series around $u = 0$ to obtain $\mathbb{E}[S] = 0$, $\mathbb{E}[S^2] = 0$. Alternatively, we can extend the argument we gave for differentiating $\psi_X(\lambda)$ to show that the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic on the entire complex plane¹. Thus by uniqueness, $\mathbb{E}[e^{uS}] = 1$ for all u . □

Definition 15.5 (Rate function). The rate function $\psi_X^* : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ is given by the *Legendre-Fenchel transform* of the log MGF:

$$\psi_X^*(\gamma) = \sup_{\lambda \in \mathbb{R}} \lambda \gamma - \psi_X(\lambda) \quad (15.5)$$

Note that the maximization (15.5) is a convex optimization problem since ψ_X is strictly convex, so we can find the maximum by taking the derivative and finding the stationary point. In fact, ψ_X^* is the precisely the convex conjugate of ψ_X ; cf. (7.78).

¹ More precisely, if we only know that $\mathbb{E}[e^{\lambda S}]$ is finite for $|\lambda| \leq 1$ then the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic in the vertical strip $\{z : |\operatorname{Re} z| < 1\}$.

15.1 Basics of large deviations theory 249

The next result describes useful properties of the rate function. See Fig. ?? for an illustration.

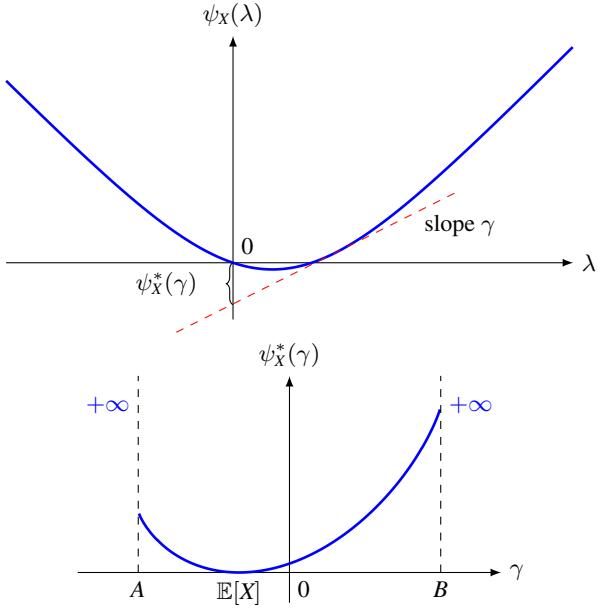


Figure 15.2 Log-MGF ψ_X and its conjugate (rate function) ψ_X^* for X taking values in $[A, B]$, continuing the example in Fig. ??.

Theorem 15.6 (Properties of ψ_X^*). *Assume that X is non-constant.*

(a) *Let $A = \text{essinf } X$ and $B = \text{esssup } X$. Then*

$$\psi_X^*(\gamma) = \begin{cases} \lambda\gamma - \psi_X(\lambda) \text{ for } \lambda \text{ s.t. } \gamma = \psi'_X(\lambda), & A < \gamma < B \\ \log \frac{1}{P(X=\gamma)} & \gamma = A \text{ or } B \\ +\infty, & \gamma < A \text{ or } \gamma > B \end{cases}$$

(b) ψ_X^* is strictly convex and strictly positive except $\psi_X^*(\mathbb{E}[X]) = 0$.

(c) ψ_X^* is decreasing when $\gamma \in (A, \mathbb{E}[X])$, and increasing when $\gamma \in [\mathbb{E}[X], B)$

Proof. By Theorem 15.3(d), since $A \leq X \leq B$ a.s., we have $A \leq \psi'_X \leq B$. When $\gamma \in (A, B)$, the strictly concave function $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ has a single stationary point which achieves the unique maximum. When $\gamma > B$ (resp. $< A$), $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ increases (resp. decreases) without bounds. When $\gamma = B$, since $X \leq B$ a.s., we have

$$\begin{aligned} \psi_X^*(B) &= \sup_{\lambda \in \mathbb{R}} \lambda B - \log(\mathbb{E}[\exp(\lambda X)]) = -\log \inf_{\lambda \in \mathbb{R}} \mathbb{E}[\exp(\lambda(X-B))] \\ &= -\log \lim_{\lambda \rightarrow \infty} \mathbb{E}[\exp(\lambda(X-B))] = -\log P(X=B), \end{aligned}$$

by the monotone convergence theorem.

By Theorem 15.3(f), since ψ_X is strictly convex, the derivative of ψ_X and ψ_X^* are inverse to each other. Hence ψ_X^* is strictly convex. Since $\psi_X(0) = 0$, we have $\psi_X^*(\gamma) \geq 0$. Moreover, $\psi_X^*(\mathbb{E}[X]) = 0$ follows from $\mathbb{E}[X] = \psi_X'(0)$. \square

15.1.2 Tilted distribution

As early as in Chapter 4, we have already introduced the concept of *tilting* in the proof of Donsker-Varadhan's variational characterization of divergence (Theorem 4.8). Let us formally define it now.

Definition 15.7 (Tilting). Given $X \sim P$ and $\lambda \in \mathbb{R}$, the tilted measure P_λ is defined by

$$P_\lambda(dx) = \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]} P(dx) = e^{\lambda x - \psi_X(\lambda)} P(dx) \quad (15.6)$$

In particular, if P has a PDF p , then the PDF of P_λ is given by $p_\lambda(x) = e^{\lambda x - \psi_X(\lambda)} p(x)$.

The set of distributions $\{P_\lambda : \lambda \in \mathbb{R}\}$ parametrized by λ is called a *standard (one-parameter) exponential family*, an important object in statistics [53]. Here are some of the examples:

- *Gaussian*: $P = \mathcal{N}(0, 1)$ with density $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$. Then P_λ has density $\frac{\exp(\lambda x)}{\exp(\lambda^2/2)} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) = \frac{1}{\sqrt{2\pi}} \exp(-(x - \lambda)^2/2)$. Hence $P_\lambda = \mathcal{N}(\lambda, 1)$.
- *Bernoulli*: $P = \text{Ber}(\frac{1}{2})$. Then $P_\lambda = \text{Ber}\left(\frac{e^\lambda}{e^\lambda + 1}\right)$ which puts more (resp. less) mass on 1 if $\lambda > 0$ (resp. < 0). Moreover, $P_\lambda \xrightarrow{d} \delta_1$ if $\lambda \rightarrow \infty$ or δ_0 if $\lambda \rightarrow -\infty$.
- *Uniform*: Let P be the uniform distribution on $[0, 1]$. Then P_λ is also supported on $[0, 1]$ with pdf $p_\lambda(x) = \frac{\lambda \exp(\lambda x)}{e^\lambda - 1}$. Therefore as λ increases, P_λ becomes increasingly concentrated near 1, and $P_\lambda \rightarrow \delta_1$ as $\lambda \rightarrow \infty$. Similarly, $P_\lambda \rightarrow \delta_0$ as $\lambda \rightarrow -\infty$.

In the above examples we see that P_λ shifts the mean of P to the right (resp. left) when $\lambda > 0$ (resp. < 0). Indeed, this is a general property of tilting.

Theorem 15.8 (Properties of P_λ).

(a) *Log MGF*:

$$\psi_{P_\lambda}(u) = \psi_X(\lambda + u) - \psi_X(\lambda)$$

(b) *Tilting trades mean for divergence*:

$$\mathbb{E}_{P_\lambda}[X] = \psi_X'(\lambda) \geq \mathbb{E}_P[X] \text{ if } \lambda \geq 0. \quad (15.7)$$

$$D(P_\lambda \| P) = \psi_X^*(\psi_X'(\lambda)) = \psi_X^*(\mathbb{E}_{P_\lambda}[X]). \quad (15.8)$$

15.2 Large-deviations exponents and KL divergence 251

(c)

$$P(X > b) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \leq b - \epsilon) \rightarrow 0 \text{ as } \lambda \rightarrow \infty$$

$$P(X < a) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \geq a + \epsilon) \rightarrow 0 \text{ as } \lambda \rightarrow -\infty$$

Therefore if $X_\lambda \sim P_\lambda$, then $X_\lambda \xrightarrow{d} \text{essinf } X = A$ as $\lambda \rightarrow -\infty$ and $X_\lambda \xrightarrow{d} \text{esssup } X = B$ as $\lambda \rightarrow \infty$.

Proof. (a) By definition.

(b) $\mathbb{E}_{P_\lambda}[X] = \frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} = \psi'_X(\lambda)$, which is strictly increasing in λ , with $\psi'_X(0) = \mathbb{E}_P[X]$.

$D(P_\lambda \| P) = \mathbb{E}_{P_\lambda} \log \frac{dP_\lambda}{dP} = \mathbb{E}_{P_\lambda} \log \frac{\exp(\lambda X)}{\mathbb{E}[\exp(\lambda X)]} = \lambda \mathbb{E}_{P_\lambda}[X] - \psi_X(\lambda) = \lambda \psi'_X(\lambda) - \psi_X(\lambda) = \psi_X^*(\psi'_X(\lambda))$, where the last equality follows from Theorem 15.7(a).

(c)

$$\begin{aligned} P_\lambda(X \leq b - \epsilon) &= \mathbb{E}_P[e^{\lambda X - \psi_X(\lambda)} \mathbf{1}_{\{X \leq b - \epsilon\}}] \\ &\leq \mathbb{E}_P[e^{\lambda(b - \epsilon) - \psi_X(\lambda)} \mathbf{1}_{\{X \leq b - \epsilon\}}] \\ &\leq e^{-\lambda\epsilon} e^{\lambda b - \psi_X(\lambda)} \\ &\leq \frac{e^{-\lambda\epsilon}}{P[X > b]} \rightarrow 0 \text{ as } \lambda \rightarrow \infty \end{aligned}$$

where the last inequality is due to the usual Chernoff bound (Theorem 15.3(g)): $P[X > b] \leq \exp(-\lambda b + \psi_X(\lambda))$. □

15.2 Large-deviations exponents and KL divergence

Large deviations problems deal with rare events by making statements about the tail probabilities of a sequence of distributions. Here, we are interested in the following special case: the speed of decay for $P\left[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma\right]$ for iid X_k .

In (15.1) we have used Chernoff bound to obtain an upper bound on the exponent via the log-MGF. Here we use a different method to give a formula for the exponent as a convex optimization problem involving the KL divergence. In the subsequent chapter (information projection). Later in Section 15.4 we shall revisit the Chernoff bound after we have computed the value of the information projection.

Theorem 15.9. Let $X_1, X_2, \dots \stackrel{i.i.d.}{\sim} P$. Then for any $\gamma \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n} \sum_{k=1}^n X_k > \gamma\right]} = \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P) \quad (15.9)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma\right]} = \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) \quad (15.10)$$

Furthermore, for every n we have the firm upper bound

$$P \left[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma \right] \leq \exp \left\{ -n \cdot \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q||P) \right\} \quad (15.11)$$

and similarly for $>$ in place of \geq .

Remark 15.2 (Subadditivity). It is possible to argue from first principles that the limits (15.9) and (15.10) exist. Indeed, note that the sequence $p_n \triangleq \log \frac{1}{P[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma]}$ satisfies $p_{n+m} \geq p_n p_m$ and hence $\log \frac{1}{p_n}$ is subadditive. As such, $\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{p_n} = \inf_n \log \frac{1}{p_n}$ by Fekete's lemma.

Proof. First note that if the events have zero probability, then both sides coincide with infinity. Indeed, if $P[\frac{1}{n} \sum_{k=1}^n X_k > \gamma] = 0$, then $P[X > \gamma] = 0$. Then $\mathbb{E}_Q[X] > \gamma \Rightarrow Q[X > \gamma] > 0 \Rightarrow Q \ll P \Rightarrow D(Q||P) = \infty$ and hence (15.9) holds trivially. The case for (15.10) is similar.

In the sequel we assume both probabilities are nonzero. We start by proving (15.9). Set $P[E_n] = P[\frac{1}{n} \sum_{k=1}^n X_k > \gamma]$.

Lower Bound on $P[E_n]$: Fix a Q such that $\mathbb{E}_Q[X] > \gamma$. Let X^n be iid. Then by WLLN,

$$Q[E_n] = Q \left[\sum_{k=1}^n X_k > n\gamma \right] \xrightarrow{\text{LLN}} 1 - o(1).$$

Now the data processing inequality (Corollary 2.2) gives

$$d(Q[E_n]||P[E_n]) \leq D(Q_{X^n}||P_{X^n}) = nD(Q||P)$$

And a lower bound for the binary divergence is

$$d(Q[E_n]||P[E_n]) \geq -h(Q[E_n]) + Q[E_n] \log \frac{1}{P[E_n]}$$

Combining the two bounds on $d(Q[E_n]||P[E_n])$ gives

$$P[E_n] \geq \exp \left(\frac{-nD(Q||P) - \log 2}{Q[E_n]} \right) \quad (15.12)$$

Optimizing over Q to give the best bound:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P[E_n]} \leq \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q||P).$$

Upper Bound on $P[E_n]$: The key observation is that given any X and any event E , $P_X(E) > 0$ can be expressed via the divergence between the conditional and unconditional distribution as: $\log \frac{1}{P_X(E)} = D(P_{X|X \in E}||P_X)$. Define $\tilde{P}_{X^n} = P_{X^n | \sum X_i > n\gamma}$, under which $\sum X_i > n\gamma$ holds a.s. Then

$$\log \frac{1}{P[E_n]} = D(\tilde{P}_{X^n}||P_{X^n}) \geq \inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] > n\gamma} D(Q_{X^n}||P_{X^n}) \quad (15.13)$$

15.2 Large-deviations exponents and KL divergence 253

We now show that the last problem ‘‘single-letterizes’’, i.e., reduces $n = 1$. Note that this is a special case of a more general phenomena – see Ex. III.7. Consider the following two steps:

$$\begin{aligned} D(Q_{X^n} \| P_{X^n}) &\geq \sum_{j=1}^n D(Q_{X_j} \| P) \\ &\geq nD(\bar{Q} \| P), \quad \bar{Q} \triangleq \frac{1}{n} \sum_{j=1}^n Q_{X_j}, \end{aligned} \quad (15.14)$$

where the first step follows from (2.26) in Theorem 2.15, after noticing that $P_{X^n} = P^n$, and the second step is by convexity of divergence (Theorem 5.1). From this argument we conclude that

$$\inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] > n\gamma} D(Q_{X^n} \| P_{X^n}) = n \cdot \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P) \quad (15.15)$$

$$\inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] \geq n\gamma} D(Q_{X^n} \| P_{X^n}) = n \cdot \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) \quad (15.16)$$

In particular, (15.12) and (15.15) imply the required lower bound in (15.9).

Next we prove (15.10). First, notice that the lower bound argument (15.12) applies equally well, so that for each n we have

$$\frac{1}{n} \log \frac{1}{P[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma]} \geq \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P).$$

To get a matching upper bound we consider two cases:

- Case I: $P[X > \gamma] = 0$. If $P[X \geq \gamma] = 0$, then both sides of (15.10) are $+\infty$. If $P[X = \gamma] > 0$, then $P[\sum X_k \geq n\gamma] = P[X_1 = \dots = X_n = \gamma] = P[X = \gamma]^n$. For the right-hand side, since $D(Q \| P) < \infty \implies Q \ll P \implies Q(X \leq \gamma) = 1$, the only possibility for $\mathbb{E}_Q[X] \geq \gamma$ is that $Q(X = \gamma) = 1$, i.e., $Q = \delta_\gamma$. Then $\inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) = \log \frac{1}{P(X=\gamma)}$.
- Case II: $P[X > \gamma] > 0$. Since $\mathbb{P}[\sum X_k \geq \gamma] \geq \mathbb{P}[\sum X_k > \gamma]$ from (15.9) we know that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P[\frac{1}{n} \sum_{k=1}^n X_k \geq \gamma]} \leq \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P).$$

We next show that in this case

$$\inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P) = \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) \quad (15.17)$$

Indeed, let $\tilde{P} = P_{X|X>\gamma}$ which is well defined since $P[X > \gamma] > 0$. For any Q such that $\mathbb{E}_Q[X] \geq \gamma$, set $\tilde{Q} = \bar{Q} + \epsilon \tilde{P}$ satisfies $\mathbb{E}_{\tilde{Q}}[X] > \gamma$. Then by convexity, $D(\tilde{Q} \| P) \leq \bar{Q} D(Q \| P) + \epsilon D(\tilde{P} \| P) = \bar{Q} D(Q \| P) + \epsilon \log \frac{1}{P[X>\gamma]}$. Sending $\epsilon \rightarrow 0$, we conclude the proof of (15.17). \square

Remark 15.3. Note that the upper bound (??) also holds for independent *non-identically* distributed X_i . Indeed, we only need to replace the step (??) with $D(Q_{X^n} \| P_{X^n}) \geq \sum_{i=1}^n D(Q_{X_i} \| P_{X_i}) \geq nD(Q \| \bar{P})$ where $\bar{P} = \frac{1}{n} \sum_{i=1}^n P_{X_i}$. This yields a bound (??) with P replaced by \bar{P} in the right-hand side.

Example 15.1 (Poisson-Binomial tails). Consider X which is a sum of n independent Bernoulli random variables so that $\mathbb{E}[X] = np$. The distribution of X is known as *Poisson-Binomial* [? ?], including $\text{Bin}(n, p)$ as a special case. Applying Theorem 15.10 (or the Remark ??), we get the following tail bounds on X :

$$\mathbb{P}[X \geq k] \leq \exp\{-nd(k/n\|p)\}, \quad \frac{k}{n} > p \quad (15.18)$$

$$\mathbb{P}[X \leq k] \leq \exp\{-nd(k/n\|p)\}, \quad \frac{k}{n} < p \quad (15.19)$$

where for (15.18) we used the fact that $\min_{Q: \mathbb{E}_Q[X] \geq k/n} D(Q\|\text{Ber}(p)) = \min_{q \geq k/n} d(q\|p) = d(\frac{k}{n}\|p)$ and similarly for (15.19). These bounds, in turn, can be used to derive various famous estimates:

- Multiplicative deviation from the mean (Bennett's inequality): We have

$$\begin{aligned} \mathbb{P}[X \geq u \mathbb{E}[X]] &\leq \exp\{-\mathbb{E}[X]f(u)\} & \forall u > 1, \\ \mathbb{P}[X \leq u \mathbb{E}[X]] &\leq \exp\{-\mathbb{E}[X]f(u)\} & \forall 0 \leq u < 1, \end{aligned}$$

where $f(u) \triangleq u \log u - (u - 1) \log e \geq 0$. These follow from (15.18)-(15.19) via the following useful estimate:

$$d(up\|p) \geq pf(u) \quad \forall p \in [0, 1], u \in [0, 1/p] \quad (15.20)$$

Indeed, consider the elementary inequality

$$x \log \frac{x}{y} \geq (x - y) \log e$$

for all $x, y \in [0, 1]$ (since the difference between the left and right side is minimized over y at $y = x$). Using $x = 1 - up$ and $y = 1 - p$ establishes (15.20).

- Bernstein's inequality:

$$\mathbb{P}[X > np + t] \leq e^{-\frac{t^2}{2(np+t)}} \quad \forall t > 0.$$

This follows from the previous bound for $u > 1$ by bounding $\frac{f(u)}{\log e} = \int_1^u \frac{u-x}{x} dx \geq \frac{1}{u} \int_1^u (u - x) dx = \frac{(u-1)^2}{2u}$.

- Okamoto's inequality: For all $0 < p < 1$ and $t > 0$,

$$\mathbb{P}[\sqrt{X} - \sqrt{np} \geq t] \leq e^{-t^2}, \quad (15.21)$$

$$\mathbb{P}[\sqrt{X} - \sqrt{np} \leq -t] \leq e^{-t^2}. \quad (15.22)$$

These simply follow from the inequality between KL divergence and Hellinger distance in (7.30). Indeed, we get $d(x\|p) \geq H^2(\text{Ber}(x), \text{Ber}(p)) \geq (\sqrt{x} - \sqrt{p})^2$. Plugging $x = \frac{(\sqrt{np}+t)^2}{n}$ into (15.18)-(15.19) we obtain the result. We note that [221, Theorem 3] shows a stronger bound of e^{-2t^2} in (15.21).

Remarkably, the bounds in (15.21) and (15.22) do not depend on n or p . This is due to the variance-stabilizing effect of the square-root transformation for binomials: $\text{Var}(\sqrt{X})$ is at most a constant for all n, p . In addition, $\sqrt{X} - \sqrt{np} = \frac{X-np}{\sqrt{X}+\sqrt{np}}$ is of a self-normalizing form: the

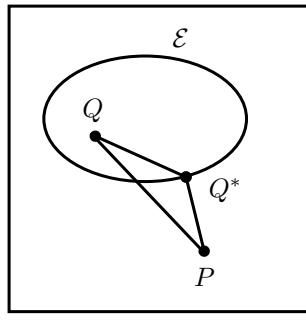
denominator is on par with the standard deviation of the numerator. For more on self-normalized sums, see [44, Problem 12.2].

15.3 Information Projection

The results of Theorem 15.10 motivate us to study the following general **information projection problem**: Let \mathcal{E} be a convex set of distributions on some abstract space Ω , then for the distribution P on Ω , we want

$$\inf_{Q \in \mathcal{E}} D(Q\|P)$$

Denote the minimizing distribution Q by Q^* . The next result shows that intuitively the “line” between P and optimal Q^* is “orthogonal” to \mathcal{E} .



Distributions on \mathcal{X}

Theorem 15.10. Suppose $\exists Q^* \in \mathcal{E}$ such that $D(Q^*\|P) = \min_{Q \in \mathcal{E}} D(Q\|P)$, then $\forall Q \in \mathcal{E}$

$$D(Q\|P) \geq D(Q\|Q^*) + D(Q^*\|P)$$

Proof. If $D(Q\|P) = \infty$, then there is nothing to prove. So we assume that $D(Q\|P) < \infty$, which also implies that $D(Q^*\|P) < \infty$. For $\lambda \in [0, 1]$, form the convex combination $Q^{(\lambda)} = \bar{\lambda}Q^* + \lambda Q \in \mathcal{E}$. Since Q^* is the minimizer of $D(Q\|P)$, then

$$0 \leq \frac{d}{d\lambda} \Big|_{\lambda=0} D(Q^{(\lambda)}\|P) = D(Q\|P) - D(Q\|Q^*) - D(Q^*\|P)$$

The rigorous analysis requires an argument for interchanging derivatives and integrals (via dominated convergence theorem) and is similar to the proof of Proposition 2.17. The details are in [82, Theorem 2.2]. \square

Remark 15.4. If we view the picture above in the Euclidean setting, the “triangle” formed by P , Q^* and Q (for Q^* , Q in a convex set, P outside the set) is always obtuse, and is a right triangle only when the convex set has a “flat face”. In this sense, the divergence is similar to the squared Euclidean distance, and the above theorem is sometimes known as a “Pythagorean” theorem.

The relevant set \mathcal{E} of Q 's that we will focus next is the “half-space” of distributions $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$, where $X : \Omega \rightarrow \mathbb{R}$ is some fixed function (random variable). This is justified by relation with the large-deviations exponent in Theorem 15.10. First, we solve this I-projection problem explicitly.

Theorem 15.11. *Given a distribution P on Ω and $X : \Omega \rightarrow \mathbb{R}$ let*

$$A = \inf \psi'_X = \text{essinf } X = \sup\{a : X \geq a \text{ } P\text{-a.s.}\} \quad (15.23)$$

$$B = \sup \psi'_X = \text{esssup } X = \inf\{b : X \leq b \text{ } P\text{-a.s.}\} \quad (15.24)$$

1 *The information projection problem over $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$ has solution*

$$\min_{Q : \mathbb{E}_Q[X] \geq \gamma} D(Q||P) = \begin{cases} 0 & \gamma < \mathbb{E}_P[X] \\ \psi_P^*(\gamma) & \mathbb{E}_P[X] \leq \gamma < B \\ \log \frac{1}{P(X=B)} & \gamma = B \\ +\infty & \gamma > B \end{cases} \quad (15.25)$$

$$= \psi_P^*(\gamma) \mathbf{1}\{\gamma \geq \mathbb{E}_P[X]\} \quad (15.26)$$

2 *Whenever the minimum is finite, the minimizing distribution is unique and equal to tilting of P along X , namely²*

$$dP_\lambda = \exp\{\lambda X - \psi(\lambda)\} \cdot dP \quad (15.27)$$

3 *For all $\gamma \in [\mathbb{E}_P[X], B]$ we have*

$$\min_{\mathbb{E}_Q[X] \geq \gamma} D(Q||P) = \inf_{\mathbb{E}_Q[X] > \gamma} D(Q||P) = \min_{\mathbb{E}_Q[X] = \gamma} D(Q||P).$$

Remark 15.5. Both Theorem 15.10 and Theorem 15.14 are stated for the right tail where the sample mean exceeds the population mean. For the left tail, simply these results to $-X_i$ to obtain for $\gamma < \mathbb{E}[X]$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n} \sum_{k=1}^n X_k < \gamma\right]} = \inf_{Q : \mathbb{E}_Q[X] < \gamma} D(Q||P) = \psi_X^*(\gamma).$$

In other words, the large deviation exponent is still given by the rate function (15.5) except that the optimal tilting parameter λ is negative.

Proof. We first prove (15.25).

- First case: Take $Q = P$.
- Fourth case: If $\mathbb{E}_Q[X] > B$, then $Q[X \geq B + \epsilon] > 0$ for some $\epsilon > 0$, but $P[X \geq B + \epsilon] = 0$, since $P(X \leq B) = 1$, by Theorem 15.3(e). Hence $Q \not\ll P \implies D(Q||P) = \infty$.

² Note that unlike the setting of Theorems 15.1 and 15.10 here P and P_λ are measures on an abstract space Ω , not necessarily on the real line.

15.3 Information Projection 257

- Third case: If $P(X = B) = 0$, then $X < B$ a.s. under P , and $Q \not\ll P$ for any Q s.t. $\mathbb{E}_Q[X] \geq B$. Then the minimum is ∞ . Now assume $P(X = B) > 0$. Since $D(Q||P) < \infty \implies Q \ll P \implies Q(X \leq B) = 1$. Therefore the only possibility for $\mathbb{E}_Q[X] \geq B$ is that $Q(X = B) = 1$, i.e., $Q = \delta_B$. Then $D(Q||P) = \log \frac{1}{P(X=B)}$.
- Second case: Fix $\mathbb{E}_P[X] \leq \gamma < B$, and find the unique λ such that $\psi'_X(\lambda) = \gamma = \mathbb{E}_{P_\lambda}[X]$ where $dP_\lambda = \exp(\lambda X - \psi_X(\lambda))dP$. This corresponds to tilting P far enough to the right to increase its mean from $\mathbb{E}_P[X]$ to γ , in particular $\lambda \geq 0$. Moreover, $\psi_X^*(\gamma) = \lambda\gamma - \psi_X(\lambda)$. Take any Q such that $\mathbb{E}_Q[X] \geq \gamma$, then

$$D(Q||P) = \mathbb{E}_Q \left[\log \frac{dQdP_\lambda}{dPdP_\lambda} \right] \quad (15.28)$$

$$\begin{aligned} &= D(Q||P_\lambda) + \mathbb{E}_Q[\log \frac{dP_\lambda}{dP}] \\ &= D(Q||P_\lambda) + \mathbb{E}_Q[\lambda X - \psi_X(\lambda)] \\ &\geq D(Q||P_\lambda) + \lambda\gamma - \psi_X(\lambda) \\ &= D(Q||P_\lambda) + \psi_X^*(\gamma) \\ &\geq \psi_X^*(\gamma), \end{aligned} \quad (15.29)$$

where the last inequality holds with equality if and only if $Q = P_\lambda$. In addition, this shows the minimizer is unique, proving the second claim. Note that even in the corner case of $\gamma = B$ (assuming $P(X = B) > 0$) the minimizer is a point mass $Q = \delta_B$, which is also a tilted measure (P_∞), since $P_\lambda \rightarrow \delta_B$ as $\lambda \rightarrow \infty$, cf. Theorem 15.9(c).

An alternative version of the solution, given by expression (15.26), follows from Theorem 15.7. For the third claim, notice that there is nothing to prove for $\gamma < \mathbb{E}_P[X]$, while for $\gamma \geq \mathbb{E}_P[X]$ we have just shown

$$\psi_X^*(\gamma) = \min_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q||P)$$

while from the next corollary we have

$$\inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q||P) = \inf_{\gamma' > \gamma} \psi_X^*(\gamma').$$

The final step is to notice that ψ_X^* is increasing and continuous by Theorem 15.7, and hence the right-hand side infimum equals $\psi_X^*(\gamma)$. The case of $\min_{Q: \mathbb{E}_Q[X] = \gamma}$ is handled similarly. \square

Corollary 15.12. *For any Q with $\mathbb{E}_Q[X] \in (A, B)$, there exists a unique $\lambda \in \mathbb{R}$ such that the tilted distribution P_λ satisfies*

$$\begin{aligned} \mathbb{E}_{P_\lambda}[X] &= \mathbb{E}_Q[X] \\ D(P_\lambda||P) &\leq D(Q||P) \end{aligned}$$

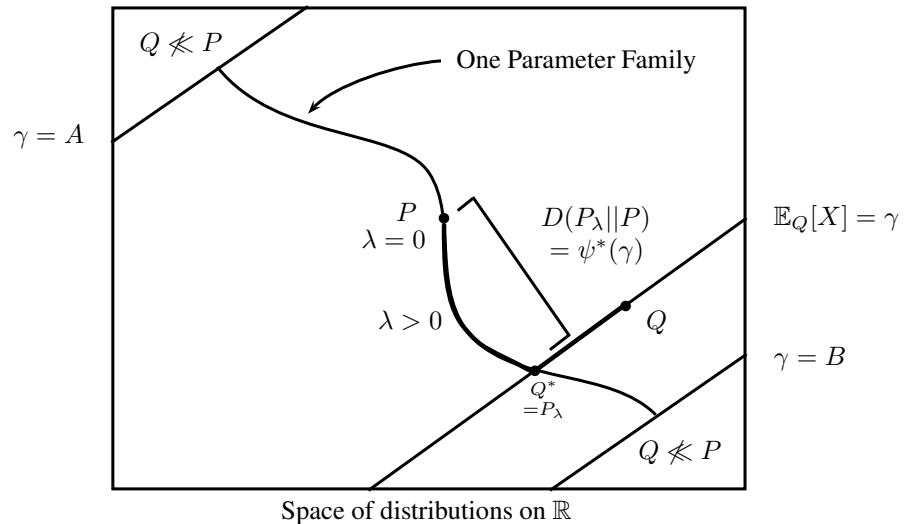
and furthermore the gap in the last inequality equals $D(Q||P_\lambda) = D(Q||P) - D(P_\lambda||P)$.

Proof. Proceed as in the proof of Theorem 15.14, and find the unique λ s.t. $\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) = \mathbb{E}_Q[X]$. Then $D(P_\lambda||P) = \psi_X^*(\mathbb{E}_Q[X]) = \lambda\mathbb{E}_Q[X] - \psi_X(\lambda)$. Repeat the steps (15.28)-(15.29) obtaining $D(Q||P) = D(Q||P_\lambda) + D(P_\lambda||P)$. \square

Remark: For any Q , this allows us to find a tilted measure P_λ that has the same mean yet smaller (or equal) divergence.

15.4 Interpretation of Information Projection

The following picture describes many properties of information projections.



- Each set $\{Q : \mathbb{E}_Q[X] = \gamma\}$ corresponds to a slice. As γ varies from A to B , the curves fill the entire space except for the corner regions.
- When $\gamma < A$ or $\gamma > B$, $Q \not\ll P$.
- As γ varies, the P_λ 's trace out a curve via $\psi^*(\gamma) = D(P_\lambda||P)$. This set of distributions is called a *one parameter family*, or *exponential family*.

Key Point: The one parameter family curve intersects each γ -slice $\mathcal{E} = \{Q : \mathbb{E}_Q[X] = \gamma\}$ “orthogonally” at the minimizing $Q^* \in \mathcal{E}$, and the distance from P to Q^* is given by $\psi^*(\lambda)$. To see this, note that applying Theorem 15.12 to the convex set \mathcal{E} gives us $D(Q||P) \geq D(Q||Q^*) + D(Q^*||P)$. Now thanks to Corollary 15.1, we in fact have *equality* $D(Q||P) = D(Q||Q^*) + D(Q^*||P)$ and $Q^* = P_\lambda$ for some tilted measure.

15.5 Generalization: Sanov's theorem

A corollary of the WLLN is that the empirical distribution of n iid observations drawn from a distribution (called population in statistics speak) converges weakly to this distribution. The following theorem due to Sanov quantifies the large-deviations behavior of this convergence.

Theorem 15.13 (Sanov's Theorem). *Consider observing n samples $X_1, \dots, X_n \sim \text{iid } P$. Let \hat{P} be the empirical distribution, i.e., $\hat{P} = \frac{1}{n} \sum_{j=1}^n \delta_{X_j}$. Let \mathcal{E} be a convex set of distributions. Then under regularity conditions on \mathcal{E} and P we have*

$$\mathbb{P}[\hat{P} \in \mathcal{E}] = \exp \left\{ -n \min_{Q \in \mathcal{E}} D(Q \| P) + o(n) \right\}.$$

Examples of regularity conditions in the above theorem include: (a) \mathcal{X} is finite and \mathcal{E} is closed with non-empty interior – see Exercise III.12 for a full proof in this case; (b) \mathcal{X} is a Polish space and the set \mathcal{E} is weakly closed and has non-empty interior.

Proof sketch. The lower bound is proved as in Theorem 15.10: Just take an arbitrary $Q \in \mathcal{E}$ and apply a suitable version of WLLN to conclude $Q^n[\hat{P} \in \mathcal{E}] = 1 + o(1)$.

For the upper bound we can again adapt the proof from Theorem 15.10. Alternatively, we can write the convex set \mathcal{E} as an intersection of half spaces. Then we have already solved the problem for half-spaces $\{Q : \mathbb{E}_Q[X] \geq \gamma\}$. The general case follows by the following consequence of Theorem 15.12: if Q^* is projection of P onto \mathcal{E}_1 and Q^{**} is projection of Q^* on \mathcal{E}_2 , then Q^{**} is also projection of P onto $\mathcal{E}_1 \cap \mathcal{E}_2$:

$$D(Q^{**} \| P) = \min_{Q \in \mathcal{E}_1 \cap \mathcal{E}_2} D(Q \| P) \Leftarrow \begin{cases} D(Q^* \| P) = \min_{Q \in \mathcal{E}_1} D(Q \| P) \\ D(Q^{**} \| Q^*) = \min_{Q \in \mathcal{E}_2} D(Q \| Q^*) \end{cases}$$

(Repeated projection property)

Indeed, by first tilting from P to Q^* we find

$$\begin{aligned} P[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] &\leq \exp(-nD(Q^* \| P)) Q^*[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] \\ &\leq \exp(-nD(Q^* \| P)) Q^*[\hat{P} \in \mathcal{E}_2] \end{aligned}$$

and from here proceed by tilting from Q^* to Q^{**} and note that $D(Q^* \| P) + D(Q^{**} \| Q^*) = D(Q^{**} \| P)$. \square

16

Hypothesis testing: error exponents

In this chapter our goal is to determine the achievable region of the exponent pairs (E_0, E_1) for the Type-I and Type-II error probabilities. Our strategy is to apply the achievability and (strong) converse bounds from Chapter 14 in conjunction with the large deviation theory developed in Chapter 15.

16.1 (E_0, E_1) -Tradeoff

Recall the setting of Chernoff regime introduced in Section 14.6, where the goal is in designing tests satisfying

$$\pi_{1|0} = 1 - \alpha \leq \exp(-nE_0), \quad \pi_{0|1} = \beta \leq \exp(-nE_1).$$

To find the best tradeoff of E_0 versus E_1 we can define the following function

$$\begin{aligned} E_1^*(E_0) &\triangleq \sup\{E_1 : \exists n_0, \forall n \geq n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1 - \exp(-nE_0), \beta < \exp(-nE_1)\} \\ &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\beta_{1-\exp(-nE_0)}(P^n, Q^n)} \end{aligned}$$

This should be compared with Stein's exponent in Definition 14.14.

Define

$$T_k = \log \frac{dQ}{dP}(X_k), \quad k = 1, \dots, n$$

which are iid copies of $T = \log \frac{dQ}{dP}(X)$. Then $\log \frac{dQ^n}{dP^n}(X^n) = \sum_{k=1}^n T_k$, which is an iid sum under both P and Q .

The log MGF of T under P (again assumed to be finite and also T is not a constant since $P \neq Q$) and the corresponding rate function are (cf. Definitions 15.2 and 15.6):

$$\psi_P(\lambda) = \log \mathbb{E}_P[\exp(\lambda T)], \quad \psi_P^*(\theta) = \sup_{\lambda \in \mathbb{R}} \theta \lambda - \psi_P(\lambda).$$

For discrete distributions, we have $\psi_P(\lambda) = \log \sum_x P(x)^{1-\lambda} Q(x)^\lambda$; in general, $\psi_P(\lambda) = \log \int d\mu \left(\frac{dP}{d\mu} \right)^{1-\lambda} \left(\frac{dQ}{d\mu} \right)^\lambda$ for some dominating measure μ .

Note that since $\psi_P(0) = \psi_P(1) = 0$, from the convexity of ψ_P (Theorem 15.3) we conclude that $\psi_P(\lambda)$ is finite on $0 \leq \lambda \leq 1$. Furthermore, assuming $P \ll Q$ and $Q \ll P$ we also have that $\lambda \mapsto \psi_P(\lambda)$ continuous everywhere on $[0, 1]$. (The continuity on $(0, 1)$ follows from convexity, but for the boundary points we need more detailed arguments.) Although all results in this section

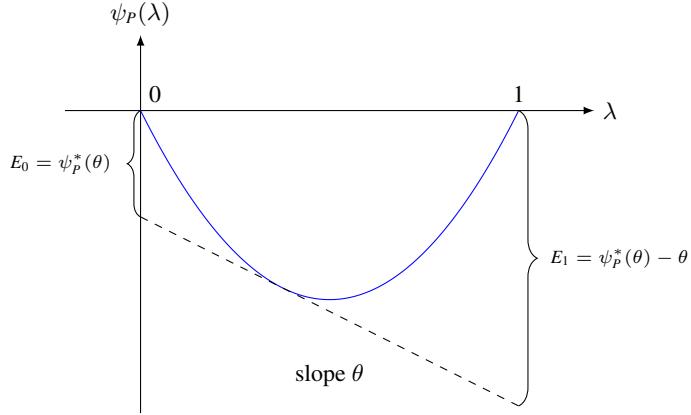
16.1 (E_0, E_1) -Tradeoff 261

Figure 16.1 Geometric interpretation of Theorem 16.1 relies on the properties of $\psi_P(\lambda)$ and $\psi_P^*(\theta)$. Note that $\psi_P(0) = \psi_P(1) = 0$. Moreover, by Theorem 15.7, $\theta \mapsto E_0(\theta)$ is increasing, $\theta \mapsto E_1(\theta)$ is decreasing.

apply under the (milder) conditions of $P \ll Q$ and $Q \ll P$, we will only present proofs under the (stronger) condition that log-MGF exists for all λ , following the convention of the previous chapter. The following result determines the optimal (E_0, E_1) -tradeoff in a parametric form. For a concrete example, see Exercise III.11 for testing two Gaussians.

Theorem 16.1. Assume $P \ll Q$ and $Q \ll P$. Then

$$E_0(\theta) = \psi_P^*(\theta), \quad E_1(\theta) = \psi_P^*(\theta) - \theta, \quad (16.1)$$

parametrized by $-D(P\|Q) \leq \theta \leq D(Q\|P)$, characterizes the upper boundary of the region of all achievable (E_0, E_1) -pairs. (See Fig. 16.1 for an illustration.)

Remark 16.1 (Rényi divergence). In Definition 7.34 we defined Rényi divergences D_λ . Note that $\psi_P(\lambda) = (\lambda - 1)D_\lambda(Q\|P) = -\lambda D_{1-\lambda}(P\|Q)$. This provides another explanation that $\psi_P(\lambda)$ is negative for λ between 0 and 1, and the slope at endpoints is: $\psi'_P(0) = -D(P\|Q)$ and $\psi'_P(1) = D(Q\|P)$. See also Ex. I.30.

Corollary 16.2 (Bayesian criterion). Fix a prior (π_0, π_1) such that $\pi_0 + \pi_1 = 1$ and $0 < \pi_0 < 1$. Denote the optimal Bayesian (average) error probability by

$$P_e^*(n) \triangleq \inf_{P_{Z|X^n}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

with exponent

$$E \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_e^*(n)}.$$

Then

$$E = \max_{\theta} \min(E_0(\theta), E_1(\theta)) = \psi_P^*(0)$$

regardless of the prior, and

$$\psi_P^*(0) = - \inf_{\lambda \in [0,1]} \psi_P(\lambda) = - \inf_{\lambda \in [0,1]} \log \int (dP)^{1-\lambda} (dQ)^\lambda \triangleq C(P, Q) \quad (16.2)$$

is called the Chernoff exponent or Chernoff information.

Notice that from (14.19) we always have

$$\log \left(1 - \frac{1}{2} H^2(P, Q) \right) \leq C(P, Q) \leq 2 \log \left(1 - \frac{1}{2} H^2(P, Q) \right)$$

and thus for small $H^2(P, Q)$ we have $C(P, Q) \asymp H^2(P, Q)$.

Remark 16.2 (Bhattacharyya distance). There is an important special case in which Chernoff exponent simplifies. Instead of i.i.d. observations, consider independent, but not identically distributed observations. Namely, suppose that two hypotheses correspond to two different strings x^n and \tilde{x}^n over a finite alphabet \mathcal{X} . The hypothesis tester observes $Y^n = (Y_1, \dots, Y_n)$ obtained by applying one of the two strings to the input of the memoryless channel $P_{Y|X}$; in other words, either $Y^n \sim \prod_{t=1}^n P_{Y|X=x_t}$ or $\prod_{t=1}^n P_{Y|X=\tilde{x}_t}$. (The alphabet \mathcal{Y} does not need to be finite, but we assume this below.) Extending Corollary 16.1 it can be shown, that in this case the optimal (average) probability of error $P_e^*(x^n, \tilde{x}^n)$ has (Chernoff) exponent¹

$$E = - \inf_{\lambda \in [0,1]} \frac{1}{n} \sum_{t=1}^n \log \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x_t)^\lambda P_{Y|X}(y|\tilde{x}_t)^{1-\lambda}.$$

If $|\mathcal{X}| = 2$ and if the compositions (types) of x^n and \tilde{x}^n are equal (!), the expression is invariant under $\lambda \leftrightarrow 1 - \lambda$ and thus from the convexity in λ we conclude that $\lambda = \frac{1}{2}$ is optimal,² yielding $E = \frac{1}{n} d_B(x^n, \tilde{x}^n)$, where

$$d_B(x^n, \tilde{x}^n) = - \sum_{t=1}^n \log \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|x_t) P_{Y|X}(y|\tilde{x}_t)} \quad (16.3)$$

is known as the *Bhattacharyya distance* between codewords x^n and \tilde{x}^n . (Compare with the Bhattacharyya coefficient defined after (7.5).) Without the two assumptions stated, $d_B(\cdot, \cdot)$ does not necessarily give the optimal error exponent. We do, however, always have the bounds, see (14.19):

$$\frac{1}{4} \exp(-2d_B(x^n, \tilde{x}^n)) \leq P_e^*(x^n, \tilde{x}^n) \leq \exp(-d_B(x^n, \tilde{x}^n)),$$

where the upper bound becomes tighter when the joint composition of (x^n, \tilde{x}^n) and that of (\tilde{x}^n, x^n) are closer.

Proof of Theorem 16.1. The idea is to apply the large deviation theory to the iid sum $\sum_{k=1}^n T_k$. Specifically, let's rewrite the achievability and converse bounds from Chapter 14 in terms of T :

¹ In short, this is because the optimal tilting parameter λ does not need to be chosen differently for different values of (x_t, \tilde{x}_t) .

² For another example where $\lambda = \frac{1}{2}$ achieves the optimal in the Chernoff information, see Exercise III.19.

16.2 Equivalent forms of Theorem 16.1 263

- Achievability (Neyman-Pearson): Applying Theorem 14.11 with $\tau = -n\theta$, the LRT achieves the following

$$\pi_{1|0} = P \left[\sum_{k=1}^n T_k \geq n\theta \right] \quad \pi_{0|1} = Q \left[\sum_{k=1}^n T_k < n\theta \right] \quad (16.4)$$

- Converse (strong): Applying Theorem 14.9 with $\gamma = \exp(-n\theta)$, any achievable $\pi_{1|0}$ and $\pi_{0|1}$ satisfy

$$\pi_{1|0} + \exp(-n\theta) \pi_{0|1} \geq P \left[\sum_{k=1}^n T_k \geq n\theta \right]. \quad (16.5)$$

For achievability, applying the nonasymptotic large deviations upper bound in Theorem 15.10 (and Theorem 15.14) to (16.4), we obtain that for any n ,

$$\begin{aligned} \pi_{1|0} &= P \left[\sum_{k=1}^n T_k \geq n\theta \right] \leq \exp(-n\psi_P^*(\theta)), \quad \text{for } \theta \geq \mathbb{E}_P T = -D(P\|Q) \\ \pi_{0|1} &= Q \left[\sum_{k=1}^n T_k < n\theta \right] \leq \exp(-n\psi_Q^*(\theta)), \quad \text{for } \theta \leq \mathbb{E}_Q T = D(Q\|P) \end{aligned}$$

Notice that by the definition of $T = \log \frac{dQ}{dP}$ we have

$$\begin{aligned} \psi_Q(\lambda) &= \log \mathbb{E}_Q[e^{\lambda T}] = \log \mathbb{E}_P[e^{(\lambda+1)T}] = \psi_P(\lambda+1) \\ \Rightarrow \psi_Q^*(\theta) &= \sup_{\lambda \in \mathbb{R}} \theta\lambda - \psi_P(\lambda+1) = \psi_P^*(\theta) - \theta. \end{aligned}$$

Thus the pair of exponents $(E_0(\theta), E_1(\theta))$ in (16.1) is achievable.

For converse, we aim to show that any achievable (E_0, E_1) pair must lie below the curve achieved by the above Neyman-Pearson test, namely $(E_0(\theta), E_1(\theta))$ parametrized by θ . Suppose $\pi_{1|0} = \exp(-nE_0)$ and $\pi_{0|1} = \exp(-nE_1)$ is achievable. Combining the strong converse bound (16.5) with the large deviations lower bound, we have: for any fixed $\theta \in [-D(P\|Q), D(Q\|P)]$,

$$\begin{aligned} \exp(-nE_0) + \exp(-n\theta) \exp(-nE_1) &\geq \exp(-n\psi_P^*(\theta) + o(n)) \\ \Rightarrow \min(E_0, E_1 + \theta) &\leq \psi_P^*(\theta) \\ \Rightarrow \text{either } E_0 &\leq \psi_P^*(\theta) \text{ or } E_1 \leq \psi_P^*(\theta) - \theta, \end{aligned}$$

proving the desired result. \square

16.2 Equivalent forms of Theorem 16.1

Alternatively, the optimal (E_0, E_1) -tradeoff can be stated in the following equivalent forms:

Theorem 16.3. (a) *The optimal exponents are given (parametrically) in terms of $\lambda \in [0, 1]$ as*

$$E_0 = D(P_\lambda \| P), \quad E_1 = D(P_\lambda \| Q) \quad (16.6)$$

where the distribution P_λ ³ is tilting of P along T given in (15.27), which moves from $P_0 = P$ to $P_1 = Q$ as λ ranges from 0 to 1:

$$dP_\lambda = (dP)^{1-\lambda} (dQ)^\lambda \exp\{-\psi_P(\lambda)\}.$$

(b) Yet another characterization of the boundary is

$$E_1^*(E_0) = \min_{Q': D(Q' \| P) \leq E_0} D(Q' \| Q), \quad 0 \leq E_0 \leq D(Q \| P) \quad (16.7)$$

Remark 16.3. The interesting consequence of this point of view is that it also suggests how typical error event looks like. Namely, consider an optimal hypothesis test achieving the pair of exponents (E_0, E_1) . Then conditioned on the error event (under either P or Q) we have that the empirical distribution of the sample will be close to P_λ . For example, if $P = \text{Bin}(m, p)$ and $Q = \text{Bin}(m, q)$, then the typical error event will correspond to a sample whose empirical distribution \hat{P}_n is approximately $\text{Bin}(m, r)$ for some $r = r(p, q, \lambda) \in (p, q)$, and not any other distribution on $\{0, \dots, m\}$.

Proof. The first part is verified trivially. Indeed, if we fix λ and let $\theta(\lambda) \triangleq \mathbb{E}_{P_\lambda}[T]$, then from (15.8) we have

$$D(P_\lambda \| P) = \psi_P^*(\theta),$$

whereas

$$D(P_\lambda \| Q) = \mathbb{E}_{P_\lambda} \left[\log \frac{dP_\lambda}{dQ} \right] = \mathbb{E}_{P_\lambda} \left[\log \frac{dP_\lambda}{dP} \frac{dP}{dQ} \right] = D(P_\lambda \| P) - \mathbb{E}_{P_\lambda}[T] = \psi_P^*(\theta) - \theta.$$

Also from (15.7) we know that as λ ranges in $[0, 1]$ the mean $\theta = \mathbb{E}_{P_\lambda}[T]$ ranges from $-D(P \| Q)$ to $D(Q \| P)$.

To prove the second claim (16.7), the key observation is the following: Since Q is itself a tilting of P along T (with $\lambda = 1$), the following two families of distributions

$$\begin{aligned} dP_\lambda &= \exp\{\lambda T - \psi_P(\lambda)\} \cdot dP \\ dQ_{\lambda'} &= \exp\{\lambda' T - \psi_Q(\lambda')\} \cdot dQ \end{aligned}$$

are in fact the same family with $Q_{\lambda'} = P_{\lambda'+1}$.

Now, suppose that Q^* achieves the minimum in (16.7) and that $Q^* \neq Q$, $Q^* \neq P$ (these cases should be verified separately). Note that we have not shown that this minimum is achieved, but it will be clear that our argument can be extended to the case of when Q'_n is a sequence achieving the infimum. Then, on one hand, obviously

$$D(Q^* \| Q) = \min_{Q': D(Q' \| P) \leq E_0} D(Q' \| Q) \leq D(P \| Q)$$

On the other hand, since $E_0 \leq D(Q \| P)$ we also have

$$D(Q^* \| P) \leq D(Q \| P).$$

³ This is called a geometric mixture of P and Q .

16.2 Equivalent forms of Theorem 16.1 265

Therefore,

$$\mathbb{E}_{Q^*}[T] = \mathbb{E}_{Q^*} \left[\log \frac{dQ^*}{dP} \frac{dQ}{dQ^*} \right] = D(Q^* \| P) - D(Q^* \| Q) \in [-D(P \| Q), D(Q \| P)]. \quad (16.8)$$

Next, we have from Corollary 15.1 that there exists a *unique* P_λ with the following three properties:⁴

$$\begin{aligned} \mathbb{E}_{P_\lambda}[T] &= \mathbb{E}_{Q^*}[T] \\ D(P_\lambda \| P) &\leq D(Q^* \| P) \\ D(P_\lambda \| Q) &\leq D(Q^* \| Q) \end{aligned}$$

Thus, we immediately conclude that minimization in (16.7) can be restricted to Q^* belonging to the family of tilted distributions $\{P_\lambda, \lambda \in \mathbb{R}\}$. Furthermore, from (16.8) we also conclude that $\lambda \in [0, 1]$. Hence, characterization of $E_1^*(E_0)$ given by (16.6) coincides with the one given by (16.7). \square

Remark 16.4. A geometric interpretation of (16.7) is given in Fig. 16.2: As λ increases from 0 to 1, or equivalently, θ increases from $-D(P \| Q)$ to $D(Q \| P)$, the optimal distribution P_λ traverses down the dotted path from P and Q . Note that there are many ways to interpolate between P and Q , e.g., by taking their (arithmetic) mixture $(1 - \lambda)P + \lambda Q$. In contrast, P_λ is a *geometric mixture* of P and Q , and this special path is in essence a geodesic connecting P to Q and the exponents E_0 and E_1 measures its respective distances to P and Q . Unlike Riemannian geometry, though, here the sum of distances to the two endpoints from an intermediate P_λ actually varies along the geodesic.

⁴ A subtlety: In Corollary 15.1 we ask $\mathbb{E}_{Q^*}[T] \in (A, B)$. But A, B – the essential range of T – depend on the distribution under which the essential range is computed, cf. (15.23). Fortunately, we have $Q \ll P$ and $P \ll Q$, so the essential range is the same under both P and Q . And furthermore (16.8) implies that $\mathbb{E}_{Q^*}[T] \in (A, B)$.

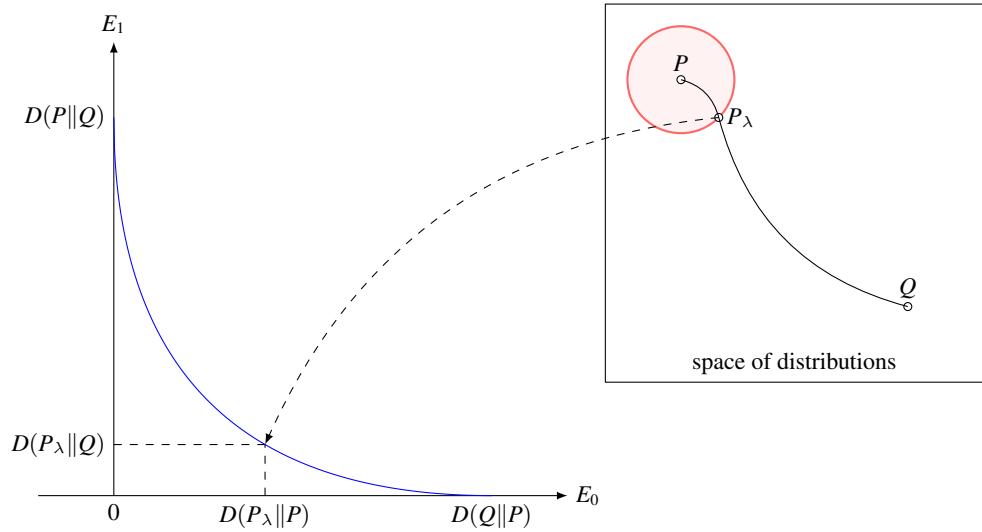


Figure 16.2 Geometric interpretation of (16.7). Here the shaded circle represents $\{Q' : D(Q'||P) \leq E_0\}$, the KL divergence “ball” of radius E_0 centered at P . The optimal $E_1^*(E_0)$ in (16.7) is given by the divergence from Q to the closest element of this ball, attained by some tilted distribution P_λ . The tilted family P_λ is the geodesic traversing from P to Q as λ increases from 0 to 1.

16.3* Sequential Hypothesis Testing

Review: Filtration and stopping time

- A sequence of nested σ -algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \dots \subset \mathcal{F}_n \dots \subset \mathcal{F}$ is called a filtration of \mathcal{F} .
- A random variable τ is called a stopping time of a filtration \mathcal{F}_n if (a) τ is valued in \mathbb{Z}_+ and (b) for every $n \geq 0$ the event $\{\tau \leq n\} \in \mathcal{F}_n$.
- The σ -algebra \mathcal{F}_τ consists of all events E such that $E \cap \{\tau \leq n\} \in \mathcal{F}_n$ for all $n \geq 0$.
- When $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$ the interpretation is that τ is a time that can be determined by causally observing the sequence X_j , and random variables measurable with respect to \mathcal{F}_τ are precisely those whose value can be determined on the basis of knowing (X_1, \dots, X_τ) .
- Let M_n be a martingale adapted to \mathcal{F}_n , i.e. M_n is \mathcal{F}_n -measurable and $\mathbb{E}[M_n | \mathcal{F}_k] = M_{\min(n,k)}$. Then $\tilde{M}_n = M_{\min(n,\tau)}$ is also a martingale. If collection $\{M_n\}$ is uniformly integrable then

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0].$$

- For more details, see [58, Chapter V].

16.3* Sequential Hypothesis Testing 267

So far we have always been working with a fixed number of observations n . However, different realizations of X^n are informative to different levels, i.e. under some realizations we are very certain about declaring the true hypothesis, whereas some other realizations leave us more doubtful. In the fixed n setting, the tester is forced to take a guess in the latter case. In the sequential setting, pioneered by Wald [?], the tester is allowed to request more samples. We show in this section that the optimal test in this setting is something known as sequential probability ratio test (SPRT) [?]. It will also be shown that the resulting tradeoff between the exponents E_0 and E_1 is much improved in the sequential setting.

We start with the concept of a **sequential test**. Informally, at each time t , upon receiving the observation X_t , a sequential test either declares H_0 , declares H_1 , or requests one more observation. The rigorous definition is as follows: a sequential hypothesis test consists of (a) a stopping time τ with respect to the filtration $\{\mathcal{F}_k, k \in \mathbb{Z}_+\}$, where $\mathcal{F}_k \triangleq \sigma\{X_1, \dots, X_n\}$ is generated by the first n observations; and (b) a random variable (decision) $Z \in \{0, 1\}$ measurable with respect to \mathcal{F}_τ . Each sequential test is associated with the following performance metrics:

$$\alpha = \mathbb{P}[Z = 0], \quad \beta = \mathbb{Q}[Z = 0] \quad (16.9)$$

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau], \quad l_1 = \mathbb{E}_{\mathbb{Q}}[\tau] \quad (16.10)$$

The easiest way to see why sequential tests may be dramatically superior to fixed-sample-size tests is the following example: Consider $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{-1}$. Since $P \not\ll Q$, we also have $P^n \not\ll Q^n$. Consequently, no finite-sample-size test can achieve zero error under both hypotheses. However, an obvious sequential test (wait for the first appearance of ± 1) achieves zero error probability with finite average number of samples (2) under both hypotheses. This advantage is also very clear in the achievable error exponents as Fig. 16.3 shows.

The following result is due to [?] (for the special case of $E_0 = D(Q||P)$ and $E_1 = D(P||Q)$) and [?] (for the generalization).

Theorem 16.4. Assume bounded LLR:⁵

$$\left| \log \frac{P(x)}{Q(x)} \right| \leq c_0, \forall x$$

where c_0 is some positive constant. Call a pair of exponents (E_0, E_1) achievable if there exist a test with $\ell_0, \ell_1 \rightarrow \infty$ and probabilities satisfy:

$$\pi_{1|0} \leq \exp(-l_0 E_0(1 + o(1))), \quad \pi_{0|1} \leq \exp(-l_1 E_1(1 + o(1)))$$

Then the set of achievable exponents must satisfy

$$E_0 E_1 \leq D(P||Q) D(Q||P).$$

Furthermore, any such (E_0, E_1) is achieved by the sequential probability ratio test SPRT(A, B) (A, B are large positive numbers) defined as follows:

$$\tau = \inf\{n : S_n \geq B \text{ or } S_n \leq -A\}$$

⁵ This assumption is satisfied for example for a pair of full support discrete distributions on finite alphabets.

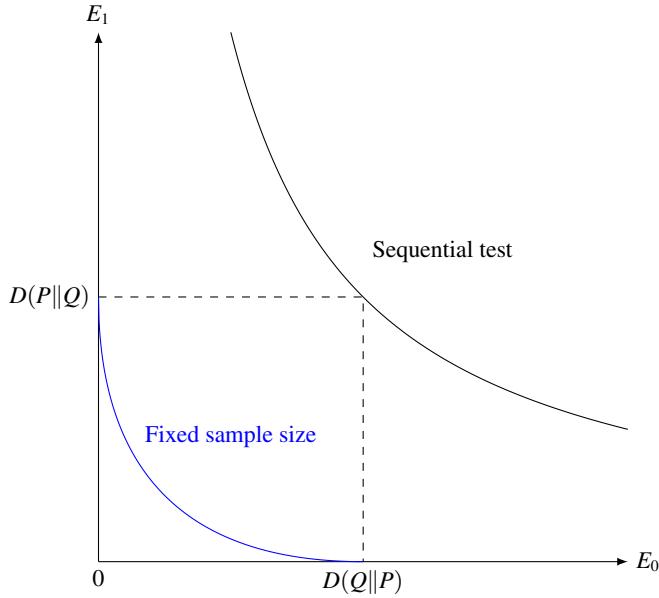


Figure 16.3 Tradeoff between Type-I and Type-II error exponents. The bottom curve corresponds to optimal tests with fixed sample size (Theorem 16.1) and the upper curve to optimal sequential tests (Theorem 16.7).

$$Z = \begin{cases} 0, & \text{if } S_\tau \geq B \\ 1, & \text{if } S_\tau < -A \end{cases}$$

where

$$S_n = \sum_{k=1}^n \log \frac{P(X_k)}{Q(X_k)}$$

is the log likelihood function of the first n observations.

Remark 16.5 (Interpretation of SPRT). Under the usual setup of hypothesis testing, we collect a sample of n iid observations, evaluate the LLR S_n , and compare it to the threshold to give the optimal test. Under the sequential setup, $\{S_n : n \geq 1\}$ is a *random walk*, which has positive (resp. negative) drift $D(P||Q)$ (resp. $-D(Q||P)$) under the null (resp. alternative)! SPRT simply declares P if the random walk crosses the upper boundary B , or Q if the random walk crosses the lower boundary $-A$. See Fig. ?? for an illustration.

Proof. As preparation we show two useful identities:

- For any stopping time with $\mathbb{E}_P[\tau] < \infty$ we have

$$\mathbb{E}_P[S_\tau] = \mathbb{E}_P[\tau]D(P||Q) \quad (16.11)$$

16.3* Sequential Hypothesis Testing 269

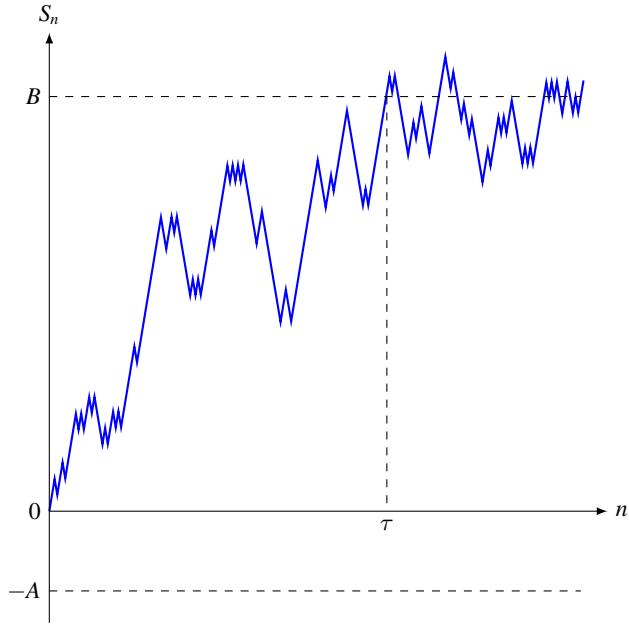


Figure 16.4 Illustration of the $\text{SPRT}(A, B)$ test. Here, at the stopping time τ , the LLR process S_n reaches B before reaching $-A$ and the decision is $Z = 1$.

and similarly, if $\mathbb{E}_Q[\tau] < \infty$ then

$$\mathbb{E}_Q[S_\tau] = -\mathbb{E}_Q[\tau]D(Q||P).$$

To prove these, notice that

$$M_n = S_n - nD(P||Q)$$

is clearly a martingale w.r.t. \mathcal{F}_n . Consequently,

$$\tilde{M}_n \triangleq M_{\min(\tau, n)}$$

is also a martingale. Thus

$$\mathbb{E}[\tilde{M}_n] = \mathbb{E}[\tilde{M}_0] = 0,$$

or, equivalently,

$$\mathbb{E}[S_{\min(\tau, n)}] = \mathbb{E}[\min(\tau, n)]D(P||Q). \quad (16.12)$$

This holds for every $n \geq 0$. From boundedness assumption we have $|S_n| \leq nc$ and thus $|S_{\min(n, \tau)}| \leq n\tau$, implying that collection $\{S_{\min(n, \tau)}, n \geq 0\}$ is uniformly integrable. Thus, we can take $n \rightarrow \infty$ in (16.12) and interchange expectation and limit safely to conclude (16.11).

- Let τ be a stopping time. Recall that Z is a Radon-Nikodym derivative of \mathbb{P} w.r.t. \mathbb{Q} on a σ -algebra \mathcal{F}_τ , denoted by $\frac{d\mathbb{P}|_{\mathcal{F}_\tau}}{d\mathbb{Q}|_{\mathcal{F}_\tau}}$, if

$$\mathbb{E}_P[1_E] = \mathbb{E}_Q[Z1_E] \quad \forall E \in \mathcal{F}_\tau. \quad (16.13)$$

We will show that it is in fact given by

$$\frac{d\mathbb{P}|_{\mathcal{F}_\tau}}{d\mathbb{Q}|_{\mathcal{F}_\tau}} = \exp\{S_\tau\}.$$

Indeed, what we need to verify is that (16.13) holds with $Z = \exp\{S_\tau\}$ and an arbitrary event $E \in \mathcal{F}_\tau$, which we decompose as

$$1_E = \sum_{n \geq 0} 1_{E \cap \{\tau=n\}}.$$

By monotone convergence theorem applied to the both sides of (16.13) it is then sufficient to verify that for every n

$$\mathbb{E}_P[1_{E \cap \{\tau=n\}}] = \mathbb{E}_Q[\exp\{S_\tau\}1_{E \cap \{\tau=n\}}]. \quad (16.14)$$

This, however, follows from the fact that $E \cap \{\tau = n\} \in \mathcal{F}_n$ and $\frac{d\mathbb{P}|_{\mathcal{F}_n}}{d\mathbb{Q}|_{\mathcal{F}_n}} = \exp\{S_n\}$ by the very definition of S_n .

We now proceed to the proof. For **achievability** we apply (16.13) to infer

$$\pi_{1|0} = \mathbb{P}[S_\tau \leq -A] = \mathbb{E}_Q[\exp\{S_\tau\}1\{S_\tau \leq -A\}] \leq e^{-A}.$$

Next, we denote $\tau_0 = \inf\{n : S_n \geq B\}$ and observe that $\tau \leq \tau_0$, whereas the expectation of τ_0 can be bounded using (16.11) as:

$$\mathbb{E}_P[\tau] \leq \mathbb{E}_P[\tau_0] = \mathbb{E}_P[S_{\tau_0}] \leq B + c_0,$$

where in the last step we used the boundedness assumption to infer

$$S_{\tau_0} \leq B + c_0.$$

Thus

$$l_0 = \mathbb{E}_P[\tau] \leq \mathbb{E}_P[\tau_0] \leq \frac{B + c_0}{D(P||Q)} \approx \frac{B}{D(P||Q)}. \text{ for large } B$$

Similarly we can show $\pi_{0|1} \leq e^{-B}$ and $l_1 \leq \frac{A}{D(Q||P)}$ for large A . Take $B = l_0 D(P||Q), A = l_1 D(Q||P)$, this shows the achievability.

Converse: Assume (E_0, E_1) achievable for large l_0, l_1 . Recall from Section 4.5* that $D(\mathbb{P}_{\mathcal{F}_\tau} || \mathbb{Q}_{\mathcal{F}_\tau})$ denotes the divergence between P and Q when viewed as measures on σ -algebra \mathcal{F}_τ . We apply the data processing inequality for divergence to obtain:

$$d(\mathbb{P}(Z=1) || \mathbb{Q}(Z=1)) \leq D(\mathbb{P}_{\mathcal{F}_\tau} || \mathbb{Q}_{\mathcal{F}_\tau}) = \mathbb{E}_P[S_\tau] \stackrel{(16.11)}{=} \mathbb{E}_P[\tau] D(P||Q) = l_0 D(P||Q),$$

16.4 Composite, robust and goodness-of-fit hypothesis testing 271

Notice that for $l_0 E_0$ and $l_1 E_1$ large, we have $d(\mathbb{P}(Z = 1) \| \mathbb{Q}(Z = 1)) = l_1 E_1(1 + o(1))$, therefore $l_1 E_1 \leq (1 + o(1))l_0 D(P \| Q)$. Similarly we can show that $l_0 E_0 \leq (1 + o(1))l_1 D(Q \| P)$. Thus taking $l_0, l_1 \rightarrow \infty$ we conclude

$$E_0 E_1 \leq D(P \| Q) D(Q \| P).$$

□

16.4 Composite, robust and goodness-of-fit hypothesis testing

In this chapter we have considered the setting of distinguishing between the two alternatives, under either of which the data distribution was specified completely. There are multiple other settings that have also been studied in the literature, which we briefly mention here for completeness.

The key departure is to replace the simple hypotheses that we started with in Chapter 14 with *composite* ones. Namely, we postulate

$$H_0 : X_i \stackrel{\text{i.i.d.}}{\sim} P, \quad P \in \mathcal{P} \quad \text{vs} \quad H_1 : X_i \stackrel{\text{i.i.d.}}{\sim} Q, \quad Q \in \mathcal{Q},$$

where \mathcal{P} and \mathcal{Q} are two families of distributions. In this case for a given test $Z = Z(X_1, \dots, X_n) \in \{0, 1\}$ we define the two types of error as before, but taking worst-case choices over the distribution:

$$1 - \alpha = \inf_{P \in \mathcal{P}} P^{\otimes n}[Z = 0], \quad \beta = \sup_{Q \in \mathcal{Q}} Q^{\otimes n}[Z = 0].$$

Unlike testing simple hypotheses for which Neyman-Pearson's test is optimal (Theorem 14.11), in general there is no explicit description for the optimal test of composite hypotheses (cf. (32.24)). The popular choice is a generalized likelihood-ratio test (GLRT) that proposes to threshold the GLR

$$T(X^n) = \frac{\sup_{P \in \mathcal{P}} P^{\otimes n}(X^n)}{\sup_{Q \in \mathcal{Q}} Q^{\otimes n}(X^n)}.$$

For examples and counterexamples of the optimality of GLRT in terms of error exponents, see, e.g. [334].

Sometimes the families \mathcal{P} and \mathcal{Q} are small balls (in some metric) surrounding the center distributions P and Q , respectively. In this case, testing \mathcal{P} against \mathcal{Q} is known as robust hypothesis testing (since the test is robust to small deviations of the data distribution). There is a notable finite-sample optimality result in this case due to Huber [157], Exercise III.20. Asymptotically, it turns out that if \mathcal{P} and \mathcal{Q} are separated in the Hellinger distance, then the probability of error can be made exponentially small: see Theorem 32.8.

Sometimes in the setting of composite testing the distance between \mathcal{P} and \mathcal{Q} is zero. This is the case, for example, for the most famous setting of a Student t -test: $\mathcal{P} = \{\mathcal{N}(0, \sigma^2) : \sigma^2 > 0\}$, $\mathcal{Q} = \{\mathcal{N}(\mu, \sigma^2) : \mu \neq 0, \sigma^2 > 0\}$. It is clear that in this case there is no way to construct a test with $\alpha + \beta < 1$, since the data distribution under H_1 can be arbitrarily close to P_0 . Here, thus, instead of minimizing worst-case β , one tries to find a test statistic $T(X_1, \dots, X_n)$ which is a) pivotal in

the sense that its distribution under the H_0 is (asymptotically) independent of the choice $P_0 \in \mathcal{P}$; and b) consistent, in the sense that $T \rightarrow \infty$ as $n \rightarrow \infty$ under any $Q \in \mathcal{Q}$. Optimality questions are studied by minimizing β as a function of $Q \in \mathcal{Q}$ (known as the power curve). The uniform most powerful tests are the gold standard in this area [193, Chapter 3], although besides a few classical settings (such as the one above) their existence is unknown.

In other settings, known as the goodness-of-fit testing [193, Chapter 14], instead of relatively low-complexity parametric families \mathcal{P} and \mathcal{Q} one is interested in a giant set of alternatives \mathcal{Q} . For example, the simplest setting is to distinguish $H_0 : X_i \stackrel{\text{i.i.d.}}{\sim} P_0$ vs $H_1 : X_i \stackrel{\text{i.i.d.}}{\sim} Q$, $\text{TV}(P_0, Q) > \delta$. If $\delta = 0$, then in this case again the worst case $\alpha + \beta = 1$ for any test and one may only ask for a statistic $T(X^n)$ with a known distribution under H_0 and $T \rightarrow \infty$ for any Q in the alternative. For $\delta > 0$ the problem is known as nonparametric detection [161, 162] and related to that of property testing [138].

Exercises for Part III

III.1 Let P_0 and P_1 be distributions on \mathcal{X} . Recall that the region of achievable pairs $(P_0[Z = 0], P_1[Z = 0])$ via randomized tests $P_{Z|X} : \mathcal{X} \rightarrow \{0, 1\}$ is denoted

$$\mathcal{R}(P_0, P_1) \triangleq \bigcup_{P_{Z|X}} (P_0[Z = 0], P_1[Z = 0]) \subseteq [0, 1]^2.$$

Let $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ be a Markov kernel, which maps P_j to Q_j according to $P_j \xrightarrow{P_{Y|X}} Q_j, j = 0, 1$. Compare the regions $\mathcal{R}(P_0, P_1)$ and $\mathcal{R}(Q_0, Q_1)$. What does this say about $\beta_\alpha(P_0, P_1)$ vs. $\beta_\alpha(Q_0, Q_1)$?

Comment: This is the most general form of data-processing, all the other ones (divergence, mutual information, f -divergence, total-variation, Rényi-divergence, etc) are corollaries.

Bonus: Prove that $\mathcal{R}(P_0, P_1) \supset \mathcal{R}(Q_0, Q_1)$ implies existence of some $P_{Y|X}$ carrying P_j to Q_j (“inclusion of \mathcal{R} is equivalent to degradation”).

III.2 Recall the total variation distance

$$\text{TV}(P, Q) \triangleq \sup_E (P[E] - Q[E]).$$

(a) Prove that

$$\text{TV}(P, Q) = \sup_{0 \leq \alpha \leq 1} \{\alpha - \beta_\alpha(P, Q)\}.$$

Explain how to read the value $\text{TV}(P, Q)$ from the region $\mathcal{R}(P, Q)$. Does it equal half the maximal vertical segment in $\mathcal{R}(P, Q)$?

(b) (Bayesian criteria) Fix a prior $\pi = (\pi_0, \pi_1)$ such that $\pi_0 + \pi_1 = 1$ and $0 < \pi_0 < 1$. Denote the optimal average error probability by

$$P_e \triangleq \inf_{P_{Z|X^\pi}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}.$$

Prove that if $\pi = (\frac{1}{2}, \frac{1}{2})$, then

$$P_e = \frac{1}{2}(1 - \text{TV}(P, Q)).$$

Find the optimal test.

- (c) Find the optimal test for general prior π (not necessarily equiprobable).
- (d) Why is it always sufficient to focus on deterministic test in order to minimize the Bayesian error probability?

274 Exercises for Part III

III.3 Let P, Q be distributions such that for all $\alpha \in [0, 1]$ we have

$$\beta_\alpha(P, Q) \triangleq \min_{P_{Z|X}: P[Z=0] \geq \alpha} Q[Z=0] = \alpha^2.$$

Find $\text{TV}(P, Q)$, $D(P\|Q)$ and $D(Q\|P)$.

III.4 Function $\alpha \mapsto \beta_\alpha(P, Q)$ is monotone and thus by Lebesgue's theorem possesses a derivative

$$\beta'_\alpha \triangleq \frac{d}{d\alpha} \beta_\alpha(P, Q).$$

almost everywhere on $[0, 1]$. Prove

$$D(P\|Q) = - \int_0^1 \log \beta'_\alpha d\alpha. \quad (\text{III.1})$$

III.5 We have shown that for testing iid products and any fixed $\epsilon \in (0, 1)$:

$$\log \beta_{1-\epsilon}(P^n, Q^n) = -nD(P\|Q) + o(n), \quad n \rightarrow \infty,$$

which is equivalent to Stein's lemma. Show furthermore that assuming $V(P\|Q) < \infty$ we have

$$\log \beta_{1-\epsilon}(P^n, Q^n) = -nD(P\|Q) + \sqrt{nV(P\|Q)}Q^{-1}(\epsilon) + o(\sqrt{n}), \quad (\text{III.2})$$

where $Q^{-1}(\cdot)$ is the functional inverse of $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ and

$$V(P\|Q) \triangleq \text{Var}_P \left[\log \frac{dP}{dQ} \right].$$

III.6 (Inverse Donsker-Varadhan) Verify for positive discrete random variables X that,

$$\log \mathbb{E}_Q[X] = \sup_P [\mathbb{E}_P[\log X] - D(P\|Q)],$$

where supremum is over all distributions P on \mathcal{X} . (Hint: It is enough to extremize $J(P) = \mathbb{E}_P[\log X] - D(P\|Q) + \lambda(\sum P(x) - 1)$).

III.7 Prove

$$\min_{Q_{Y^n} \in \mathcal{F}} D(Q_{Y^n} \| \prod_{j=1}^n P_{Y_j}) = \min \sum_{j=1}^n D(Q_{Y_j} \| P_{Y_j})$$

whenever the constraint set \mathcal{F} is marginals-based, i.e.:

$$Q_{Y^n} \in \mathcal{F} \iff (Q_{Y_1}, \dots, Q_{Y_n}) \in \mathcal{F}'$$

for some \mathcal{F}' .

Conclude that in the case when $P_{Y_j} = P$ and

$$\mathcal{F} = \left\{ Q_{Y^n} : \mathbb{E}_Q \left[\sum_{j=1}^n f(Y_j) \right] \geq n\gamma \right\}$$

we have (*single-letterization*)

$$\min_{Q^n} D(Q^n \| P^n) = n \min_{Q_Y: \mathbb{E}_{Q_Y}[f(Y)] \geq \gamma} D(Q_Y \| P),$$

of which (15.15) is a special case. Hint: Convexity of divergence.

- III.8** Fix a distribution P_X on a finite set \mathcal{X} and a channel $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$. Consider a sequence x^n with composition P_X , i.e.

$$\#\{j : x_j = a\} = nP_X(a) \pm 1 \quad \forall a \in \mathcal{X}.$$

Let Y^n be generated according to $P_{Y|X}^n(\cdot | x^n)$. Show that

$$\log \mathbb{P} \left[\sum_{j=1}^n f(X_j, Y_j) \geq n\gamma \middle| X^n = x^n \right] = -n \min_{\mathbb{E}_Q[f(X, Y)] \geq \gamma} D(Q_{Y|X} \| P_{Y|X}|P_X) + o(n),$$

where minimum is over all Q_{XY} with $Q_X = P_X$.

- III.9** (Large deviations on the boundary) Recall that $A = \inf_\lambda \Psi'(\lambda)$ and $B = \sup_\lambda \Psi'(\lambda)$ were shown to be the boundaries of the support of P_X :⁶

$$B = \sup\{b : \mathbb{P}[X > b] > 0\}.$$

- (a) Show by example that $\Psi^*(B)$ can be finite or infinite.
- (b) Show by example that asymptotic behavior of

$$\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n X_i \geq B \right], \tag{III.3}$$

can be quite different depending on the distribution of P_X .

- (c) Compare $\Psi^*(B)$ and the exponent in (III.3) for your examples. Prove a general statement.

- III.10** (Small-ball probability I.) Let $Z \sim \mathcal{N}(0, I_d)$. Without using the χ^2 -density, show the following bound on $\mathbb{P}[\|Z\|_2 \leq \epsilon]$.

- (a) Using the Chernoff bound, show that for all $\epsilon > \sqrt{d}$,

$$\mathbb{P}[\|Z\|_2 \leq \epsilon] \leq \left(\frac{e\epsilon^2}{d} \right)^{d/2} e^{-\epsilon^2/2}.$$

- (b) Prove the lower bound

$$\mathbb{P}[\|Z\|_2 \leq \epsilon] \geq \left(\frac{\epsilon^2}{2\pi d} \right)^{d/2} e^{-\epsilon^2/2}.$$

- (c) Extend the results to $Z \sim \mathcal{N}(0, \Sigma)$.

See Exercise V.10 for an example in infinite dimensions.

⁶ In this exercise and the next, you may assume all log's and exp's are to the natural basis and that MGF exists for all λ .

276 Exercises for Part III

III.11 Consider the hypothesis testing problem:

$$H_0 : X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P = \mathcal{N}(0, 1),$$

$$H_1 : X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} Q = \mathcal{N}(\mu, 1).$$

- (a) Show that the Stein exponent is $V = \frac{\log e}{2}\mu^2$.
- (b) Show that the optimal tradeoff between achievable error-exponent pairs (E_0, E_1) is given by

$$E_1 = \frac{\log e}{2}(\mu - \sqrt{2E_0})^2, \quad 0 \leq E_0 \leq \frac{\log e}{2}\mu^2,$$

- (c) Show that the Chernoff exponent is $C(P, Q) = \frac{\log e}{8}\mu^2$.

III.12 Let X_j be i.i.d. exponential with unit mean. Since the log-MGF $\psi_X(\lambda) \triangleq \log \mathbb{E}[\exp\{\lambda X\}]$ does not exist for all $\lambda > 1$, the large deviations result

$$\mathbb{P}\left[\sum_{j=1}^n X_j \geq n\gamma\right] = \exp\{-n\psi_X^*(\gamma) + o(n)\} \quad (\text{III.4})$$

does not apply. Show (III.4) via the following steps:

- (a) Apply Chernoff argument directly to prove an upper bound:

$$\mathbb{P}\left[\sum_{j=1}^n X_j \geq n\gamma\right] \leq \exp\{-n\psi_X^*(\gamma)\} \quad (\text{III.5})$$

- (b) Fix an arbitrary $A > 0$ and prove

$$\mathbb{P}\left[\sum_{j=1}^n X_j \geq n\gamma\right] \geq \mathbb{P}\left[\sum_{j=1}^n (X_j \wedge A) \geq n\gamma\right], \quad (\text{III.6})$$

where $u \wedge v = \min(u, v)$.

- (c) Apply the results shown in Chapter to investigate the asymptotics of the right-hand side of (III.6).
- (d) Conclude the proof of (III.4) by taking $A \rightarrow \infty$.

III.13 Baby version of Sanov's theorem. Let \mathcal{X} be a finite set. Let \mathcal{E} be a convex set of probability distributions on \mathcal{X} . Assume that \mathcal{E} has non-empty interior. Let $X^n = (X_1, \dots, X_n)$ be iid drawn from some distribution P and let π_n denote the empirical distribution, i.e., $\pi_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. Our goal is to show that

$$E \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P(\pi_n \in \mathcal{E})} = \inf_{Q \in \mathcal{E}} D(Q \| P). \quad (\text{III.7})$$

- (a) Define the following set of joint distributions $\mathcal{E}_n \triangleq \{Q_{X^n} : Q_{X_i} \in \mathcal{E}, i = 1, \dots, n\}$. Show that

$$\inf_{Q_{X^n} \in \mathcal{E}_n} D(Q_{X^n} \| P_{X^n}) = n \inf_{Q \in \mathcal{E}} D(Q \| P),$$

where $P_{X^n} = P^n$.

- (b) Consider the conditional distribution $\tilde{P}_{X^n} = P_{X^n|\pi_n \in \mathcal{E}}$. Show that $\tilde{P}_{X^n} \in \mathcal{E}_n$.
(c) Prove the following nonasymptotic upper bound:

$$P(\pi_n \in \mathcal{E}) \leq \exp\left(-n \inf_{Q \in \mathcal{E}} D(Q||P)\right), \quad \forall n.$$

- (d) For any Q in the interior of \mathcal{E} , show that

$$P(\pi_n \in \mathcal{E}) \geq \exp(-nD(Q||P) + o(n)), \quad n \rightarrow \infty.$$

(Hint: Use data processing as in the proof of the large deviations theorem.)

- (e) Conclude (III.7).

III.14 Error exponents of data compression. Let X^n be iid according to P on a finite alphabet \mathcal{X} . Let $\epsilon_n^*(R)$ denote the minimal probability of error achieved by fixed-length compressors and decompressors for X^n of compression rate R . We have learned that if $R < H(P)$, then $\epsilon_n^*(R)$ tends to zero. The goal of this exercise is to show it converges exponentially fast and find the best exponent.

- (a) For any sequence x^n , denote by $\pi(x^n)$ its empirical distribution and by $\hat{H}(x^n)$ its empirical entropy, i.e., the entropy of the empirical distribution.⁷ For each $R > 0$, define the set $T = \{x^n : \hat{H}(x^n) < R\}$. Show that

$$|T| \leq \exp(nR)(n+1)^{|\mathcal{X}|}.$$

- (b) Show that for any $R > H(P)$,

$$\epsilon_n^*(R) \leq \exp\left(-n \inf_{Q:H(Q)>R} D(Q||P)\right).$$

Specify the achievable scheme. (Hint: Use Sanov's theorem in Exercise III.13.)

- (c) Prove that the above exponent is asymptotically optimal:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\epsilon_n^*(R)} \leq \inf_{Q:H(Q)>R} D(Q||P).$$

(Hint: Recall that any compression scheme for memoryless source with rate below the entropy fails with probability tending to one. Use data processing inequality.)

III.15 Denote by $N(\mu, \sigma^2)$ the one-dimensional Gaussian distribution with mean μ and variance σ^2 . Let $a > 0$. All logarithms below are natural.

- (a) Show that

$$\min_{Q:\mathbb{E}_Q[X] \geq a} D(Q||N(0, 1)) = \frac{a^2}{2}.$$

- (b) Let X_1, \dots, X_n be drawn iid from $N(0, 1)$. Using part (a) show that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\mathbb{P}[X_1 + \dots + X_n \geq na]} = \frac{a^2}{2}. \quad (\text{III.8})$$

⁷ For example, for the binary sequence $x^n = (010110)$, the empirical distribution is $\text{Ber}(1/2)$ and the empirical entropy is 1 bit.

278 Exercises for Part III

- (c) Let $\bar{\Phi}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ denote the complementary CDF of the standard Gaussian distribution. Express $\mathbb{P}[X_1 + \dots + X_n \geq na]$ in terms of the $\bar{\Phi}$ function. Using the fact that $\bar{\Phi}(x) = e^{-x^2/2+o(x^2)}$ as $x \rightarrow \infty$, reprove (III.8).
- (d) Let Y be a continuous random variable with zero mean and unit variance. Show that

$$\min_{\mu, \sigma} D(P_Y \| N(\mu, \sigma^2)) = D(P_Y \| N(0, 1)).$$

III.16 (Gibbs distribution) Let \mathcal{X} be finite alphabet, $f: \mathcal{X} \rightarrow \mathbb{R}$ some function and $E_{\min} = \min f(x)$.

- (a) Using I -projection show that for any $E \geq E_{\min}$ the solution of

$$H^*(E) = \max \{H(X) : \mathbb{E}[f(X)] \leq E\}$$

is given by $P_X(x) = \frac{1}{Z(\beta)} e^{-\beta f(x)}$ for some $\beta = \beta(E)$.

Comment: In statistical physics x is state of the system (e.g. locations and velocities of all molecules), $f(x)$ is energy of the system in state x , P_X is the Gibbs distribution and $\beta = \frac{1}{T}$ is the inverse temperature of the system. In thermodynamic equilibrium, $P_X(x)$ gives fraction of time system spends in state x .

- (b) Show that $\frac{dH^*(E)}{dE} = \beta(E)$.
- (c) Next consider two functions f_0, f_1 (i.e. two types of molecules with different state-energy relations). Show that for $E \geq \min_{x_0} f(x_0) + \min_{x_1} f(x_1)$ we have

$$\max_{\mathbb{E}[f_0(X_0) + f_1(X_1)] \leq E} H(X_0, X_1) = \max_{E_0 + E_1 \leq E} H_0^*(E_0) + H_1^*(E_1) \quad (\text{III.9})$$

where $H_j^*(E) = \max_{\mathbb{E}[f_j(X)] \leq E} H(X)$.

- (d) Further, show that for the optimal choice of E_0 and E_1 in (III.9) we have

$$\beta_0(E_0) = \beta_1(E_1) \quad (\text{III.10})$$

or equivalently that the optimal distribution P_{X_0, X_1} is given by

$$P_{X_0, X_1}(a, b) = \frac{1}{Z_0(\beta)Z_1(\beta)} e^{-\beta(f_0(a) + f_1(b))} \quad (\text{III.11})$$

Remark: (III.11) also just follows from part 1 by taking $f(x_0, x_1) = f_0(x_0) + f_1(x_1)$. The point here is relation (III.10): when two thermodynamical systems are brought in contact with each other, the energy distributes among them in such a way that β parameters (temperatures) equalize.

III.17 (Importance Sampling [62]) Let μ and ν be two probability measures on set \mathcal{X} . Assume that $\nu \ll \mu$. Let $L = D(\nu \| \mu)$ and $\rho = \frac{d\nu}{d\mu}$ be the Radon-Nikodym derivative. Let $f: \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function. We would like to estimate $\mathbb{E}_\nu f$ using samples from μ .

Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mu$ and $I_n(f) = \frac{1}{n} \sum_{1 \leq i \leq n} f(X_i) \rho(X_i)$. Prove the following.

- (a) For $n \geq \exp(L+t)$ with $t \geq 0$, we have

$$\mathbb{E}|I_n(f) - \mathbb{E}_\nu f| \leq \|f\|_{L^2(\nu)} \left(\exp(-t/4) + 2\sqrt{\mathbb{P}_\mu(\log \rho > L+t/2)} \right).$$

Hint: Let $h = f \mathbf{1}\{\rho \leq \exp(L+t/2)\}$. Use triangle inequality and bound $\mathbb{E}|I_n(h) - \mathbb{E}_\nu h|$, $\mathbb{E}|I_n(h) - I_n(f)|$, $|\mathbb{E}_\nu f - \mathbb{E}_\nu h|$ separately.

(b) On the other hand, for $n \leq \exp(L - t)$ with $t \geq 0$, we have

$$\mathbb{P}(I_n(1) \geq 1 - \delta) \leq \exp(-t/2) + \frac{\mathbb{P}_\mu(\log \rho \leq L - t/2)}{1 - \delta},$$

for all $\delta \in (0, 1)$, where 1 is the constant-1 function.

Hint: Divide into two cases depending on whether $\max_{1 \leq i \leq n} \rho(X_i) \leq \exp(L - t/2)$.

This shows that a sample of size $\exp(D(\nu\|\mu) + \Theta(1))$ is both necessary and sufficient for accurate estimation by importance sampling.

III.18 *M-ary hypothesis testing.*⁸ The following result [190] generalizes Corollary 16.1 on the best average probability of error for testing two hypotheses to multiple hypotheses.

Fix a collection of distributions $\{P_1, \dots, P_M\}$. Conditioned on θ , which takes value i with probability $\pi_i > 0$ for $i = 1, \dots, M$, let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$. Denote the optimal average probability of error by $p_n^* = \inf \mathbb{P}[\hat{\theta} \neq \theta]$, where the infimum is taken over all decision rules $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$.

(a) Show that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{p_n^*} = \min_{1 \leq i < j \leq M} C(P_i, P_j), \quad (\text{III.12})$$

where C is the Chernoff information defined in (16.2).

(b) It is clear that the optimal decision rule is the Maximum a Posteriori (MAP) rule. Does maximum likelihood rule also achieve the optimal exponent (III.12)? Prove or disprove it.

III.19 Given n observations $(X_1, Y_1), \dots, (X_n, Y_n)$, where each observation consists of a pair of random variables, we want to test the following hypothesis:

$$\begin{aligned} H_0 : (X_i, Y_i) &\stackrel{\text{i.i.d.}}{\sim} P \times Q \\ H_1 : (X_i, Y_i) &\stackrel{\text{i.i.d.}}{\sim} Q \times P \end{aligned}$$

where \times denotes product distribution as usual.

(a) Show that the Stein exponent $D(P \times Q \| Q \times P)$ is equal to $D(P \| Q) + D(Q \| P)$.

(b) Show that the Chernoff exponent $C(P \times Q, Q \times P)$ is equal to $-2 \log(1 - \frac{H^2(P, Q)}{2}) = -2 \log \int \sqrt{dP dQ}$, where $H(P, Q)$ is the Hellinger distance – cf. (7.5).

Comment: This type of hypothesis testing arises in the context of community detection and stochastic block model, where n nodes indexed by $i \in [n]$ are partitioned into two communities (labeled by $\sigma_i = +$ and $\sigma_i = -$ uniformly and independently). The task is to classify the nodes based on the pairwise observations $W = (W_{ij} : 1 \leq i < j \leq n)$ are independent conditioned on σ_i 's and $W_{ij} \sim P$ if $\sigma_i = \sigma_j$ and Q otherwise. As a means to prove the impossibility result [322], consider the setting where an oracle reveals all labels except for σ_1 . Define $S_+ = \{j = 2, \dots, n : \sigma_j = +\}$ and similarly S_- . If $\sigma_1 = +$, $\{W_{1,j} : j \in S_+\} \stackrel{\text{i.i.d.}}{\sim} P$ and $\{W_{1,j} : j \in S_-\} \stackrel{\text{i.i.d.}}{\sim} Q$ and vice versa if $\sigma_1 = -$.

⁸ Not to be confused with multiple testing in the statistics literature, which refers to testing multiple pairs of binary hypotheses simultaneously.

280 Exercises for Part III

III.20 (Stochastic dominance and robust LRT) Let $\mathcal{P}_0, \mathcal{P}_1$ be two families of probability distributions on \mathcal{X} . Suppose that there is a *least favorable pair* (LFP) $(Q_0, Q_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ such that

$$\begin{aligned} Q_0[\pi > t] &\geq Q'_0[\pi > t] \\ Q_1[\pi > t] &\leq Q'_1[\pi > t], \end{aligned}$$

for all $t \geq 0$ and $Q'_i \in \mathcal{P}_i$, where $\pi = dQ_1/dQ_0$. Prove that (Q_0, Q_1) simultaneously minimizes all f -divergences between \mathcal{P}_0 and \mathcal{P}_1 , i.e.

$$D_f(Q_1 \| Q_0) \leq D_f(Q'_1 \| Q'_0) \quad \forall Q'_0 \in \mathcal{P}_0, Q'_1 \in \mathcal{P}_1. \quad (\text{III.13})$$

Hint: Interpolate between (Q_0, Q_1) and (Q'_0, Q'_1) and differentiate.

Remark: For the case of two TV-balls, i.e. $\mathcal{P}_i = \{Q : \text{TV}(Q, P_i) \leq \epsilon\}$, the existence of LFP is shown in [157], in which case $\pi = \min(c', \max(c'', \frac{dP_0}{dP_1}))$ for some $0 \leq c' < c'' \leq \infty$ giving the *robust likelihood-ratio test*.

Part IV

Channel coding



In this Part we study a new type of problem known as “channel coding”. Historically, this was the first application area of information theory that lead to widely recognized and surprising results [268]. To explain the relation of this Part to others, let us revisit what problems we have studied so far.

In Part II our objective was data compression. The main object there was a single distribution P_X and the fundamental limit $\mathbb{E}[\ell(f^*(X))]$ – the minimal compression length. The main result was connection between the fundamental limit and an information quantity, that we can summarize as

$$\mathbb{E}[\ell(f^*(X))] \approx H(X)$$

In Part III we studied binary hypothesis testing. There the main object was a pair of distributions (P, Q) , the fundamental limit was the Neyman-Pearson curve $\beta_{1-\epsilon}(P^n, Q^n)$ and the main result

$$\beta_{1-\epsilon}(P^n, Q^n) \approx \exp\{-nD(P||Q)\},$$

again connecting an operational quantity to an information measure.

In channel coding – the topic of this Part – the main object is going to be a channel $P_{Y|X}$. The fundamental limit is $M^*(\epsilon)$, the maximum number of messages that can be transmitted with probability of error at most ϵ , which we rigorously define in this chapter. Our main result in this part is to show the celebrated Shannon’s noisy channel coding theorem:

$$\log M^*(\epsilon) \approx \max_{P_X} I(X; Y).$$

17 Error correcting codes

17.1 Codes and probability of error

We start with a simple definition of a code.

Definition 17.1. An M -code for $P_{Y|X}$ is an encoder/decoder pair (f, g) of (randomized) functions¹

- encoder $f : [M] \rightarrow \mathcal{X}$
- decoder $g : \mathcal{Y} \rightarrow [M] \cup \{\text{e}\}$

In most cases f and g are deterministic functions, in which case we think of them, equivalently, in terms of codewords, codebooks, and decoding regions (see Fig. 17.1 for an illustration)

- $\forall i \in [M] : c_i \triangleq f(i)$ are *codewords*, the collection $\mathcal{C} = \{c_1, \dots, c_M\}$ is called a *codebook*.
- $\forall i \in [M], D_i \triangleq g^{-1}(\{i\})$ is the *decoding region* for i .

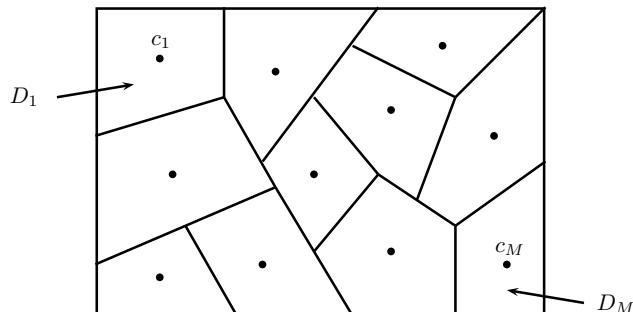


Figure 17.1 When $\mathcal{X} = \mathcal{Y}$, the decoding regions can be pictured as a partition of the space, each containing one codeword.

Given an M -code we can define a probability space, underlying all the subsequent developments in this Part. For that we chain the three objects – message W , the encoder and the decoder – together

¹ For randomized encoder/decoders, we identify f and g as probability transition kernels $P_{X|W}$ and $P_{\hat{W}|Y}$.

17.1 Codes and probability of error 285

into the following Markov chain:

$$W \xrightarrow{f} X \xrightarrow{P_{Y|X}} Y \xrightarrow{g} \hat{W} \quad (17.1)$$

where we set $W \sim \text{Unif}([M])$. In the case of discrete spaces, we can explicitly write out the joint distribution of these variables as follows:

$$\begin{aligned} (\text{general}) \quad P_{W,X,Y,\hat{W}}(m, a, b, \hat{m}) &= \frac{1}{M} P_{X|W}(a|m) P_{Y|X}(b|a) P_{\hat{W}|Y}(\hat{m}|b) \\ (\text{deterministic } f, g) \quad P_{W,X,Y,\hat{W}}(m, c_m, b, \hat{m}) &= \frac{1}{M} P_{Y|X}(b|c_m) \mathbb{1}\{b \in D_{\hat{m}}\} \end{aligned}$$

Throughout these sections, these random variables will be referred to by their traditional names: W – original (true) message, X - (induced) channel input, Y - channel output and \hat{W} - decoded message.

Although any pair (f, g) is called an M -code, in reality we are only interested in those that satisfy certain “error-correcting” properties. To assess their quality we define the following *performance metrics*:

- 1 *Maximum error probability*: $P_{e,\max}(f, g) \triangleq \max_{m \in [M]} \mathbb{P}[\hat{W} \neq m | W = m]$.
- 2 *Average error probability*: $P_e(f, g) \triangleq \mathbb{P}[W \neq \hat{W}]$.

Note that, clearly, $P_e \leq P_{e,\max}$. Therefore, requirement of the small maximum error probability is a more stringent criterion, and offers uniform protection for all codewords. Some codes (such as linear codes, see Section 18.6) have the property of $P_e = P_{e,\max}$ by construction, but generally these two metrics could be very different.

Having defined the concept of an M -code and the performance metrics, we can finally define the *fundamental limits* for a given channel $P_{Y|X}$.

Definition 17.2. A code (f, g) is an (M, ϵ) -code for $P_{Y|X}$ if $P_e(f, g) \leq \epsilon$. Similarly, an $(M, \epsilon)_{\max}$ -code must satisfy $P_{e,\max} \leq \epsilon$. The fundamental limits of channel coding are defined as

$$\begin{aligned} M^*(\epsilon; P_{Y|X}) &= \max\{M : \exists (M, \epsilon)\text{-code}\} \\ M_{\max}^*(\epsilon; P_{Y|X}) &= \max\{M : \exists (M, \epsilon)_{\max}\text{-code}\} \end{aligned}$$

The argument $P_{Y|X}$ will be omitted when $P_{Y|X}$ is clear from the context.

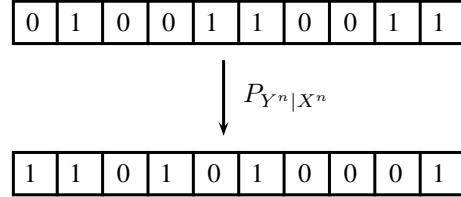
In other words, the quantity $\log_2 M^*(\epsilon)$ gives the maximum number of bits that we can push through a noisy transformation $P_{Y|X}$, while still guaranteeing the error probability in the appropriate sense to be at most ϵ .

Example 17.1. The channel $\text{BSC}_{\delta}^{\otimes n}$ (recall from Example 3.5 that BSC stands for binary symmetric channel) acts between $\mathcal{X} = \{0, 1\}^n$ and $\mathcal{Y} = \{0, 1\}^n$, where the input X^n is contaminated by additive noise $Z^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\delta)$ independent of X^n , resulting in the channel output

$$Y^n = X^n \oplus Z^n.$$

286

In other words, the $\text{BSC}_{\delta}^{\otimes n}$ channel takes a binary sequence length n and flips each bit independently with probability δ ; pictorially,



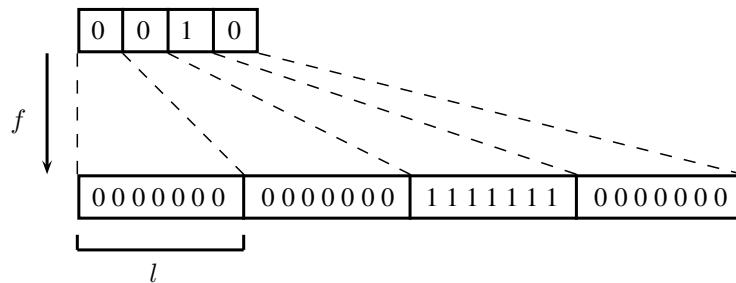
In the next section we discuss coding for the BSC channel in more detail.

17.2 Coding for Binary Symmetric Channels

To understand the problem of designing the encoders and decoders, let us consider the BSC transformation with $\delta = 0.11$ and $n = 1000$. The problem of studying $\log_2 M^*(\epsilon)$ attempts to answer what is the maximum number k of bits you can send with $P_e \leq \epsilon$? For concreteness, let us fix $\epsilon = 10^{-3}$ and discuss some of the possible ideas.

Perhaps our first attempt would be to try sending $k = 1000$ bits with one data bit mapped to one channel input position. However, a simple calculation shows that in this case we get $P_e = 1 - (1 - \delta)^n \approx 1$. In other word, the *uncoded transmission* does not meet our objective of small P_e and some form of coding is necessary. This incurs a fundamental tradeoff: reduce the number of bits to send (and use the freed channel inputs for sending redundant copies) in order to increase the probability of success.

So let us consider the next natural idea: the *repetition coding*. We take each of the input data bits and repeat it ℓ times:



Decoding can be done by taking a majority vote inside each ℓ -block. Thus, each data bit is decoded with probability of bit error $P_b = \mathbb{P}[\text{Binom}(l, \delta) > l/2]$. However, the probability of block error of this scheme is $P_e \leq k\mathbb{P}[\text{Binom}(l, \delta) > l/2]$. (This bound is essentially tight in the current regime). Consequently, to satisfy $P_e \leq 10^{-3}$ we must solve for k and ℓ satisfying $kl \leq n = 1000$ and also

$$k\mathbb{P}[\text{Binom}(l, \delta) > l/2] \leq 10^{-3}.$$

17.2 Coding for Binary Symmetric Channels 287

This gives $l = 21$, $k = 47$ bits. So we can see that using repetition coding we can send 47 data bits by using 1000 channel uses.

Repetition coding is a natural idea. It also has a very natural tradeoff: if you want better reliability, then the number ℓ needs to increase and hence the ratio $\frac{k}{n} = \frac{1}{\ell}$ should drop. Before Shannon's groundbreaking work, it was almost universally accepted that *this is fundamentally unavoidable: vanishing error probability should imply vanishing communication rate $\frac{k}{n}$* .

Before delving into optimal codes let us offer a glimpse of more sophisticated ways of injecting redundancy into the channel input n -sequence than simple repetition. For that, consider the so-called first-order Reed-Muller codes $(1, r)$. We interpret a sequence of r data bits $a_0, \dots, a_{r-1} \in \mathbb{F}_2^r$ as a degree-1 polynomial in $(r - 1)$ variables:

$$a = (a_0, \dots, a_{r-1}) \mapsto f_a(x) \triangleq \sum_{i=1}^{r-1} a_i x_i + a_0.$$

In order to transmit these r bits of data we simply evaluate $f_a(\cdot)$ at all possible values of the variables $x^{r-1} \in \mathbb{F}_2^{r-1}$. This code, which maps r bits to 2^{r-1} bits, has minimum distance $d_{min} = 2^{r-2}$. That is, for two distinct $a \neq a'$ the number of positions in which f_a and $f_{a'}$ disagree is at least 2^{r-2} . In coding theory notation $[n, k, d_{min}]$ we say that the first-order Reed-Muller code $(1, 7)$ is a $[64, 7, 32]$ code. It can be shown that the optimal decoder for this code achieves over the $\text{BSC}_{0.11} \otimes 64$ channel a probability of error at most $6 \cdot 10^{-6}$. Thus, we can use 16 such blocks (each carrying 7 data bits and occupying 64 bits on the channel) over the $\text{BSC}_{\delta} \otimes 1024$, and still have (by the union bound) overall probability of block error $P_e \lesssim 10^{-4} < 10^{-3}$. Thus, with the help of Reed-Muller codes we can send $7 \cdot 16 = 112$ bits in 1024 channel uses, more than doubling that of the repetition code.

Shannon's noisy channel coding theorem (Theorem 19.8) – a crown jewel of information theory – tells us that over memoryless channel $P_{Y^n|X^n} = (P_{Y|X})^n$ of blocklength n the fundamental limit satisfies

$$\log M^*(\epsilon; P_{Y^n|X^n}) = nC + o(n) \quad (17.2)$$

as $n \rightarrow \infty$ and for arbitrary $\epsilon \in (0, 1)$. Here $C = \max_{P_{X_1}} I(X_1; Y_1)$ is the capacity of the single-letter channel. In our case of BSC we have

$$C = \log 2 - h(\delta) \approx \frac{1}{2} \text{ bit},$$

since the optimal input distribution is uniform (from symmetry) – see Section 19.3. Shannon's expansion (17.2) can be used to predict (not completely rigorously, of course, because of the $o(n)$ residual) that it should be possible to send around 500 bits reliably. As it turns out, for the blocklength $n = 1000$ this is not quite possible.

Note that computing M^* exactly requires iterating over all possible encoders and decoder – an impossible task even for small values of n . However, there exist rigorous and computationally tractable finite blocklength bounds [231] that demonstrate for our choice of $n = 1000, \delta = 0.11, \epsilon = 10^{-3}$:

$$414 \leq \log_2 M^* \leq 416 \text{ bits} \quad (17.3)$$

Thus we can see that Shannon's prediction is about 20% too optimistic. We will see below some of such finite-length bounds. Notice, however, that while the guarantee existence of an encoder-decoder pair achieving a prescribed performance, building an actual f and g implementable with a modern software/hardware is a different story.

It took about 60 years after Shannon's discovery of (17.2) to construct practically implementable codes achieving that performance. The first codes that approach the bounds on $\log M^*$ are called *Turbo codes* [30] (after the turbocharger engine, where the exhaust is fed back in to power the engine). This class of codes is known as *sparse graph codes*, of which the low-density parity check (LDPC) codes invented by Gallager are particularly well studied [255]. As a rule of thumb, these codes typically approach 80...90% of $\log M^*$ when $n \approx 10^3 \dots 10^4$. For shorter blocklengths in the range of $n = 100 \dots 1000$ there is an exciting alternative to LDPC codes: the polar codes of Arikan [15], which are most typically used together with the list-decoding idea of Tal and Vardy [293]. And of course, the story is still evolving today as new channel models become relevant and new hardware possibilities open up.

We wanted to point out a subtle but very important conceptual paradigm shift introduced by Shannon's insistence on coding over many (information) bits together. Indeed, consider the situation discussed above, where we constructed a powerful code with $M \approx 2^{400}$ codewords and $n = 1000$. Now, one might imagine this code as a constellation of 2^{400} points carefully arranged inside a hypercube $\{0, 1\}^{1000}$ to guarantee some degree of separation between them, cf. (17.6). Next, suppose one was using this code every second for the lifetime of the universe ($\approx 10^{18}$ sec). Yet, even after this laborious process she will have explored at most 2^{60} different codewords from among an overwhelmingly large codebook 2^{400} . So a natural question arises: why did we need to carefully place all these many codewords if majority of them will never be used by anyone? The answer is at the heart of the concept of information: to transmit information is to convey a selection of one element (W) from a collection of possibilities ($[M]$). The fact that we do not know which W will be selected forces us to apriori prepare for every one of the possibilities. This simple idea, proposed in the first paragraph of [268], is now tacitly assumed by everyone, but was one of the subtle ways in which Shannon revolutionized scientific approach to the study of information exchange.

17.3 Optimal decoder

Given any encoder $f : [M] \rightarrow \mathcal{X}$, the decoder that minimizes P_e is the *Maximum A Posteriori (MAP)* decoder, or equivalently, the *Maximal Likelihood (ML)* decoder, since the codewords are equiprobable (W is uniform):

$$\begin{aligned} g^*(y) &= \operatorname{argmax}_{m \in [M]} \mathbb{P}[W = m | Y = y] \\ &= \operatorname{argmax}_{m \in [M]} \mathbb{P}[Y = y | W = m] \end{aligned} \tag{17.4}$$

Notice that the optimal decoder is deterministic. For the special case of deterministic encoder, where we can identify the encoder with its image \mathcal{C} the minimal (MAP) probability of error for

the codebook \mathcal{C} can be written as

$$P_{e,MAP}(\mathcal{C}) = 1 - \frac{1}{M} \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{C}} P_{Y|X}(y|x), \quad (17.5)$$

with a similar extension to non-discrete \mathcal{Y} .

Remark 17.1. For the case of $\text{BSC}_{\delta}^{\otimes n}$ MAP decoder has a nice geometric interpretation. Indeed, if $d_H(x^n, y^n) = |\{i : x_i \neq y_i\}|$ denotes the Hamming distance and if f (the encoder) is deterministic with codewords $\mathcal{C} = \{c_i, i \in [M]\}$ then

$$g^*(y^n) = \operatorname{argmin}_{m \in [M]} d_H(c_m, y^n). \quad (17.6)$$

Consequently, the optimal decoding regions – see Fig. 17.1 – become the *Voronoi cells* tesselating the Hamming space $\{0, 1\}^n$. Similarly, the MAP decoder for the AWGN channel induces a Voronoi tessellation of \mathbb{R}^n – see Section 20.3.

So we have seen that the optimal decoder is without loss of generality can be assumed to be deterministic. Similarly, we can represent any randomized encoder f as a function of two arguments: the true message W and an external randomness $U \perp\!\!\!\perp W$, so that $X = f(W, U)$ where this time f is a deterministic function. Then we have

$$\mathbb{P}[W \neq \hat{W}] = \mathbb{E}[\mathbb{P}[W \neq \hat{W}|U]],$$

which implies that if $P[W \neq \hat{W}] \leq \epsilon$ then there must exist some choice u_0 such that $\mathbb{P}[W \neq \hat{W}|U = u_0] \leq \epsilon$. In other words, the fundamental limit $M^*(\epsilon)$ is unchanged if we restrict our attention to deterministic encoders and decoders only.

Note, however, that neither of the above considerations apply to the maximal probability of error $P_{e,max}$. Indeed, the fundamental limit $M_{\max}^*(\epsilon)$ does indeed require considering randomized encoders and decoders. For example, when $M = 2$ from the decoding point of view we are back to the setting of binary hypotheses testing in Part III. The optimal decoder (test) that minimizes the maximal Type-I and II error probability, i.e., $\max\{1 - \alpha, \beta\}$, will not be deterministic if $\max\{1 - \alpha, \beta\}$ is not achieved at a vertex of the Neyman-Pearson region $\mathcal{R}(P_{Y|W=1}, P_{Y|W=2})$.

17.4 Weak converse bound

The main focus of both theory and practice of channel coding lies in showing existence (or constructing explicit) (M, ϵ) codes with large M and small ϵ . To understand how close the constructed code is to the fundamental limit, one needs to prove an “impossibility result” bounding M from the above or ϵ from below. Such results are known as “converse bounds”, with the name coming from the fact that classically such bounds followed right after the existential results and were preceded with the words “Conversely, ...”.

Theorem 17.3 (Weak converse). Any (M, ϵ) -code for $P_{Y|X}$ satisfies

$$\log M \leq \frac{\sup_{P_X} I(X; Y) + h(\epsilon)}{1 - \epsilon},$$

where $h(x) = H(\text{Ber}(x))$ is the binary entropy function.

Proof. This can be derived as a one-line application of Fano's inequality (Theorem 6.4), but we proceed slightly differently. Consider an M -code with probability of error P_e and its corresponding probability space: $W \rightarrow X \rightarrow Y \rightarrow \hat{W}$. We want to show that this code can be used as a hypothesis test between distributions $P_{X,Y}$ and $P_X P_Y$. Indeed, given a pair (X, Y) we can sample (W, \hat{W}) from $P_{W,\hat{W}|X,Y} = P_{W|X} P_{\hat{W}|Y}$ and compute the binary value $Z = 1\{W \neq \hat{W}\}$. (Note that in the most interesting cases when encoder and decoder are deterministic and the encoder is injective, the value Z is a deterministic function of (X, Y) .) Let us compute performance of this binary hypothesis test under two hypotheses. First, when $(X, Y) \sim P_X P_Y$ we have that $\hat{W} \perp\!\!\!\perp W \sim \text{Unif}([M])$ and therefore:

$$P_X P_Y[Z = 1] = \frac{1}{M}.$$

Second, when $(X, Y) \sim P_{X,Y}$ then by definition we have

$$P_{X,Y}[Z = 1] = 1 - P_e.$$

Thus, we can now apply the data-processing inequality for divergence to conclude: Since $W \rightarrow X \rightarrow Y \rightarrow \hat{W}$, we have the following chain of inequalities (cf. Fano's inequality Theorem 6.4):

$$\begin{aligned} D(P_{X,Y} \| P_X P_Y) &\stackrel{\text{DPI}}{\geq} d(1 - P_e \| \frac{1}{M}) \\ &\geq -h(\mathbb{P}[W \neq \hat{W}]) + (1 - P_e) \log M \end{aligned}$$

By noticing that the left-hand side is $I(X; Y) \leq \sup_{P_X} I(X; Y)$ we obtain:

$$\log M \leq \frac{\sup_{P_X} I(X; Y) + h(P_e)}{1 - P_e},$$

and the proof is completed by checking that $p \mapsto \frac{h(p)}{1-p}$ is monotonically increasing. \square

Remark 17.2. The bound can be significantly improved by considering other divergence measures in the data-processing step. In particular, we will see below how one can get “strong” converse (explaining the term “weak” converse here as well) in Section 22.1. The proof technique is known as meta-converse, see Section 22.3.

18 Random and maximal coding

So far our discussion of channel coding was mostly following the same lines as the M -ary hypothesis testing (HT) in statistics. In this chapter we introduce the key departure: the principal and most interesting goal in information theory is the design of the encoder $f : [M] \rightarrow \mathcal{X}$ or the codebook $\{c_i \triangleq f(i), i \in [M]\}$. Once the codebook is chosen, the problem indeed becomes that of M -ary HT and can be tackled by the standard statistical methods. However, the task of choosing the encoder f has no exact analogs in statistical theory (the closest being design of experiments). Each f gives rise to a different HT problem and the goal is to choose these M hypotheses $P_{X|c_1}, \dots, P_{X|c_M}$ to ensure maximal testability. It turns out that the problem of choosing a good f will be much simplified if we adopt a suboptimal way of testing M -ary HT. Namely, roughly speaking we will run M binary HTs testing $P_{Y|X=c_m}$ against P_Y , which tries to distinguish the channel output induced by the message m from an “average background noise” P_Y . An optimal such test, as we know from Neyman-Pearson (Theorem 14.11), thresholds the following quantity

$$\log \frac{P_{Y|X=x}}{P_Y}$$

This explains the central role played by the information density (see below) in these achievability bounds.

In this chapter it will be convenient to introduce the following *independent pairs* $(X, Y) \perp\!\!\!\perp (\bar{X}, \bar{Y})$ with their joint distribution given by:

$$P_{X,Y,\bar{X},\bar{Y}}(a, b, \bar{a}, \bar{b}) = P_X(a)P_{Y|X}(b|a)P_X(\bar{a})P_{Y|X}(\bar{b}|\bar{a}). \quad (18.1)$$

We will often call X the sent codeword and \bar{X} the unsent codeword.

18.1 Information density

A crucial object for the subsequent development is the information density. Informally speaking, we simply want to define $i(x; y) = \log \frac{dP_{X,Y}}{dP_X P_Y}(x, y)$. However, we want to make this definition sufficiently general so as to take into account both the possibility of $P_{X,Y} \ll P_X P_Y$ (in which case the value under the log can equal $+\infty$) and the possibility of argument of the log being equal to 0. The definition below is similar to what we did in Definition 14.5 and (2.10), but we repeat it below for convenience.

Definition 18.1 (Information density). Let $P_{X,Y} \ll \mu$ and $P_X P_Y \ll \mu$ for some dominating measure μ , and denote by $f(x,y) = \frac{dP_{X,Y}}{d\mu}$ and $\bar{f}(x,y) = \frac{dP_X P_Y}{d\mu}$ the Radon-Nikodym derivatives of $P_{X,Y}$ and $P_X P_Y$ with respect to μ , respectively. Then recalling the Log definition (2.10) we set

$$i_{P_{X,Y}}(x; y) \triangleq \text{Log} \frac{f(x,y)}{\bar{f}(x,y)} = \begin{cases} \log \frac{f(x,y)}{\bar{f}(x,y)}, & f(x,y) > 0, \bar{f}(x,y) > 0 \\ +\infty, & f(x,y) > 0, \bar{f}(x,y) = 0 \\ -\infty, & f(x,y) = 0, \bar{f}(x,y) > 0 \\ 0, & f(x,y) = \bar{f}(x,y) = 0, \end{cases} \quad (18.2)$$

Note that when $P_{X,Y} \ll P_X P_Y$ we have simply

$$i_{P_{X,Y}}(x; y) = \log \frac{dP_{X,Y}}{dP_X P_Y}(x, y),$$

with $\log 0 = -\infty$.

Notice that the information density as a function depends on the underlying $P_{X,Y}$. Throughout this Part, however, the $P_{Y|X}$ is going to be a fixed channel (fixed by the problem at hand), and thus information density only depends on the choice of P_X . Most of the time P_X (and, correspondingly, the $P_{X,Y}$) used to define information density will be apparent from the context. Thus for the benefit of the reader as well as our own, we will write $i(x; y)$ dropping the subscript $P_{X,Y}$.

Information density is a natural concept for understanding the decoding process. We will see shortly that what our decoders will do is threshold information density. We wanted to briefly give intuition for this idea. First, consider, for simplicity, the case of discrete alphabets and $P_{X,Y} \ll P_X P_Y$. Then we have an equivalent expression

$$i(x; y) = \log \frac{P_{Y|X}(y|x)}{P_Y(y)}.$$

Therefore, the optimal (maximum likelihood) decoder can be written in terms of the information density:

$$\begin{aligned} g^*(y) &= \underset{m \in [M]}{\text{argmax}} P_{X|Y}(c_m|y) \\ &= \underset{m \in [M]}{\text{argmax}} P_{Y|X}(y|c_m) \\ &= \underset{m \in [M]}{\text{argmax}} i(c_m; y). \end{aligned} \quad (18.3)$$

Note an important observation: (18.3) holds regardless of the input distribution P_X used for the definition of $i(x; y)$, in particular we do not have to use the code-induced distribution $P_X = \frac{1}{M} \sum_{i=1}^M \delta_{c_i}$. However, if we are to threshold information density, different choices of P_X will result in different decoders, so we need to justify the choice of P_X .

To that end, recall that to distinguish between two codewords c_i and c_j , one can apply (as we learned in Part III for binary HT) the likelihood ratio test, namely thresholding the LLR $\log \frac{P_{Y|X=c_i}}{P_{Y|X=c_j}}$. As we explained at the beginning of this Part, a (possibly suboptimal) approach in M -ary HT is to run binary tests by thresholding each information density $i(c_i; y)$. This, loosely speaking,

18.1 Information density 293

evaluates the likelihood of c_i against the average distribution of the other $M - 1$ codewords, which we approximate by P_Y (as opposed to the more precise form $\frac{1}{M-1} \sum_{j \neq i} P_{Y|X=c_j}$). Putting these ideas together we can propose the decoder as

$$g(y) = \text{any } m \text{ s.t. } i(c_m; y) > \gamma,$$

where λ is a threshold and P_X is judiciously chosen (to maximize $I(X; Y)$ as we will see soon).

We proceed to show some elementary properties of the information density. The next result explains the name “information density”¹

Proposition 18.2. *The expectation $\mathbb{E}[i(X; Y)]$ is well-defined and non-negative (but possibly infinite). In any case, we have $I(X; Y) = \mathbb{E}[i(X; Y)]$.*

Proof. This follows from (2.12) and the definition of $i(x; y)$ as log-ratio. \square

Being defined as log-likelihood, information density possesses the standard properties of the latter, cf. Theorem 14.6. However, because it's defined in terms of two variables (X, Y) , there are also very useful conditional expectation versions. To illustrate the meaning of the next proposition, let us consider the case of discrete X, Y and $P_{X,Y} \ll P_X P_Y$. Then we have *for every* x :

$$\sum_y f(x, y) P_X(x) P_Y(y) = \sum_y f(x, y) \exp\{-i(x; y)\} P_{X,Y}(x, y).$$

The general case requires a little more finesse.

Proposition 18.3 (Conditioning-unconditioning trick). *Let $\bar{X} \perp\!\!\!\perp (X, Y)$ be a copy of X . We have the following:*

1 *For any function $f: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$*

$$\mathbb{E}[f(\bar{X}, Y) \mathbf{1}\{i(\bar{X}; Y) > -\infty\}] = \mathbb{E}[f(X, Y) \exp\{-i(X; Y)\}]. \quad (18.4)$$

2 *Let f_+ be a non-negative function. Then for P_X -almost every x we have*

$$\mathbb{E}[f_+(\bar{X}, Y) \mathbf{1}\{i(\bar{X}; Y) > -\infty\} | \bar{X} = x] = \mathbb{E}[f_+(X, Y) \exp\{-i(X; Y)\} | X = x] \quad (18.5)$$

Proof. The first part (18.4) is simply a restatement of (14.5). For the second part, let us define

$$a(x) \triangleq \mathbb{E}[f_+(\bar{X}, Y) \mathbf{1}\{i(\bar{X}; Y) > -\infty\} | \bar{X} = x], \quad b(x) \triangleq \mathbb{E}[f_+(X, Y) \exp\{-i(X; Y)\} | X = x]$$

We first additionally assume that f is bounded. Fix $\epsilon > 0$ and denote $S_\epsilon = \{x : a(x) \geq b(x) + \epsilon\}$. As $\epsilon \rightarrow 0$ we have $S_\epsilon \nearrow \{x : a(x) > b(x)\}$ and thus if we show $P_X[S_\epsilon] = 0$ this will imply that $a(x) \leq b(x)$ for P_X -a.e. x . The symmetric argument shows $b(x) \leq a(x)$ and completes the proof of the equality.

¹ Still an unfortunate name for a quantity that can be negative, though.

To show $P_X[S_\epsilon] = 0$ let us apply (18.4) to the function $f(x, y) = f_+(x, y)1\{x \in S_\epsilon\}$. Then we get

$$\mathbb{E}[f_+(X, Y)1\{X \in S_\epsilon\} \exp\{-i(X; Y)\}] = \mathbb{E}[f_+(\bar{X}, Y)1\{i(\bar{X}; Y) > -\infty\}1\{\bar{X} \in S_\epsilon\}] .$$

Let us re-express both sides of this equality by taking the conditional expectations over Y to get:

$$\mathbb{E}[b(X)1\{X \in S_\epsilon\}] = \mathbb{E}[a(\bar{X})1\{\bar{X} \in S_\epsilon\}] .$$

But from the definition of S_ϵ we have

$$\mathbb{E}[b(X)1\{X \in S_\epsilon\}] \geq \mathbb{E}[(b(\bar{X}) + \epsilon)1\{\bar{X} \in S_\epsilon\}] .$$

Recall that $X \stackrel{(d)}{=} \bar{X}$ and hence

$$\mathbb{E}[b(X)1\{X \in S_\epsilon\}] \geq \mathbb{E}[b(X)1\{X \in S_\epsilon\}] + \epsilon P_X[S_\epsilon] .$$

Since f_+ (and therefore b) was assumed to be bounded we can cancel the common term from both sides and conclude $P_X[S_\epsilon] = 0$ as required.

Finally, to show (18.5) in full generality, given an unbounded f_+ we define $f_n(x, y) = \min(f_+(x, y), n)$. Since (18.5) holds for f_n we can take limit as $n \rightarrow \infty$ on both sides of it:

$$\lim_{n \rightarrow \infty} \mathbb{E}[f_n(\bar{X}, Y)1\{i(\bar{X}; Y) > -\infty\}|\bar{X} = x] = \lim_{n \rightarrow \infty} \mathbb{E}[f_n(X, Y) \exp\{-i(X; Y)\}|X = x]$$

By the monotone convergence theorem (for conditional expectations!) we can take the limits inside the expectations to conclude the proof. \square

Corollary 18.4. *For P_X -almost every x we have*

$$\mathbb{P}[i(x; Y) > t] \leq \exp(-t), \quad (18.6)$$

$$\mathbb{P}[i(\bar{X}; Y) > t] \leq \exp(-t) \quad (18.7)$$

Proof. Pick $f_+(x, y) = 1\{i(x; y) > t\}$ in (18.5). \square

Remark 18.1. This estimate has been used by us several times before. In the hypothesis testing part we used (Corollary 14.2):

$$Q\left[\log \frac{dP}{dQ} \geq t\right] \leq \exp(-t). \quad (18.8)$$

In data compression, we used the fact that $|\{x : \log P_X(x) \geq t\}| \leq \exp(-t)$, which is also of the form (18.8) with Q being the counting measure.

18.2 Shannon's random coding bound

In this section we present perhaps the most virtuous technical result of Shannon. As we discussed before, good error correcting code is supposed to be a geometrically elegant constellation in a high-dimensional space. Its chief goal is to push different codewords as far apart as possible, so as to reduce the deleterious effects of channel noise. However, in early 1940's there were no codes

18.2 Shannon's random coding bound 295

and no tools for constructing them available to Shannon. So facing the problem of understanding if error-correction is even possible, Shannon decided to check if placing codewords randomly in space will somehow result in favorable geometric arrangement. To everyone's astonishment, which is still producing aftershocks today, this method not only produced reasonable codes, but in fact turned out to be optimal asymptotically (and almost-optimal non-asymptotically [231]). We also remark that the method of proving existence of certain combinatorial objects by random selection is known as Erdős's *probabilistic method* [10], which Shannon apparently discovered independently and, perhaps, earlier.

Theorem 18.5 (Shannon's achievability bound). *Fix a channel $P_{Y|X}$ and an arbitrary input distribution P_X . Then for every $\tau > 0$ there exists an (M, ϵ) -code with*

$$\epsilon \leq \mathbb{P}[i(X; Y) \leq \log M + \tau] + \exp(-\tau). \quad (18.9)$$

Proof. Recall that for a given codebook $\{c_1, \dots, c_M\}$, the optimal decoder is MAP and is equivalent to maximizing information density, cf. (18.3). The step of maximizing the $i(c_m; Y)$ makes analyzing the error probability difficult. Similar to what we did in almost loss compression, cf. Theorem 11.6, the first important step for showing the achievability bound is to consider a suboptimal decoder. In Shannon's bound, we consider a threshold-based suboptimal decoder $g(y)$ as follows:

$$g(y) = \begin{cases} m, & \exists! c_m \text{ s.t. } i(c_m; y) \geq \log M + \tau \\ e, & \text{o.w.} \end{cases} \quad (18.10)$$

In words, decoder g reports m as decoded message if and only if codeword c_m is a unique one with information density exceeding the threshold $\log M + \tau$. If there are multiple or none such codewords, then decoder outputs a special value of e , which always results in error since $W \neq e$ ever. (We could have decreased probability of error slightly by allowing the decoder to instead output a random message, or to choose any one of the messages exceeding the threshold, or any other clever ideas. The point, however, is that even the simplistic resolution of outputting e already achieves all qualitative goals, while simplifying the analysis considerably.)

For a given codebook (c_1, \dots, c_M) , the error probability is:

$$P_e(c_1, \dots, c_M) = \mathbb{P}[\{i(c_W; Y) \leq \log M + \tau\} \cup \{\exists \bar{m} \neq W, i(c_{\bar{m}}; Y) > \log M + \tau\}]$$

where W is uniform on $[M]$ and the probability space is as in (17.1).

The second (and most ingenious) step proposed by Shannon was to forego the complicated discrete optimization of the codebook. His proposal is to generate the codebook (c_1, \dots, c_M) randomly with $c_m \sim P_X$ i.i.d. for $m \in [M]$ and then try to reason about the average $\mathbb{E}[P_e(c_1, \dots, c_M)]$. By symmetry, this averaged error probability over all possible codebooks is unchanged if we condition on $W = 1$. Considering also the random variables (X, Y, \bar{X}) as in (18.1), we get the following chain:

$$\begin{aligned} & \mathbb{E}[P_e(c_1, \dots, c_M)] \\ &= \mathbb{E}[P_e(c_1, \dots, c_M)|W = 1] \\ &= \mathbb{P}[\{i(c_1; Y) \leq \log M + \tau\} \cup \{\exists \bar{m} \neq 1, i(c_{\bar{m}}, Y) > \log M + \tau\}|W = 1] \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{P}[i(c_1; Y) \leq \log M + \tau | W = 1] + \sum_{\bar{m}=2}^M \mathbb{P}[i(c_{\bar{m}}; Y) > \log M + \tau | W = 1] \quad (\text{union bound}) \\
&\stackrel{(a)}{=} \mathbb{P}[i(X; Y) \leq \log M + \tau] + (M - 1) \mathbb{P}[i(\bar{X}; Y) > \log M + \tau] \\
&\leq \mathbb{P}[i(X; Y) \leq \log M + \tau] + (M - 1) \exp(-(\log M + \tau)) \quad (\text{by Corollary 18.1}) \\
&\leq \mathbb{P}[i(X; Y) \leq \log M + \tau] + \exp(-\tau),
\end{aligned}$$

where the crucial step (a) follows from the fact that given $W = 1$ and $\bar{m} \neq 1$ we have

$$(c_1, c_{\bar{m}}, Y) \stackrel{d}{=} (X, \bar{X}, Y)$$

with the latter triple defined in (18.1).

The last expression does indeed conclude the proof of existence of the (M, ϵ) code: it shows that the average of $P_e(c_1, \dots, c_M)$ satisfies the required bound on probability of error, and thus there must exist at least *one* choice of c_1, \dots, c_M satisfying the same bound. \square

Remark 18.2 (Joint typicality). Shortly in Chapter 19, we will apply this theorem for the case of $P_X = P_{X_1}^{\otimes n}$ (the iid input) and $P_{Y|X} = P_{Y_1|X_1}^{\otimes n}$ (the memoryless channel). Traditionally, cf. [80], decoders in such settings were defined with the help of so called ‘‘joint typicality’’. Those decoders given $y = y^n$ search for the codeword x^n (both of which are n -letter vectors) such that the empirical joint distribution is close to the true joint distribution, i.e., $\hat{P}_{x^n, y^n} \approx P_{X_1, Y_1}$, where

$$\hat{P}_{x^n, y^n}(a, b) = \frac{1}{n} \cdot |\{j \in [n] : x_j = a, y_j = b\}|$$

is the joint empirical distribution of (x^n, y^n) . This definition is used for the case when random coding is done with $c_j \sim \text{uniform}$ on the type class $\{x^n : \hat{P}_{x^n} \approx P_X\}$. Another alternative, ‘‘entropic typicality’’, cf. [75], is to search for a codeword with $\sum_{j=1}^n \log \frac{1}{P_{X_1, Y_1}(x_j, y_j)} \approx H(X, Y)$. We think of our requirement, $\{i(x^n; y^n) \geq n\gamma_1\}$, as a version of ‘‘joint typicality’’ that is applicable to much wider generality of channels (not necessarily over product alphabets, or memoryless).

18.3 Dependence-testing bound

The following result is a slight refinement of Theorem 18.5, that results in a bound that is free from the auxiliary parameters and is provably stronger.

Theorem 18.6 (DT bound). *Fix a channel $P_{Y|X}$ and an arbitrary input distribution P_X . Then for every $\tau > 0$ there exists an (M, ϵ) -code with*

$$\epsilon \leq \mathbb{E} \left[\exp \left\{ - \left(i(X; Y) - \log \frac{M-1}{2} \right)^+ \right\} \right] \quad (18.11)$$

where $x^+ \triangleq \max(x, 0)$.

18.3 Dependence-testing bound 297

Proof. For a fixed γ , consider the following suboptimal decoder:

$$g(y) = \begin{cases} m, & \text{for the smallest } m \text{ s.t. } i(c_m; y) \geq \gamma \\ e, & \text{o/w} \end{cases}$$

Setting $\hat{W} = g(Y)$ we note that given a codebook $\{c_1, \dots, c_M\}$, we have by union bound

$$\begin{aligned} \mathbb{P}[\hat{W} \neq j | W = j] &= \mathbb{P}[i(c_j; Y) \leq \gamma | W = j] + \mathbb{P}[i(c_j; Y) > \gamma, \exists k \in [j-1], \text{s.t. } i(c_k; Y) > \gamma] \\ &\leq \mathbb{P}[i(c_j; Y) \leq \gamma | W = j] + \sum_{k=1}^{j-1} \mathbb{P}[i(c_k; Y) > \gamma | W = j]. \end{aligned}$$

Averaging over the randomly generated codebook, the expected error probability is upper bounded by:

$$\begin{aligned} \mathbb{E}[P_e(c_1, \dots, c_M)] &= \frac{1}{M} \sum_{j=1}^M \mathbb{P}[\hat{W} \neq j | W = j] \\ &\leq \frac{1}{M} \sum_{j=1}^M \left(\mathbb{P}[i(X; Y) \leq \gamma] + \sum_{k=1}^{j-1} \mathbb{P}[i(\bar{X}; Y) > \gamma] \right) \\ &= \mathbb{P}[i(X; Y) \leq \gamma] + \frac{M-1}{2} \mathbb{P}[i(\bar{X}; Y) > \gamma] \\ &= \mathbb{P}[i(X; Y) \leq \gamma] + \frac{M-1}{2} \mathbb{E}[\exp(-i(X; Y)) \mathbf{1}\{i(X; Y) > \gamma\}] \quad (\text{by (18.4)}) \\ &= \mathbb{E}\left[\mathbf{1}\{i(X; Y) \leq \gamma\} + \frac{M-1}{2} \exp(-i(X; Y)) \mathbf{1}\{i(X; Y) > \gamma\} \right] \end{aligned}$$

To optimize over γ , note the simple observation that $U1_E + V1_{E^c} \geq \min\{U, V\}$, with equality iff $U \geq V$ on E . Therefore for any x, y , $1[i(x; y) \leq \gamma] + \frac{M-1}{2}e^{-i(x; y)}1[i(x; y) > \gamma] \geq \min(1, \frac{M-1}{2}e^{-i(x; y)})$, achieved by $\gamma = \log \frac{M-1}{2}$ regardless of x, y . Thus, we continue the bounding as follows

$$\begin{aligned} \inf_{\gamma} \mathbb{E}[P_e(c_1, \dots, c_M)] &\leq \inf_{\gamma} \mathbb{E}\left[\mathbf{1}\{i(X; Y) \leq \gamma\} + \frac{M-1}{2} \exp(-i(X; Y)) \mathbf{1}\{i(X; Y) > \gamma\} \right] \\ &= \mathbb{E}\left[\min\left(1, \frac{M-1}{2} \exp(-i(X; Y))\right) \right] \\ &= \mathbb{E}\left[\exp\left\{-\left(i(X; Y) - \log \frac{M-1}{2}\right)^+\right\} \right]. \end{aligned}$$

□

Remark 18.3 (Dependence testing interpretation). The RHS of (18.11) equals to $\frac{M+1}{2}$ multiple of the minimum error probability of the following Bayesian hypothesis testing problem:

$$\begin{aligned} H_0 : X, Y &\sim P_{X,Y} \text{ versus } H_1 : X, Y \sim P_X P_Y \\ \text{prior prob.: } \pi_0 &= \frac{2}{M+1}, \pi_1 = \frac{M-1}{M+1}. \end{aligned}$$

Note that $X, Y \sim P_{X,Y}$ and $\bar{X}, Y \sim P_X P_Y$, where X is the sent codeword and \bar{X} is the unsent codeword. As we know from binary hypothesis testing, the best threshold for the likelihood ratio test (minimizing the weighted probability of error) is $\log \frac{\pi_1}{\pi_0}$, as we indeed found out.

One of the immediate benefits of Theorem 18.7 compared to Theorem 18.5 is precisely the fact that we do not need to perform a cumbersome minimization over τ in (18.9) to get the minimum upper bound in Theorem 18.5. Nevertheless, it can be shown that the DT bound is stronger than Shannon's bound with optimized τ . See also Exc. IV.30.

Finally, we remark (and will develop this below in our treatment of linear codes) that DT bound and Shannon's bound both hold without change if we generate $\{c_i\}$ by any other (non-iid) procedure with a prescribed marginal *and pairwise independent codewords* – see Theorem 18.18 below.

18.4 Feinstein's maximal coding bound

The previous achievability results are obtained using *probabilistic* methods (random coding). In contrast, the following achievability bound due to Feinstein uses a *greedy* construction. One immediate advantage of Feinstein's method is that it shows existence of codes satisfying **maximal** probability of error criterion.²

Theorem 18.7 (Feinstein's lemma). *Fix a channel $P_{Y|X}$ and an arbitrary input distribution P_X . Then for every $\gamma > 0$ and for every $\epsilon \in (0, 1)$ there exists an $(M, \epsilon)_{\max}$ -code with*

$$M \geq \gamma(\epsilon - \mathbb{P}[i(X; Y) < \log \gamma]) \quad (18.12)$$

Remark 18.4 (Comparison with Shannon's bound). We can also interpret (18.12) differently: for any fixed M , there exists an $(M, \epsilon)_{\max}$ -code that achieves the maximal error probability bounded as follows:

$$\epsilon \leq \mathbb{P}[i(X; Y) < \log \gamma] + \frac{M}{\gamma}$$

If we take $\log \gamma = \log M + \tau$, this gives the bound of exactly the same form as Shannon's (18.9). It is rather surprising that two such different methods of proof produced essentially the same bound (modulo the difference between maximal and average probability of error). We will discuss the reason for this phenomenon in Section 18.7.

Proof. From the definition of $(M, \epsilon)_{\max}$ -code, we recall that our goal is to find codewords $c_1, \dots, c_M \in \mathcal{X}$ and disjoint subsets (decoding regions) $D_1, \dots, D_M \subset \mathcal{Y}$, s.t.

$$P_{Y|X}(D_i | c_i) \geq 1 - \epsilon, \forall i \in [M].$$

Feinstein's idea is to construct a codebook of size M in a sequential greedy manner.

² Nevertheless, we should point out that this is not a serious advantage: from any (M, ϵ) code we can extract an (M', ϵ') -subcode with a smaller M' and larger ϵ' – see Theorem 19.4.

18.4 Feinstein's maximal coding bound 299

For every $x \in \mathcal{X}$, associate it with a preliminary decoding region E_x defined as follows:

$$E_x \triangleq \{y \in \mathcal{Y} : i(x; y) \geq \log \gamma\}$$

Notice that the preliminary decoding regions $\{E_x\}$ may be overlapping, and we will trim them into final decoding regions $\{D_x\}$, which will be disjoint. Next, we apply Corollary 18.1 and find out that there is a set $F \subset \mathcal{X}$ with two properties: a) $P_X[F] = 1$ and b) for every $x \in F$ we have

$$P_Y(E_x) \leq \frac{1}{\gamma}. \quad (18.13)$$

We can assume that $\mathbb{P}[i(X; Y) < \log \gamma] \leq \epsilon$, for otherwise the RHS of (18.12) is negative and there is nothing to prove. We first claim that there exists some $c \in F$ such that $\mathbb{P}[Y \in E_c | X = c] = P_{Y|X}(E_c | c) \geq 1 - \epsilon$. Indeed, assume (for the sake of contradiction) that $\forall c \in F$, $\mathbb{P}[i(c; Y) \geq \log \gamma | X = c] < 1 - \epsilon$. Note that since $P_X(F) = 1$ we can average this inequality over $c \sim P_X$. Then we arrive at $\mathbb{P}[i(X; Y) \geq \log \gamma] < 1 - \epsilon$, which is a contradiction.

With these preparations we construct the codebook in the following way:

- 1 Pick c_1 to be any codeword in F such that $P_{Y|X}(E_{c_1} | c_1) \geq 1 - \epsilon$, and set $D_1 = E_{c_1}$;
- 2 Pick c_2 to be any codeword in F such that $P_{Y|X}(E_{c_2} \setminus D_1 | c_2) \geq 1 - \epsilon$, and set $D_2 = E_{c_2} \setminus D_1$;
- ...
- 3 Pick c_M to be any codeword in F such that $P_{Y|X}(E_{c_M} \setminus \bigcup_{j=1}^{M-1} D_j | c_M) \geq 1 - \epsilon$, and set $D_M = E_{c_M} \setminus \bigcup_{j=1}^{M-1} D_j$.

We stop if c_{M+1} codeword satisfying the requirement cannot be found. Thus, M is determined by the stopping condition:

$$\forall c \in F, P_{Y|X}(E_c \setminus \bigcup_{j=1}^M D_j | c) < 1 - \epsilon$$

Averaging the stopping condition over $c \sim P_X$ (which is permissible due to $P_X(F) = 1$), we obtain

$$\mathbb{P} \left[i(X; Y) \geq \log \gamma \text{ and } Y \notin \bigcup_{j=1}^M D_j \right] < 1 - \epsilon,$$

or, equivalently,

$$\epsilon < \mathbb{P} \left[i(X; Y) < \log \gamma \text{ or } Y \in \bigcup_{j=1}^M D_j \right].$$

Applying the union bound to the right hand side yields

$$\begin{aligned} \epsilon &< \mathbb{P}[i(X; Y) < \log \gamma] + \sum_{j=1}^M P_Y(D_j) \\ &\leq \mathbb{P}[i(X; Y) < \log \gamma] + \sum_{j=1}^M P_Y(E_{c_j}) \end{aligned}$$

$$\leq \mathbb{P}[i(X; Y) < \log \gamma] + \frac{M}{\gamma}$$

where the last step makes use of (18.13). Evidently, this completes the proof. \square

18.5 RCU and Gallager's bound

Although the bounds we demonstrated so far will be sufficient for recovering the noisy channel coding theorem later, they are not the best possible. Namely, for a given M one can show much smaller upper bounds on the probability of error. Two such bounds are the so-called random-coding union (RCU) and the Gallager's bound, which we prove here. The main new ingredient is that instead of using suboptimal (threshold) decoders as before, we will analyze the optimal maximum likelihood decoder.

Theorem 18.8 (RCU bound). *Fix a channel $P_{Y|X}$ and an arbitrary input distribution P_X . Then for every integer $M \geq 1$ there exists an (M, ϵ) -code with*

$$\epsilon \leq \mathbb{E} [\min \{1, (M-1)\mathbb{P}[i(\bar{X}; Y) \geq i(X; Y) | X, Y]\}] , \quad (18.14)$$

where the joint distribution of (X, \bar{X}, Y) is as in (18.1).

Proof. For a given codebook (c_1, \dots, c_M) the average probability of error for the maximum likelihood decoder, cf. (18.3), is upper bounded by

$$\epsilon \leq \frac{1}{M} \sum_{m=1}^M \mathbb{P} \left[\bigcup_{j=1; j \neq m}^M \{i(c_j; Y) \geq i(c_m; Y)\} | X = c_m \right].$$

Note that we do not necessarily have equality here, since the maximum likelihood decoder will resolve ties (i.e. the cases when multiple codewords maximize information density) in favor of the correct codeword, whereas in the expression above we pessimistically assume that all ties are resolved incorrectly. Now, similar to Shannon's bound in Theorem 18.5 we prove existence of a good code by averaging the last expression over $c_j \sim P_X^{\text{i.i.d.}}$.

To that end, notice that expectations of each term in the sum coincide (by symmetry). To evaluate this expectation, let us take $m = 1$ condition on $W = 1$ and observe that under this conditioning we have

$$(c_1, Y, c_2, \dots, c_M) \sim P_{X,Y} \prod_{j=2}^M P_X.$$

With this observation in mind we have the following chain:

$$\mathbb{P} \left[\bigcup_{j=2}^M \{i(c_j; Y) \geq i(c_1; Y)\} \middle| W = 1 \right]$$

18.5 RCU and Gallager's bound 301

$$\begin{aligned} &\stackrel{(a)}{=} \mathbb{E}_{(x,y) \sim P_{X,Y}} \left[\mathbb{P} \left[\bigcup_{j=2}^M \{i(c_j; Y) \geq i(c_1; Y)\} \mid c_1 = x, Y = y, W = 1 \right] \right] \\ &\stackrel{(b)}{\leq} \mathbb{E} [\min\{1, (M-1)\mathbb{P}[i(\bar{X}; Y) \geq i(X; Y) \mid X, Y]\}] \end{aligned}$$

where (a) is just expressing the probability by first conditioning on the values of (c_1, Y) ; and (b) corresponds to applying the union bound but capping the result by 1. This completes the proof of the bound. We note that the step (b) is the essence of the RCU bound and corresponds to the self-evident fact that for any collection of events E_j we have

$$\mathbb{P}[\cup E_j] \leq \min\{1, \sum_j \mathbb{P}[E_j]\}.$$

What makes its application clever is that we first conditioned on (c_1, Y) . If we applied the union bound right from the start without conditioning, the resulting estimate on ϵ would have been much weaker (in particular, would not have lead to achieving capacity). \square

It turns out that Shannon's bound Theorem 18.5 is just a weakening of (18.14) obtained by splitting the expectation according to whether or not $i(X; Y) \leq \log \beta$ and upper bounding $\min\{x, 1\}$ by 1 when $i(X; Y) \leq \log \beta$ and by x otherwise. Another such weakening is a famous Gallager's bound [129]:

Theorem 18.9 (Gallager's bound). *Fix a channel $P_{Y|X}$, an arbitrary input distribution P_X and $\rho \in [0, 1]$. Then there exists an (M, ϵ) code such that*

$$\epsilon \leq M^\rho \mathbb{E} \left[\left(\mathbb{E} \left[\exp \frac{i(\bar{X}; Y)}{1+\rho} \mid Y \right] \right)^{1+\rho} \right] \quad (18.15)$$

where again $(\bar{X}, Y) \sim P_X P_Y$ as in (18.1).

Proof. We first notice that by Proposition 18.3 applied with $f_+(x, y) = \exp\{\frac{i(x; y)}{1+\rho}\}$ and interchanged X and Y we have for P_Y -almost every y

$$\mathbb{E}[\exp\{-i(X; Y) \frac{\rho}{1+\rho}\} \mid Y = y] = \mathbb{E}[\exp\{i(X; \bar{Y}) \frac{1}{1+\rho}\} \mid \bar{Y} = y] = \mathbb{E}[\exp\{i(\bar{X}; Y) \frac{1}{1+\rho}\} \mid Y = y], \quad (18.16)$$

where we also used the fact that $(X, \bar{Y}) \stackrel{d}{=} (\bar{X}, Y)$ under (18.1).

Now, consider the bound (18.14) and replace the min via the bound

$$\min\{t, 1\} \leq t^\rho \quad \forall t \geq 0. \quad (18.17)$$

this results in

$$\epsilon \leq M^\rho \mathbb{E} [\mathbb{P}[i(\bar{X}; Y) > i(X; Y) \mid X, Y]^\rho]. \quad (18.18)$$

We apply the Chernoff bound

$$\mathbb{P}[i(\bar{X}; Y) > i(X; Y) \mid X, Y] \leq \exp\{-\frac{1}{1+\rho} i(X; Y)\} \mathbb{E}[\exp\{\frac{1}{1+\rho} i(\bar{X}; Y)\} \mid Y].$$

Raising this inequality to ρ -power and taking expectation $\mathbb{E}[\cdot|Y]$ we obtain

$$\mathbb{E} [\mathbb{P}[i(\bar{X}; Y) > i(X; Y)|X, Y]^\rho | Y] \leq \mathbb{E}^\rho [\exp\{\frac{1}{1+\rho}i(\bar{X}; Y)|Y\} \mathbb{E}[\exp\{-\frac{\rho}{1+\rho}i(X; Y)\}|Y].$$

The last term can be now re-expressed via (18.16) to obtain

$$\mathbb{E} [\mathbb{P}[i(\bar{X}; Y) > i(X; Y)|X, Y]^\rho | Y] \leq \mathbb{E}^{1+\rho} [\exp\{\frac{1}{1+\rho}i(\bar{X}; Y)|Y\}].$$

Applying this estimate to (18.18) completes the proof.

Gallager's bound (18.15) can also be obtained by analyzing the average behavior of random coding and maximum-likelihood decoding. In fact, it is easy to verify that we can weaken (18.14) to recover (18.15) using $\max\{0, x\} \leq x^{1/(1+\rho)}$ and $\min\{x, 1\} \leq x^\rho$.

□

The key innovation of Gallager – a step (18.17), which became known as the ρ -trick – corresponds to the following version of the union bound: For any events E_j and $0 \leq \rho \leq 1$ we have

$$\mathbb{P}[\cup E_j] \leq \min\{1, \sum_j \mathbb{P}[E_j]\} \leq \left(\sum_j \mathbb{P}[E_j]\right)^\rho.$$

Now to understand properly the significance of Gallager's bound we need to first define the concept of the memoryless channels (see (19.1) below). For such channels and using the iid inputs, the expression (18.15) turns, after optimization over ρ , into

$$\epsilon \leq \exp\{-nE_r(R)\},$$

where $R = \frac{\log M}{n}$ is the rate and $E_r(R)$ is the Gallager's random coding exponent. This shows that not only the error probability at a fixed rate can be made to vanish, but in fact it can be made to vanish exponentially fast in the blocklength. We will discuss such exponential estimates in more detail in Section 22.4*.

18.6 Linear codes

So far in this Chapter we have shown existence of good error-correcting codes by either doing the random or maximal coding. The constructed codes have little structure. At the same time, most codes used in practice are so-called linear codes and a natural question whether restricting to linear codes leads to loss in performance. In this section we show that there exist good linear codes as well. A pleasant property of linear codes is that $P_e = P_{e,\max}$ and, therefore, bounding average probability of error (as in Shannon's bound) automatically yields control of the maximum probability of error as well.

Definition 18.10 (Linear codes). Let \mathbb{F}_q denote the finite field of cardinality q (cf. Definition 11.11). Let the input and output space of the channel be $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^n$. We say a codebook $\mathcal{C} = \{c_u : u \in \mathbb{F}_q^k\}$ of size $M = q^k$ is a **linear code** if \mathcal{C} is a k -dimensional *linear subspace* of \mathbb{F}_q^n .

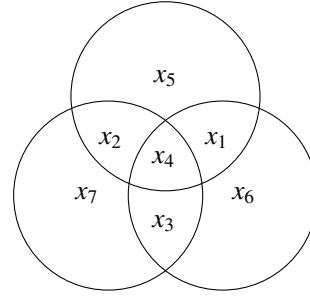
A linear code can be equivalently described by:

- *Generator matrix* $G \in \mathbb{F}_q^{k \times n}$, so that the codeword for each $u \in \mathbb{F}_q^k$ is given by $c_u = uG$ (row-vector convention) and the codebook \mathcal{C} is the row-span of G , denoted by $\text{Im}(G)$;
- *Parity-check matrix* $H \in \mathbb{F}_q^{(n-k) \times n}$, so that each codeword $c \in \mathcal{C}$ satisfies $Hc^\top = 0$. Thus \mathcal{C} is the nullspace of H , denoted by $\text{Ker}(H)$. We have $HG^\top = 0$.

Example 18.1 (Hamming code). The $[7, 4, 3]_2$ Hamming code over \mathbb{F}_2 is a linear code with the following generator and parity check matrices:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

In particular, G and H are of the form $G = [I; P]$ and $H = [-P^\top; I]$ (systematic codes) so that $HG^\top = 0$. The following picture helps to visualize the parity check operation:



Note that all four bits in each circle (corresponding to a row of H) sum up to zero. One can verify that the minimum distance of this code is 3 bits. As such, it can correct 1 bit of error and detect 2 bits of error.

Linear codes are almost always examined with channels of additive noise, a precise definition of which is given below:

Definition 18.11 (Additive noise). A channel $P_{Y|X}$ with input and output space \mathbb{F}_q^n is called additive-noise if

$$P_{Y|X}(y|x) = P_Z(y - x)$$

for some random vector Z taking values in \mathbb{F}_q^n . In other words, $Y = X + Z$, where $Z \perp\!\!\!\perp X$.

Given a linear code and an additive-noise channel $P_{Y|X}$, it turns out that there is a special “syndrome decoder” that is optimal.

Theorem 18.12. Any $[k, n]_{\mathbb{F}_q}$ linear code over an additive-noise $P_{Y|X}$ has a maximum likelihood (ML) decoder $g : \mathbb{F}_q^n \rightarrow \mathbb{F}_q^k$ such that:

- 1 $g(y) = y - g_{\text{synd}}(Hy^\top)$, i.e., the decoder is a function of the “syndrome” Hy^\top only. Here $g_{\text{synd}} : \mathbb{F}_q^{n-k} \rightarrow \mathbb{F}_q^k$, defined by $g_{\text{synd}}(s) \triangleq \operatorname{argmax}_{z:Hz^\top=s} P_Z(z)$, is called the “syndrome decoder”, which decodes the most likely realization of the noise.
- 2 (Geometric uniformity) Decoding regions are translates of $D_0 = \operatorname{Im}(g_{\text{synd}})$: $D_u = c_u + D_0$ for any $u \in \mathbb{F}_q^k$.
- 3 $P_{e,\max} = P_e$.

In other words, syndrome is a *sufficient statistic* (Definition 3.12) for decoding a linear code.

Proof. 1 The maximum likelihood decoder for a linear code is

$$g(y) = \operatorname{argmax}_{c \in \mathcal{C}} P_{Y|X}(y|c) = \operatorname{argmax}_{c: Hc^\top=0} P_Z(y - c) = y - \operatorname{argmax}_{z: Hz^\top=Hy^\top} P_Z(z) = y - g_{\text{synd}}(Hy^\top).$$

2 For any u , the decoding region

$$D_u = \{y : g(y) = c_u\} = \{y : y - g_{\text{synd}}(Hy^\top) = c_u\} = \{y : y - c_u = g_{\text{synd}}(H(y - c_u)^\top)\} = c_u + D_0,$$

where we used $Hc_u^\top = 0$ and $c_0 = 0$.

3 For any u ,

$$\mathbb{P}[\hat{W} \neq u | W = u] = \mathbb{P}[g(c_u + Z) \neq c_u] = \mathbb{P}[c_u + Z - g_{\text{synd}}(Hc_u^\top + HZ^\top) \neq c_u] = \mathbb{P}[g_{\text{synd}}(HZ^\top) \neq Z].$$

□

Remark 18.5. As a concrete example, consider the binary symmetric channel $\text{BSC}_\delta^{\otimes n}$ previously considered in Example 17.1 and Section 17.2. This is an additive-noise channel over \mathbb{F}_2^n , where $Y = X + Z$ and $Z = (Z_1, \dots, Z_n) \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\delta)$. Assuming $\delta < 1/2$, the syndrome decoder aims to find the noise realization with the fewest number of flips that is compatible with the received codeword, namely $g_{\text{synd}}(s) = \operatorname{argmin}_{z:Hz^\top=s} w_H(z)$, where w_H denotes the Hamming weight. In this case elements of the image of g_{synd} , which we denote by D_0 , are known as “minimal weight coset leaders”. Counting how many of them occur at each Hamming weight is a difficult open problem even for the most well-studied codes such as Reed-Muller ones. In Hamming space D_0 looks like a Voronoi region of a lattice and D_u ’s constitute a Voronoi tessellation of \mathbb{F}_q^n .

Remark 18.6. Overwhelming majority of practically used codes are in fact linear codes. Early in the history of coding, linearity was viewed as a way towards fast and *low-complexity encoding* (just binary matrix multiplication) and slightly lower complexity of the maximum-likelihood decoding (via the syndrome decoder). As codes became longer and longer, though, the syndrome decoding became impractical and today only those codes are used in practice for which there are fast and low-complexity (suboptimal) decoders.

18.6 Linear codes 305

Theorem 18.13 (DT bound for linear codes). *Let $P_{Y|X}$ be an additive noise channel over \mathbb{F}_q^n . For all integers $k \geq 1$ there exists a linear code $f: \mathbb{F}_q^k \rightarrow \mathbb{F}_q^n$ with error probability:*

$$P_{e,\max} = P_e \leq \mathbb{E} \left[q^{-\left(n-k - \log_q \frac{1}{P_{Z|X}} \right)^+} \right]. \quad (18.19)$$

Remark 18.7. The bound above is the same as Theorem 18.7 evaluated with $P_X = \text{Unif}(\mathbb{F}_q^n)$. The analogy between Theorems 18.7 and 18.18 is the same as that between Theorems 11.6 and 11.12 (full random coding vs random linear codes).

Proof. Recall that in proving DT bound (Theorem 18.7), we selected the codewords $c_1, \dots, c_M \stackrel{\text{i.i.d.}}{\sim} P_X$ and showed that

$$\mathbb{E}[P_e(c_1, \dots, c_M)] \leq \mathbb{P}[i(X; Y) \leq \gamma] + \frac{M-1}{2} \mathbb{P}[i(\bar{X}; Y) \geq \gamma]$$

Here we will adopt the same approach and take $P_X = \text{Unif}(\mathbb{F}_q^n)$ and $M = q^k$.

By Theorem 18.15 the optimal decoding regions are translational invariant, i.e. $D_u = c_u + D_0, \forall u$, and therefore:

$$P_{e,\max} = P_e = \mathbb{P}[\hat{W} \neq u | W = u], \forall u.$$

Step 1: Random linear coding with dithering: Let codewords be chosen as

$$c_u = uG + h, \quad \forall u \in \mathbb{F}_q^k$$

where random G and h are drawn as follows: the $k \times n$ entries of G and the $1 \times n$ entries of h are i.i.d. uniform over \mathbb{F}_q . We add the dithering to eliminate the special role that the all-zero codeword plays (since it is contained in any linear codebook).

Step 2: We claim that the codewords are pairwise independent and uniform, i.e. $\forall u \neq u'$, $(c_u, c_{u'}) \sim (X, \bar{X})$, where $P_{X, \bar{X}}(x, \bar{x}) = 1/q^{2n}$. To see this note that

$$\begin{aligned} c_u &\sim \text{uniform on } \mathbb{F}_q^n \\ c_{u'} &= u'G + h = uG + h + (u' - u)G = c_u + (u' - u)G \end{aligned}$$

We claim that $c_u \perp\!\!\!\perp G$ because conditioned on the generator matrix $G = G_0$, $c_u \sim \text{uniform on } \mathbb{F}_q^n$ due to the dithering h .

We also claim that $c_u \perp\!\!\!\perp c_{u'}$ because conditioned on c_u , $(u' - u)G \sim \text{uniform on } \mathbb{F}_q^n$.

Thus random linear coding with dithering indeed gives codewords $c_u, c_{u'}$ pairwise independent and are uniformly distributed.

Step 3: Repeat the same argument in proving DT bound for the symmetric and pairwise independent codewords, we have

$$\mathbb{E}[P_e(c_1, \dots, c_M)] \leq \mathbb{E}[\exp\{-\left(i(X; Y) - \log \frac{M-1}{2}\right)^+\}] = \mathbb{E}[q^{-\left(i(X; Y) - \log_q \frac{q^k - 1}{2}\right)^+}] \leq \mathbb{E}[q^{-\left(i(X; Y) - k\right)^+}]$$

where we used $M = q^k$ and picked the base of log to be q .

Step 4: compute $i(X; Y)$:

$$i(a; b) = \log_q \frac{P_Z(b - a)}{q^{-n}} = n - \log_q \frac{1}{P_Z(b - a)}$$

therefore

$$P_e \leq \mathbb{E}[q^{-(n-k-\log_q \frac{1}{P_Z(Z)})^+}] \quad (18.20)$$

Step 5: Remove dithering h . We claim that there exists a linear code without dithering such that (18.20) is satisfied. The intuition is that shifting a codebook has no effect on its performance. Indeed,

- Before, with dithering, the encoder maps u to $uG + h$, the channel adds noise to produce $Y = uG + h + Z$, and the decoder g outputs $g(Y)$.
- Now, without dithering, we encode u to uG , the channel adds noise to produce $Y = uG + Z$, and we apply decode g' defined by $g'(Y) = g(Y + h)$.

By doing so, we “simulate” dithering at the decoder end and the probability of error remains the same as before. Note that this is possible thanks to the additivity of the noisy channel. \square

We see that random coding can be done with different ensembles of codebooks. For example, we have

- Shannon ensemble: $\mathcal{C} = \{c_1, \dots, c_M\} \stackrel{\text{i.i.d.}}{\sim} P_X$ – fully random ensemble.
- Elias ensemble [112]: $\mathcal{C} = \{uG : u \in \mathbb{F}_q^k\}$, with the $k \times n$ generator matrix G drawn uniformly at random from the set of all matrices. (This ensemble is used in the proof of Theorem 18.18.)
- Gallager ensemble: $\mathcal{C} = \{c : Hc^\top = 0\}$, with the $(n-k) \times n$ parity-check matrix H drawn uniformly at random. Note this is not the same as the Elias ensemble.
- One issue with Elias ensemble is that with some non-zero probability G may fail to be full rank. (It is a good exercise to find $\mathbb{P}[\text{rank}(G) < k]$ as a function of n, k, q .) If G is not full rank, then there are two identical codewords and hence $P_{e,\max} \geq 1/2$. To fix this issue, one may let the generator matrix G be uniform on the set of all $k \times n$ matrices of full (row) rank.
- Similarly, we may modify Gallager’s ensemble by taking the parity-check matrix H to be uniform on all $n \times (n-k)$ full rank matrices.

For the modified Elias and Gallager’s ensembles, we could still do the analysis of random coding. A small modification would be to note that this time (X, \bar{X}) would have distribution

$$P_{X, \bar{X}} = \frac{1}{q^{2n} - q^n} \mathbf{1}_{\{X \neq X'\}}$$

uniform on all pairs of *distinct* codewords and are *not* pairwise independent.

Finally, we note that the Elias ensemble with dithering, $c_u = uG + h$, has pairwise independence property and its joint entropy $H(c_1, \dots, c_M) = H(G) + H(h) = (nk + n) \log q$. This is significantly smaller than for Shannon’s fully random ensemble that we used in Theorem 18.5. Indeed, when

18.7 Why random and maximal coding work well? 307

$c_j \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{F}_q^n)$ we have $H(c_1, \dots, c_M) = q^k n \log q$. An interesting question, thus, is to find

$$\min H(c_1, \dots, c_M)$$

where the minimum is over all distributions with $P[c_i = a, c_j = b] = q^{-2n}$ when $i \neq j$ (pairwise independent, uniform codewords). Note that $H(c_1, \dots, c_M) \geq H(c_1, c_2) = 2n \log q$. Similarly, we may ask for (c_i, c_j) to be uniform over all pairs of *distinct* elements. In this case, the Wozencraft ensemble (see Exercise IV.13) for the case of $n = 2k$ achieves $H(c_1, \dots, c_{q^k}) \approx 2n \log q$, which is essentially our lower bound.

18.7 Why random and maximal coding work well?

As we will see later the bounds developed in this chapter are very tight both asymptotically and non-asymptotically. That is, the codes constructed by the apparently rather naive processes of randomly selecting codewords or a greedily growing the codebook turn out to be essentially optimal in many ways. An additional mystery is that the bounds we obtained via these two rather different processes are virtually the same. These questions have puzzled researchers since the early days of information theory.

A rather satisfying reason was finally given in an elegant work of Barman and Fawzi [23]. Before going into the details, we want to vocalize explicitly the two questions we want to address:

- 1 Why is greedy procedure close to optimal?
- 2 Why is random coding procedure (with a simple P_X) close to optimal?

In short, we will see that the answer is that both of these methods are well-known to be (almost) optimal for submodular function maximization, and this is exactly what channel coding is about.

Before proceeding, we notice that in the second question it is important to qualify that P_X is simple, since taking P_X to be supported on the optimal $M^*(\epsilon)$ -achieving codebook would of course result in very good performance. However, instead we will see that choosing rather simple P_X already achieves a rather good lower bound on $M^*(\epsilon)$. More explicitly, by simple we mean a product distribution for the memoryless channel. Or, as an even better example to have in mind, consider an additive-noise vector channel:

$$Y^n = X^n + Z^n$$

with addition over a product abelian group and arbitrary (even non-memoryless) noise Z^n . In this case the choice of uniform P_X in random coding bound works, and is definitely “simple”.

The key observation of [23] is *submodularity* of the function mapping a codebook $\mathcal{C} \subset \mathcal{X}$ to the $|\mathcal{C}|(1 - P_{e,\text{MAP}}(\mathcal{C}))$, where $P_{e,\text{MAP}}(\mathcal{C})$ is the probability of error under the MAP decoder (17.5). (Recall (1.7) for the definition of submodularity.) More explicitly, consider a discrete \mathcal{Y} and define

$$S(\mathcal{C}) \triangleq \sum_{y \in \mathcal{Y}} \max_{x \in \mathcal{C}} P_{Y|X}(y|x), \quad S(\emptyset) = 0$$

It is clear that $S(\mathcal{C})$ is submodular non-decreasing as a sum of submodular non-decreasing functions \max (i.e. $T \mapsto \max_{x \in T} \phi(x)$ is submodular for any ϕ). On the other hand, $P_{e,\text{MAP}}(\mathcal{C}) = 1 - \frac{1}{|\mathcal{C}|} S(\mathcal{C})$, and thus search for the minimal error codebook is equivalent to maximizing the set-function S .

The question of finding

$$S^*(M) \triangleq \max_{|\mathcal{C}| \leq M} S(\mathcal{C})$$

was algorithmically resolved in a groundbreaking work of [218] showing (approximate) optimality of a greedy process. Consider, the following natural greedy process of constructing a sequence of good sets \mathcal{C}_t . Start with $\mathcal{C}_0 = \emptyset$. At each step find any

$$x_{t+1} \in \operatorname{argmax}_{x \notin \mathcal{C}_t} S(\mathcal{C}_t \cup \{x\})$$

and set

$$\mathcal{C}_{t+1} = \mathcal{C}_t \cup \{x_{t+1}\}.$$

They showed that

$$S(\mathcal{C}_t) \geq (1 - 1/e) \max_{|\mathcal{C}|=t} S(\mathcal{C}).$$

In other words, the probability of successful (MAP) decoding for the greedily constructed codebook is at most a factor $(1 - 1/e)$ away from the largest possible probability of success among all codebooks of the same cardinality. Since we are mostly interested in success probabilities very close to 1, this result may not appear very exciting. However, a small modification of the argument yields the following (see [181, Theorem 1.5] for the proof):

Theorem 18.14 ([218]). *For any non-negative submodular set-function f and a greedy sequence \mathcal{C}_t we have for all ℓ, t :*

$$f(\mathcal{C}_\ell) \geq (1 - e^{-\ell/t}) \max_{|\mathcal{C}|=t} f(\mathcal{C}).$$

Applying this to the special case of $f(\cdot) = S(\cdot)$ we obtain the result of [23]: The greedily constructed codebook \mathcal{C}' with M' codewords satisfies

$$1 - P_{e,\text{MAP}}(\mathcal{C}') \geq \frac{M}{M'} (1 - e^{-M'/M}) (1 - \epsilon^*(M)).$$

In particular, the greedily constructed code with $M' = M2^{-10}$ achieves probability of success that is $\geq 0.9995(1 - \epsilon^*(M))$. In other words, compared to the best possible code a greedy code carrying 10 bits fewer of data suffers at most $5 \cdot 10^{-4}$ worse probability of error. *This is a very compelling evidence for why greedy construction is so good.* We do note, however, that Feinstein's bound does greedy construction not with the MAP decoder, but with a suboptimal one.

Next we address the question of *random coding*. Recall that our goal is to explain how can selecting codewords uniformly at random from a “simple” distribution P_X be any good. The key

18.7 Why random and maximal coding work well? 309

idea is again contained in [218]. The set-function $S(\mathcal{C})$ can also be understood as a function with domain $\{0, 1\}^{|\mathcal{X}|}$. Here is a natural extension of this function to the entire solid hypercube $[0, 1]^{|\mathcal{X}|}$:

$$S_{LP}(\pi) = \sup \left\{ \sum_{x,y} P_{Y|X}(y|x) r_{x,y} : 0 \leq r_{x,y} \leq \pi_x, \sum_x r_{x,y} \leq 1 \right\}. \quad (18.21)$$

Indeed, it is easy to see that $S_{LP}(1_{\mathcal{C}}) = S(\mathcal{C})$ and that S_{LP} is a concave function.³

Since S_{LP} is an extension of S it is clear that

$$S^*(M) \leq S_{LP}^*(M) \triangleq \max \{S_{LP}(\pi) : 0 \leq \pi_x \leq 1, \sum_x \pi_x \leq M\}. \quad (18.22)$$

In fact, we will see later in Section 22.3 that this bound coincides with the bound known as meta-converse. Surprisingly, [218] showed that the greedy construction not only achieves a large multiple of $S^*(M)$ but also of $S_{LP}^*(M)$:

$$S(\mathcal{C}_M) \geq (1 - e^{-1})S_{LP}^*(M). \quad (18.23)$$

The importance of this result (which is specific to submodular functions $\mathcal{C} \mapsto \sum_y \max_{x \in \mathcal{C}} g(x, y)$) is that it gave one of the first integrality gap results relating the value of combinatorial optimization $S^*(M)$ and a linear program relaxatioa $S_{LP}^*(M)$: $(1 - e^{-1})S_{LP}^*(M) \leq S^*(M) \leq S_{LP}^*(M)$.

An extension of (18.23) similar to the preceding theorem can also be shown: for all M', M we have

$$S(\mathcal{C}_{M'}) \geq (1 - e^{-M'/M})S_{LP}^*(M).$$

To connect to the concept of random coding, though, we need the following result of [23]:⁴

Theorem 18.15. Fix $\pi \in [0, 1]^{|\mathcal{X}|}$ and let $M = \sum_{x \in \mathcal{X}} \pi_x$. Let $\mathcal{C} = \{c_1, \dots, c_{M'}\}$ with $c_j \stackrel{i.i.d.}{\sim} P_X(x) = \frac{\pi_x}{M}$. Then we have

$$\mathbb{E}[S(\mathcal{C})] \geq (1 - e^{-M'/M})S_{LP}(\pi).$$

The proof of this result trivially follows from applying the following lemma with $g(x) = P_{Y|X}(y|x)$, summing over y and recalling the definition of S_{LP} in (18.21).

Lemma 18.16. Let π and \mathcal{C} be as in Theorem. Let $g : \mathcal{X} \rightarrow \mathbb{R}$ be any function and denote $T(\pi, g) = \max \{\sum_x r_x g(x) : 0 \leq r_x \leq \pi_x, \sum_x r_x = 1\}$. Then

$$\mathbb{E}[\max_{x \in \mathcal{C}} g(x)] \geq (1 - e^{-M'/M})T(\pi, g).$$

³ There are a number of standard extensions of a submodular function f to a hypercube. The largest convex interpolant f_+ , also known as Lovász extension, the least concave interpolant f_- , and multi-linear extension [55]. However, S_{LP} does not coincide with any of these and in particular strictly larger (in general) than f_- .

⁴ There are other ways of doing “random coding” to produce an integer solution from a fractional one. For example, see the multi-linear extension based one in [55].

Proof. Without loss of generality we take $\mathcal{X} = [m]$ and $g(1) \geq g(2) \geq \dots \geq g(m) \geq g(m+1) \triangleq 0$. Denote for convenience $a = 1 - (1 - \frac{1}{M})^{M'} \geq 1 - e^{-M'/M}$, $b(j) \triangleq \mathbb{P}[\{1, \dots, j\} \cap \mathcal{C} \neq \emptyset]$. Then

$$\mathbb{P}[\max_{x \in \mathcal{C}} g(x) = g(j)] = b(j) - b(j-1),$$

and from the summation by parts we get

$$\mathbb{E}[\max_{x \in \mathcal{C}} g(x)] = \sum_{j=1}^m (g(j) - g(j+1))b(j). \quad (18.24)$$

On the other hand, denoting $c(j) = \min(\sum_{i \leq j} \pi_i, 1)$. Now from the definition of $b(j)$ we have

$$b(j) = 1 - (1 - \frac{\pi_1 + \dots + \pi_j}{M})^\ell \geq 1 - (1 - \frac{c(j)}{M})^{M'}.$$

From the simple inequality $(1 - \frac{x}{M})^{M'} \leq 1 - ax$ (valid for any $x \in [0, 1]$) we get

$$b(j) \geq ac(j).$$

Plugging this into (18.24) we conclude the proof by noticing that $r_j = c(j) - c(j-1)$ attains the maximum in the definition of $T(\pi, g)$. \square

Theorem 18.21 completes this section's goal and shows that the random coding (as well as the greedy/maximal coding) attains an almost optimal value of $S^*(M)$. Notice also that the random coding distribution that we should be using is the one that attains the definition of $S_{LP}^*(M)$. For input symmetric channels (such as additive noise ones) it is easy to show that the optimal $\pi \in [0, 1]^\mathcal{X}$ is a constant vector, and hence the codewords are to be generated iid uniformly on \mathcal{X} .

19 Channel capacity

In this chapter we apply methods developed in the previous chapters (namely the weak converse and the random/maximal coding achievability) to compute the channel capacity. This latter notion quantifies the maximal amount of (data) bits that can be reliably communicated *per single channel use* in the limit of using the channel many times. Formalizing the latter statement will require introducing the concept of a communication channel. Then for special kinds of channels (the memoryless and the information stable ones) we will show that computing the channel capacity reduces to maximizing the (sequence of the) mutual informations. This result, known as Shannon's noisy channel coding theorem, is very special as it relates the value of a (discrete, combinatorial) optimization problem over codebooks to that of a (convex) optimization problem over information measures. It builds a bridge between the abstraction of Information Measures (Part I) and the practical engineering problems.

Information theory as a subject is sometimes accused of “asymptopia”, or the obsession with asymptotic results and computing various limits. Although in this book we mostly refrain from asymptopia, the topic of this chapter requires committing this sin *ipso facto*.

19.1 Channels and channel capacity

As we discussed in Chapter 17 the main information-theoretic question of data transmission is the following: How many bits can one transmit reliably if one is allowed to use a given noisy channel n times? The normalized quantity equal to the number of message bits per channel use is known as *rate*, and *capacity* refers to the highest achievable rate under a small probability of decoding error. However, what does it mean to “use channel many times”? How do we formalize the concept of a channel use? To that end, we need to change the meaning of the term “channel”. So far in this book we have used the term *channel* as a synonym of the Markov kernel (Definition 2.9). More correctly, however, this term should be used to refer to the following notion.

Definition 19.1. Fix an input alphabet \mathcal{A} and an output alphabet \mathcal{B} . A sequence of Markov kernels $P_{Y^n|X^n} : \mathcal{A}^n \rightarrow \mathcal{B}^n$ indexed by the integer $n = 1, 2, \dots$ is called a *channel*. The length of the input n is known as *blocklength*.

To give this abstract notion more concrete form one should recall Section 17.2, in which we described the BSC channel. Note, however, that despite this definition, it is customary to use the term *channel* to refer to a single Markov kernel (as we did before in this book). An even worse,

yet popular, abuse of terminology is to refer to n -th element of the sequence, the kernel $P_{Y^n|X^n}$, as the n -letter channel.

Although we have not imposed any requirements on the sequence of kernels $P_{Y^n|X^n}$, one is never interested in channels at this level of generality. Most of the time the elements of the channel input $X^n = (X_1, \dots, X_n)$ are thought as indexed by time. That is the X_t corresponds to the letter that is transmitted at time t , while Y_t is the letter received at time t . The channel's action is that of “adding noise” to X_t and outputting Y_t . However, the generality of the previous definition allows to model situations where the channel has internal state, so that the amount and type of noise added to X_t depends on the previous inputs and in principle even on the future inputs. The interpretation of t as time, however, is not exclusive. In storage (magnetic, non-volatile or flash) t indexes space. In those applications, the noise may have a rather complicated structure with transformation $X_t \rightarrow Y_t$ depending on both the “past” $X_{<t}$ and the “future” $X_{>t}$.

Almost all channels of interest satisfy one or more of the restrictions that we list next:

- A channel is called *non-anticipatory* if it has the following extension property. Under the n -letter kernel $P_{Y^n|X^n}$, the conditional distribution of the first k output symbols Y^k only depends on X^k (and not on X_{k+1}^n) and coincides with the kernel $P_{Y^k|X^k}$ (the k -th element of the channel sequence) the k -th channel transition kernel in the sequence. This requirement models the scenario wherein channel outputs depend causally on the inputs.
- A channel is *discrete* if \mathcal{A} and \mathcal{B} are finite.
- A channel is *additive-noise* if $\mathcal{A} = \mathcal{B}$ are abelian group and $Y^n = X^n + Z^n$ for some Z^n independent of X^n (see Definition 18.14). Thus

$$P_{Y^n|X^n}(y^n|x^n) = P_{Z^n}(y^n - x^n).$$

- A channel is *memoryless* if $P_{Y^n|X^n}$ factorizes into a product distribution. Namely,

$$P_{Y^n|X^n} = \prod_{k=1}^n P_{Y_k|X_k}. \quad (19.1)$$

where each $P_{Y_k|X_k} : \mathcal{A} \rightarrow \mathcal{B}$; in particular, $P_{Y^n|X^n}$ are compatible at different blocklengths n .

- A channel is *stationary memoryless* if (19.1) is satisfied with $P_{Y_k|X_k}$ not depending on k , denoted commonly by $P_{Y|X}$. In other words,

$$P_{Y^n|X^n} = (P_{Y|X})^{\otimes n}. \quad (19.2)$$

Thus, in discrete cases, we have

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i). \quad (19.3)$$

The interpretation is that each coordinate of the transmitted codeword X^n is corrupted by noise independently with the same noise statistic.

- Discrete memoryless stationary channel (DMC): A DMC is a channel that is both discrete and stationary memoryless. It can be specified in two ways:

19.1 Channels and channel capacity 313

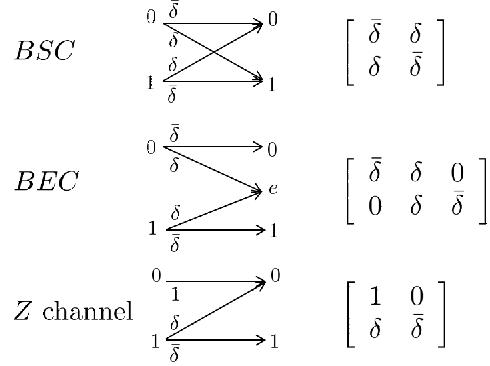


Figure 19.1 Examples of DMCs.

- an $|\mathcal{A}| \times |\mathcal{B}|$ -dimensional (row-stochastic) matrix $P_{Y|X}$ where elements specify the transition probabilities;
- a bipartite graph with edge weight specifying the transition probabilities.

Fig. 19.1 lists some common binary-input binary-output DMCs.

Let us recall the example of the AWGN channel Example 3.3: the alphabets $\mathcal{A} = \mathcal{B} = \mathbb{R}$ and $Y^n = X^n + Z^n$, with $X^n \perp\!\!\!\perp Z^n \sim \mathcal{N}(0, \sigma^2 I_n)$. This channel is a non-discrete, stationary memoryless, additive-noise channel.

Having defined the notion of the channel, we can define next the operational problem that the communication engineer faces when tasked with establishing a data link across the channel. Since the channel is noisy, the data is not going to pass unperturbed and the error-correcting codes are naturally to be employed. To send one of $M = 2^k$ messages (or k data bits) with low probability of error, it is often desirable to use the shortest possible length of the input sequence. This desire explains the following definitions, which extend the fundamental limits in Definition 17.2 to involve the blocklength n .

Definition 19.2 (Fundamental Limits of Channel Coding).

- An (n, M, ϵ) -code is an (M, ϵ) -code for $P_{Y^n|X^n}$, consisting of an encoder $f : [M] \rightarrow \mathcal{A}^n$ and a decoder $g : \mathcal{B}^n \rightarrow [M] \cup \{\epsilon\}$.
- An $(n, M, \epsilon)_{\max}$ -code is analogously defined for the maximum probability of error.

The (non-asymptotic) fundamental limits are

$$M^*(n, \epsilon) = \max\{M : \exists (n, M, \epsilon)\text{-code}\}, \quad (19.4)$$

$$M_{\max}^*(n, \epsilon) = \max\{M : \exists (n, M, \epsilon)_{\max}\text{-code}\}. \quad (19.5)$$

How to understand the behaviour of $M^*(n, \epsilon)$? Recall that blocklength n measures the amount of time or space resource used by the code. Thus, it is natural to maximize the ratio of the data

transmitted to the resource used, and that leads us to the notion of *the transmission rate* defined as $R = \frac{\log_2 M}{n}$ and equal to the number of bits transmitted per channel use. Consequently, instead of studying $M^*(n, \epsilon)$ one is lead to the study of $\frac{1}{n} \log M^*(n, \epsilon)$. A natural first question is to determine the first-order asymptotics of this quantity and this motivates the final definition of the Section.

Definition 19.3 (Channel capacity). The ϵ -capacity C_ϵ and **Shannon capacity** C are defined as follows

$$C_\epsilon \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon);$$

$$C = \lim_{\epsilon \rightarrow 0+} C_\epsilon.$$

The *operational* meaning of C_ϵ (resp. C) is the maximum achievable rate at which one can communicate through a noisy channel with probability of error at most ϵ (resp. $o(1)$). In other words, for any $R < C$, there exists an $(n, \exp(nR), \epsilon_n)$ -code, such that $\epsilon_n \rightarrow 0$. In this vein, C_ϵ and C can be equivalently defined as follows:

$$C_\epsilon = \sup\{R : \forall \delta > 0, \exists n_0(\delta), \forall n \geq n_0(\delta), \exists (n, \exp(n(R - \delta)), \epsilon)\text{-code}\}$$

$$C = \sup\{R : \forall \epsilon > 0, \forall \delta > 0, \exists n_0(\delta, \epsilon), \forall n \geq n_0(\delta, \epsilon), \exists (n, \exp(n(R - \delta)), \epsilon)\text{-code}\}$$

The reason that capacity is defined as a large- n limit (as opposed to a supremum over n) is because we are concerned with rate limit of transmitting large amounts of data without errors (such as in communication and storage).

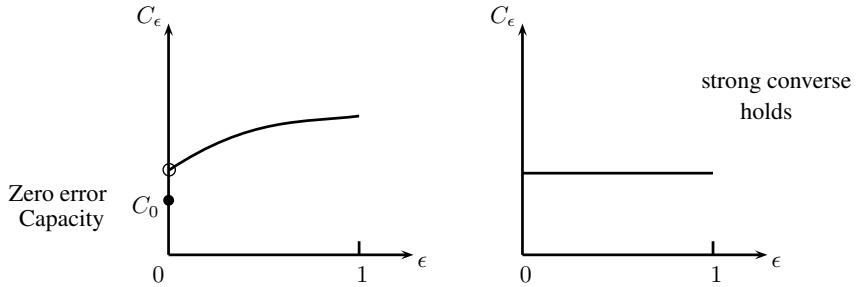
The case of zero-error ($\epsilon = 0$) is so different from $\epsilon > 0$ that the topic of $\epsilon = 0$ constitutes a separate subfield of its own (cf. the survey [178]). Introduced by Shannon in 1956 [?], the value

$$C_0 \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, 0) \tag{19.6}$$

is known as the *zero-error capacity* and represents the maximal achievable rate with no error whatsoever. Characterizing the value of C_0 is often a hard combinatorial problem. However, for many practically relevant channels it is quite trivial to show $C_0 = 0$. This is the case, for example, for the DMCs we considered before: the BSC or BEC. Indeed, for them we have $\log M^*(n, 0) = 0$ for all n , meaning transmitting any amount of information across these channels requires accepting some (perhaps vanishingly small) probability of error. Nevertheless, there are certain interesting and important channels for which C_0 is positive, cf. Section 23.3.1 for more.

As a function of ϵ the C_ϵ could (most generally) behave like the plot below on the left-hand side below. It may have a discontinuity at $\epsilon = 0$ and may be monotonically increasing (possibly even with jump discontinuities) in ϵ . Typically, however, C_ϵ is zero at $\epsilon = 0$ and stays constant for all $0 < \epsilon < 1$ and, hence, coincides with C (see the plot on the right-hand side). In such cases we say that the **strong converse** holds (more on this later in Section 22.1).

19.1 Channels and channel capacity 315



In Definition 19.3, the capacities C_ϵ and C are defined with respect to the average probability of error. By replacing M^* with M_{\max}^* , we can define, analogously, the capacities $C_\epsilon^{(\max)}$ and $C^{(\max)}$ with respect to the maximal probability of error. It turns out that these two definitions are equivalent, as the next theorem shows.

Theorem 19.4. $\forall \tau \in (0, 1)$,

$$\tau M^*(n, \epsilon(1 - \tau)) \leq M_{\max}^*(n, \epsilon) \leq M^*(n, \epsilon)$$

Proof. The second inequality is obvious, since any code that achieves a maximum error probability ϵ also achieves an average error probability of ϵ .

For the first inequality, take an $(n, M, \epsilon(1 - \tau))$ -code, and define the error probability for the j^{th} codeword as

$$\lambda_j \triangleq \mathbb{P}[\hat{W} \neq j | W = j]$$

Then

$$M(1 - \tau)\epsilon \geq \sum \lambda_j = \sum \lambda_j 1_{\{\lambda_j \leq \epsilon\}} + \sum \lambda_j 1_{\{\lambda_j > \epsilon\}} \geq \epsilon |\{j : \lambda_j > \epsilon\}|.$$

Hence $|\{j : \lambda_j > \epsilon\}| \leq (1 - \tau)M$. (Note that this is exactly Markov inequality.) Now by removing those codewords¹ whose λ_j exceeds ϵ , we can extract an $(n, \tau M, \epsilon)_{\max}$ -code. Finally, take $M = M^*(n, \epsilon(1 - \tau))$ to finish the proof. \square

Corollary 19.5 (Capacity under maximal probability of error). $C_\epsilon^{(\max)} = C_\epsilon$ for all $\epsilon > 0$ such that In particular, $C^{(\max)} = C$.

Proof. Using the definition of M^* and the previous theorem, for any fixed $\tau > 0$

$$C_\epsilon \geq C_\epsilon^{(\max)} \geq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \tau M^*(n, \epsilon(1 - \tau)) \geq C_{\epsilon(1-\tau)}$$

Sending $\tau \rightarrow 0$ yields $C_\epsilon \geq C_\epsilon^{(\max)} \geq C_{\epsilon-}$. \square

¹ This operation is usually referred to as *expurgation* which yields a smaller code by killing part of the codebook to reach a desired property.

19.2 Shannon's noisy channel coding theorem

Now that we have the basic definitions for Shannon capacity, we define another type of capacity, and show that for a *stationary memoryless* channels, these two notions (“operational” and “information”) of capacity coincide.

Definition 19.6. The *information capacity* of a channel is

$$C^{(I)} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n),$$

where for each n the supremum is taken over all joint distributions P_{X^n} on \mathcal{A}^n .

Note that information capacity $C^{(I)}$ so defined is not the same as the Shannon capacity C in Definition 19.3; as such, from first principles it has no direct interpretation as an operational quantity related to coding. Nevertheless, they are related by the following *coding theorems*. We start with a converse result:

Theorem 19.7 (Upper Bound for C_ϵ). *For any channel, $\forall \epsilon \in [0, 1]$, $C_\epsilon \leq \frac{C^{(I)}}{1-\epsilon}$ and $C \leq C^{(I)}$.*

Proof. Applying the general weak converse bound in Theorem 17.4 to $P_{Y^n|X^n}$ yields

$$\log M^*(n, \epsilon) \leq \frac{\sup_{P_{X^n}} I(X^n; Y^n) + h(\epsilon)}{1 - \epsilon}$$

Normalizing this by n and taking the \liminf as $n \rightarrow \infty$, we have

$$C_\epsilon = \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \frac{\sup_{P_{X^n}} I(X^n; Y^n) + h(\epsilon)}{1 - \epsilon} = \frac{C^{(I)}}{1 - \epsilon}.$$

□

Next we give an achievability bound:

Theorem 19.8 (Lower Bound for C_ϵ). *For a stationary memoryless channel, $C_\epsilon \geq \sup_{P_X} I(X; Y)$, for any $\epsilon \in (0, 1]$.*

Proof. Fix an arbitrary P_X on \mathcal{A} and let $P_{X^n} = P_X^{\otimes n}$ be an iid product of a single-letter distribution P_X . Recall Shannon's achievability bound Theorem 18.5 (or any other one would work just as well). From that result we know that for any n, M and any $\tau > 0$, there exists an (n, M, ϵ_n) -code with

$$\epsilon_n \leq \mathbb{P}[i(X^n; Y^n) \leq \log M + \tau] + \exp(-\tau)$$

Here the information density is defined with respect to the distribution $P_{X^n, Y^n} = P_{X, Y}^{\otimes n}$ and, therefore,

$$i(X^n; Y^n) = \sum_{k=1}^n \log \frac{dP_{X,Y}}{dP_X P_Y}(X_k, Y_k) = \sum_{k=1}^n i(X_k; Y_k),$$

19.2 Shannon's noisy channel coding theorem 317

where $i(x; y) = i_{P_{X,Y}}(x; y)$ and $i(x^n; y^n) = i_{P_{X^n,Y^n}}(x^n; y^n)$. What is important is that under P_{X^n,Y^n} the random variable $i(X^n; Y^n)$ is a sum of iid random variables with mean $I(X; Y)$. Thus, by the weak law of large numbers we have

$$\mathbb{P}[i(X^n; Y^n) < n(I(X; Y) - \delta)] \rightarrow 0$$

for any $\delta > 0$.

With this in mind, let us set $\log M = n(I(X; Y) - 2\delta)$ for some $\delta > 0$, and take $\tau = \delta n$ in Shannon's bound. Then for the error bound we have

$$\epsilon_n \leq \mathbb{P} \left[\sum_{k=1}^n i(X_k; Y_k) \leq nI(X; Y) - \delta n \right] + \exp(-\delta n) \xrightarrow{n \rightarrow \infty} 0, \quad (19.7)$$

Since the bound converges to 0, we have shown that there exists a sequence of (n, M_n, ϵ_n) -codes with $\epsilon_n \rightarrow 0$ and $\log M_n = n(I(X; Y) - 2\delta)$. Hence, for all n such that $\epsilon_n \leq \epsilon$

$$\log M^*(n, \epsilon) \geq n(I(X; Y) - 2\delta)$$

And so

$$C_\epsilon = \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon) \geq I(X; Y) - 2\delta$$

Since this holds for all P_X and all $\delta > 0$, we conclude $C_\epsilon \geq \sup_{P_X} I(X; Y)$. \square

The following result follows from pairing the upper and lower bounds on C_ϵ .

Theorem 19.9 (Shannon's Noisy Channel Coding Theorem [268]). *For a stationary memoryless channel,*

$$C = C^{(I)} = \sup_{P_X} I(X; Y). \quad (19.8)$$

As we mentioned several times already this result is among the most significant results in information theory. From the engineering point of view, the major surprise was that $C > 0$, i.e. communication over a channel is possible with strictly positive rate for any arbitrarily small probability of error. The way to achieve this is to encode the input data jointly (i.e. over many input bits together). This is drastically different from the pre-1948 methods, which operated on a letter-by-letter bases (such as Morse code). This theoretical result gave impetus (and still gives guidance) to the evolution of practical communication systems – quite a rare achievement for an asymptotic mathematical fact.

Proof. Statement (19.8) contains two equalities. The first one follows automatically from the second and Theorems 19.6 and 19.7. To show the second equality $C^{(I)} = \sup_{P_X} I(X; Y)$, we note that for stationary memoryless channels $C^{(I)}$ is in fact easy to compute. Indeed, rather than solving a sequence of optimization problems (one for each n) and taking the limit of $n \rightarrow \infty$, memorylessness of the channel implies that only the $n = 1$ problem needs to be solved. This type of result is known as “single-letterization” in information theory and we show it formally in the following proposition, which concludes the proof. \square

Proposition 19.10 (Memoryless input is optimal for memoryless channels).

- For memoryless channels,

$$\sup_{P_{X^n}} I(X^n; Y^n) = \sum_{i=1}^n \sup_{P_{X_i}} I(X_i; Y_i).$$

- For stationary memoryless channels,

$$C^{(I)} = \sup_{P_X} I(X; Y).$$

Proof. Recall that from Theorem 6.1 we know that for product kernels $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$, mutual information satisfies $I(X^n; Y^n) \leq \sum_{k=1}^n I(X_k; Y_k)$ with equality whenever X_i 's are independent. Then

$$C^{(I)} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n) = \liminf_{n \rightarrow \infty} \sup_{P_X} I(X; Y) = \sup_{P_X} I(X; Y).$$

□

Shannon's noisy channel theorem shows that by employing codes of large blocklength, we can approach the channel capacity arbitrarily close. Given the asymptotic nature of this result (or any other asymptotic result), a natural question is understanding the price to pay for reaching capacity. This can be understood in two ways:

- 1 The **complexity** of achieving capacity: Is it possible to find low-complexity encoders and decoders with polynomial number of operations in the blocklength n which achieve the capacity? This question was resolved by Forney [?] who showed that this is possible in *linear* time with exponentially small error probability.

Note that if we are content with polynomially small probability of error, e.g., $P_e = O(n^{-100})$, then we can construct polynomial-time decodable codes as follows. First, it can be shown that with rate strictly below capacity, the error probability of optimal codes decays exponentially w.r.t. the blocklength. Now divide the block of length n into shorter block of length $c \log n$ and apply the optimal code for blocklength $c \log n$ with error probability n^{-101} . Then by the union bound, the whole block has error with probability at most n^{-100} . The encoding and exhaustive-search decoding are obviously polynomial time.

- 2 The **speed** of achieving capacity: Suppose we want to achieve 90% of the capacity, we want to know how long do we need wait? The blocklength is a good proxy for delay. In other words, we want to know how fast the gap to capacity vanish as blocklength grows. Shannon's theorem shows that the gap $C - \frac{1}{n} \log M^*(n, \epsilon) = o(1)$. Next theorem shows that under proper conditions, the $o(1)$ term is in fact $O(\frac{1}{\sqrt{n}})$.

The main tool in the proof of Theorem 19.7 was the law of large numbers. The lower bound $C_\epsilon \geq C^{(I)}$ in Theorem 19.7 shows that $\log M^*(n, \epsilon) \geq nC + o(n)$ (this just restates the fact that normalizing by n and taking the liminf must result in something $\geq C$). If instead we apply

19.2 Shannon's noisy channel coding theorem 319

a more careful analysis using the central-limit theorem (CLT), we obtain the following refined achievability result.

Theorem 19.11. *Consider a stationary memoryless channel with a capacity-achieving input distribution. Namely, $C = \max_{P_X} I(X; Y) = I(X_*; Y_*)$ is attained at P_X^* , which induces $P_{X^*Y^*} = P_{X^*}P_{Y|X}$. Assume that $V = \text{Var}[i(X^*; Y^*)] < \infty$. Then*

$$\log M^*(n, \epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n}),$$

where $Q(\cdot)$ is the complementary Gaussian CDF and $Q^{-1}(\cdot)$ is its functional inverse.

Proof. Writing the little-o notation in terms of \liminf , our goal is

$$\liminf_{n \rightarrow \infty} \frac{\log M^*(n, \epsilon) - nC}{\sqrt{nV}} \geq -Q^{-1}(\epsilon) = \Phi^{-1}(\epsilon),$$

where $\Phi(t)$ is the standard normal CDF.

Recall Feinstein's bound

$$\exists(n, M, \epsilon)_{\max} : M \geq \beta (\epsilon - \mathbb{P}[i(X^n; Y^n) \leq \log \beta])$$

Take $\log \beta = nC + \sqrt{nVt}$, then applying the CLT gives

$$\begin{aligned} \log M &\geq nC + \sqrt{nVt} + \log \left(\epsilon - \mathbb{P} \left[\sum i(X_k; Y_k) \leq nC + \sqrt{nVt} \right] \right) \\ &\implies \log M \geq nC + \sqrt{nVt} + \log(\epsilon - \Phi(t)) \quad \forall \Phi(t) < \epsilon \\ &\implies \frac{\log M - nC}{\sqrt{nV}} \geq t + \frac{\log(\epsilon - \Phi(t))}{\sqrt{nV}}, \end{aligned}$$

where $\Phi(t)$ is the standard normal CDF. Taking the \liminf of both sides

$$\liminf_{n \rightarrow \infty} \frac{\log M^*(n, \epsilon) - nC}{\sqrt{nV}} \geq t,$$

for all t such that $\Phi(t) < \epsilon$. Finally, taking $t \nearrow \Phi^{-1}(\epsilon)$, and writing the \liminf in little-oh notation completes the proof

$$\log M^*(n, \epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n}).$$

□

Remark 19.1. Theorem 19.8 implies that for any $R < C$, there exists a sequence of $(n, \exp(nR), \epsilon_n)$ -codes such that the probability of error ϵ_n vanishes as $n \rightarrow \infty$. Examining the upper bound (19.7), we see that the probability of error actually vanishes exponentially fast, since the event in the first term is of large-deviations type (recall Chapter 15) so that both terms are exponentially small. Finding the value of the optimal exponent (or even the existence thereof) has a long history (but remains generally open) in information theory, see Section 22.4*. Recently, however, it was understood that a practically more relevant, and also much easier to analyze, is the regime of fixed (non-vanishing) error ϵ , in which case the main question is to bound the speed of convergence of $R \rightarrow C_\epsilon = C$. Previous theorem shows one bound on this speed of convergence. See Sections 22.5

and 22.6 for more. In particular, we will show that the bound on the \sqrt{n} term in Theorem 19.10 is often tight.

19.3 Examples of computing capacity

We compute the capacities of the simple DMCs from Fig. 19.1 and plot them in Fig. 19.2.

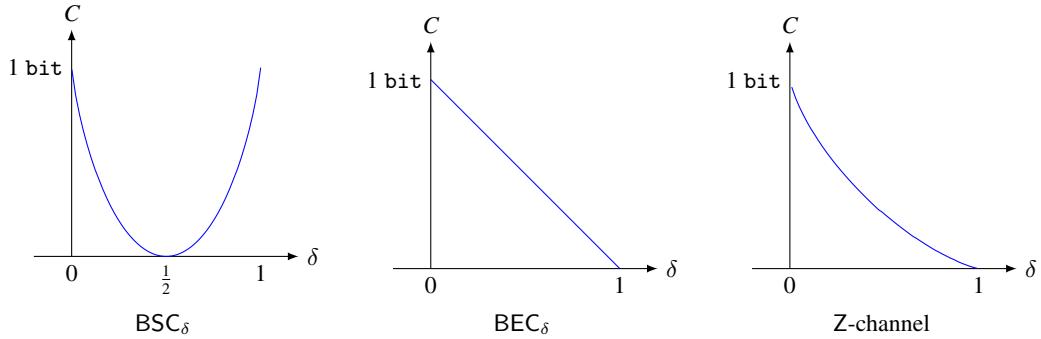


Figure 19.2 Capacities of three simple DMCs.

First for the BSC_δ we have the following description of the input-output law:

$$Y = X + Z \pmod{2}, \quad Z \sim \text{Ber}(\delta) \perp\!\!\!\perp X$$

To compute the capacity, let us notice

$$I(X; X+Z) = H(X+Z) - H(X+Z|X) = H(X+Z) - H(Z) \leq \log 2 - h(\delta)$$

with equality iff $X \sim \text{Ber}(1/2)$. Hence we have shown

$$C = \sup_{P_X} I(X; Y) = \log 2 - h(\delta)$$

More generally, for all additive-noise channel over a finite abelian group G , $C = \sup_{P_X} I(X; X+Z) = \log |G| - H(Z)$, achieved by $X \sim \text{Unif}(G)$.

Next we consider the binary erasure channel (BEC). BEC_δ is a *multiplicative* channel. Indeed, if we define the input $X \in \{\pm 1\}$ and output $Y \in \{\pm 1, 0\}$, then BEC relation can be written as

$$Y = XZ, \quad Z \sim \text{Ber}(\delta) \perp\!\!\!\perp X.$$

To compute the capacity, we first notice that even without evaluating Shannon's formula, it is clear that $C \leq 1 - \delta$ (bit), because for a large blocklength n about δ -fraction of the message is completely lost (even if the encoder knows a priori where the erasures are going to occur, the rate still cannot exceed $1 - \delta$). More formally, we notice that $\mathbb{P}[X = +1|Y = e] = \frac{\mathbb{P}[X=1]\delta}{\delta} = P[X = 1]$ and therefore

$$I(X; Y) = H(X) - H(X|Y) = H(X) - H(X|Y = e) \leq (1 - \delta)H(X) \leq (1 - \delta)\log 2,$$

with equality iff $X \sim \text{Ber}(1/2)$. Thus we have shown

$$C = \sup_{P_X} I(X; Y) = 1 - \delta \text{ bits}$$

Finally, the Z-channel can be thought of as a multiplicative channel with transition law

$$Y = XZ, \quad X \in \{0, 1\} \perp Z \sim \text{Ber}(1 - \delta),$$

so that $\mathbb{P}[Z = 0] = \delta$. For this channel if $X \sim \text{Ber}(p)$ we have

$$I(X; Y) = H(Y) - H(Y|X) = h(p(1 - \delta)) - ph(\delta).$$

Optimizing over p we get that the optimal input is given by

$$p^*(\delta) = \frac{1}{1 - \delta} \frac{1}{1 + \exp\left\{\frac{h(\delta)}{1 - \delta}\right\}}.$$

The capacity-achieving input distribution $p^*(\delta)$ monotonically decreases from $\frac{1}{2}$ when $\delta = 0$ to $\frac{1}{e}$ when $\delta \rightarrow 1$. (Infamously, there is no “explanation” for this latter limiting value.) For the capacity, thus, we get

$$C = h(p^*(\delta)(1 - \delta)) - p^*(\delta)h(\delta).$$

19.4* Symmetric channels

Definition 19.12. A pair of measurable maps $f = (f_i, f_o)$ is a symmetry of $P_{Y|X}$ if

$$P_{Y|X}(f_o^{-1}(E)|f_i(x)) = P_{Y|X}(E|x),$$

for all measurable $E \subset \mathcal{Y}$ and $x \in \mathcal{X}$. Two symmetries f and g can be composed to produce another symmetry as

$$(g_i, g_o) \circ (f_i, f_o) \triangleq (g_i \circ f_i, g_o \circ f_o). \quad (19.9)$$

A symmetry group G of $P_{Y|X}$ is any collection of invertible symmetries (automorphisms) closed under the group operation (19.9).

Note that both components of an automorphism $f = (f_i, f_o)$ are bimeasurable bijections, that is $f_i, f_i^{-1}, f_o, f_o^{-1}$ are all measurable and well-defined functions.

Naturally, every symmetry group G possesses a canonical left action on $\mathcal{X} \times \mathcal{Y}$ defined as

$$g \cdot (x, y) \triangleq (g_i(x), g_o^{-1}(y)). \quad (19.10)$$

Since the action on $\mathcal{X} \times \mathcal{Y}$ splits into actions on \mathcal{X} and \mathcal{Y} , we will abuse notation slightly and write

$$g \cdot (x, y) \triangleq (gx, gy).$$

For the cases of infinite \mathcal{X}, \mathcal{Y} we need to impose certain additional regularity conditions:

Definition 19.13. A symmetry group G is called regular if it possesses a left-invariant Haar probability measure ν such that the group action (19.10)

$$G \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X} \times \mathcal{Y}$$

is measurable.

Note that under the regularity assumption the action (19.10) also defines a left action of G on $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ according to

$$(gP_X)[E] \triangleq P_X[g^{-1}E], \quad (19.11)$$

$$(gQ_Y)[E] \triangleq Q_Y[g^{-1}E], \quad (19.12)$$

or, in words, if $X \sim P_X$ then $gX \sim gP_X$, and similarly for Y and gY . For every distribution P_X we define an averaged distribution \bar{P}_X as

$$\bar{P}_X[E] \triangleq \int_G P_X[g^{-1}E]\nu(dg), \quad (19.13)$$

which is the distribution of random variable gX when $g \sim \nu$ and $X \sim P_X$. The measure \bar{P}_X is G -invariant, in the sense that $g\bar{P}_X = \bar{P}_X$. Indeed, by left-invariance of ν we have for every bounded function f

$$\int_G f(g)\nu(dg) = \int_G f(hg)\nu(dg) \quad \forall h \in G,$$

and therefore

$$\bar{P}_X[h^{-1}E] = \int_G P_X[(hg)^{-1}E]\nu(dg) = \bar{P}_X[E].$$

Similarly one defines \bar{Q}_Y :

$$\bar{Q}_Y[E] \triangleq \int_G Q_Y[g^{-1}E]\nu(dg), \quad (19.14)$$

which is also G -invariant: $g\bar{Q}_Y = \bar{Q}_Y$.

The main property of the action of G may be rephrased as follows: For arbitrary $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ we have

$$\begin{aligned} & \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi(x, y) P_{Y|X}(dy|x) (gP_X)(dx) \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \phi(gx, gy) P_{Y|X}(dy|x) P_X(dx). \end{aligned} \quad (19.15)$$

In other words, if the pair (X, Y) is generated by taking $X \sim P_X$ and applying $P_{Y|X}$, then the pair (gX, gY) has marginal distribution gP_X but conditional kernel is still $P_{Y|X}$. For finite \mathcal{X}, \mathcal{Y} this is equivalent to

$$P_{Y|X}(gy|gx) = P_{Y|X}(y|x), \quad (19.16)$$

19.4* Symmetric channels 323

which may also be taken as the definition of the automorphism. In terms of the G -action on $\mathcal{P}(\mathcal{Y})$ we may also say:

$$gP_{Y|X=x} = P_{Y|X=gx} \quad \forall g \in G, x \in \mathcal{X}. \quad (19.17)$$

It is not hard to show that for any channel and a regular group of symmetries G the capacity-achieving output distribution must be G -invariant, and capacity-achieving input distribution can be chosen to be G -invariant. That is, the saddle point equation

$$\inf_{P_X} \sup_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) = \sup_{Q_Y} \inf_{P_X} D(P_{Y|X} \| Q_Y | P_X),$$

can be solved in the class of G -invariant distribution. Often, the action of G is transitive on \mathcal{X} (\mathcal{Y}), in which case the capacity-achieving input (output) distribution can be taken to be uniform.

Below we systematize many popular notions of channel symmetry and explain relationship between them.

- $P_{Y|X}$ is called input-symmetric (output-symmetric) if there exists a regular group of symmetries G acting transitively on \mathcal{X} (\mathcal{Y}).
- An input-symmetric channel with a binary \mathcal{X} is known as BMS (for Binary Memoryless Symmetric). These channels possess a rich theory [255, Section 4.1].
- $P_{Y|X}$ is called weakly input-symmetric if there exists an $x_0 \in \mathcal{X}$ and a channel $T_x : \mathcal{B} \rightarrow \mathcal{B}$ for each $x \in \mathcal{X}$ such that $T_x \circ P_{Y|X=x_0} = P_{Y|X=x}$ and $T_x \circ P_Y^*$, where P_Y^* is the caod. In [230, Section 3.4.5] it is shown that the allowing for a randomized maps T_x is essential and that not all $P_{Y|X}$ are weakly input-symmetric.
- DMC $P_{Y|X}$ is a *group-noise channel* if $\mathcal{X} = \mathcal{Y}$ is a group and $P_{Y|X}$ acts by composing X with a noise variable Z :

$$Y = X \circ Z,$$

where \circ is a group operation and Z is independent of X .

- DMC $P_{Y|X}$ is called *Dobrushin-symmetric* if every row of $P_{Y|X}$ is a permutation of the first one and every column of $P_{Y|X}$ is a permutation of the first one; see [96].
- DMC $P_{Y|X}$ is called *Gallager-symmetric* if the output alphabet \mathcal{Y} can be split into a disjoint union of sub-alphabets such that restricted to each sub-alphabet $P_{Y|X}$ has the Dobrushin property: every row (every column) is a permutation of the first row (column); see [130, Section 4.5].
- for convenience, say that the channel is *square* if $|\mathcal{X}| = |\mathcal{Y}|$.

We demonstrate some of the relationship between these various notions of symmetry:

- Note that it is an easy consequence of the definitions that any input-symmetric (resp. output-symmetric) channel's $P_{Y|X}$ has all rows (resp. columns) – permutations of the first row (resp. column). Hence,

$$\text{input-symmetric, output-symmetric} \implies \text{Dobrushin} \quad (19.18)$$

- Group-noise channels satisfy all other definitions of symmetry:

$$\text{group-noise} \implies \text{square, input/output-symmetric} \quad (19.19)$$

$$\implies \text{Dobrushin, Gallager} \quad (19.20)$$

- Since Gallager symmetry implies all rows are permutations of the first one, while output symmetry implies the same statement for columns we have

$$\text{Gallager, output-symmetric} \implies \text{Dobrushin}$$

- Clearly, not every Dobrushin-symmetric channel is square. One may wonder, however, whether every square Dobrushin channel is a group-noise channel. This is not so. Indeed, according to [276] the latin squares that are Cayley tables are precisely the ones in which composition of two rows (as permutations) gives another row. An example of the latin square which is not a Cayley table is the following:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 1 & 3 \\ 3 & 1 & 2 & 5 & 4 \\ 4 & 3 & 5 & 2 & 1 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}. \quad (19.21)$$

Thus, by multiplying this matrix by $\frac{1}{15}$ we obtain a counter-example:

$$\text{Dobrushin, square} \not\implies \text{group-noise}$$

In fact, this channel is not even input-symmetric. Indeed, suppose there is $g \in G$ such that $g4 = 1$ (on \mathcal{X}). Then, applying (19.16) with $x = 4$ we figure out that on \mathcal{Y} the action of g must be:

$$1 \mapsto 4, 2 \mapsto 3, 3 \mapsto 5, 4 \mapsto 2, 5 \mapsto 1.$$

But then we have

$$gP_{Y|X=1} = (5 \ 4 \ 2 \ 1 \ 3) \cdot \frac{1}{15},$$

which by a simple inspection does not match any of the rows in (19.21). Thus, (19.17) cannot hold for $x = 1$. We conclude:

$$\text{Dobrushin, square} \not\implies \text{input-symmetric}$$

Similarly, if there were $g \in G$ such that $g2 = 1$ (on \mathcal{Y}), then on \mathcal{X} it would act as

$$1 \mapsto 2, 2 \mapsto 5, 3 \mapsto 1, 4 \mapsto 3, 5 \mapsto 4,$$

which implies via (19.16) that $P_{Y|X}(g1|x)$ is not a column of (19.21). Thus:

$$\text{Dobrushin, square} \not\implies \text{output-symmetric}$$

19.5* Information Stability 325

- Clearly, not every input-symmetric channel is Dobrushin (e.g., BEC). One may even find a counter-example in the class of square channels:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \\ 4 & 2 & 3 & 1 \\ 4 & 3 & 2 & 1 \end{pmatrix} \cdot \frac{1}{10} \quad (19.22)$$

This shows:

$$\text{input-symmetric, square} \not\Rightarrow \text{Dobrushin}$$

- Channel (19.22) also demonstrates:

$$\text{Gallager-symmetric, square} \not\Rightarrow \text{Dobrushin}.$$

- Example (19.22) naturally raises the question of whether every input-symmetric channel is Gallager symmetric. The answer is positive: by splitting \mathcal{Y} into the orbits of G we see that a subchannel $\mathcal{X} \rightarrow \{\text{orbit}\}$ is input and output symmetric. Thus by (19.18) we have:

$$\text{input-symmetric} \Rightarrow \text{Gallager-symmetric} \Rightarrow \text{weakly input-symmetric} \quad (19.23)$$

(The second implication is evident).

- However, not all weakly input-symmetric channels are Gallager-symmetric. Indeed, consider the following channel

$$W = \begin{pmatrix} 1/7 & 4/7 & 1/7 & 1/7 \\ 4/7 & 1/7 & 0 & 4/7 \\ 0 & 0 & 4/7 & 2/7 \\ 2/7 & 2/7 & 2/7 & 0 \end{pmatrix}. \quad (19.24)$$

Since $\det W \neq 0$, the capacity achieving input distribution is unique. Since $H(Y|X = x)$ is independent of x and $P_X = [1/4, 1/4, 3/8, 1/8]$ achieves uniform P_Y^* it must be the unique optimum. Clearly any permutation T_x fixes a uniform P_Y^* and thus the channel is weakly input-symmetric. At the same time it is not Gallager-symmetric since no row is a permutation of another.

- For more on the properties of weakly input-symmetric channels see [230, Section 3.4.5].

A pictorial representation of these relationships between the notions of symmetry is given schematically on Fig. 19.3.

19.5* Information Stability

We saw that $C = C^{(I)}$ for stationary memoryless channels, but what other channels does this hold for? And what about non-stationary channels? To answer this question, we introduce the notion of *information stability*.

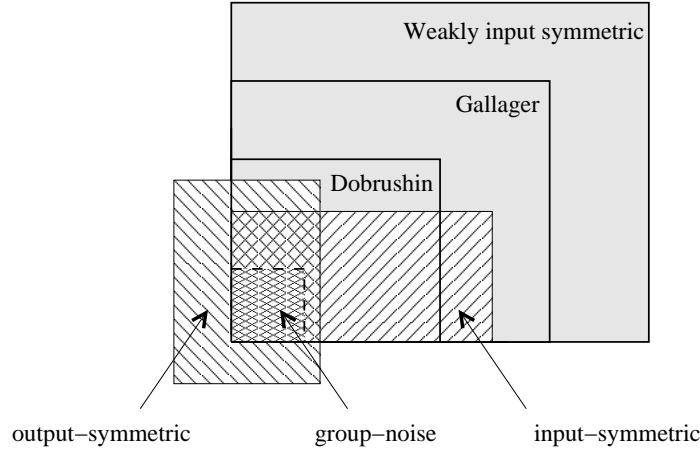


Figure 19.3 Schematic representation of inclusions of various classes of channels.

Definition 19.14. A channel is called *information stable* if there exists a sequence of input distributions $\{P_{X^n}, n = 1, 2, \dots\}$ such that

$$\frac{1}{n} i(X^n; Y^n) \xrightarrow{\mathbb{P}} C^{(I)} .$$

For example, we can pick $P_{X^n} = (P_X^*)^n$ for stationary memoryless channels. Therefore stationary memoryless channels are information stable.

The purpose for defining information stability is the following theorem.

Theorem 19.15. For an information stable channel, $C = C^{(I)}$.

Proof. Like the stationary, memoryless case, the upper bound comes from the general converse Theorem 17.4, and the lower bound uses a similar strategy as Theorem 19.7, except utilizing the definition of information stability in place of WLLN. \square

The next theorem gives conditions to check for information stability in memoryless channels which are *not* necessarily stationary.

Theorem 19.16. A memoryless channel is information stable if there exists $\{X_k^* : k \geq 1\}$ such that both of the following hold:

$$\frac{1}{n} \sum_{k=1}^n I(X_k^*; Y_k^*) \rightarrow C^{(I)} \quad (19.25)$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \text{Var}[i(X_n^*; Y_n^*)] < \infty . \quad (19.26)$$

In particular, this is satisfied if

$$|\mathcal{A}| < \infty \text{ or } |\mathcal{B}| < \infty \quad (19.27)$$

Proof. To show the first part, it is sufficient to prove

$$\mathbb{P} \left[\frac{1}{n} \left| \sum_{k=1}^n i(X_k^*; Y_k^*) - I(X_k^*, Y_k^*) \right| > \delta \right] \rightarrow 0$$

So that $\frac{1}{n} i(X^n; Y^n) \rightarrow C^{(I)}$ in probability. We bound this by Chebyshev's inequality

$$\mathbb{P} \left[\frac{1}{n} \left| \sum_{k=1}^n i(X_k^*; Y_k^*) - I(X_k^*, Y_k^*) \right| > \delta \right] \leq \frac{\frac{1}{n^2} \sum_{k=1}^n \text{Var}[i(X_k^*; Y_k^*)]}{\delta^2} \rightarrow 0,$$

where convergence to 0 follows from Kronecker lemma (Lemma 19.17 to follow) applied with $b_n = n^2$, $x_n = \text{Var}[i(X_n^*; Y_n^*)]/n^2$.

The second part follows from the first. Indeed, notice that

$$C^{(I)} = \liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \sup_{P_{X_k}} I(X_k; Y_k).$$

Now select $P_{X_k^*}$ such that

$$I(X_k^*; Y_k^*) \geq \sup_{P_{X_k}} I(X_k; Y_k) - 2^{-k}.$$

(Note that each $\sup_{P_{X_k}} I(X_k; Y_k) \leq \log \min\{|\mathcal{A}|, |\mathcal{B}|\} < \infty$.) Then, we have

$$\sum_{k=1}^n I(X_k^*; Y_k^*) \geq \sum_{k=1}^n \sup_{P_{X_k}} I(X_k; Y_k) - 1,$$

and hence normalizing by n we get (19.25). We next show that for any joint distribution $P_{X,Y}$ we have

$$\text{Var}[i(X; Y)] \leq 2 \log^2(\min(|\mathcal{A}|, |\mathcal{B}|)). \quad (19.28)$$

The argument is symmetric in X and Y , so assume for concreteness that $|\mathcal{B}| < \infty$. Then

$$\begin{aligned} & \mathbb{E}[i^2(X; Y)] \\ & \triangleq \int_{\mathcal{A}} dP_X(x) \sum_{y \in \mathcal{B}} P_{Y|X}(y|x) \left[\log^2 P_{Y|X}(y|x) + \log^2 P_Y(y) - 2 \log P_{Y|X}(y|x) \cdot \log P_Y(y) \right] \\ & \leq \int_{\mathcal{A}} dP_X(x) \sum_{y \in \mathcal{B}} P_{Y|X}(y|x) \left[\log^2 P_{Y|X}(y|x) + \log^2 P_Y(y) \right] \end{aligned} \quad (19.29)$$

$$\begin{aligned} & = \int_{\mathcal{A}} dP_X(x) \left[\sum_{y \in \mathcal{B}} P_{Y|X}(y|x) \log^2 P_{Y|X}(y|x) \right] + \left[\sum_{y \in \mathcal{B}} P_Y(y) \log^2 P_Y(y) \right] \\ & \leq \int_{\mathcal{A}} dP_X(x) g(|\mathcal{B}|) + g(|\mathcal{B}|) \end{aligned} \quad (19.30)$$

$$=2g(|\mathcal{B}|),$$

where (19.29) is because $2 \log P_{Y|X}(y|x) \cdot \log P_Y(y)$ is always non-negative, and (19.30) follows because each term in square-brackets can be upper-bounded using the following optimization problem:

$$g(n) \triangleq \sup_{a_j \geq 0: \sum_{j=1}^n a_j = 1} \sum_{j=1}^n a_j \log^2 a_j. \quad (19.31)$$

Since the $x \log^2 x$ has unbounded derivative at the origin, the solution of (19.31) is always in the interior of $[0, 1]^n$. Then it is straightforward to show that for $n > e$ the solution is actually $a_j = \frac{1}{n}$. For $n = 2$ it can be found directly that $g(2) = 0.5629 \log^2 2 < \log^2 2$. In any case,

$$2g(|\mathcal{B}|) \leq 2 \log^2 |\mathcal{B}|.$$

Finally, because of the symmetry, a similar argument can be made with $|\mathcal{B}|$ replaced by $|\mathcal{A}|$. \square

Lemma 19.17 (Kronecker Lemma). *Let a sequence $0 < b_n \nearrow \infty$ and a non-negative sequence $\{x_n\}$ such that $\sum_{n=1}^{\infty} x_n < \infty$, then*

$$\frac{1}{b_n} \sum_{j=1}^n b_j x_j \rightarrow 0$$

Proof. Since b_n 's are strictly increasing, we can split up the summation and bound them from above

$$\sum_{k=1}^n b_k x_k \leq b_m \sum_{k=1}^m x_k + \sum_{k=m+1}^n b_k x_k$$

Now throw in the rest of the x_k 's in the summation

$$\begin{aligned} &\Rightarrow \frac{1}{b_n} \sum_{k=1}^n b_k x_k \leq \frac{b_m}{b_n} \sum_{k=1}^{\infty} x_k + \sum_{k=m+1}^n \frac{b_k}{b_n} x_k \leq \frac{b_m}{b_n} \sum_{k=1}^{\infty} x_k + \sum_{k=m+1}^{\infty} x_k \\ &\Rightarrow \lim_{n \rightarrow \infty} \frac{1}{b_n} \sum_{k=1}^n b_k x_k \leq \sum_{k=m+1}^{\infty} x_k \rightarrow 0 \end{aligned}$$

Since this holds for any m , we can make the last term arbitrarily small. \square

How to show information stability? One important class of channels with memory for which information stability can be shown easily are Gaussian channels. The complete details will be shown below (see Sections 20.5* and 20.6*), but here we demonstrate a crucial fact.

For jointly Gaussian (X, Y) we always have bounded variance:

$$\text{Var}[i(X; Y)] = \rho^2(X, Y) \log^2 e \leq \log^2 e, \quad \rho(X, Y) = \frac{\text{cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}. \quad (19.32)$$

19.6 Capacity under bit error rate 329

Indeed, first notice that we can always represent $Y = \tilde{X} + Z$ with $\tilde{X} = aX \perp\!\!\!\perp Z$. On the other hand, we have

$$i(\tilde{x}; y) = \frac{\log e}{2} \left[\frac{\tilde{x}^2 + 2\tilde{x}z}{\sigma_Y^2} - \frac{\sigma^2}{\sigma_Y^2 \sigma_Z^2} z^2 \right], \quad z \triangleq y - \tilde{x}.$$

From here by using $\text{Var}[\cdot] = \text{Var}[\mathbb{E}[\cdot|\tilde{X}]] + \text{Var}[\cdot|\tilde{X}]$ we need to compute two terms separately:

$$\mathbb{E}[i(\tilde{X}; Y)|\tilde{X}] = \frac{\log e}{2} \left[\frac{\tilde{X}^2 - \frac{\sigma_X^2}{\sigma_Z^2}}{\sigma_Y^2} \right],$$

and hence

$$\text{Var}[\mathbb{E}[i(\tilde{X}; Y)|\tilde{X}]] = \frac{2 \log^2 e}{4 \sigma_Y^4} \sigma_X^4.$$

On the other hand,

$$\text{Var}[i(\tilde{X}; Y)|\tilde{X}] = \frac{2 \log^2 e}{4 \sigma_Y^4} [4\sigma_X^2 \sigma_Z^2 + 2\sigma_X^4].$$

Putting it all together we get (19.32). Inequality (19.32) justifies information stability of all sorts of Gaussian channels (memoryless and with memory), as we will see shortly.

19.6 Capacity under bit error rate

In most cases of interest the space $[M]$ of messages can be given additional structure by identifying $[M] = \{0, 1\}^k$, which is, of course, only possible for $M = 2^k$. In these cases, in addition to P_e and $P_{e,\max}$ every code (f, g) has another important figure of merit – the so called *Bit Error Rate (BER)*, denoted as P_b . In fact, we have already defined a similar concept in Section 6.5:

$$P_b \triangleq \frac{1}{k} \sum_{j=1}^k \mathbb{P}[S_j \neq \hat{S}_j] = \frac{1}{k} \mathbb{E}[d_H(S^k, \hat{S}^k)], \quad (19.33)$$

where we represented W and \hat{W} as k -tuples: $W = (S_1, \dots, S_k)$, $\hat{W} = (\hat{S}_1, \dots, \hat{S}_k)$, and d_H denotes the Hamming distance (6.14). In words, P_b is the average fraction of errors in a decoded k -bit block.

In addition to constructing codes minimizing block error probability P_e or $P_{e,\max}$ as we studied above, the problem of minimizing the BER P_b is also practically relevant. Here, we discuss some simple facts about this setting. In particular, we will see that the capacity value for memoryless channels does not increase even if one insists only on a vanishing P_b – a much weaker criterion compared to vanishing P_e .

First, we give a bound on the average probability of error (block error rate) in terms of the bit error rate.

Theorem 19.18. *For all (f, g) , $M = 2^k \implies P_b \leq P_e \leq kP_b$*

Proof. Recall that $M = 2^k$ gives us the interpretation of $W = S^k$ sequence of bits.

$$\frac{1}{k} \sum_{i=1}^k 1\{S_i \neq \hat{S}_i\} \leq 1\{S^k \neq \hat{S}^k\} \leq \sum_{i=1}^k 1\{S_i \neq \hat{S}_i\},$$

where the first inequality is obvious and the second follow from the union bound. Taking expectation of the above expression gives the theorem. \square

Next, the following pair of results is often useful for lower bounding P_b for some specific codes.

Theorem 19.19 (Assouad). *If $M = 2^k$ then*

$$P_b \geq \min\{\mathbb{P}[\hat{W} = c' | W = c] : c, c' \in \{0, 1\}^k, d_H(c, c') = 1\}.$$

Proof. Let e_i be a length k vector that is 1 in the i -th position, and zero everywhere else. Then

$$\sum_{i=1}^k 1\{S_i \neq \hat{S}_i\} \geq \sum_{i=1}^k 1\{S^k = \hat{S}^k + e_i\}$$

Dividing by k and taking expectation gives

$$\begin{aligned} P_b &\geq \frac{1}{k} \sum_{i=1}^k \mathbb{P}[S^k = \hat{S}^k + e_i] \\ &\geq \min\{\mathbb{P}[\hat{W} = c' | W = c] : c, c' \in \{0, 1\}^k, d_H(c, c') = 1\}. \end{aligned}$$

\square

Similarly, we can prove the following generalization:

Theorem 19.20. *If $A, B \in \{0, 1\}^k$ (with arbitrary marginals!) then for every $r \geq 1$ we have*

$$P_b = \frac{1}{k} \mathbb{E}[d_H(A, B)] \geq \binom{k-1}{r-1} P_{r,\min} \quad (19.34)$$

$$P_{r,\min} \triangleq \min\{\mathbb{P}[B = c' | A = c] : c, c' \in \{0, 1\}^k, d_H(c, c') = r\} \quad (19.35)$$

Proof. First, observe that

$$\mathbb{P}[d_H(A, B) = r | A = a] = \sum_{b: d_H(a, b)=r} P_{B|A}(b|a) \geq \binom{k}{r} P_{r,\min}.$$

Next, notice

$$d_H(x, y) \geq r 1\{d_H(x, y) = r\}$$

and take the expectation with $x \sim A, y \sim B$. \square

In statistics, Assouad's Lemma is a useful tool for obtaining lower bounds on the minimax risk of an estimator. See Section 31.2 for more.

The following is a converse bound for channel coding under BER constraint.

19.7 Joint Source Channel Coding 331

Theorem 19.21 (Converse under BER). Any M -code with $M = 2^k$ and bit-error rate P_b satisfies

$$\log M \leq \frac{\sup_{P_X} I(X; Y)}{\log 2 - h(P_b)}.$$

Proof. Note that $S^k \rightarrow X \rightarrow Y \rightarrow \hat{S}^k$, where $S^k \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. Recall from Theorem 6.1 that for iid S^n , $\sum I(S_i; \hat{S}_i) \leq I(S^k; \hat{S}^k)$. This gives us

$$\begin{aligned} \sup_{P_X} I(X; Y) &\geq I(X; Y) \geq \sum_{i=1}^k I(S_i; \hat{S}_i) \\ &\geq k \frac{1}{k} \sum d\left(\mathbb{P}[S_i = \hat{S}_i] \middle\| \frac{1}{2}\right) \\ &\geq kd\left(\frac{1}{k} \sum_{i=1}^k \mathbb{P}[S_i = \hat{S}_i] \middle\| \frac{1}{2}\right) \\ &= kd\left(1 - P_b \middle\| \frac{1}{2}\right) = k(\log 2 - h(P_b)) \end{aligned}$$

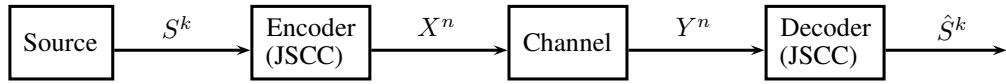
where the second line used Fano's inequality (Theorem 6.4) for binary random variables (or data processing inequality for divergence), and the third line used the convexity of divergence.² \square

Pairing this bound with Proposition 19.9 shows that any sequence of codes with $P_b \rightarrow 0$ (for a memoryless channel) must have rate $R < C$. In other words, relaxing the constraint from P_e to P_b does not yield any higher rates.

Later in Section 26.3 we will see that channel coding under BER constraint is a special case of a more general paradigm known as lossy joint source channel coding so that Theorem 19.21 follows from Theorem 26.8. Furthermore, this converse bound is in fact achievable asymptotically for stationary memoryless channels.

19.7 Joint Source Channel Coding

Now we will examine a slightly different data transmission scenario called *Joint Source Channel Coding* (JSCC):



Formally, a JSCC code consists of an encoder $f : \mathcal{A}^k \rightarrow \mathcal{X}^n$ and a decoder $g : \mathcal{Y}^n \rightarrow \mathcal{A}^k$. The goal is to maximize the transmission rate $R = \frac{k}{n}$ (symbol per channel use) while ensuring the

² Note that this last chain of inequalities is similar to the proof of Proposition 6.7.

probability of error $\mathbb{P}[S^k \neq \hat{S}^k]$ is small. The fundamental limit (optimal probability of error) is defined as

$$\epsilon_{\text{JSCC}}^*(k, n) = \inf_{f,g} \mathbb{P}[S^k \neq \hat{S}^k]$$

In channel coding we are interested in transmitting M messages and all messages are born equal. Here we want to convey the source realizations which might not be equiprobable (has redundancy). Indeed, if S^k is uniformly distributed on, say, $\{0, 1\}^k$, then we are back to the channel coding setup with $M = 2^k$ under average probability of error, and $\epsilon_{\text{JSCC}}^*(k, n)$ coincides with $\epsilon^*(n, 2^k)$ defined in Section 22.1.

Here, we look for a clever scheme to directly encode k symbols from \mathcal{A} into a length n channel input such that we achieve a small probability of error over the channel. This feels like a mix of two problems we've seen: compressing a source and coding over a channel. The following theorem shows that compressing and channel coding separately is optimal. This is a relief, since it implies that we do not need to develop any new theory or architectures to solve the Joint Source Channel Coding problem. As far as the leading term in the asymptotics is concerned, the following two-stage scheme is optimal: First use the optimal compressor to *eliminate all the redundancy* in the source, then use the optimal channel code to *add redundancy to combat the noise* in the data transmission.

Theorem 19.22. *Let the source $\{S_k\}$ be stationary memoryless on a finite alphabet with entropy H . Let the channel be stationary memoryless with finite capacity C . Then*

$$\epsilon_{\text{JSCC}}^*(nR, n) \begin{cases} \rightarrow 0 & R < C/H \\ \not\rightarrow 0 & R > C/H \end{cases} \quad n \rightarrow \infty.$$

The interpretation of this result is as follows: Each source symbol has information content (entropy) H bits. Each channel use can convey C bits. Therefore to reliably transmit k symbols over n channel uses, we need $kH \leq nC$.

Proof. (Achievability.) The idea is to separately compress our source and code it for transmission. Since this is a feasible way to solve the JSCC problem, it gives an achievability bound. This separated architecture is

$$S^k \xrightarrow{f_1} W \xrightarrow{f_2} X^n \xrightarrow{P_{Y^n|X^n}} Y^n \xrightarrow{g_2} \hat{W} \xrightarrow{g_1} \hat{S}^k$$

Where we use the optimal compressor (f_1, g_1) and optimal channel code (*maximum* probability of error) (f_2, g_2). Let W denote the output of the compressor which takes at most M_k values. Then from Corollary 11.1 and Theorem 19.8 we get:

$$(\text{From optimal compressor}) \frac{1}{k} \log M_k > H + \delta \implies \mathbb{P}[\hat{S}^k \neq S^k(W)] \leq \epsilon \quad \forall k \geq k_0$$

$$(\text{From optimal channel code}) \frac{1}{n} \log M_k < C - \delta \implies \mathbb{P}[\hat{W} \neq m | W = m] \leq \epsilon \quad \forall m, \forall k \geq k_0$$

19.7 Joint Source Channel Coding 333

Using both of these,

$$\begin{aligned}\mathbb{P}[S^k \neq \hat{S}^k(\hat{W})] &\leq \mathbb{P}[S^k \neq \hat{S}^k, W = \hat{W}] + \mathbb{P}[W \neq \hat{W}] \\ &\leq \mathbb{P}[S^k \neq \hat{S}^k(W)] + \mathbb{P}[W \neq \hat{W}] \leq \epsilon + \epsilon\end{aligned}$$

And therefore if $R(H + \delta) < C - \delta$, then $\epsilon^* \rightarrow 0$. By the arbitrariness of $\delta > 0$, we conclude the weak converse for any $R > C/H$.

(Converse.) To prove the converse notice that any JSCC encoder/decoder induces a Markov chain

$$S^k \rightarrow X^n \rightarrow Y^n \rightarrow \hat{S}^k.$$

Applying data processing for mutual information

$$I(S^k; \hat{S}^k) \leq I(X^n; Y^n) \leq \sup_{P_{X^n}} I(X^n; Y^n) = nC.$$

On the other hand, since $\mathbb{P}[S^k \neq \hat{S}^k] \leq \epsilon_n$, Fano's inequality (Theorem 6.4) yields

$$I(S^k; \hat{S}^k) = H(S^k) - H(S^k | \hat{S}^k) \geq kH - \epsilon_n \log |\mathcal{A}|^k - \log 2.$$

Combining the two gives

$$nC \geq kH - \epsilon_n \log |\mathcal{A}|^k - \log 2. \quad (19.36)$$

Since $R = \frac{k}{n}$, dividing both sides by n and sending $n \rightarrow \infty$ yields

$$\liminf_{n \rightarrow \infty} \epsilon_n \geq \frac{RH - C}{R \log |\mathcal{A}|}.$$

Therefore ϵ_n does not vanish if $R > C/H$. □

We remark that instead of using Fano's inequality we could have lower bounded $I(S^k; \hat{S}^k)$ as in the proof of Theorem 17.4 by defining $Q_{S^k \hat{S}^k} = U_{S^k} P_{\hat{S}^k}$ (with $U_{S^k} = \text{Unif}(\{0, 1\}^k)$) and applying the data processing inequality to the map $(S^k, \hat{S}^k) \mapsto 1\{S^k = \hat{S}^k\}$:

$$D(P_{S^k \hat{S}^k} \| Q_{S^k \hat{S}^k}) = D(P_{S^k} \| U_{S^k}) + D(P_{\hat{S}^k | S^k} \| P_{\hat{S}^k} | P_{S^k}) \geq d(1 - \epsilon_n \|\mathcal{A}\|^{-k})$$

Rearranging terms yields (19.36). As we discussed in Remark 17.5, replacing D with other f -divergences can be very fruitful.

In a very similar manner, by invoking Corollary 12.1 and Theorem 19.15 we obtain:

Theorem 19.23. *Let source $\{S_k\}$ be ergodic on a finite alphabet, and have entropy rate H . Let the channel have capacity C and be information stable. Then*

$$\lim_{n \rightarrow \infty} \epsilon_{\text{JSCC}}^*(nR, n) \begin{cases} = 0 & R > H/C \\ > 0 & R < H/C \end{cases}$$

334

We leave the proof as an exercise.

The figure illustrates the power allocation via water-filling. In this particular case, the second branch is too noisy (σ_2 too big) such that it is better be discarded, i.e., the assigned power is zero.

20

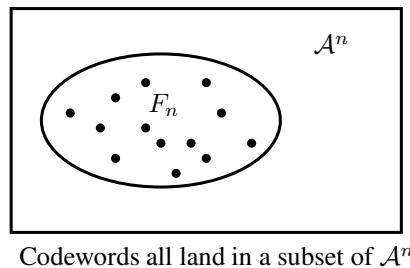
Channels with input constraints. Gaussian channels.

In this chapter we study data transmission with constraints on the channel input. Namely, in previous chapter the encoder for blocklength n code was permitted to produce arbitrary sequences of inputs, i.e. elements of \mathcal{A}^n . However, in many practical problem only a subset of \mathcal{A}^n is allowed to be used. As a motivation, consider the setting of the AWGN channel Example 3.3. Without restricting the input, i.e. allowing arbitrary elements of \mathbb{R}^n as input, the channel capacity is infinite: $\sup_{P_X} I(X; X + Z) = \infty$ (for example, take $X \sim \mathcal{N}(0, P)$ and $P \rightarrow \infty$). Indeed, one can transmit arbitrarily many messages with arbitrarily small error probability by choosing elements of \mathbb{R}^n with giant pairwise distance. In reality, however, one is limited by the available power. In other words, only the elements $x^n \in \mathbb{R}^n$ are allowed satisfying

$$\frac{1}{n} \sum_{t=1}^n x_t^2 \leq P,$$

where $P > 0$ is known as the power constraint. How many bits per channel use can we transmit under this constraint on the codewords? To answer this question in general, we need to extend the setup and coding theorems to *channels with input constraints*. After doing that we will apply these results to compute capacities of various Gaussian channels (memoryless, with inter-symbol interference and subject to fading).

20.1 Channel coding with input constraints



We will say that an (n, M, ϵ) -code satisfies the input constraint $F_n \subset \mathcal{A}^n$ if the encoder maps $[M]$ into F_n , i.e. $f: [M] \rightarrow F_n$. What subsets F_n are of interest?

In the context of Gaussian channels, we have $\mathcal{A} = \mathbb{R}$. Then one often talks about the following constraints:

- Average power constraint:

$$\frac{1}{n} \sum_{i=1}^n |x_i|^2 \leq P \Leftrightarrow \|x^n\|_2 \leq \sqrt{nP}.$$

In other words, codewords must lie in a ball of radius \sqrt{nP} .

- Peak power constraint :

$$\max_{1 \leq i \leq n} |x_i| \leq A \Leftrightarrow \|x^n\|_\infty \leq A$$

Notice that the second type of constraint does not introduce any new problems: we can simply restrict the input space from $\mathcal{A} = \mathbb{R}$ to $\mathcal{A} = [-A, A]$ and be back into the setting of input-unconstrained coding. The first type of the constraint is known as a *separable cost-constraint*. We will restrict our attention from now on to it exclusively.

Definition 20.1. A channel with a *separable* cost constraint is specified as follows:

- 1 \mathcal{A}, \mathcal{B} : input/output spaces
- 2 $P_{Y^n|X^n} : \mathcal{A}^n \rightarrow \mathcal{B}^n, n = 1, 2, \dots$
- 3 Cost function $c : \mathcal{A} \rightarrow \mathbb{R} \cup \{\pm\infty\}$.

We extend the per-letter cost to n -sequences as follows:

$$c(x^n) \triangleq \frac{1}{n} \sum_{k=1}^n c(x_k)$$

We next extend the channel coding notions to such channels.

Definition 20.2. • A code is an (n, M, ϵ, P) -code if it is an (n, M, ϵ) -code satisfying input constraint $F_n \triangleq \{x^n : \frac{1}{n} \sum_{k=1}^n c(x_k) \leq P\}$
 • Finite- n fundamental limits:

$$\begin{aligned} M^*(n, \epsilon, P) &= \max \{M : \exists (n, M, \epsilon, P)\text{-code}\} \\ M_{\max}^*(n, \epsilon, P) &= \max \{M : \exists (n, M, \epsilon, P)_{\max}\text{-code}\} \end{aligned}$$

- ϵ -capacity and Shannon capacity

$$\begin{aligned} C_\epsilon(P) &= \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon, P) \\ C(P) &= \lim_{\epsilon \downarrow 0} C_\epsilon(P) \end{aligned}$$

- Information capacity

$$C^{(I)}(P) = \liminf_{n \rightarrow \infty} \frac{1}{n} \sup_{P_{X^n} : \mathbb{E}[\sum_{k=1}^n c(X_k)] \leq nP} I(X^n; Y^n)$$

20.1 Channel coding with input constraints 337

- Information stability: Channel is information stable if for all (admissible) P , there exists a sequence of channel input distributions P_{X^n} such that the following two properties hold:

$$\frac{1}{n} i_{P_{X^n}, Y^n}(X^n; Y^n) \xrightarrow{\mathbb{P}} C^{(I)}(P) \quad (20.1)$$

$$\mathbb{P}[c(X^n) > P + \delta] \rightarrow 0 \quad \forall \delta > 0. \quad (20.2)$$

These definitions clearly parallel those of Definitions 19.3 and 19.5 for channels without input constraints. A notable and crucial exception is the definition of the information capacity $C^{(I)}(P)$. Indeed, under input constraints instead of maximizing $I(X^n; Y^n)$ over distributions supported on F_n we extend maximization to a richer set of distributions, namely, those satisfying

$$\mathbb{E}\left[\sum_{k=1}^n c(X_k)\right] \leq nP.$$

This will be crucial for single-letterization of $C^{(I)}(P)$ soon.

Definition 20.3 (Admissible constraint). We say P is an admissible constraint if $\exists x_0 \in \mathcal{A}$ s.t. $c(x_0) \leq P$, or equivalently, $\exists P_X : \mathbb{E}[c(X)] \leq P$. The set of admissible P 's is denoted by \mathcal{D}_c , and can be either in the form (P_0, ∞) or $[P_0, \infty)$, where $P_0 \triangleq \inf_{x \in \mathcal{A}} c(x)$.

Clearly, if $P \notin \mathcal{D}_c$, then there is no code (even a useless one, with 1 codeword) satisfying the input constraint. So in the remaining we always assume $P \in \mathcal{D}_c$.

Proposition 20.4. Define $\phi(P) \triangleq \sup_{P_X : \mathbb{E}[c(X)] \leq P} I(X; Y)$. Then

- 1 ϕ is concave and non-decreasing. The domain of ϕ is $\text{dom } \phi \triangleq \{x : f(x) > -\infty\} = \mathcal{D}_c$.
- 2 One of the following is true: $\phi(P)$ is continuous and finite on (P_0, ∞) , or $\phi = \infty$ on (P_0, ∞) .

Both of these properties extend to the function $P \mapsto C^{(I)}(P)$.

Proof. In the first part all statements are obvious, except for concavity, which follows from the concavity of $P_X \mapsto I(X; Y)$. For any P_{X_i} such that $\mathbb{E}[c(X_i)] \leq P_i, i = 0, 1$, let $X \sim \bar{\lambda}P_{X_0} + \lambda P_{X_1}$. Then $\mathbb{E}[c(X)] \leq \bar{\lambda}P_0 + \lambda P_1$ and $I(X; Y) \geq \bar{\lambda}I(X_0; Y_0) + \lambda I(X_1; Y_1)$. Hence $\phi(\bar{\lambda}P_0 + \lambda P_1) \geq \bar{\lambda}\phi(P_0) + \lambda\phi(P_1)$. The second claim follows from concavity of $\phi(\cdot)$.

To extend these results to $C^{(I)}(P)$ observe that for every n

$$P \mapsto \frac{1}{n} \sup_{P_{X^n} : \mathbb{E}[c(X^n)] \leq P} I(X^n; Y^n)$$

is concave. Then taking $\liminf_{n \rightarrow \infty}$ the same holds for $C^{(I)}(P)$. \square

An immediate consequence is that memoryless input is optimal for memoryless channel with separable cost, which gives us the single-letter formula of the information capacity:

Corollary 20.5 (Single-letterization). *Information capacity of stationary memoryless channel with separable cost:*

$$C^{(I)}(P) = \phi(P) = \sup_{\mathbb{E}[\mathbf{c}(X)] \leq P} I(X; Y).$$

Proof. $C^{(I)}(P) \geq \phi(P)$ is obvious by using $P_{X^n} = (P_X)^{\otimes n}$. For “ \leq ”, fix any P_{X^n} satisfying the cost constraint. Consider the chain

$$I(X^n; Y^n) \stackrel{(a)}{\leq} \sum_{j=1}^n I(X_j; Y_j) \stackrel{(b)}{\leq} \sum_{j=1}^n \phi(\mathbb{E}[\mathbf{c}(X_j)]) \stackrel{(c)}{\leq} n\phi\left(\frac{1}{n} \sum_{j=1}^n \mathbb{E}[\mathbf{c}(X_j)]\right) \leq n\phi(P),$$

where (a) follows from Theorem 6.1; (b) from the definition of ϕ ; and (c) from Jensen’s inequality and concavity of ϕ . \square

20.2 Channel capacity under separable cost constraints

We start with a straightforward extension of the weak converse to the case of input constraints.

Theorem 20.6 (Weak converse).

$$C_\epsilon(P) \leq \frac{C^{(I)}(P)}{1 - \epsilon}$$

Proof. The argument is the same as we used in Theorem 17.4. Take any (n, M, ϵ, P) -code, $W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W}$. Applying Fano’s inequality and the data-processing, we get

$$-h(\epsilon) + (1 - \epsilon) \log M \leq I(W; \hat{W}) \leq I(X^n; Y^n) \leq \sup_{P_{X^n}: \mathbb{E}[\mathbf{c}(X^n)] \leq P} I(X^n; Y^n)$$

Normalizing both sides by n and taking $\liminf_{n \rightarrow \infty}$ we obtain the result. \square

Next we need to extend one of the coding theorems to the case of input constraints. We do so for the Feinstein’s lemma (Theorem 18.9). Note that when $F = \mathcal{X}$, it reduces to the original version.

Theorem 20.7 (Extended Feinstein’s lemma). *Fix a Markov kernel $P_{Y|X}$ and an arbitrary P_X . Then for any measurable subset $F \subset \mathcal{X}$, every $\gamma > 0$ and any integer $M \geq 1$, there exists an $(M, \epsilon)_{\max}$ -code such that*

- Encoder satisfies the input constraint: $f: [M] \rightarrow F \subset \mathcal{X}$;
- Probability of error bound:

$$\epsilon P_X(F) \leq \mathbb{P}[i(X; Y) < \log \gamma] + \frac{M}{\gamma}$$

20.2 Channel capacity under separable cost constraints 339

Proof. Similar to the proof of the original Feinstein's lemma, define the preliminary decoding regions $E_c = \{y : i(c; y) \geq \log \gamma\}$ for all $c \in \mathcal{X}$. Next, we apply Corollary 18.1 and find out that there is a set $F_0 \subset \mathcal{X}$ with two properties: a) $P_X[F_0] = 1$ and b) for every $x \in F_0$ we have $P_Y(E_x) \leq \frac{1}{\gamma}$. We now let $F' = F \cap F_0$ and notice that $P_X[F'] = P_X[F]$.

We sequentially pick codewords $\{c_1, \dots, c_M\}$ from the set F' (!) and define the decoding regions $\{D_1, \dots, D_M\}$ as $D_j \triangleq E_{c_j} \setminus \bigcup_{k=1}^{j-1} D_k$. The stopping criterion is that M is maximal, i.e.,

$$\begin{aligned} & \forall x_0 \in F', P_Y[E_{x_0} \setminus \bigcup_{j=1}^M D_j | X = x_0] < 1 - \epsilon \\ \Leftrightarrow & \forall x_0 \in \mathcal{X}, P_Y[E_{x_0} \setminus \bigcup_{j=1}^M D_j | X = x_0] < (1 - \epsilon) \mathbb{1}[x_0 \in F'] + \mathbb{1}[x_0 \in F'^c] \end{aligned}$$

Now average the last inequality over $x_0 \sim P_X$ to obtain

$$\mathbb{P}[\{i(X; Y) \geq \log \gamma\} \setminus \bigcup_{j=1}^M D_j] \leq (1 - \epsilon) P_X[F'] + P_X[F'^c] = 1 - \epsilon P_X[F]$$

From here, we complete the proof by following the same steps as in the proof of original Feinstein's lemma (Theorem 18.9). \square

Given the coding theorem we can establish a lower bound on capacity

Theorem 20.8 (Capacity lower bound). *For any information stable channel with input constraints and $P > P_0$ we have*

$$C(P) \geq C^{(I)}(P). \quad (20.3)$$

Proof. Let us consider a special case of the stationary memoryless channel (the proof for general information stable channel follows similarly). Thus, we assume $P_{Y^n|X^n} = (P_{Y|X})^{\otimes n}$.

Fix $n \geq 1$. Choose a P_X such that $\mathbb{E}[c(X)] < P$, Pick $\log M = n(I(X; Y) - 2\delta)$ and $\log \gamma = n(I(X; Y) - \delta)$.

With the input constraint set $F_n = \{x^n : \frac{1}{n} \sum c(x_k) \leq P\}$, and iid input distribution $P_{X^n} = P_X^{\otimes n}$, we apply the extended Feinstein's lemma. This shows existence of an $(n, M, \epsilon_n, P)_{\max}$ -code with the encoder satisfying input constraint F_n and vanishing (maximal) error probability

$$\epsilon_n \underbrace{P_{X^n}[F_n]}_{\rightarrow 1} \leq \underbrace{\mathbb{P}[i(X^n; Y^n) \leq n(I(X; Y) - \delta)]}_{\rightarrow 0 \text{ as } n \rightarrow \infty \text{ by WLLN and stationary memoryless assumption}} + \underbrace{\exp(-n\delta)}_{\rightarrow 0}$$

Indeed, the first term is vanishing by the weak law of large numbers: since $\mathbb{E}[c(X)] < P$, we have $P_{X^n}(F_n) = \mathbb{P}[\frac{1}{n} \sum c(x_k) \leq P] \rightarrow 1$. Since $\epsilon_n \rightarrow 0$ this implies that for every $\epsilon > 0$ we have

$$\begin{aligned} C_\epsilon(P) & \geq \frac{1}{n} \log M = I(X; Y) - 2\delta, \quad \forall \delta > 0, \forall P_X \text{ s.t. } \mathbb{E}[c(X)] < P \\ \Rightarrow C_\epsilon(P) & \geq \sup_{P_X: \mathbb{E}[c(X)] < P} \lim_{\delta \rightarrow 0} (I(X; Y) - 2\delta) \\ \Rightarrow C_\epsilon(P) & \geq \sup_{P_X: \mathbb{E}[c(X)] < P} I(X; Y) = C^{(I)}(P-) = C^{(I)}(P) \end{aligned}$$

where the last equality is from the continuity of $C^{(I)}$ on (P_0, ∞) by Proposition 20.4.

For a general information stable channel, we just need to use the definition to show that $\mathbb{P}[i(X^n; Y^n) \leq n(C^{(I)} - \delta)] \rightarrow 0$, and the rest of the proof follows similarly. \square

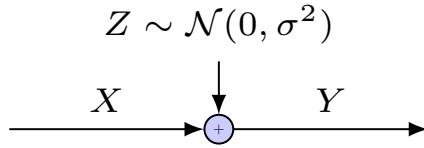
Theorem 20.9 (Channel capacity under cost constraint). *For an information stable channel with cost constraint and for any admissible constraint P we have*

$$C(P) = C^{(I)}(P).$$

Proof. The boundary case of $P = P_0$ is treated in Ex. IV.10, which shows that $C(P_0) = C^{(I)}(P_0)$ even though $C^{(I)}(P)$ may be discontinuous at P_0 . So assume $P > P_0$ next. Theorem 20.5 shows $C_\epsilon(P) \leq \frac{C^{(I)}(P)}{1-\epsilon}$, thus $C(P) \leq C^{(I)}(P)$. On the other hand, from Theorem 20.7 we have $C(P) \geq C^{(I)}(P)$. \square

20.3 Stationary AWGN channel

We start our applications with perhaps the most important channel (from the point of view of communication engineering).



Definition 20.10 (The stationary AWGN channel). The Additive White Gaussian Noise (AWGN) channel is a stationary memoryless additive-noise channel with separable cost constraint: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $c(x) = x^2$, and a single-letter kernel $P_{Y|X}$ given by $Y = X + Z$, where $Z \sim \mathcal{N}(0, \sigma^2) \perp\!\!\!\perp X$. The n -letter kernel is given by a product extension, i.e. $Y^n = X^n + Z^n$ with $Z^n \sim \mathcal{N}(0, I_n)$. When the power constraint is $\mathbb{E}[c(X)] \leq P$ we say that the signal-to-noise ratio (SNR) equals $\frac{P}{\sigma^2}$.

The terminology white noise refers to the fact that the noise variables are uncorrelated across time. This makes the power spectral density of the process $\{Z_j\}$ constant in frequency (or “white”). We often drop the word stationary when referring to this channel. The definition we gave above is more correctly should be called the *real* AWGN, or \mathbb{R} -AWGN, channel. The complex AWGN, or \mathbb{C} -channel is defined similarly: $\mathcal{A} = \mathcal{B} = \mathbb{C}$, $c(x) = |x|^2$, and $Y^n = X^n + Z^n$, with $Z^n \sim \mathcal{N}_c(0, I_n)$ being the circularly symmetric complex gaussian.

Theorem 20.11. *For the stationary AWGN channel, the channel capacity is equal to information capacity, and is given by:*

$$\begin{aligned} C(P) &= C^{(I)}(P) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) && \text{for } \mathbb{R}\text{-AWGN} \\ C(P) &= C^{(I)}(P) = \log \left(1 + \frac{P}{\sigma^2} \right) && \text{for } \mathbb{C}\text{-AWGN} \end{aligned} \quad (20.4)$$

20.3 Stationary AWGN channel 341

Proof. By Corollary 20.1,

$$C^{(1)} = \sup_{P_X: \mathbb{E}X^2 \leq P} I(X; X + Z)$$

Then using Theorem 5.9 (the Gaussian saddle point) to conclude $X \sim \mathcal{N}(0, P)$ (or $\mathcal{N}_c(0, P)$) is the unique capacity-achieving input distribution. \square

At this point it is also instructive to revisit Section 6.2* which shows that Gaussian capacity can in fact be derived essentially without solving the maximization of mutual information: the Euclidean rotational symmetry implies the optimal input should be Gaussian.

There is a great deal of deep knowledge embedded in the simple looking formula of Shannon (20.4). First, from the engineering point of view we immediately see that to transmit information faster (per unit time) one needs to pay with radiating at higher power. Second, the amount of energy spent per transmitted information bit is minimized by solving

$$\inf_{P>0} \frac{P \log 2}{C(P)} = 2\sigma^2 \log_e 2 \quad (20.5)$$

and is achieved by taking $P \rightarrow 0$. (We will discuss the notion of energy-per-bit more in Section 21.1.) Thus, we see that in order to maximize communication *rate* we need to send powerful, high-power waveforms. But in order to minimize energy-per-bit we need to send in very quiet “whisper” and at very low communication rate.¹ In either case the waveforms of good error-correcting codes should look like samples of the white gaussian process.

Third, from the mathematical point of view, formula (20.4) reveals certain properties of high-dimensional Euclidean geometry as follows. Since $Z^n \sim \mathcal{N}(0, \sigma^2)$, then with high probability, $\|Z^n\|_2$ concentrates around $\sqrt{n\sigma^2}$. Similarly, due to the power constraint and the fact that $Z^n \perp\!\!\!\perp X^n$, we have $\mathbb{E}[\|Y^n\|^2] = \mathbb{E}[\|Y^n\|^2] + \mathbb{E}[\|Z^n\|^2] \leq n(P + \sigma^2)$ and the received vector Y^n lies in an ℓ_2 -ball of radius approximately $\sqrt{n(P + \sigma^2)}$. Since the noise can at most perturb the codeword by $\sqrt{n\sigma^2}$ in Euclidean distance, if we can pack M balls of radius $\sqrt{n\sigma^2}$ into the ℓ_2 -ball of radius $\sqrt{n(P + \sigma^2)}$ centered at the origin, this yields a good codebook and decoding regions – see Fig. 20.1 for an illustration. So how large can M be? Note that the volume of an ℓ_2 -ball of radius r in \mathbb{R}^n is given by $c_n r^n$ for some constant c_n . Then $\frac{c_n(n(P+\sigma^2))^{n/2}}{c_n(n\sigma^2)^{n/2}} = \left(1 + \frac{P}{\sigma^2}\right)^{n/2}$. Taking the log and dividing by n , we get $\frac{1}{n} \log M^* \approx \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2}\right)$. This tantalizingly convincing reasoning, however, is flawed in at least two ways. (a) Computing the volume ratio only gives an upper bound on the maximal number of disjoint balls (See Section 27.2 for an extensive discussion on this topic.) (b) Codewords need not correspond to centers of *disjoint* ℓ_2 -balls. Indeed, the fact that we allow some vanishing (but non-zero) probability of error means that the $\sqrt{n\sigma^2}$ -balls are slightly overlapping and Shannon’s formula establishes the maximal number of such partially overlapping balls that we can pack so that they are (mostly) inside a larger ball.

Theorem 20.10 applies to Gaussian noise. What if the noise is non-Gaussian and how sensitive is the capacity formula $\frac{1}{2} \log(1 + \text{SNR})$ to the Gaussian assumption? Recall the Gaussian

¹ This explains why, for example, the deep space probes communicate with earth via very low-rate codes and very long blocklengths.

342

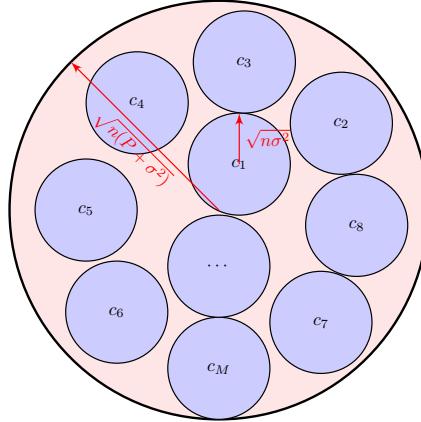


Figure 20.1 Interpretation of the AWGN capacity formula as “soft” packing.

saddlepoint result we have studied in Chapter 5 where we showed that for the same variance, Gaussian noise is the worst which shows that the capacity of any non-Gaussian noise is at least $\frac{1}{2} \log(1 + \text{SNR})$. Conversely, it turns out the increase of the capacity can be controlled by how non-Gaussian the noise is (in terms of KL divergence). The following result is due to Ihara [159].

Theorem 20.12 (Additive Non-Gaussian noise). *Let Z be a real-valued random variable independent of X and $\mathbb{E}Z^2 < \infty$. Let $\sigma^2 = \text{Var } Z$. Then*

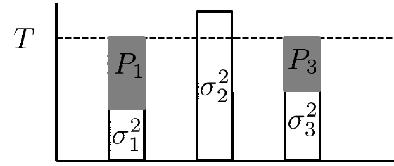
$$\frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) \leq \sup_{P_X: \mathbb{E}X^2 \leq P} I(X; X + Z) \leq \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) + D(P_Z \| \mathcal{N}(\mathbb{E}Z, \sigma^2)).$$

Proof. See Ex. IV.11. □

Remark 20.1. The quantity $D(P_Z \| \mathcal{N}(\mathbb{E}Z, \sigma^2))$ is sometimes called the *non-Gaussianity* of Z , where $\mathcal{N}(\mathbb{E}Z, \sigma^2)$ is a Gaussian with the same mean and variance as Z . So if Z has a non-Gaussian density, say, Z is uniform on $[0, 1]$, then the capacity can only differ by a constant compared to AWGN, which still scales as $\frac{1}{2} \log \text{SNR}$ in the high-SNR regime. On the other hand, if Z is discrete, then $D(P_Z \| \mathcal{N}(\mathbb{E}Z, \sigma^2)) = \infty$ and indeed in this case one can show that the capacity is infinite because the noise is “too weak”.

20.4 Parallel AWGN channel

Definition 20.13 (Parallel AWGN). A parallel AWGN channel with L branches is a stationary memoryless channel whose single-letter kernel is defined as follows: alphabets $\mathcal{A} = \mathcal{B} = \mathbb{R}^L$, the cost $c(x) = \sum_{k=1}^L |x_k|^2$; and the kernel $P_{Y^L|X^L} : Y_k = X_k + Z_k$, for $k = 1, \dots, L$, and $Z_k \sim \mathcal{N}(0, \sigma_k^2)$ are independent for each branch.



waterfilling across 3 parallel channels

Figure 20.2 Power allocation via water-filling. Here, the second branch is too noisy (σ_2 too big) for the amount of available power P and the optimal coding should discard (input zeros to) this branch altogether.

Theorem 20.14 (Waterfilling). *The capacity of L -parallel AWGN channel is given by*

$$C = \frac{1}{2} \sum_{j=1}^L \log^+ \frac{T}{\sigma_j^2}$$

where $\log^+(x) \triangleq \max(\log x, 0)$, and $T \geq 0$ is determined by

$$P = \sum_{j=1}^L |T - \sigma_j^2|^+$$

Proof.

$$\begin{aligned} C^{(I)}(P) &= \sup_{P_{X^L}: \sum \mathbb{E}[X_i^2] \leq P} I(X^L; Y^L) \\ &\leq \sup_{\sum P_k \leq P, P_k \geq 0} \sum_{k=1}^L \sup_{\mathbb{E}[X_k^2] \leq P_k} I(X_k; Y_k) \\ &= \sup_{\sum P_k \leq P, P_k \geq 0} \sum_{k=1}^L \frac{1}{2} \log\left(1 + \frac{P_k}{\sigma_k^2}\right) \end{aligned}$$

with equality if $X_k \sim \mathcal{N}(0, P_k)$ are independent. So the question boils down to the last maximization problem – **power allocation**: Denote the Lagrangian multipliers for the constraint $\sum P_k \leq P$ by λ and for the constraint $P_k \geq 0$ by μ_k . We want to solve $\max \sum \frac{1}{2} \log\left(1 + \frac{P_k}{\sigma_k^2}\right) - \mu_k P_k + \lambda(P - \sum P_k)$. First-order condition on P_k gives that

$$\frac{1}{2} \frac{1}{\sigma_k^2 + P_k} = \lambda - \mu_k, \quad \mu_k P_k = 0$$

therefore the optimal solution is

$$P_k = |T - \sigma_k^2|^+, \quad T \text{ is chosen such that } P = \sum_{k=1}^L |T - \sigma_k^2|^+$$

□

On Fig. 20.2 we give a visual interpretation of the waterfilling solution. It has a number of practically important conclusions. First, it gives a precise recipe for how much power to allocate to different frequency bands. This solution, simple and elegant, was actually pivotal for bringing high-speed internet to many homes (via cable modems): initially, before information theorists had a say, power allocations were chosen on the basis of costly and imprecise simulations of real codes. Simplicity of the waterfilling power allocation allowed to make power allocation dynamic and enable instantaneous reaction to changing noise environments.

Second, there is a very important consequence for multiple-antenna (MIMO) communication. Given n_r receive antennas and n_t transmit antennas, very often one gets as a result a parallel AWGN with $L = \min(n_r, n_t)$ branches. For a single-antenna system the capacity then scales as $\frac{1}{2} \log P$ with increasing power (Theorem 20.10), while the capacity for a MIMO AWGN channel is approximately $\frac{L}{2} \log(\frac{P}{L}) \approx \frac{L}{2} \log P$ for large P . This results in a L -fold increase in capacity at high SNR. This is the basis of a powerful technique of spatial multiplexing in MIMO, largely behind much of advance in 4G, 5G cellular (3GPP) and post-802.11n WiFi systems.

Notice that spatial diversity (requiring both receive and transmit antennas) is different from a so-called multipath diversity (which works even if antennas are added on just one side). Indeed, if a single stream of data is sent through every parallel channel simultaneously, then sufficient statistic would be to average the received vectors, resulting in a the effective noise level reduced by $\frac{1}{L}$ factor. The result is capacity increase from $\frac{1}{2} \log P$ to $\frac{1}{2} \log(LP)$ – a far cry from the L -fold increase of spatial multiplexing. These exciting topics are explored in excellent books [303, 186].

20.5* Non-stationary AWGN

Definition 20.15 (Non-stationary AWGN). A non-stationary AWGN channel is a non-stationary memoryless channel with j -th single-letter kernel defined as follows: alphabets $\mathcal{A} = \mathcal{B} = \mathbb{R}$, cost constraint $c(x) = x^2$, $P_{Y_j|X_j} : Y_j = X_j + Z_j$, where $Z_j \sim \mathcal{N}(0, \sigma_j^2)$. The n -letter channel $P_{Y^n|X^n} = \prod_{j=1}^n P_{Y_j|X_j}$.

Theorem 20.16. Assume that for every T the following limits exist:

$$\tilde{C}^{(I)}(T) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \frac{1}{2} \log^+ \frac{T}{\sigma_j^2}$$

$$\tilde{P}(T) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n |T - \sigma_j^2|^+$$

Then the capacity of the non-stationary AWGN channel is given by the parameterized form: $C(T) = \tilde{C}^{(I)}(T)$ with input power constraint $\tilde{P}(T)$.

Proof. Fix $T > 0$. Then it is clear from the waterfilling solution that

$$\sup I(X^n; Y^n) = \sum_{j=1}^n \frac{1}{2} \log^+ \frac{T}{\sigma_j^2}, \quad (20.6)$$

20.6* Additive colored Gaussian noise channel 345

where supremum is over all P_{X^n} such that

$$\mathbb{E}[\mathbf{c}(X^n)] \leq \frac{1}{n} \sum_{j=1}^n |T - \sigma_j^2|^+. \quad (20.7)$$

Now, by assumption, the LHS of (20.7) converges to $\tilde{P}(T)$. Thus, we have that for every $\delta > 0$

$$C^{(I)}(\tilde{P}(T) - \delta) \leq \tilde{C}^{(I)}(T) \quad (20.8)$$

$$C^{(I)}(\tilde{P}(T) + \delta) \geq \tilde{C}^{(I)}(T) \quad (20.9)$$

Taking $\delta \rightarrow 0$ and invoking continuity of $P \mapsto C^{(I)}(P)$, we get that the information capacity satisfies

$$C^{(I)}(\tilde{P}(T)) = \tilde{C}^{(I)}(T).$$

The channel is information stable. Indeed, from (19.32)

$$\text{Var}(i(X_j; Y_j)) = \frac{\log^2 e}{2} \frac{P_j}{P_j + \sigma_j^2} \leq \frac{\log^2 e}{2}$$

and thus

$$\sum_{j=1}^n \frac{1}{n^2} \text{Var}(i(X_j; Y_j)) < \infty.$$

From here information stability follows via Theorem 19.16. \square

Non-stationary AWGN is primarily interesting due to its relationship to the additive colored gaussian noise channel in the following section.

20.6* Additive colored Gaussian noise channel

Definition 20.17. The Additive Colored Gaussian Noise (ACGN) channel is a channel with memory defined as follows. The single-letter alphabets are $\mathcal{A} = \mathcal{B} = \mathbb{R}$ and the separable cost is $\mathbf{c}(x) = x^2$. The channel acts on the input vector X^n by addition $Y^n = X^n + Z^n$, where Z_j is a stationary Gaussian process with power spectral density $f_Z(\omega) \geq 0, \omega \in [-\pi, \pi]$ (recall Example 6.3).

Theorem 20.18. *The capacity of the ACGN channel with $f_Z(\omega) > 0$ for almost every $\omega \in [-\pi, \pi]$ is given by the following parametric form:*

$$C(T) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega,$$

$$P(T) = \frac{1}{2\pi} \int_0^{2\pi} |T - f_Z(\omega)|^+ d\omega.$$

346

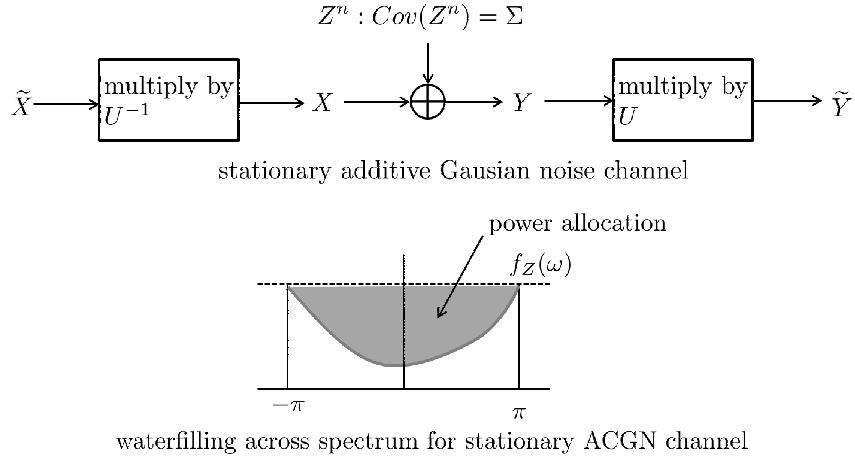


Figure 20.3 The ACGN channel: the “whitening” process used in the capacity proof and the waterfilling solution.

Proof. Take $n \geq 1$, consider the diagonalization of the covariance matrix of Z^n :

$$\text{Cov}(Z^n) = \Sigma = U^* \tilde{\Sigma} U, \text{ such that } \tilde{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_n)$$

Since $\text{Cov}(Z^n)$ is positive semi-definite, U is a unitary matrix. Define $\tilde{X}^n = UX^n$ and $\tilde{Y}^n = UY^n$, the channel between \tilde{X}^n and \tilde{Y}^n is thus

$$\begin{aligned} \tilde{Y}^n &= \tilde{X}^n + UZ^n, \\ \text{Cov}(UZ^n) &= U \cdot \text{Cov}(Z^n) \cdot U^* = \tilde{\Sigma} \end{aligned}$$

Therefore we have the equivalent channel as follows:

$$\tilde{Y}^n = \tilde{X}^n + \tilde{Z}^n, \quad \tilde{Z}_j^n \sim \mathcal{N}(0, \sigma_j^2) \text{ independent across } j.$$

By Theorem 20.16, we have that

$$\begin{aligned} \tilde{C} &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \log^+ \frac{T}{\sigma_j^2} = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega. \text{ (Szegő's Theorem, see (6.20))} \\ &\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n |T - \sigma_j^2|^+ = P(T). \end{aligned}$$

Finally since U is unitary, $C = \tilde{C}$. □

The idea used in the proof as well as the waterfilling power allocation are illustrated on Fig. 20.3. Note that most of the time the noise that impacts real-world systems is actually “born” white (because it’s a thermal noise). However, between the place of its injection and the processing there are usually multiple circuit elements. If we model them linearly then their action can equivalently be described as the ACGN channel, since the effective noise added becomes colored. In fact, this

20.7* Additive White Gaussian Noise channel with Intersymbol Interference 347

filtering can be inserted deliberately in order to convert the actual channel into an additive noise one. This is the content of the next section.

20.7* Additive White Gaussian Noise channel with Intersymbol Interference

Oftentimes in wireless communication systems a signal is propagating through a rich scattering environment. Thus, reaching the receiver are multiple delayed and attenuated copies of the initial signal. Such situation is formally called *intersymbol interference (ISI)*. A similar effect also occurs when the cable modem attempts to send signals across telephone or TV wires due to the presence of various linear filters, transformers and relays. The mathematical model for such channels is the following.

Definition 20.19 (AWGN with ISI). An AWGN channel with ISI is a channel with memory that is defined as follows: the alphabets are $\mathcal{A} = \mathcal{B} = \mathbb{R}$, and the separable cost is $c(x) = x^2$. The channel law $P_{Y^n|X^n}$ is given by

$$Y_k = \sum_{j=1}^n h_{k-j} X_j + Z_k, \quad k = 1, \dots, n$$

where $Z_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is white Gaussian noise, $\{h_k, k = -\infty, \dots, \infty\}$ are coefficients of a discrete-time channel filter.

The coefficients $\{h_k\}$ describe the action of the environment. They are often learned by the receiver during the “handshake” process of establishing a communication link.

Theorem 20.20. Suppose that the sequence $\{h_k\}$ is the inverse Fourier transform of a frequency response $H(\omega)$:

$$h_k = \frac{1}{2\pi} \int_0^{2\pi} e^{i\omega k} H(\omega) d\omega.$$

Assume also that $H(\omega)$ is a continuous function on $[0, 2\pi]$. Then the capacity of the AWGN channel with ISI is given by

$$\begin{aligned} C(T) &= \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ (T|H(\omega)|^2) d\omega \\ P(T) &= \frac{1}{2\pi} \int_0^{2\pi} \left| T - \frac{1}{|H(\omega)|^2} \right|^+ d\omega \end{aligned}$$

Proof sketch. At the decoder apply the inverse filter with frequency response $\omega \mapsto \frac{1}{H(\omega)}$. The equivalent channel then becomes a stationary colored-noise Gaussian channel:

$$\tilde{Y}_j = X_j + \tilde{Z}_j,$$

where \tilde{Z}_j is a stationary Gaussian process with spectral density

$$f_{\tilde{Z}}(\omega) = \frac{1}{|H(\omega)|^2}.$$

Then apply Theorem 20.18 to the resulting channel.

To make the above argument rigorous one must carefully analyze the non-zero error introduced by truncating the deconvolution filter to finite n . This would take us too much outside of the scope of this book. \square

20.8* Gaussian channels with amplitude constraints

We have examined some classical results of additive Gaussian noise channels. In the following, we will list some more recent results without proof.

Theorem 20.21 (Amplitude-constrained AWGN channel capacity). *For an AWGN channel $Y_i = X_i + Z_i$ with amplitude constraint $|X_i| \leq A$, we denote the capacity by:*

$$C(A) = \max_{P_X: |X| \leq A} I(X; X + Z).$$

The capacity achieving input distribution P_X^ is discrete, with finitely many atoms on $[-A, A]$. The number of atoms is $\Omega(A)$ and $O(A^2)$ as $A \rightarrow \infty$. Moreover,*

$$\frac{1}{2} \log \left(1 + \frac{2A^2}{e\pi} \right) \leq C(A) \leq \frac{1}{2} \log (1 + A^2)$$

Theorem 20.22 (Amplitude-and-power-constrained AWGN channel capacity). *For an AWGN channel $Y_i = X_i + Z_i$ with amplitude constraint $|X_i| \leq A$ and power constraint $\sum_{i=1}^n X_i^2 \leq nP$, we denote the capacity by:*

$$C(A, P) = \max_{P_X: |X| \leq A, \mathbb{E}|X|^2 \leq P} I(X; X + Z).$$

Capacity achieving input distribution P_X^ is discrete, with finitely many atoms on $[-A, A]$. Moreover, the convergence speed of $\lim_{A \rightarrow \infty} C(A, P) = \frac{1}{2} \log(1 + P)$ is of the order $e^{-O(A^2)}$.*

For details, see [280], [238, Section III] and [108, 244] for the $O(A^2)$ bound on the number of atoms.

20.9* Gaussian channels with fading

So far we assumed that the channel is either additive (as in AWGN or ACGN) or has known multiplicative gains (as in ISI). However, in practice the channel gain is a random variable. This situation is called “fading” and is often used to model the urban signal propagation with multipath

20.9* Gaussian channels with fading 349

or shadowing. The received signal at time i , Y_i , is affected by multiplicative fading coefficient H_i and additive noise Z_i as follows:

$$Y_i = H_i X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, \sigma^2)$$

This is illustrated in Fig. 20.4.

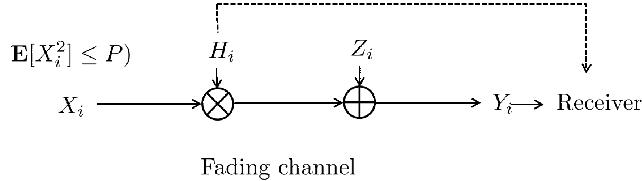


Figure 20.4 AWGN with fading.

There are two drastically different cases of fading channels, depending on the presence or absence of the dashed link on Fig. 20.4. In the first case, known as the coherent case or the CSIR case (for channel state information at the receiver), the receiver is assumed to have perfect estimate of the channel state information H_i at every time i . In other words, the channel output is effectively (Y_i, H_i) . This situation occurs, for example, when there are pilot signals sent periodically and are used at the receiver to estimate the channel. in some cases, the i then references different frequencies or sub-channels of an OFDM frame).

Whenever H_j is a stationary ergodic process, we have the channel capacity given by:

$$C(P) = \mathbb{E} \left[\frac{1}{2} \log \left(1 + \frac{P|H|^2}{\sigma^2} \right) \right]$$

and the capacity achieving input distribution is the usual $P_X = \mathcal{N}(0, P)$. Note that the capacity $C(P)$ is in the order of $\log P$ and we call the channel “energy efficient”.

In the second case, known as non-coherent or no-CSIR, the receiver does not have any knowledge of the H_i . In this case, there is no simple expression for the channel capacity. Most of the known results were shown for the case of $H_i \stackrel{\text{i.i.d.}}{\sim}$ according to the Rayleigh distribution. In this case, the capacity achieving input distribution is discrete [2], and the capacity scales as [295, 187]

$$C(P) = O(\log \log P), \quad P \rightarrow \infty \quad (20.10)$$

This channel is said to be “energy inefficient” since increase in communication rate requires dramatic expenditures in power.

Further generalization of the Gaussian channel models requires introducing multiple input and output antennas. In this case, the single-letter input $X_i \in \mathbb{C}^{n_t}$ and the output $Y_i \in \mathbb{C}^{n_r}$ are related by

$$Y_i = H_i X_i + Z_i, \quad (20.11)$$

where $Z_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \sigma^2 I_{n_r})$, n_t and n_r are the number of transmit and receive antennas, and $H_i \in \mathbb{C}^{n_t \times n_r}$ is a matrix-valued channel gain process. An incredible effort in the 1990s and 2000s was

350

invested by the information-theoretic and communication-theoretic researchers to understand this channel model. Some of the highlights include a beautiful transmit-diversity 2x2 code of Alamouti [5]; generalization of which lead to the discovery of space-time coding [297, 296]; the result of Telatar [299] showing that under coherent fading the capacity scales as $\min(n_t, n_r) \log P$; and the result of Zheng and Tse [337] showing a different pre-log in the scaling for the non-coherent (block-fading) case. It is not possible to cover these and many other beautiful results in any detail here, unfortunately. We suggest textbook [303] as an introduction to this field of MIMO communication.

21

Energy-per-bit, continuous-time channels

In this chapter we will consider an interesting variation of the channel coding problem. Instead of constraining the blocklength (i.e. the number of channel uses), we will constrain the total cost incurred by the codewords. The motivation is the following. Consider a deep space probe which has a k bit message that needs to be delivered to Earth (or a satellite orbiting it). The duration of transmission is of little worry for the probe, but what is really limited is the amount of energy it has stored in its battery. In this chapter we will learn how to study this question abstractly, how coding over large number of bits $k \rightarrow \infty$ reduces the energy spent (per bit), and how this fundamental limit is related to communication over continuous-time channels.

21.1 Energy per bit

In this chapter we will consider Markov kernels $P_{Y^\infty|X^\infty}$ acting between two spaces of infinite sequences. The prototypical example is again the AWGN channel:

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, N_0/2). \quad (21.1)$$

Note that in this chapter we have denoted the noise level for Z_i to be $\frac{N_0}{2}$. There is a long tradition for such a notation. Indeed, most of the noise in communication systems is a white thermal noise at the receiver. The power spectral density of that noise is flat and denoted by N_0 (in Joules per second per Hz). However, recall that received signal is complex-valued and, thus, each real component has power $\frac{N_0}{2}$. Note also that thermodynamics suggests that $N_0 = kT$, where $k = 1.38 \times 10^{-23}$ is the Boltzmann constant, and T is the absolute temperature in Kelvins.

In previous chapter, we analyzed the maximum number of information messages ($M^*(n, \epsilon, P)$) that can be sent through this channel for a given n number of channel uses and under the power constraint P . We have also hinted that in (20.5) that there is a fundamental minimal required cost to send each (data) bit. Here we develop this question more rigorously. Everywhere in this chapter for $v \in \mathbb{R}^\infty$ or $u \in \mathbb{C}^\infty$ we define

$$\|v\|_2^2 = \sum_{j=1}^{\infty} v_j^2, \quad \|u\|_2^2 = \sum_{j=1}^{\infty} |u_j|^2.$$

Definition 21.1 $((E, 2^k, \epsilon)$ -code). Given a Markov kernel with input space \mathbb{R}^∞ or \mathbb{C}^∞ we define an $(E, 2^k, \epsilon)$ -code to be an encoder-decoder pair, $f : [2^k] \rightarrow \mathbb{R}^\infty$ and $g : \mathbb{R}^\infty \rightarrow [2^k]$ (or similar

randomized versions), such that for all messages $m \in [2^k]$ we have $\|f(m)\|_2^2 \leq E$ and

$$\mathbb{P}[g(Y^\infty) \neq W] \leq \epsilon,$$

where as usual the probability space is $W \rightarrow X^\infty \rightarrow Y^\infty \rightarrow \hat{W}$ with $W \sim \text{Unif}([2^k])$, $X^\infty = f(W)$ and $\hat{W} = g(Y^\infty)$. The fundamental limit is defined to be

$$E^*(k, \epsilon) = \min\{E : \exists(E, 2^k, \epsilon) \text{ code}\}$$

The operational meaning of $E^*(k, \epsilon)$ should be apparent: it is the minimal amount of energy the space probe needs to draw from the battery in order to send k bits of data.

Theorem 21.2 ($(E_b/N_0)_{\min} = -1.6\text{dB}$). *For the AWGN channel we have*

$$\lim_{\epsilon \rightarrow 0} \limsup_{k \rightarrow \infty} \frac{E^*(k, \epsilon)}{k} = \frac{N_0}{\log_2 e}. \quad (21.2)$$

Remark 21.1. This result, first obtained by Shannon [268], is colloquially referred to as: minimal E_b/N_0 (pronounced “eebee over enzero” or “ebno”) is -1.6 dB. The latter value is simply $10 \log_{10}(\frac{1}{\log_2 e}) \approx -1.592$. Achieving this value of the ebno was an ultimate quest for coding theory, first resolved by the turbo codes [30]. See [72] for a review of this long conquest.

Proof. We start with a lower bound (or the “converse” part). As usual, we have the working probability space

$$W \rightarrow X^\infty \rightarrow Y^\infty \rightarrow \hat{W}.$$

Then consider the following chain:

$$\begin{aligned} -h(\epsilon) + \bar{\epsilon}k &\leq d((1-\epsilon)\|\frac{1}{M}) && \text{Fano's inequality} \\ &\leq I(W; \hat{W}) && \text{data processing for divergence} \\ &\leq I(X^\infty; Y^\infty) && \text{data processing for mutual information} \\ &\leq \sum_{i=1}^{\infty} I(X_i; Y_i) && \lim_{n \rightarrow \infty} I(X^n; U) = I(X^\infty; U) \text{ by (4.30)} \\ &\leq \sum_{i=1}^{\infty} \frac{1}{2} \log(1 + \frac{\mathbb{E}X_i^2}{N_0/2}) && \text{Gaussian capacity, Theorem 5.9} \\ &\leq \frac{\log e}{2} \sum_{i=1}^{\infty} \frac{\mathbb{E}X_i^2}{N_0/2} && \text{linearization of log} \\ &\leq \frac{E}{N_0} \log e \end{aligned}$$

Thus, we have shown

$$\frac{E^*(k, \epsilon)}{k} \geq \frac{N_0}{\log e} (\bar{\epsilon} - \frac{h(\epsilon)}{k})$$

21.1 Energy per bit 353

and taking the double limit in n then in ϵ completes the proof.

Next, for the upper bound (the “achievability” part). We first give a traditional existential proof. Notice that a $(n, 2^k, \epsilon, P)$ -code for the AWGN channel is also a $(nP, 2^k, \epsilon)$ -code for the energy problem without time constraint. Therefore,

$$\log_2 M^*(n, \epsilon, P) \geq k \Rightarrow E^*(k, \epsilon) \leq nP.$$

Take $k_n = \lfloor \log_2 M^*(n, \epsilon, P) \rfloor$, so that we have $\frac{E^*(k_n, \epsilon)}{k_n} \leq \frac{nP}{k_n}$ for all $n \geq 1$. Taking the limit then we get

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{E^*(k_n, \epsilon)}{k_n} &\leq \limsup_{n \rightarrow \infty} \frac{nP}{\log M^*(n, \epsilon, P)} \\ &= \frac{P}{\liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon, P)} \\ &= \frac{P}{\frac{1}{2} \log(1 + \frac{P}{N_0/2})}, \end{aligned}$$

where in the last step we applied Theorem 20.10. Now the above statement holds for every $P > 0$, so let us optimize it to get the best bound:

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{E^*(k_n, \epsilon)}{k_n} &\leq \inf_{P \geq 0} \frac{P}{\frac{1}{2} \log(1 + \frac{P}{N_0/2})} \\ &= \lim_{P \rightarrow 0} \frac{P}{\frac{1}{2} \log(1 + \frac{P}{N_0/2})} \\ &= \frac{N_0}{\log_2 e} \end{aligned} \tag{21.3}$$

Note that the fact that minimal energy per bit is attained at $P \rightarrow 0$ implies that in order to send information reliably at the Shannon limit of $-1.6 dB$, infinitely many time slots are needed. In other words, the information rate (also known as spectral efficiency) should be vanishingly small. Conversely, in order to have non-zero spectral efficiency, one necessarily has to step away from the $-1.6 dB$. This tradeoff is known as spectral efficiency vs. energy-per-bit.

We next can give a simpler and more explicit construction of the code, not relying on the random coding implicit in Theorem 20.10. Let $M = 2^k$ and consider the following code, known as the pulse-position modulation (PPM):

$$\text{PPM encoder: } \forall m, f(m) = \mathbf{c}_m \triangleq (0, 0, \dots, \underbrace{\sqrt{E}}_{m\text{-th location}}, \dots) \tag{21.4}$$

It is not hard to derive an upper bound on the probability of error that this code achieves [234, Theorem 2]:

$$\epsilon \leq \mathbb{E} \left[\min \left\{ MQ \left(\sqrt{\frac{2E}{N_0}} + Z \right), 1 \right\} \right], \quad Z \sim \mathcal{N}(0, 1). \tag{21.5}$$

Indeed, our orthogonal codebook under a maximum likelihood decoder has probability of error equal to

$$P_e = 1 - \frac{1}{\sqrt{\pi N_0}} \int_{-\infty}^{\infty} \left[1 - Q\left(\sqrt{\frac{2}{N_0}}z\right) \right]^{M-1} e^{-\frac{(z-\sqrt{E})^2}{N_0}} dz, \quad (21.6)$$

which is obtained by observing that conditioned on $(W = j, Z_j)$ the events $\{||\mathbf{c}_j + \mathbf{z}||^2 \leq ||\mathbf{c}_j + \mathbf{z} - \mathbf{c}_i||^2\}$, $i \neq j$ are independent. A change of variables $x = \sqrt{\frac{2}{N_0}}z$ and application of the bound $1 - (1-y)^{M-1} \leq \min\{My, 1\}$ weakens (21.6) to (21.5).

To see that (21.5) implies (21.3), fix $c > 0$ and condition on $|Z| \leq c$ in (21.5) to relax it to

$$\epsilon \leq MQ\left(\sqrt{\frac{2E}{N_0}} - c\right) + 2Q(c).$$

Recall the expansion for the Q -function [313, (3.53)]:

$$\log Q(x) = -\frac{x^2 \log e}{2} - \log x - \frac{1}{2} \log 2\pi + o(1), \quad x \rightarrow \infty \quad (21.7)$$

Thus, choosing $\tau > 0$ and setting $E = (1 + \tau)k \frac{N_0}{\log_2 e}$ we obtain

$$2^k Q\left(\sqrt{\frac{2E}{N_0}} - c\right) \rightarrow 0$$

as $k \rightarrow \infty$. Thus choosing $c > 0$ sufficiently large we obtain that $\limsup_{k \rightarrow \infty} E^*(k, \epsilon) \leq (1 + \tau) \frac{N_0}{\log_2 e}$ for every $\tau > 0$. Taking $\tau \rightarrow 0$ implies (21.3). \square

Remark 21.2 (Simplex conjecture). The code (21.4) in fact achieves the first three terms in the large- k expansion of $E^*(k, \epsilon)$, cf. [234, Theorem 3]. In fact, the code can be further slightly optimized by subtracting the common center of gravity $(2^{-k}\sqrt{E}, \dots, 2^{-k}\sqrt{E}, \dots)$ and rescaling each codeword to satisfy the power constraint. The resulting constellation is known as the *simplex code*. It is conjectured to be the actual optimal code achieving $E^*(k, \epsilon)$ for a fixed k and ϵ , see [74, Section 3.16] and [286] for more.

21.2 Capacity per unit cost

Generalizing the energy-per-bit setting of the previous section, we get the problem of *capacity per unit cost*.

Definition 21.3. Given a channel $P_{Y^\infty | X^\infty} : \mathcal{X}^\infty \rightarrow \mathcal{Y}^\infty$ and a cost function $c : \mathcal{X} \rightarrow \mathbb{R}_+$, we define (E, M, ϵ) -code to be an encoder $f : [M] \rightarrow \mathcal{X}^\infty$ and a decoder $g : \mathcal{Y}^\infty \rightarrow [M]$ s.t. a) for every m the output of the encoder $x^\infty \triangleq f(m)$ satisfies

$$\sum_{t=1}^{\infty} c(x_t) \leq E. \quad (21.8)$$

21.2 Capacity per unit cost 355

and b) the probability of error is small

$$\mathbb{P}[g(Y^\infty) \neq W] \leq \epsilon,$$

where as usual we operate on the space $W \rightarrow X^\infty \rightarrow Y^\infty \rightarrow \hat{W}$ with $W \sim \text{Unif}([M])$. We let

$$M^*(E, \epsilon) = \max\{M : (E, M, \epsilon)\text{-code}\},$$

and define capacity per unit cost as

$$C_{puc} \triangleq \lim_{\epsilon \rightarrow 0} \liminf_{E \rightarrow \infty} \frac{1}{E} \log M^*(E, \epsilon).$$

Let $C(P)$ be the capacity-cost function of the channel (in the usual sense of capacity, as defined in (20.1)). Assuming $P_0 = 0$ and $C(0) = 0$ it is not hard to show (basically following the steps of Theorem 21.2) that:

$$C_{puc} = \sup_P \frac{C(P)}{P} = \lim_{P \rightarrow 0} \frac{C(P)}{P} = \left. \frac{d}{dP} \right|_{P=0} C(P).$$

The surprising discovery of Verdú [312] is that one can avoid computing $C(P)$ and derive the C_{puc} directly. This is a significant help, as for many practical channels $C(P)$ is unknown. Additionally, this gives a yet another fundamental meaning to divergence.

Theorem 21.4. *For a stationary memoryless channel $P_{Y^\infty|X^\infty} = \prod P_{Y|X}$ with $P_0 = c(x_0) = 0$ (i.e. there is a symbol of zero cost), we have*

$$C_{puc} = \sup_{x \neq x_0} \frac{D(P_{Y|X=x} \| P_{Y|X=x_0})}{c(x)}.$$

In particular, $C_{puc} = \infty$ if there exists $x_1 \neq x_0$ with $c(x_1) = 0$.

Proof. Let

$$C_V = \sup_{x \neq x_0} \frac{D(P_{Y|X=x} \| P_{Y|X=x_0})}{c(x)}.$$

Converse: Consider a (E, M, ϵ) -code $W \rightarrow X^\infty \rightarrow Y^\infty \rightarrow \hat{W}$. Introduce an auxiliary distribution $Q_{W, X^\infty, Y^\infty, \hat{W}}$, where a channel is a useless one

$$Q_{Y^\infty|X^\infty} = Q_{Y^\infty} \triangleq P_{Y|X=x_0}^\infty.$$

That is, the overall factorization is

$$Q_{W, X^\infty, Y^\infty, \hat{W}} = P_W P_{X^\infty|W} Q_{Y^\infty} P_{\hat{W}|Y^\infty}.$$

Then, as usual we have from the data-processing for divergence

$$\begin{aligned} (1 - \epsilon) \log M + h(\epsilon) &\leq d(1 - \epsilon \| \frac{1}{M}) \\ &\leq D(P_{W, X^\infty, Y^\infty, \hat{W}} \| Q_{W, X^\infty, Y^\infty, \hat{W}}) \end{aligned}$$

$$\begin{aligned}
&= D(P_{Y^\infty|X^\infty} \| Q_{Y^\infty|P_{X^\infty}}) \\
&= \mathbb{E} \left[\sum_{t=1}^{\infty} d(X_t) \right], \tag{21.9}
\end{aligned}$$

where we denoted for convenience $d(x) \triangleq D(P_{Y|X=x} \| P_{Y|X=x_0})$. By the definition of C_V we have

$$d(x) \leq c(x)C_V.$$

Thus, continuing (21.9) we obtain

$$(1 - \epsilon) \log M + h(\epsilon) \leq C_V \mathbb{E} \left[\sum_{t=1}^{\infty} c(X_t) \right] \leq C_V \cdot E,$$

where the last step is by the cost constraint (21.8). Thus, dividing by E and taking limits we get

$$C_{puc} \leq C_V.$$

Achievability: We generalize the PPM code (21.4). For each $x_1 \in \mathcal{X}$ and $n \in \mathbb{Z}_+$ we define the encoder f as follows:

$$f(1) = (\underbrace{x_1, x_1, \dots, x_1}_{n\text{-times}}, \underbrace{x_0, \dots, x_0}_{n(M-1)\text{-times}}) \tag{21.10}$$

$$f(2) = (\underbrace{x_0, x_0, \dots, x_0}_{n\text{-times}}, \underbrace{x_1, \dots, x_1}_{n\text{-times}}, \underbrace{x_0, \dots, x_0}_{n(M-2)\text{-times}}) \tag{21.11}$$

$$\dots \tag{21.12}$$

$$f(M) = (\underbrace{x_0, \dots, x_0}_{n(M-1)\text{-times}}, \underbrace{x_1, x_1, \dots, x_1}_{n\text{-times}}) \tag{21.13}$$

Now, by Stein's lemma (Theorem 14.15) there exists a subset $S \subset \mathcal{Y}^n$ with the property that

$$\mathbb{P}[Y^n \in S | X^n = (x_1, \dots, x_1)] \geq 1 - \epsilon_1 \tag{21.14}$$

$$\mathbb{P}[Y^n \in S | X^n = (x_0, \dots, x_0)] \leq \exp\{-nD(P_{Y|X=x_1} \| P_{Y|X=x_0}) + o(n)\}. \tag{21.15}$$

Therefore, we propose the following (suboptimal!) decoder:

$$Y^n \in S \implies \hat{W} = 1 \tag{21.16}$$

$$Y_{n+1}^{2n} \in S \implies \hat{W} = 2 \tag{21.17}$$

$$\dots \tag{21.18}$$

From the union bound we find that the overall probability of error is bounded by

$$\epsilon \leq \epsilon_1 + M \exp\{-nD(P_{Y|X=x_1} \| P_{Y|X=x_0}) + o(n)\}.$$

At the same time the total cost of each codeword is given by $nc(x_1)$. Thus, taking $n \rightarrow \infty$ and after straightforward manipulations, we conclude that

$$C_{puc} \geq \frac{D(P_{Y|X=x_1} \| P_{Y|X=x_0})}{c(x_1)}.$$

21.3 Energy-per-bit for the fading channel 357

This holds for any symbol $x_1 \in \mathcal{X}$, and so we are free to take supremum over x_1 to obtain $C_{puc} \geq C_V$, as required. \square

21.3 Energy-per-bit for the fading channel

We can now apply the results of Theorem 21.6 to the case of a channel whose capacity (as a function of power) is unknown. Namely, we consider a stationary memoryless Gaussian channel with fading H_j unknown at the receiver (i.e. non-coherent fading channel, see Section 20.9*):

$$Y_j = H_j X_j + Z_j, \quad H_j \sim \mathcal{N}_c(0, 1) \perp Z_j \sim \mathcal{N}_c(0, N_0)$$

(we use here a more convenient \mathbb{C} -valued fading channel, the $H_j \sim \mathcal{N}_c$ is known as the Rayleigh fading). The cost function is the usual quadratic one $c(x) = |x|^2$. As we discussed previously, cf. (20.10), the capacity-cost function $C(P)$ is unknown in closed form, but is known to behave drastically different from the case of non-fading AWGN (i.e. when $H_j = 1$). So here Theorem 21.6 comes handy. Let us perform a simple computation required, cf. (2.9):

$$C_{puc} = \sup_{x \neq 0} \frac{D(\mathcal{N}_c(0, |x|^2 + N_0) \| \mathcal{N}_c(0, N_0))}{|x|^2} \quad (21.19)$$

$$= \frac{1}{N_0} \sup_{x \neq 0} \left(\log e - \frac{\log(1 + \frac{|x|^2}{N_0})}{\frac{|x|^2}{N_0}} \right) \quad (21.20)$$

$$= \frac{\log e}{N_0} \quad (21.21)$$

Comparing with Theorem 21.2 we discover that surprisingly, the capacity-per-unit-cost is unaffected by the presence of fading. In other words, the random multiplicative noise which is so detrimental at high SNR, appears to be much more benign at low SNR (recall that $C_{puc} = C'(0)$ and thus computing C_{puc} corresponds to computing $C(P)$ at $P \rightarrow 0$). There is one important difference: the supremization over x in (21.20) is solved at $x = \infty$. Following the proof of the converse bound, we conclude that any code hoping to achieve optimal C_{puc} must satisfy a strange constraint:

$$\sum_t |x_t|^2 \mathbf{1}\{|x_t| \geq A\} \approx \sum_t |x_t|^2 \quad \forall A > 0$$

i.e. the total energy expended by each codeword must be almost entirely concentrated in very large spikes. Such a coding method is called “flash signalling”. Thus, we can see that unlike the non-fading AWGN (for which due to rotational symmetry all codewords can be made relatively non-spiky), the only hope of achieving full C_{puc} in the presence of fading is by signalling in short bursts of energy.

This effect manifests itself in the speed of convergence to C_{puc} with increasing constellation sizes. Namely, the energy-per-bit $\frac{E^*(k, \epsilon)}{k}$ behaves asymptotically as

$$\frac{E^*(k, \epsilon)}{k} = (-1.59 \text{ dB}) + \sqrt{\frac{\text{const}}{k}} Q^{-1}(\epsilon) \quad (\text{AWGN}) \quad (21.22)$$

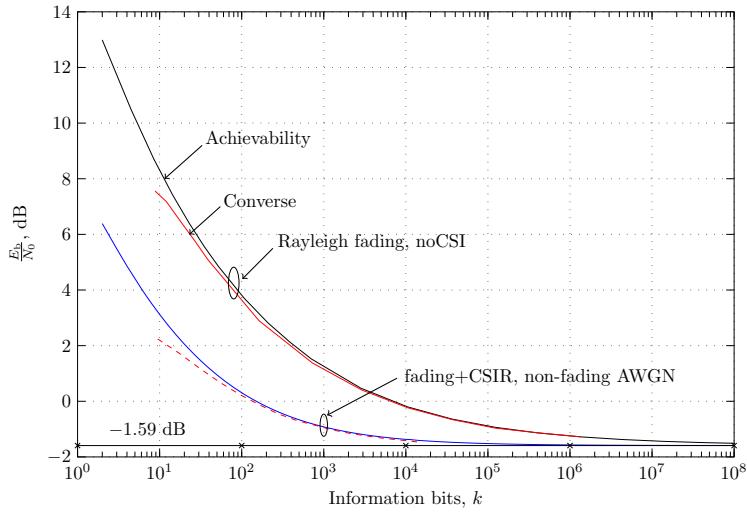


Figure 21.1 Comparing the energy-per-bit required to send a packet of k -bits for different channel models (curves represent upper and lower bounds on the unknown optimal value $\frac{E^*(k,\epsilon)}{k}$). As a comparison: to get to -1.5 dB one has to code over $6 \cdot 10^4$ data bits when the channel is non-fading AWGN or fading AWGN with H_j known perfectly at the receiver. For fading AWGN without knowledge of H_j (noCSI), one has to code over at least $7 \cdot 10^7$ data bits to get to the same -1.5 dB . Plot generated via [281].

$$\frac{E^*(k,\epsilon)}{k} = (-1.59 \text{ dB}) + \sqrt[3]{\frac{\log k}{k} (Q^{-1}(\epsilon))^2} \quad (\text{non-coherent fading}) \quad (21.23)$$

That is we see that the speed of convergence to Shannon limit is much slower under fading. Fig. 21.1 shows this effect numerically by plotting evaluation of (the upper and lower bounds for) $E^*(k,\epsilon)$ for the fading and non-fading channels. See [328] for details.

21.4 Capacity of the continuous-time AWGN channel

We now briefly discuss the topic of continuous-time channels. We would like to define the channel as acting on waveforms $x(t)$, $t \geq 0$ by adding white Gaussian noise as this:

$$Y(t) = X(t) + N(t),$$

where the $N(t)$ is a (generalized) Gaussian process with covariance function

$$\mathbb{E}[N(t)N(s)] = \frac{N_0}{2} \delta(t-s),$$

where δ is the Dirac δ -function. Defining the channel in this way requires careful understanding of the nature of $N(t)$ (in particular, it is not a usual stochastic process, since its value at any point $N(t) = \infty$), but is preferred by engineers. Mathematicians prefer to define the continuous-time

21.4 Capacity of the continuous-time AWGN channel 359

channel by introducing the standard Wiener process W_t and setting

$$Y_{int}(t) = \int_0^t X(\tau) d\tau + \sqrt{\frac{N_0}{2}} W_t,$$

where W_t is the zero-mean Gaussian process with covariance function

$$\mathbb{E}[W_s W_t] = \min(s, t).$$

Let $M^*(T, \epsilon, P)$ the maximum number of messages that can be sent through this channel such that given an encoder $f: [M] \rightarrow L_2[0, T]$ for each $m \in [M]$ the waveform $x(t) \triangleq f(m)$

- 1 is non-zero only on $[0, T]$;
- 2 input energy constrained to $\int_{t=0}^T x^2(t) dt \leq TP$;

and the decoding error probability $P[\hat{W} \neq W] \leq \epsilon$. This is a natural extension of the previously defined $\log M^*$ functions to continuous-time setting.

We prove the capacity result for this channel next.

Theorem 21.5. *The maximal reliable rate of communication across the continuous-time AWGN channel is $\frac{P}{N_0} \log e$ (per unit of time). More formally, we have*

$$\lim_{\epsilon \rightarrow 0} \liminf_{T \rightarrow \infty} \frac{1}{T} \log M^*(T, \epsilon, P) = \frac{P}{N_0} \log e \quad (21.24)$$

Proof. Note that the space of all square-integrable functions on $[0, T]$, denoted $L_2[0, T]$ has countable basis (e.g. sinusoids). Thus, by expanding our input and output waveforms in that basis we obtain an equivalent channel model:

$$\tilde{Y}_j = \tilde{X}_j + \tilde{Z}_j, \quad \tilde{Z}_j \sim \mathcal{N}(0, \frac{N_0}{2}),$$

and energy constraint (dependent upon duration T):

$$\sum_{j=1}^{\infty} \tilde{X}_j^2 \leq PT.$$

But then the problem is equivalent to the energy-per-bit for the (discrete-time) AWGN channel (see Theorem 21.2) and hence

$$\log_2 M^*(T, \epsilon, P) = k \iff E^*(k, \epsilon) = PT.$$

Thus,

$$\lim_{\epsilon \rightarrow 0} \liminf_{n \rightarrow \infty} \frac{1}{T} \log_2 M^*(T, \epsilon, P) = \frac{P}{\lim_{\epsilon \rightarrow 0} \limsup_{k \rightarrow \infty} \frac{E^*(k, \epsilon)}{k}} = \frac{P}{N_0} \log_2 e,$$

where the last step is by Theorem 21.2. \square

21.5* Capacity of the continuous-time band-limited AWGN channel

An engineer looking at the previous theorem will immediately point out an issue with the definition of an error-correcting code. Namely, we allowed the waveforms $x(t)$ to have bounded duration and bounded power, but did not constrain its frequency content. In practice, waveforms are also required to occupy a certain limited band of B Hz. What is the capacity of the AWGN channel subject to both the power p and the bandwidth B constraints?

Unfortunately, answering this question rigorously requires a long and delicate digression into functional analysis and prolate spheroidal functions. We thus only sketch the main result, without stating it as a rigorous theorem. For a full treatment, consider the monograph of Ihara [160].

Let us again define $M_{CT}^*(T, \epsilon, P, B)$ to be the maximum number of waveforms that can be sent with probability of error ϵ in time T , power P and so that each waveform in addition to those two constraints also has Fourier spectrum entirely contained in $[f_c - B/2, f_c + B/2]$, where f_c is a certain “carrier” frequency.¹

We claim that

$$C_B(P) \triangleq \lim_{\epsilon \rightarrow 0} \liminf_{n \rightarrow \infty} \frac{1}{T} \log M_{CT}^*(T, \epsilon, P, B) = B \log\left(1 + \frac{P}{N_0 B}\right), \quad (21.25)$$

In other words, the capacity of this channel is $B \log\left(1 + \frac{P}{N_0 B}\right)$. To understand the idea of the proof, we need to recall the concept of modulation first. Every signal $X(t)$ that is required to live in $[f_c - B/2, f_c + B/2]$ frequency band can be obtained by starting with a complex-valued signal $X_B(t)$ with frequency content in $[-B/2, B/2]$ and mapping it to $X(t)$ via the modulation:

$$X(t) = \operatorname{Re}(X_B(t)\sqrt{2}e^{j\omega_c t}),$$

where $\omega_c = 2\pi f_c$. Upon receiving the sum $Y(t) = X(t) + N(t)$ of the signal and the white noise $N(t)$ we may demodulate Y by computing

$$Y_B(t) = \sqrt{2} \operatorname{LPF}(Y(t)e^{-j\omega_c t}),$$

where the LPF is a low-pass filter removing all frequencies beyond $[-B/2, B/2]$. The important fact is that converting from $Y(t)$ to $Y_B(t)$ does not lose information.

Overall we have the following input-output relation:

$$Y_B(t) = X_B(t) + \tilde{N}(t),$$

where all processes are \mathbb{C} -valued and $\tilde{N}(t)$ is a complex Gaussian white noise and

$$\mathbb{E}[\tilde{N}(t)\tilde{N}(s)^*] = N_0 \delta(t-s).$$

¹ Here we already encounter a major issue: the waveform $x(t)$ supported on a finite interval $(0, T]$ cannot have spectrum supported on a compact. The requirements of finite duration and finite spectrum are only satisfied by the zero waveform. Rigorously, one should relax the bandwidth constraint to requiring that the signal have a vanishing out-of-band energy as $T \rightarrow \infty$. As we said, rigorous treatment of this issue lead to the theory of prolate spheroidal functions [277].

21.5* Capacity of the continuous-time band-limited AWGN channel 361

(Notice that after demodulation, the power spectral density of the noise is $N_0/2$ with $N_0/4$ in the real part and $N_0/4$ in the imaginary part, and after the $\sqrt{2}$ amplifier the spectral density of the noise is restored to $N_0/2$ in both real and imaginary part.)

Next, we do Nyquist sampling to convert from continuous-time to discrete time. Namely, the input waveform $X_B(t)$ is going to be represented by its values at an equispaced grid of time instants, separated by $\frac{1}{B}$. Similar representation is done to $Y_B(t)$. It is again known that these two operations do not lead to either restriction of the space of input waveforms (since every band limited signal can be uniquely represented by its samples at Nyquist rate) or loss of the information content in $Y_B(t)$ (again, Nyquist samples represent the signal Y_B completely). Mathematically, what we have done is

$$\begin{aligned} X_B(t) &= \sum_{i=-\infty}^{\infty} X_i \operatorname{sinc}_B(t - \frac{i}{B}) \\ Y_i &= \int_{t=-\infty}^{\infty} Y_B(t) \operatorname{sinc}_B(t - \frac{i}{B}) dt, \end{aligned}$$

where $\operatorname{sinc}_B(x) = \frac{\sin(Bx)}{x}$ and $X_i = X_B(i/B)$. After the Nyquist sampling on X_B and Y_B we get the following equivalent input-output relation:

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}_c(0, N_0) \quad (21.26)$$

where the noise $Z_i = \int_{t=-\infty}^{\infty} \tilde{N}(t) \operatorname{sinc}_B(t - \frac{i}{B}) dt$. Finally, given that $X_B(t)$ is only non-zero for $t \in (0, T]$ we see that the \mathbb{C} -AWGN channel (21.26) is only allowed to be used for $i = 1, \dots, TB$. This fact is known in communication theory as “bandwidth B and duration T signal has BT complex degrees of freedom”.

Let us summarize what we obtained so far:

- After sampling the equivalent channel model is that of discrete-time \mathbb{C} -AWGN.
- Given time T and bandwidth B the discrete-time equivalent channel has blocklength $n = BT$.
- The power constraint in the discrete-time model corresponds to:

$$\sum_{i=1}^{BT} |X_i|^2 = \|X(t)\|_2^2 \leq PT,$$

Thus the effective discrete-time power constraint becomes $P_d = \frac{P}{B}$.

Hence, we have established the following fact:

$$\frac{1}{T} \log M_{CT}^*(T, \epsilon, P, B) = \frac{1}{T} \log M_{\mathbb{C}\text{-AWGN}}^*(BT, \epsilon, P_d),$$

where $M_{\mathbb{C}\text{-AWGN}}^*$ denotes the fundamental limit of the \mathbb{C} -AWGN channel from Theorem 20.10. Thus, taking $T \rightarrow \infty$ we get (21.25).

Note also that in the limit of large bandwidth B the capacity formula (21.25) yields

$$C_{B=\infty}(P) = \lim_{B \rightarrow \infty} B \log(1 + \frac{P}{N_0 B}) = \frac{P}{N_0} \log e,$$

362

agreeing with (21.24).

22

Strong converse. Channel dispersion and error exponents. Finite Blocklength Bounds.

In previous chapters our main object of study was the fundamental limit of blocklength- n coding:

$$M^*(n, \epsilon) = \max\{M : \exists(n, M, \epsilon)\text{-code}\}$$

Equivalently, we can define it in terms of the smallest probability of error at a given M :

$$\epsilon^*(n, M) = \inf\{\epsilon : \exists(n, M, \epsilon)\text{-code}\}$$

What we learned so far is that for stationary memoryless channels we have

$$\lim_{\epsilon \rightarrow 0} \liminf_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, \epsilon) = C,$$

or, equivalently,

$$\limsup_{n \rightarrow \infty} \epsilon^*(n, \exp\{nR\}) = \begin{cases} 0, & R < C \\ > 0, & R > C. \end{cases}$$

These results were proved 75 years ago by Shannon. What happened in the ensuing 75 years is that we obtained much more detailed information about M^* and ϵ^* . For example, the strong converse says that in the previous limit the > 0 can be replaced with 1. The error-exponents show that convergence of $\epsilon^*(n, \exp\{nR\})$ to zero or one happens exponentially fast (with partially known exponents). The channel dispersion refines the asymptotic description to

$$\log M^*(n, \epsilon) = nC - \sqrt{nVQ^{-1}(\epsilon)} + O(\log n).$$

Finally, the finite blocklength information theory strives to prove the sharpest possible *computational* bounds on $\log M^*(n, \epsilon)$ at finite n , which allows evaluating real-world codes' performance taking their latency n into account. These results are surveyed in this chapter.

22.1 Strong Converse

We begin by stating the main theorem.

Theorem 22.1. *For any stationary memoryless channel with either $|\mathcal{A}| < \infty$ or $|\mathcal{B}| < \infty$ we have $C_\epsilon = C$ for $0 < \epsilon < 1$. Equivalently, for every $0 < \epsilon < 1$ we have*

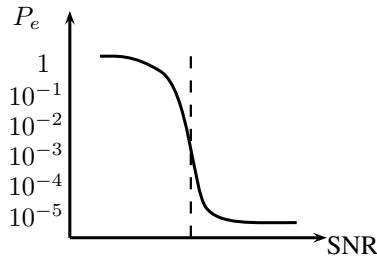
$$\log M^*(n, \epsilon) = nC + o(n), \quad n \rightarrow \infty.$$

Previously in Theorem 19.6, we showed that $C \leq C_\epsilon \leq \frac{C}{1-\epsilon}$. Now we are asserting that equality holds for every ϵ . Our previous converse arguments (Theorem 17.4 based on Fano's inequality) showed that communication with an arbitrarily small error probability is possible only when using rate $R < C$; the strong converse shows that when communicating at any rate above capacity $R > C$, the probability of error in fact goes to 1 (in fact, with speed exponential in n). In other words,

$$\epsilon^*(n, \exp(nR)) \rightarrow \begin{cases} 0 & R < C \\ 1 & R > C \end{cases} \quad (22.1)$$

where $\epsilon^*(n, M)$ is the inverse of $M^*(n, \epsilon)$ defined in (19.4).

In practice, engineers observe this effect differently. They fix a code and then allow the channel parameter (SNR for the AWGN channel, or δ for BSC_δ) vary. This typically results in a *waterfall plot* for the probability of error:



In other words, below a certain critical SNR, the probability of error quickly approaches 1, so that the receiver cannot decode anything meaningful. Above the critical SNR the probability of error quickly approaches 0 (unless there is an effect known as the *error floor*, in which case probability of error decreases reaches that floor value and stays there regardless of the further SNR increase).

Proof. We will improve the method used in the proof of Theorem 17.4. Take an (n, M, ϵ) -code and consider the usual probability space

$$W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W},$$

where $W \sim \text{Unif}([M])$. Note that P_{X^n} is the empirical distribution induced by the encoder at the channel input. Our goal is to replace this probability space with a different one where the true channel $P_{Y^n|X^n} = P_{Y|X}^{\otimes n}$ is replaced with a “dummy” channel:

$$Q_{Y^n|X^n} = (Q_Y)^{\otimes n}.$$

We will denote this new measure by \mathbb{Q} . Note that for communication purposes, $Q_{Y^n|X^n}$ is a useless channel since it ignores the input and randomly picks a member of the output space according to $Y_i \stackrel{\text{i.i.d.}}{\sim} Q_Y$, so that X^n and Y^n are independent (under \mathbb{Q}). Therefore, for the probability of success under each channel we have

$$\begin{aligned} \mathbb{Q}[\hat{W} = W] &= \frac{1}{M} \\ \mathbb{P}[\hat{W} = W] &\geq 1 - \epsilon \end{aligned}$$

22.1 Strong Converse 365

Therefore, the random variable $1_{\{\hat{W}=W\}}$ is likely to be 1 under \mathbb{P} and likely to be 0 under \mathbb{Q} . It thus looks like a rather good choice for a binary hypothesis test statistic distinguishing the two distributions, $P_{W,X^n,Y^n,\hat{W}}$ and $Q_{W,X^n,Y^n,\hat{W}}$. Since no hypothesis test can beat the optimal (Neyman-Pearson) test, we get the upper bound

$$\beta_{1-\epsilon}(P_{W,X^n,Y^n,\hat{W}}, Q_{W,X^n,Y^n,\hat{W}}) \leq \frac{1}{M} \quad (22.2)$$

(Recall the definition of β from (14.3).) The likelihood ratio is a sufficient statistic for this hypothesis test, so let us compute it:

$$\frac{P_{W,X^n,Y^n,\hat{W}}}{Q_{W,X^n,Y^n,\hat{W}}} = \frac{P_W P_{X^n|W} P_{Y^n|X^n} P_{\hat{W}|Y^n}}{P_W P_{X^n|W} (Q_Y)^{\otimes n} P_{\hat{W}|Y^n}} = \frac{P_{W|X^n} P_{X^n,Y^n} P_{\hat{W}|Y^n}}{P_{W|X^n} P_{X^n} (Q_Y)^{\otimes n} P_{\hat{W}|Y^n}} = \frac{P_{X^n,Y^n}}{P_{X^n} (Q_Y)^{\otimes n}}$$

Therefore, inequality above becomes

$$\beta_{1-\epsilon}(P_{X^n,Y^n}, P_{X^n} (Q_Y)^{\otimes n}) \leq \frac{1}{M} \quad (22.3)$$

Computing the LHS of this bound may appear to be impossible because the distribution P_{X^n} depends on the unknown code. However, it will turn out that a judicious choice of Q_Y will make knowledge of P_{X^n} unnecessary. Before presenting a formal argument, let us consider a special case of the BSC_δ channel. It will show that for symmetric channels we can select Q_Y to be the capacity achieving output distribution (recall, that it is unique by Corollary 5.1). To treat the general case later we will (essentially) decompose the channel into symmetric subchannels (corresponding to “composition” of the input).

Special case: BSC_δ . So let us take $P_{Y^n|X^n} = \text{BSC}_\delta^{\otimes n}$ and for Q_Y we will take the capacity achieving output distribution which is simply $Q_Y = \text{Ber}(1/2)$.

$$\begin{aligned} P_{Y^n|X^n}(y^n|x^n) &= P_Z^n(y^n - x^n), \quad Z^n \sim \text{Ber}(\delta)^n \\ (Q_Y)^{\otimes n}(y^n) &= 2^{-n} \end{aligned}$$

From the Neyman Pearson test, the optimal HT takes the form

$$\beta_\alpha \left(\underbrace{P_{X^n Y^n}}_{\mathbb{P}}, \underbrace{P_{X^n} (Q_Y)^{\otimes n}}_{\mathbb{Q}} \right) = \mathbb{Q} \left[\log \frac{P_{X^n Y^n}}{P_{X^n} (Q_Y)^{\otimes n}} \geq \gamma \right] \quad \text{where } \alpha = \mathbb{P} \left[\log \frac{P_{X^n Y^n}}{P_{X^n} (Q_Y)^{\otimes n}} \geq \gamma \right]$$

For the BSC, this becomes

$$\log \frac{P_{X^n Y^n}}{P_{X^n} (P_Y^*)^n} = \log \frac{P_{Z^n}(y^n - x^n)}{2^{-n}}.$$

Notice that the effect of unknown P_{X^n} completely disappeared, and so we can compute β_α :

$$\begin{aligned} \beta_\alpha(P_{X^n Y^n}, P_{X^n} (Q_Y)^{\otimes n}) &= \beta_\alpha(\text{Ber}(\delta)^{\otimes n}, \text{Ber}(\frac{1}{2})^{\otimes n}) \\ &= \exp\{-nD(\text{Ber}(\delta)\|\text{Ber}(\frac{1}{2})) + o(n)\} \quad (\text{by Stein's Lemma: Theorem 14.15}) \end{aligned} \quad (22.4)$$

Putting this together with our main bound (22.3), we see that any (n, M, ϵ) code for the BSC satisfies

$$\log M \leq nD(\text{Ber}(\delta) \parallel \text{Ber}(\frac{1}{2})) + o(n) = nC + o(n).$$

Clearly, this implies the strong converse for the BSC.

For the **general channel**, let us denote by P_Y^* the capacity achieving output distribution. Recall that by Corollary 5.1 it is unique and by (5.1) we have for every $x \in \mathcal{A}$:

$$D(P_{Y|X=x} \parallel P_Y^*) \leq C. \quad (22.5)$$

This property will be very useful. We next consider two cases separately:

1 If $|\mathcal{B}| < \infty$ we take $Q_Y = P_Y^*$ and note that from (19.31) we have

$$\sum_y P_{Y|X}(y|x_0) \log^2 P_{Y|X}(y|x_0) \leq \log^2 |\mathcal{B}| \quad \forall x_0 \in \mathcal{A}$$

and since $\min_y P_Y^*(y) > 0$ (without loss of generality), we conclude that for some constant $K > 0$ and for all $x_0 \in \mathcal{A}$ we have

$$\text{Var} \left[\log \frac{P_{Y|X}(Y|X=x_0)}{Q_Y(Y)} | X = x_0 \right] \leq K < \infty.$$

Thus, if we let

$$S_n = \sum_{i=1}^n \log \frac{P_{Y|X}(Y_i|X_i)}{P_Y^*(Y_i)},$$

then we have

$$\mathbb{E}[S_n|X^n] \leq nC, \quad \text{Var}[S_n|X^n] \leq Kn. \quad (22.6)$$

Hence from Chebyshev inequality (applied conditional on X^n), we have

$$\mathbb{P}[S_n > nC + \lambda\sqrt{Kn}] \leq \mathbb{P}[S_n > \mathbb{E}[S_n|X^n] + \lambda\sqrt{Kn}] \leq \frac{1}{\lambda^2}. \quad (22.7)$$

2 If $|\mathcal{A}| < \infty$, then first we recall that without loss of generality the encoder can be taken to be deterministic. Then for each codeword $c \in \mathcal{A}^n$ we define its *composition* as

$$\hat{P}_c(x) \triangleq \frac{1}{n} \sum_{j=1}^n 1\{c_j = x\}.$$

By simple counting it is clear that from any (n, M, ϵ) code, it is possible to select an (n, M', ϵ) subcode, such that a) all codeword have the same composition P_0 ; and b) $M' > \frac{M}{(n+1)^{|\mathcal{A}|-1}}$. Note that, $\log M = \log M' + O(\log n)$ and thus we may replace M with M' and focus on the analysis of the chosen subcode. Then we set $Q_Y = P_{Y|X} \circ P_0$. From now on we also assume that $P_0(x) > 0$ for all $x \in \mathcal{A}$ (otherwise just reduce \mathcal{A}). Let $i(x; y)$ denote the information density with respect

22.1 Strong Converse 367

to $P_0 P_{Y|X}$. If $X \sim P_0$ then $I(X; Y) = D(P_{Y|X} \| Q_Y | P_0) \leq \log |\mathcal{A}| < \infty$ and we conclude that $P_{Y|X=x} \ll Q_Y$ for each x and thus

$$i(x; y) = \log \frac{dP_{Y|X=x}}{dQ_Y}(y).$$

From (19.28) we have

$$\text{Var}[i(X; Y)|X] \leq \text{Var}[i(X; Y)] \leq K < \infty$$

Furthermore, we also have

$$\mathbb{E}[i(X; Y)|X] = D(P_{Y|X} \| Q_Y | P_0) = I(X; Y) \leq C \quad X \sim P_0.$$

So if we define

$$S_n = \sum_{i=1}^n \log \frac{dP_{Y|X=X_i}(Y_i|X_i)}{dQ_Y}(Y_i) = \sum_{i=1}^n i(X_i; Y_i),$$

we again first get the estimates (22.6) and then (22.7).

To proceed with (22.3) we apply the lower bound on β from (14.9):

$$\gamma \beta_{1-\epsilon}(P_{X^n, Y^n}, P_{X^n}(Q_Y)^{\otimes n}) \geq 1 - \epsilon - \mathbb{P}[S_n > \log \gamma],$$

where γ is arbitrary. We set $\log \gamma = nC + \lambda \sqrt{Kn}$ and $\lambda^2 = \frac{2}{1-\epsilon}$ to obtain (via (22.7)):

$$\gamma \beta_{1-\epsilon}(P_{X^n, Y^n}, P_{X^n}(Q_Y)^{\otimes n}) \geq \frac{1-\epsilon}{2},$$

which then implies

$$\log \beta_{1-\epsilon}(P_{X^n, Y^n}, P_{X^n}(Q_Y)^n) \geq -nC + O(\sqrt{n}).$$

Consequently, from (22.3) we conclude that

$$\log M \leq nC + O(\sqrt{n}),$$

implying the strong converse. \square

We note several lessons from this proof. First, we basically followed the same method as in the proof of the weak converse, except instead of invoking data-processing inequality for divergence, we analyzed the hypothesis testing problem explicitly. Second, the bound on variance of the information density is important. Thus, while the AWGN channel is excluded by the assumptions of the Theorem, the strong converse for it does hold as well (see Ex. IV.9). Third, this method of proof is also known as ‘‘sphere-packing’’, for the reason that becomes clear if we do the example of the BSC slightly differently (see Ex. IV.8).

22.2 Stationary memoryless channel without strong converse

In the proof above we basically only used the fact that the sum of independent random variables concentrates around its expectation (we used second moment to show that, but it could have been done more generally, when the second moment does not exist). Thus one may wonder whether the strong converse should hold for *all* stationary memoryless channels (it was only showed in Theorem 22.1 for those with finite input or output spaces). In this section we construct a counter-example.

Let the output alphabet be $\mathcal{B} = [0, 1]$. The input \mathcal{A} is going to be countably infinite. It will be convenient to define it as

$$\mathcal{A} = \{(j, m) : j, m \in \mathbb{Z}_+, 0 \leq j \leq m\}.$$

The single-letter channel $P_{Y|X}$ is defined in terms of probability density function as

$$p_{Y|X}(y|(j, m)) = \begin{cases} a_m, & \frac{j}{m} \leq y \leq \frac{j+1}{m}, \\ b_m, & \text{otherwise,} \end{cases}$$

where a_m, b_m are chosen to satisfy

$$\frac{1}{m}a_m + (1 - \frac{1}{m})b_m = 1 \quad (22.8)$$

$$\frac{1}{m}a_m \log a_m + (1 - \frac{1}{m})b_m \log b_m = C, \quad (22.9)$$

where $C > 0$ is an arbitrary fixed constant. Note that for large m we have

$$a_m = \frac{mC}{\log m} \left(1 + O\left(\frac{1}{\log m}\right)\right), \quad (22.10)$$

$$b_m = 1 - \frac{C}{\log m} + O\left(\frac{1}{\log^2 m}\right) \quad (22.11)$$

It is easy to see that $P_Y^* = \text{Unif}(0, 1)$ is the capacity-achieving output distribution and

$$\sup_{P_X} I(X; Y) = C.$$

Thus by Theorem 19.8 the capacity of the corresponding stationary memoryless channel is C . We next show that nevertheless the ϵ -capacity can be strictly greater than C .

Indeed, fix blocklength n and consider a *single letter* distribution P_X assigning equal weights to all atoms (j, m) with $m = \exp\{2nC\}$. It can be shown that in this case, the distribution of a single-letter information density is given by

$$i(X; Y) = \begin{cases} \log a_m, & \text{w.p. } \frac{a_m}{m} \\ \log b_m, & \text{w.p. } 1 - \frac{a_m}{m} \end{cases} = \begin{cases} 2nC + O(\log n), & \text{w.p. } \frac{a_m}{m} \\ O(\frac{1}{n}), & \text{w.p. } 1 - \frac{a_m}{m} \end{cases}.$$

Thus, for blocklength- n density we have

$$\frac{1}{n}i(X^n; Y^n) = \frac{1}{n} \sum_{i=1}^n i(X_i; Y_i) \stackrel{d}{=} O\left(\frac{1}{n}\right) + (2C + O(\frac{1}{n} \log n)) \cdot \text{Bin}(n, \frac{a_m}{m}) \xrightarrow{d} 2C \cdot \text{Poisson}(1/2),$$

22.3 Meta-converse 369

where we used the fact that $\frac{a_m}{m} = (1+o(1))\frac{1}{2n}$ and invoked the Poisson limit theorem for Binomial. Therefore, from Theorem 18.5 we get that for $\epsilon > e^{-1/2}$ there exist (n, M, ϵ) -codes with

$$\log M \geq 2nC(1+o(1)).$$

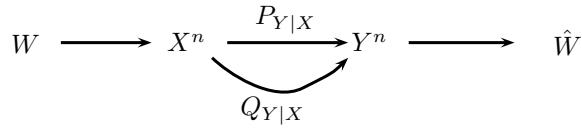
In particular,

$$C_\epsilon \geq 2C \quad \forall \epsilon > e^{-1/2}$$

22.3 Meta-converse

We have seen various ways in which one can derive upper (impossibility or converse) bounds on the fundamental limits such as $\log M^*(n, \epsilon)$. In Theorem 17.4 we used data-processing and Fano's inequalities. In the proof of Theorem 22.1 we reduced the problem to that of hypothesis testing. There are many other converse bounds that were developed over the years. It turns out that there is a very general approach that encompasses all of them. For its versatility it is sometimes referred to as the “meta-converse”.

To describe it, let us fix a Markov kernel $P_{Y|X}$ (usually, it will be the n -letter channel $P_{Y^n|X^n}$, but in the spirit of “one-shot” approach, we avoid introducing blocklength). We are also given a certain (M, ϵ) code and the goal is to show that there is an upper bound on M in terms of $P_{Y|X}$ and ϵ . The essence of the meta-converse is described by the following diagram:



Here the $W \rightarrow X$ and $Y \rightarrow \hat{W}$ represent encoder and decoder of our fixed (M, ϵ) code. The upper arrow $X \rightarrow Y$ corresponds to the actual channel, whose fundamental limits we are analyzing. The lower arrow is an *auxiliary channel* that we are free to select.

The $P_{Y|X}$ or $Q_{Y|X}$ together with P_X (distribution induced by the code) define two distributions: $P_{X,Y}$ and $Q_{X,Y}$. Consider a map $(X, Y) \mapsto Z \triangleq 1\{W \neq \hat{W}\}$ defined by the encoder and decoder pair (if decoders are randomized or $W \rightarrow X$ is not injective, we consider a Markov kernel $P_{Z|X,Y}(1|x, y) = \mathbb{P}[Z = 1|X = x, Y = y]$ instead). We have

$$P_{X,Y}[Z = 0] = 1 - \epsilon, \quad Q_{X,Y}[Z = 0] = 1 - \epsilon',$$

where ϵ and ϵ' are the average probabilities of error of the given code under the $P_{Y|X}$ and $Q_{Y|X}$ respectively. This implies the following relation for the binary HT problem of testing $P_{X,Y}$ vs $Q_{X,Y}$:

$$\beta_{1-\epsilon}(P_{X,Y}, Q_{X,Y}) \leq 1 - \epsilon'.$$

The high-level idea of the meta-converse is to select a convenient $Q_{Y|X}$, bound $1 - \epsilon'$ from above (i.e. prove a converse result for the $Q_{Y|X}$), and then use the Neyman-Pearson β -function to *lift the Q-channel converse to P-channel*.

How one chooses $Q_{Y|X}$ is a matter of art. For example, in the proof of Case 2 of Theorem 22.1 we used the trick of reducing to the constant-composition subcode. This can instead be done by taking $Q_{Y^n|X^n=c} = (P_{Y|X} \circ \hat{P}_c)^{\otimes n}$. Since there are at most $(n+1)^{|\mathcal{A}| - 1}$ different output distributions, we can see that

$$1 - \epsilon' \leq \frac{(n+1)^{|\mathcal{A}| - 1}}{M},$$

and bounding of β can be done similar to Case 2 proof of Theorem 22.1. For channels with $|\mathcal{A}| = \infty$ the technique of reducing to constant-composition codes is not available, but the meta-converse can still be applied. Examples include proof of parallel AWGN channel's dispersion [230, Theorem 78] and the study of the properties of good codes [237, Theorem 21].

However, the most common way of using meta-converse is to apply it with the trivial channel $Q_{Y|X} = Q_Y$. We have already seen this idea in Section 22.1. Indeed, with this choice the proof of the converse for the Q -channel is trivial, because we always have: $1 - \epsilon' = \frac{1}{M}$. Therefore, we conclude that any (M, ϵ) code must satisfy

$$\frac{1}{M} \geq \beta_{1-\epsilon}(P_{X,Y}, P_X Q_Y). \quad (22.12)$$

Or, after optimization we obtain

$$\frac{1}{M^*(\epsilon)} \geq \inf_{P_X} \sup_{Q_Y} \beta_{1-\epsilon}(P_{X,Y}, P_X Q_Y).$$

This is a special case of the meta-converse known as the *minimax meta-converse*. It has a number of interesting properties. First, the minimax problem in question possesses a saddle-point and is of convex-concave type [239]. It, thus, can be seen as a stronger version of the capacity saddle-point result for divergence in Theorem 5.6.

Second, the bound given by the minimax meta-converse coincides with the bound we obtained before via linear programming relaxation (18.22), as discovered by [207]. To see this connection, instead of writing the meta-converse as an upper bound M (for a given ϵ) let us think of it as an upper bound on $1 - \epsilon$ (for a given M).

We have seen that existence of an (M, ϵ) -code for $P_{Y|X}$ implies existence of the (stochastic) map $(X, Y) \mapsto Z \in \{0, 1\}$, denoted by $P_{Z|X,Y}$, with the following property:

$$P_{X,Y}[Z = 0] \geq 1 - \epsilon, \quad \text{and} \quad P_X Q_Y[Z = 0] \leq \frac{1}{M} \forall Q_Y.$$

That is $P_{Z|X,Y}$ is a test of a simple null hypothesis $(X, Y) \sim P_{X,Y}$ against a composite alternative $(X, Y) \sim P_X Q_Y$ for an arbitrary Q_Y . In other words every (M, ϵ) code must satisfy

$$1 - \epsilon \leq \tilde{\alpha}(M; P_X),$$

where (we are assuming finite \mathcal{X}, \mathcal{Y} for simplicity)

$$\tilde{\alpha}(M; P_X) \triangleq \sup_{P_{Z|X,Y}} \left\{ \sum_{x,y} P_{X,Y}(x,y) P_{Z|X,Y}(0|x,y) : \sum_{x,y} P_X(x) Q_Y(y) P_{Z|X,Y}(0|x,y) \leq \frac{1}{M} \quad \forall Q_Y \right\}.$$

We can simplify the constraint by rewriting it as

$$\sup_{Q_Y} \sum_{x,y} P_X(x) Q_Y(y) P_{Z|X,Y}(0|x,y) \leq \frac{1}{M},$$

and further simplifying it to

$$\sum_x P_X(x) P_{Z|X,Y}(0|x,y) \leq \frac{1}{M}, \quad \forall y \in \mathcal{Y}.$$

Let us now replace P_X with a $\pi_x \triangleq M P_X(x), x \in \mathcal{X}$. It is clear that $\pi \in [0, 1]^{\mathcal{X}}$. Let us also replace the optimization variable with $r_{x,y} \triangleq M P_{Z|X,Y}(0|x,y) P_X(x)$. With these notational changes we obtain

$$\tilde{\alpha}(M; P_X) = \frac{1}{M} \sup \left\{ \sum_{x,y} P_{Y|X}(y|x) r_{x,y} : 0 \leq r_{x,y} \leq \pi_x, \sum_x r_{x,y} \leq 1 \right\}.$$

It is now obvious that $\tilde{\alpha}(M; P_X) = S_{LP}(\pi)$ defined in (18.21). Optimizing over the choice of P_X (or equivalently π with $\sum_x \pi_x \leq M$) we obtain

$$1 - \epsilon \leq \frac{1}{M} S_{LP}(\pi) \leq \frac{1}{M} \sup \{ S_{LP}(\pi) : \sum_x \pi_x \leq M \} = \frac{S_{LP}^*(M)}{M}.$$

Now recall that in (18.23) we showed that a greedy procedure (essentially, the same as the one we used in the Feinstein's bound Theorem 18.9) produces a code with probability of success

$$1 - \epsilon \geq \left(1 - \frac{1}{e}\right) \frac{S_{LP}^*(M)}{M}.$$

This indicates that in the regime of a fixed ϵ the bound based on minimax metaconverse should be very sharp. This, of course, provided we can select the best Q_Y in applying it. Fortunately, for symmetric channels optimal Q_Y can be guessed fairly easily, cf. [239] for more.

22.4* Error exponents

We have studied the question of optimal error exponents in hypothesis testing before (Chapter 16). The corresponding topic for channel coding is much harder and full of open problems.

We motivate the question by trying to understand the speed of convergence in the strong converse (22.1). If we return to the proof of Theorem 19.8, namely the step (19.7), we see that by applying large-deviations Theorem 15.10 we can prove that for some $\tilde{E}(R)$ and any $R < C$ we have

$$\epsilon^*(n, \exp\{nR\}) \leq \exp\{-n\tilde{E}(R)\}.$$

What is the best value of $\tilde{E}(R)$ for each R ? This is perhaps the most famous open question in all of channel coding. More formally, let us define what is known as *reliability function* of a channel as

$$E(R) = \begin{cases} \lim_{n \rightarrow \infty} -\frac{1}{n} \log \epsilon^*(n, \exp\{nR\}) & R < C \\ \lim_{n \rightarrow \infty} -\frac{1}{n} \log(1 - \epsilon^*(n, \exp\{nR\})) & R > C. \end{cases}$$

We leave $E(R)$ as undefined if the limit does not exist. Unfortunately, there is no general argument showing that this limit exists. The only way to show its existence is to prove an achievability bound

$$\liminf_{n \rightarrow \infty} -\frac{1}{n} \log \epsilon^*(n, \exp\{nR\}) \geq E_{lower}(R),$$

a converse bound

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \epsilon^*(n, \exp\{nR\}) \leq E_{upper}(R),$$

and conclude that the limit exists whenever $E_{lower} = E_{upper}$. It is common to abuse notation and write such pair of bounds as

$$E_{lower}(R) \leq E(R) \leq E_{upper}(R),$$

even though, as we said, the $E(R)$ is not known to exist unless the two bounds match unless the two bounds match.

From now on we restrict our discussion to the case of a DMC. An important object to define is the Gallager's E_0 function, which is nothing else than the right-hand side of Gallager's bound (18.15). For the DMC it has the following expression:

$$\begin{aligned} E_0(\rho, P_X) &= -\log \sum_{y \in \mathcal{B}} \left(\sum_{x \in \mathcal{A}} P_X(x) P_{Y|X}^{\frac{1}{1+\rho}}(y|x) \right)^{1+\rho} \\ E_0(\rho) &= \max_{P_X} E_0(\rho, P_X), \quad \rho \geq 0 \\ E_0(\rho) &= \min_{P_X} E_0(\rho, P_X), \quad \rho \leq 0. \end{aligned}$$

This expression is defined in terms of the single-letter channel $P_{Y|X}$. It is not hard to see that E_0 function for the n -letter extension evaluated with $P_X^{\otimes n}$ just equals $nE_0(\rho, P_X)$, i.e. it tensorizes similar to mutual information.¹ From this observation we can apply Gallager's random coding bound (Theorem 18.12) with $P_X^{\otimes n}$ to obtain

$$\epsilon^*(n, \exp\{nR\}) \leq \exp\{n(\rho R - E_0(\rho, P_X))\} \quad \forall P_X, \rho \in [0, 1]. \quad (22.13)$$

Optimizing the choice of P_X we obtain our first estimate on the reliability function

$$E(R) \leq E_r(R) \triangleq \sup_{\rho \in [0, 1]} E_0(\rho) - \rho R.$$

¹ There is one more very pleasant analogy with mutual information: the optimization problems in the definition of $E_0(\rho)$ also tensorize. That is, the optimal distribution for the n -letter channel is just $P_X^{\otimes n}$, where P_X is optimal for a single-letter one.

An analysis, e.g. [130, Section 5.6], shows that the function $E_r(R)$ is a convex, decreasing and strictly positive on $0 \leq R < C$. Therefore, Gallager's bound provides a non-trivial estimate of the reliability function for the entire range of rates below capacity. At rates $R \rightarrow C$ the optimal choice of $\rho \rightarrow 0$. As R departs further away from the capacity the optimal ρ reaches 1 at a certain rate $R = R_{cr}$ known as the critical rate, so that for $R < R_{cr}$ we have $E_r(R) = E_0(1) - R$ behaving linearly.

Going to the upper bounds, taking Q_Y to be the iid product distribution in (22.12) and optimizing yields the bound [269] known as the sphere-packing bound:

$$E(R) \leq E_{sp}(R) \triangleq \sup_{\rho \geq 0} E_0(\rho) - \rho R.$$

Comparing the definitions of E_{sp} and E_r we can see that for $R_{cr} < R < C$ we have

$$E_{sp}(R) = E(R) = E_r(R)$$

thus establishing reliability function value for high rates. However, for $R < R_{cr}$ we have $E_{sp}(R) > E_r(R)$, so that $E(R)$ remains unknown.

Both upper and lower bounds have classical improvements. The random coding bound can be improved via technique known as *expurgation* showing

$$E(R) \geq E_{ex}(R),$$

and $E_{ex}(R) > E_r(R)$ for rates $R < R_x$ where $R_x \leq R_{cr}$ is the second critical rate; see Exc. IV.18. The sphere packing bound can also be improved at low rates by analyzing a combinatorial packing problem by showing that any code must have a pair of codewords which are close (in terms of Hellinger distance between the induced output distributions) and concluding that confusing these two leads to a lower bound on probability of error (via (16.3)). This class of bounds is known as "minimum distance" based bounds. The straight-line bound [130, Theorem 5.8.2] allows to interpolate between any minimum distance bound and the $E_{sp}(R)$. Unfortunately, these (classical) improvements tightly bound $E(R)$ at only one additional rate point $R = 0$. This state of affairs remains unchanged (for a general DMC) since 1967. As far as we know, the common belief is that $E_{ex}(R)$ is in fact the true value of $E(R)$ for all rates.

We demonstrate these bounds (and some others, but not the straight-line bound) on the reliability function on Fig. 22.1 for the case of the BSC_δ . For this channel, there is an interesting interpretation of the expurgated bound. To explain it, let us recall the different ensembles of random codes that we discussed in Section 18.6. In particular, we had the Shannon ensemble (as used in Theorem 18.5) and the random linear code (either Elias or Gallager ensembles, we do not need to make a distinction here).

For either ensemble, it is known [131] that $E_r(R)$ is not just an estimate, but in fact the exact value of the exponent of the average probability of error (averaged over a code in the ensemble). For either ensemble, however, for low rates the average is dominated by few bad codes, whereas a typical (high probability) realization of the code has a much better error exponent. For Shannon ensemble this happens at $R < \frac{1}{2}R_x$ and for the linear ensemble it happens at $R < R_x$. Furthermore, the typical linear code in fact has error exponent exactly equal to the expurgated exponent $E_{ex}(R)$, see [21].

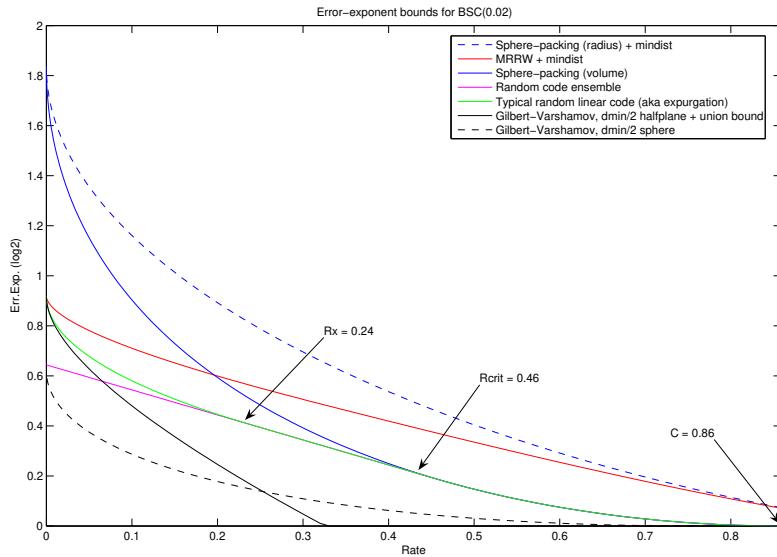


Figure 22.1 Comparison of bounds on the error exponent of the BSC. The MMRW stands for the upper bound on the minimum distance of a code [209] and Gilbert-Varshamov is a lower bound of Theorem 27.7.

There is a famous conjecture in combinatorics stating that the best possible minimum pairwise Hamming distance of a code with rate R is given by the Gilbert-Varshamov bound (Theorem 27.7). If true, this would imply that $E(R) = E_{ex}(R)$ for $R < R_x$, see e.g. [197].

The most outstanding development in the error exponents since 1967 was a sequence of papers starting from [197], which proposed a new technique for bounding $E(R)$ from above. Litsyn's idea was to first prove a geometric result (that any code of a given rate has a large number of pairs of codewords at a given distance) and then use de Caen's inequality to convert it into a lower bound on the probability of error. The resulting bound was very cumbersome. Thus, it was rather surprising when Barg and MacGregor [22] were able to show that the new upper bound on $E(R)$ matched $E_r(R)$ for $R_{cr} - \epsilon < R < R_{cr}$ for some small $\epsilon > 0$. This, for the first time since [269] extended the range of knowledge of the reliability function. Their amazing result (together with Gilbert-Varshamov conjecture) reinforced the belief that the typical linear codes achieve optimal error exponent in the whole range $0 \leq R \leq C$.

Regarding $E(R)$ for $R > C$ the situation is much simpler. We have

$$E(R) = \sup_{\rho \in (-1, 0)} E_0(\rho) - \rho R.$$

The lower (achievability) bound here is due to [105] (see also [223]), while the harder (converse) part is by Arimoto [16]. It was later discovered that Arimoto's converse bound can be derived by a simple modification of the weak converse (Theorem 17.4): instead of applying data-processing to

the KL divergence, one uses Rényi divergence of order $\alpha = \frac{1}{1+\rho}$; see [235] for details. This suggests a general conjecture that replacing Shannon information measures with Rényi ones upgrades the (weak) converse proofs to a strong converse.

22.5 Channel dispersion

Historically, first error-correcting codes had rather meager rates R very far from channel capacity. As we have seen in Section 22.4* the best codes at any rate $R < C$ have probability of error that behaves as

$$P_e = \exp\{-nE(R) + o(n)\}.$$

Therefore, for a while the question of non-asymptotic characterization of $\log M^*(n, \epsilon)$ and $\epsilon^*(n, M)$ was equated with establishing the sharp value of the error exponent $E(R)$. However, as codes became better and started having rates approaching the channel capacity, the question has changed to that of understanding behavior of $\log M^*(n, \epsilon)$ in the regime of fixed ϵ and large n (and, thus, rates $R \rightarrow C$). It was soon discovered by [231] that the next-order terms in the asymptotic expansion of $\log M^*$ give surprisingly sharp estimates on the true value of the $\log M^*$. Since then, the work on channel coding focused on establishing sharp upper and lower bounds on $\log M^*(n, \epsilon)$ for finite n (the topic of Section 22.6) and refining the classical results on the asymptotic expansions, which we discuss here.

We have already seen that the strong converse (Theorem 22.1) can be stated in the asymptotic expansion form as: for every fixed $\epsilon \in (0, 1)$,

$$\log M^*(n, \epsilon) = nC + o(n), \quad n \rightarrow \infty.$$

Intuitively, though, the smaller values of ϵ should make convergence to capacity slower. This suggests that the term $o(n)$ hides some interesting dependence on ϵ . What is it?

This topic has been investigated since the 1960s, see [95, 287, 231, 230], and resulted in the concept of *channel dispersion*. We first present the rigorous statement of the result and then explain its practical uses.

Theorem 22.2. *Consider one of the following channels:*

- 1 DMC
- 2 DMC with cost constraint
- 3 AWGN
- 4 Parallel AWGN

Let (X^, Y^*) be the input-output of the channel under the capacity achieving input distribution, and $i(x; y)$ be the corresponding (single-letter) information density. The following expansion holds for a fixed $0 < \epsilon < 1/2$ and $n \rightarrow \infty$*

$$\log M^*(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n) \tag{22.14}$$

where Q^{-1} is the inverse of the complementary standard normal CDF, the channel capacity is $C = I(X^*; Y^*) = \mathbb{E}[i(X^*; Y^*)]$, and the channel dispersion² is $V = \text{Var}[i(X^*; Y^*)|X^*]$.

Proof. The full proofs of these results are somewhat technical, even for the DMC.³ Here we only sketch the details.

First, in the absence of cost constraints the achievability (lower bound on $\log M^*$) part has already been done by us in Theorem 19.10, where we have shown that $\log M^*(n, \epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$ by refining the proof of the noisy channel coding theorem and using the CLT. Replacing the CLT with its non-asymptotic version (Berry-Esseen inequality [121, Theorem 2, Chapter XVI.5]) improves $o(\sqrt{n})$ to $O(\log n)$. In the presence of cost constraints, one is inclined to attempt to use an appropriate version of the achievability bound such as Theorem 20.6. However, for the AWGN this would require using input distribution that is uniform on the sphere. Since this distribution is non-product, the information density ceases to be a sum of iid, and CLT is harder to justify. Instead, there is a different achievability bound known as the κ - β bound [231, Theorem 25] that has become the workhorse of achievability proofs for cost-constrained channels with continuous input spaces.

The upper (converse) bound requires various special methods depending on the channel. However, the high-level idea is to always apply the meta-converse bound from (22.12) with an appropriate choice of Q_Y . Most often, Q_Y is taken as the n -th power of the capacity achieving output distribution for the channel. We illustrate the details for the special case of the BSC. In (22.4) we have shown that

$$\log M^*(n, \epsilon) \leq -\log \beta_\alpha(\text{Ber}(\delta)^{\otimes n}, \text{Ber}(\frac{1}{2})^{\otimes n}). \quad (22.15)$$

On the other hand, Exc. III.5 shows that

$$-\log \beta_{1-\epsilon}(\text{Ber}(\delta)^{\otimes n}, \text{Ber}(\frac{1}{2})^{\otimes n}) = nd(\delta \parallel \frac{1}{2}) + \sqrt{nv}Q^{-1}(\epsilon) + o(\sqrt{n}),$$

where v is just the variance of the (single-letter) log-likelihood ratio:

$$v = \text{Var}_{Z \sim \text{Ber}(\delta)} \left[Z \log \frac{\delta}{\frac{1}{2}} + (1-Z) \log \frac{1-\delta}{\frac{1}{2}} \right] = \text{Var}[Z \log \frac{\delta}{1-\delta}] = \delta(1-\delta) \log^2 \frac{\delta}{1-\delta}.$$

Upon inspection we notice that $v = V$ – the channel dispersion of the BSC, which completes the proof of the upper bound:

$$\log M^*(n, \epsilon) \leq nC - \sqrt{nv}Q^{-1}(\epsilon) + o(\sqrt{n})$$

Improving the $o(\sqrt{n})$ to $O(\log n)$ is done by applying the Berry-Esseen inequality in place of the CLT, similar to the upper bound. Many more details on these proofs are contained in [230]. \square

² There could be multiple capacity-achieving input distributions, in which case P_{X^*} should be chosen as the one that minimizes $\text{Var}[i(X^*; Y^*)|X^*]$. See [231] for more details.

³ Recently, subtle gaps in [287] and [231] in the treatment of DMCs with non-unique capacity-achieving input distributions were found and corrected in [56].

22.5 Channel dispersion 377

Remark 22.1 (Zero dispersion). We notice that $V = 0$ is entirely possible. For example, consider an additive-noise channel $Y = X + Z$ over some abelian group G with Z being uniform on some subset of G , e.g. channel in Exc. IV.13. Among the zero-dispersion channels there is a class of *exotic* channels [231], which for $\epsilon > 1/2$ have asymptotic expansions of the form [230, Theorem 51]:

$$\log M^*(n, \epsilon) = nC + \Theta_\epsilon(n^{1/3}).$$

Existence of this special case is why we restricted the theorem above to $\epsilon < \frac{1}{2}$.

Remark 22.2. The expansion (22.14) only applies to certain channels (as described in the theorem). If, for example, $\text{Var}[i(X^*; Y^*)] = \infty$, then the theorem need not hold and there might be other stable (non-Gaussian) distributions that the n -letter information density will converge to. Also notice that in the absence of cost constraints we have

$$\text{Var}[i(X^*; Y^*)|X^*] = \text{Var}[i(X^*; Y^*)]$$

since, by capacity saddle-point (Corollary 5.3), $\mathbb{E}[i(X^*; Y^*)|X^* = x] = C$ for P_{X^*} -almost all x .

As an example, we have the following dispersion formulas for the common channels that we discussed so far:

$$\begin{aligned} \text{BEC}_\delta : V(\delta) &= \delta \bar{\delta} \log^2 2 \\ \text{BSC}_\delta : V(\delta) &= \delta \bar{\delta} \log^2 \frac{\bar{\delta}}{\delta} \\ \text{AWGN: } V(P) &= \frac{P(P+2)}{2(P+1)^2} \log^2 e \text{ (Real)} \quad \frac{P(P+2)}{(P+1)^2} \log^2 e \text{ (Complex)} \end{aligned}$$

What about channel dispersion for other channels? Discrete channels with memory have seen some limited success in [232], which expresses dispersion in terms of the Fourier spectrum of the information density process. The compound DMC (Ex. IV.23) has a much more delicate dispersion formula (and the remainder term is not $O(\log n)$, but $O(n^{1/4})$), see [240]. For non-discrete channels (other than the AWGN and Poisson) new difficulties appear in the proof of the converse part. For example, the dispersion of a (coherent) fading channel is known only if one additionally restricts the input codewords to have limited peak values, cf. [70, Remark 1]. In particular, dispersion of the following *Gaussian erasure channel* is unknown:

$$Y_i = H_i(X_i + Z_i),$$

where we have $\mathcal{N}(0, 1) \sim Z_i \perp\!\!\!\perp H_i \sim \text{Ber}(1/2)$ and the usual quadratic cost constraint $\sum_{i=1}^n x_i^2 \leq nP$.

Multi-antenna (MIMO) channels (20.11) present interesting new challenges as well. For example, for coherent channels the capacity achieving input distribution is non-unique [70]. The quasi-static channels are similar to fading channels but the $H_1 = H_2 = \dots$, i.e. the channel gain matrix in (20.11) is not changing with time. This channel model is often used to model cellular networks. By leveraging an unexpected amount of differential geometry, it was shown in [327]

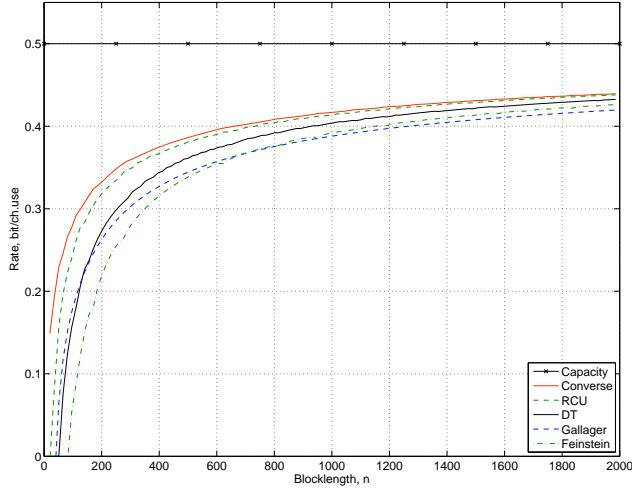


Figure 22.2 Comparing various lower (achievability) bounds on $\frac{1}{n} \log M^*(n, \epsilon)$ for the BSC_δ channel ($\delta = 0.11$, $\epsilon = 10^{-3}$).

that they have *zero-dispersion*, or more specifically:

$$\log M^*(n, \epsilon) = nC_\epsilon + O(\log n),$$

where the ϵ -capacity C_ϵ is known as outage capacity in this case (and depends on ϵ). The main implication is that C_ϵ is a good predictor of the ultimate performance limits for these practically-relevant channels (better than C is for the AWGN channel, for example). But some caution must be taken in approximating $\log M^*(n, \epsilon) \approx nC_\epsilon$, nevertheless. For example, in the case where H matrix is known at the transmitter, the same paper demonstrated that the standard water-filling power allocation (Theorem 20.14) that maximizes C_ϵ is rather sub-optimal at finite n .

22.6 Finite blocklength bounds and normal approximation

As stated earlier, direct computation of $M^*(n, \epsilon)$ by exhaustive search doubly exponential in complexity, and thus is infeasible in most cases. However, the bounds we have developed so far can often help sandwich the unknown value pretty tightly. Less rigorously, we may also evaluate the *normal approximation* which simply suggests dropping unknown terms in the expansion (22.14):

$$\log M^*(n, \epsilon) \approx nC - \sqrt{nV}Q^{-1}(\epsilon)$$

(The $\log n$ term in (22.14) is known to be equal to $O(1)$ for the BEC, and $\frac{1}{2} \log n$ for the BSC, AWGN and binary-input AWGN. For these latter channels, normal approximation is typically defined with $+\frac{1}{2} \log n$ added to the previous display.)

For example, considering the $\text{BEC}_{1/2}$ channel we can easily compute the capacity and dispersion to be $C = (1 - \delta)$ and $V = \delta(1 - \delta)$ (in bits and bits², resp.). Detailed calculation in Ex. IV.31

22.6 Finite blocklength bounds and normal approximation 379

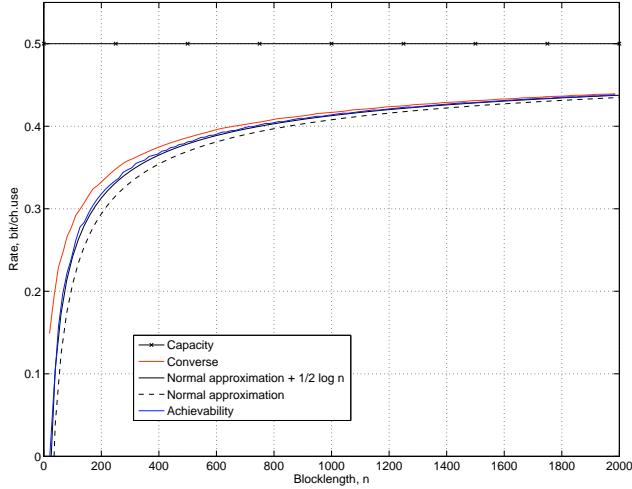


Figure 22.3 Comparing the normal approximation against the best upper and lower bounds on $\frac{1}{n} \log M^*(n, \epsilon)$ for the BSC_δ channel ($\delta = 0.11$, $\epsilon = 10^{-3}$).

lead to the following rigorous estimates:

$$213 \leq \log_2 M^*(500, 10^{-3}) \leq 217.$$

At the same time the normal approximation yields

$$\log M^*(500, 10^{-3}) \approx n\bar{\delta} - \sqrt{n\delta\bar{\delta}}Q^{-1}(10^{-3}) \approx 215.5 \text{ bits}$$

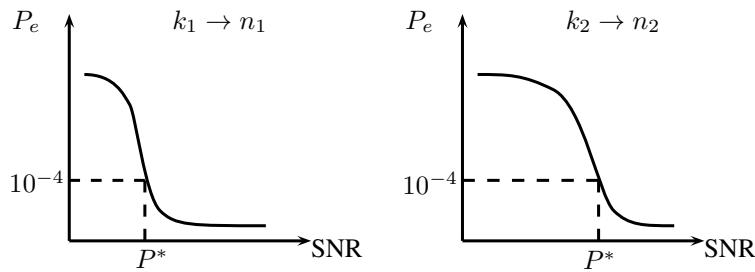
This tightness is preserved across wide range of n, ϵ, δ .

As another example, we can consider the BSC_δ channel. We have already presented numerical results for this channel in (17.3). Here, we evaluate all the lower bounds that were discussed in Chapter 18. We show the results in Fig. 22.2 together with the upper bound (22.15). We conclude that (unsurprisingly) the RCU bound is the tightest and is impressively close to the converse bound, as we have already seen in (17.3). The normal approximation (with and without the $1/2 \log n$ term) is compared against the rigorous bounds on Fig. 22.3. The excellent precision of the approximation should be contrasted with a fairly loose estimate arising from the error-exponent approximation (which coincides with the “Gallager” curve on Fig. 22.2).

We can see that for the simple cases of the BEC and the BSC, the solution to the incredibly complex combinatorial optimization problem $\log M^*(n, \epsilon)$ can be rather well approximated by considering the first few terms in the expansion (22.14). This justifies further interest in computing channel dispersion and establishing such expansions for other channels.

22.7 Normalized Rate

Suppose we are considering two different codes. One has $M = 2^{k_1}$ and blocklength n_1 (and so, in engineering language is a k_1 -to- n_1 code) and another is a k_2 -to- n_2 code. How can we compare the two of them fairly? A traditional way of presenting the code performance is in terms of the “waterfall plots” showing dependence of the probability of error on the SNR (or crossover probability) of the channel. These two codes could have waterfall plots of the following kind:



After inspecting these plots, one may believe that the $k_1 \rightarrow n_1$ code is better, since it requires a smaller SNR to achieve the same error probability. However, this ignores the fact that the rate of this code $\frac{k_1}{n_1}$ might be much smaller as well. The concept of *normalized rate* allows us to compare the codes of different blocklengths and coding rates.

Specifically, suppose that a $k \rightarrow n$ code is given. Fix $\epsilon > 0$ and find the value of the SNR P for which this code attains probability of error ϵ (for example, by taking a horizontal intercept at level ϵ on the waterfall plot). The normalized rate is defined as

$$R_{\text{norm}}(\epsilon) = \frac{k}{\log_2 M^*(n, \epsilon, P)} \approx \frac{k}{nC(P) - \sqrt{nV(P)}Q^{-1}(\epsilon)},$$

where $\log M^*$, capacity and dispersion correspond to the channel over which evaluation is being made (most often the AWGN, BI-AWGN or the fading channel). We also notice that, of course, the value of $\log M^*$ is not possible to compute exactly and thus, in practice, we use the normal approximation to evaluate it.

This idea allows us to clearly see how much different ideas in coding theory over the decades were driving the value of normalized rate upward to 1. This comparison is shown on Fig. 22.4. A short summary is that at blocklengths corresponding to “data stream” channels in cellular networks ($n \sim 10^4$) the LDPC codes and non-binary LDPC codes are already achieving 95% of the information-theoretic limit. At blocklengths corresponding to “control plane” ($n \lesssim 10^3$) the polar codes and LDPC codes are at similar performance and at 90% of the fundamental limits.

22.7 Normalized Rate 381

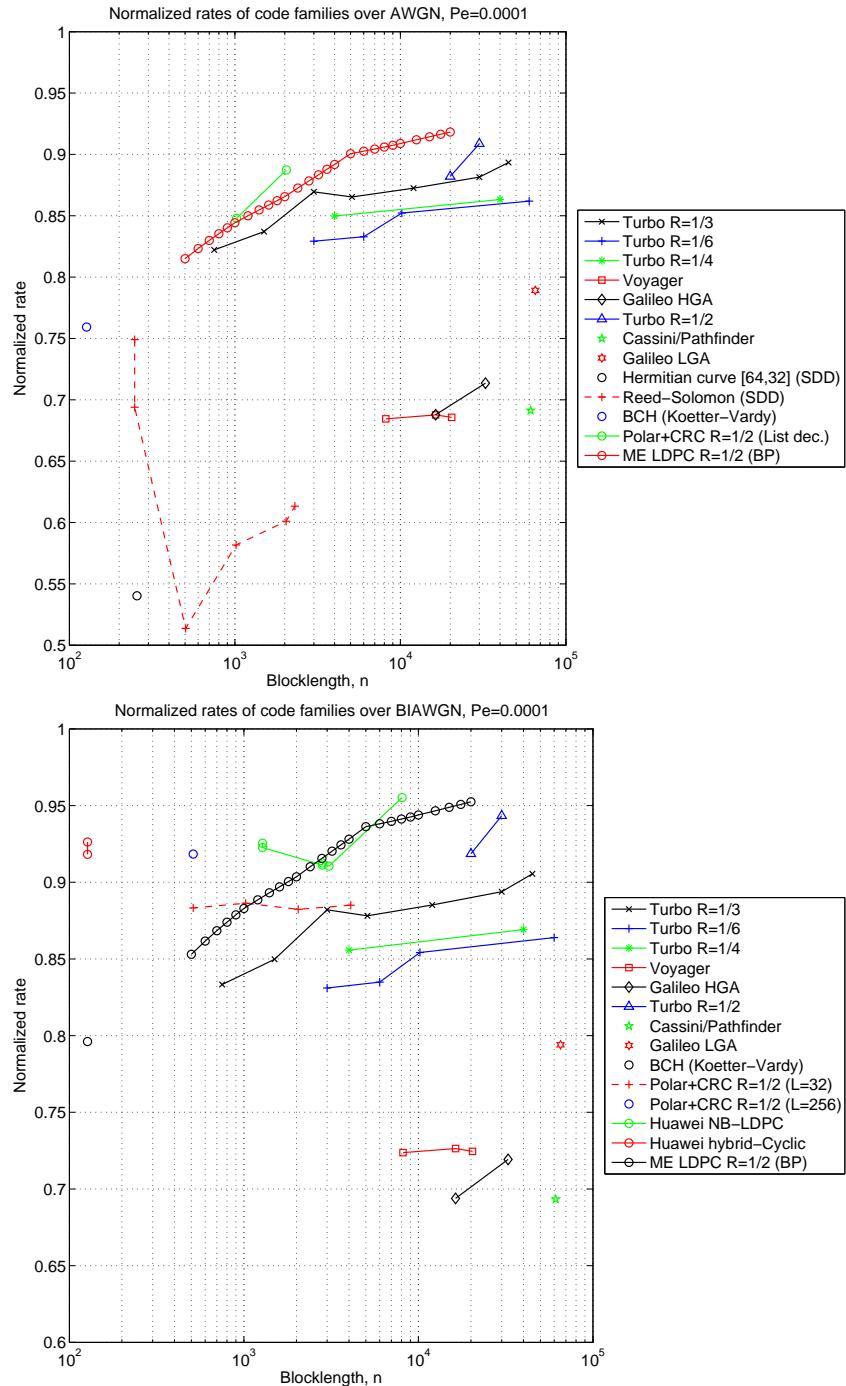


Figure 22.4 Normalized rates for various codes. Plots generated via [281] (color version recommended)

23 Channel coding with feedback

So far we have been focusing on the paradigm for one-way communication: data are mapped to codewords and transmitted, and later decoded based on the received noisy observations. In most practical settings (except for storage), frequently the communication goes in both ways so that the receiver can provide certain *feedback* to the transmitter. As a motivating example, consider the communication channel of the downlink transmission from a satellite to earth. Downlink transmission is very expensive (power constraint at the satellite), but the uplink from earth to the satellite is cheap which makes virtually noiseless feedback readily available at the transmitter (satellite). In general, channel with noiseless feedback is interesting when such asymmetry exists between uplink and downlink. Even in less ideal settings, noisy or partial feedbacks are commonly available that can potentially improve the reliability or complexity of communication.

In the first half of our discussion, we shall follow Shannon to show that even with noiseless feedback “nothing” can be gained in the conventional setup, while in the second half, we examine situations where feedback is extremely helpful.

23.1 Feedback does not increase capacity for stationary memoryless channels

Definition 23.1 (Code with feedback). An (n, M, ϵ) -code with feedback is specified by the encoder-decoder pair (f, g) as follows:

- Encoder: (time varying)

$$\begin{aligned} f_1 &: [M] \rightarrow \mathcal{A} \\ f_2 &: [M] \times \mathcal{B} \rightarrow \mathcal{A} \\ &\vdots \\ f_n &: [M] \times \mathcal{B}^{n-1} \rightarrow \mathcal{A} \end{aligned}$$

- Decoder:

$$g : \mathcal{B}^n \rightarrow [M]$$

such that $\mathbb{P}[W \neq \hat{W}] \leq \epsilon$.

23.1 Feedback does not increase capacity for stationary memoryless channels 383

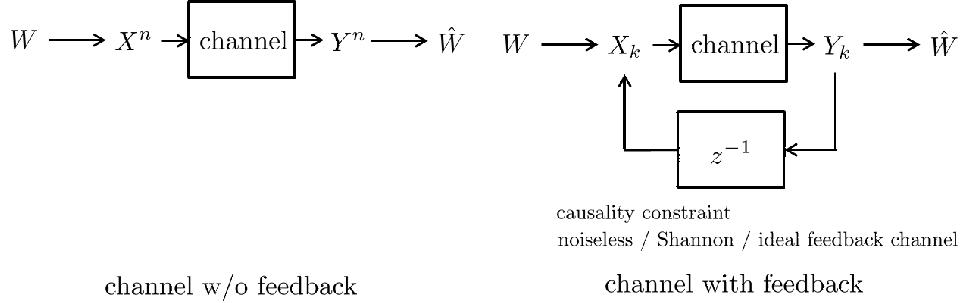


Figure 23.1 Schematic representation of coding without feedback (left) and with full noiseless feedback (right).

Here the symbol transmitted at time t depends on both the message and the history of received symbols:

$$X_t = f_t(W, Y_1^{t-1}).$$

Hence the probability space is as follows:

$$\begin{aligned} W &\sim \text{uniform on } [M] \\ X_1 &= f_1(W) \xrightarrow{P_{Y|X}} Y_1 \\ &\vdots \\ X_n &= f_n(W, Y_1^{n-1}) \xrightarrow{P_{Y|X}} Y_n \end{aligned} \quad \left. \right\} \longrightarrow \hat{W} = g(Y^n)$$

Fig. 23.1 compares the settings of channel coding without feedback and with full feedback:

Definition 23.2 (Fundamental limits).

$$M_{fb}^*(n, \epsilon) = \max \{M : \exists (n, M, \epsilon) \text{ code with feedback.}\}$$

$$C_{fb, \epsilon} = \liminf_{n \rightarrow \infty} \frac{1}{n} \log M_{fb}^*(n, \epsilon)$$

$$C_{fb} = \lim_{\epsilon \rightarrow 0} C_{fb, \epsilon} \quad (\text{Shannon capacity with feedback})$$

Theorem 23.3 (Shannon 1956). *For a stationary memoryless channel,*

$$C_{fb} = C = C^{(I)} = \sup_{P_X} I(X; Y).$$

Proof. *Achievability:* Although it is obvious that $C_{fb} \geq C$, we wanted to demonstrate that in fact constructing codes achieving capacity with *full feedback* can be done directly, without appealing to a (much harder) problem of non-feedback codes. Let $\pi_t(\cdot) \triangleq P_{W|Y^t}(\cdot | Y^t)$ with the (random) posterior distribution after t steps. It is clear that due to the knowledge of Y^t on both ends, transmitter and receiver have perfectly synchronized knowledge of π_t . Now consider how the transmission progresses:

- 1 Initialize $\pi_0(\cdot) = \frac{1}{M}$
- 2 At $(t+1)$ -th step, having knowledge of π_t all messages are partitioned into classes \mathcal{P}_a , according to the values $f_{t+1}(\cdot, Y^t)$:

$$\mathcal{P}_a \triangleq \{j \in [M] : f_{t+1}(j, Y^t) = a\} \quad a \in \mathcal{A}.$$

Then transmitter, possessing the knowledge of the true message W , selects a letter $X_{t+1} = f_{t+1}(W, Y^t)$.

- 3 Channel perturbs X_{t+1} into Y_{t+1} and both parties compute the updated posterior:

$$\pi_{t+1}(j) \triangleq \pi_t(j) B_{t+1}(j), \quad B_{t+1}(j) \triangleq \frac{P_{Y|X}(Y_{t+1}|f_{t+1}(j, Y^t))}{\sum_{a \in \mathcal{A}} \pi_t(\mathcal{P}_a)}.$$

Notice that (this is the crucial part!) the random multiplier satisfies:

$$\mathbb{E}[\log B_{t+1}(W)|Y^t] = \sum_{a \in \mathcal{A}} \sum_{y \in \mathcal{B}} \pi_t(\mathcal{P}_a) \log \frac{P_{Y|X}(y|a)}{\sum_{a \in \mathcal{A}} \pi_t(\mathcal{P}_a)a} = I(\tilde{\pi}_t, P_{Y|X}) \quad (23.1)$$

where $\tilde{\pi}_t(a) \triangleq \pi_t(\mathcal{P}_a)$ is a (random) distribution on \mathcal{A} .

The goal of the code designer is to come up with such a partitioning $\{\mathcal{P}_a : a \in \mathcal{A}\}$ that the speed of growth of $\pi_t(W)$ is maximal. Now, analyzing the speed of growth of a random-multiplicative process is best done by taking logs:

$$\log \pi_t(j) = \sum_{s=1}^t \log B_s + \log \pi_0(j).$$

Intuitively, we expect that the process $\log \pi_t(W)$ resembles a random walk starting from $-\log M$ and having a positive drift. Thus to estimate the time it takes for this process to reach value 0 we need to estimate the upward drift. Appealing to intuition and the law of large numbers we approximate

$$\log \pi_t(W) - \log \pi_0(W) \approx \sum_{s=1}^t \mathbb{E}[\log B_s].$$

Finally, from (23.1) we conclude that the best idea is to select partitioning at each step in such a way that $\tilde{\pi}_t \approx P_X^*$ (capacity-achieving input distribution) and this obtains

$$\log \pi_t(W) \approx tC - \log M,$$

implying that the transmission terminates in time $\approx \frac{\log M}{C}$. The important lesson here is the following: *The optimal transmission scheme should map messages to channel inputs in such a way that the induced input distribution $P_{X_{t+1}|Y^t}$ is approximately equal to the one maximizing $I(X; Y)$.* This idea is called *posterior matching* and explored in detail in [272].¹

¹ Note that the magic of Shannon's theorem is that this optimal partitioning can also be done blindly. That is, it is possible to preselect partitions \mathcal{P}_a in a way that is independent of π_t (but dependent on t) and so that $\pi_t(\mathcal{P}_a) \approx P_X^*(a)$ with overwhelming probability and for all $t \in [n]$.

23.1 Feedback does not increase capacity for stationary memoryless channels 385

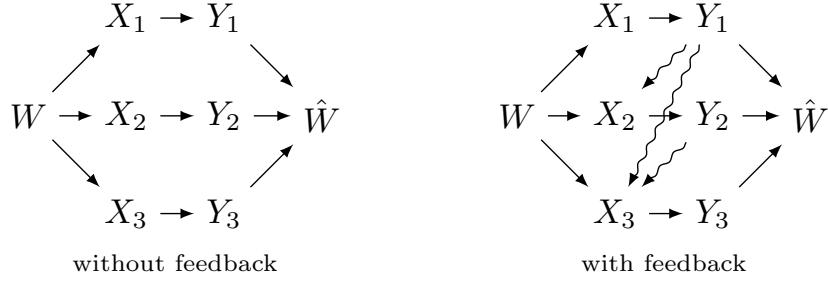
Converse: We are left to show that $C_{fb} \leq C^{(I)}$. Recall the key in proving weak converse for channel coding without feedback: Fano's inequality plus the graphical model

$$W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W}. \quad (23.2)$$

Then

$$-h(\epsilon) + \bar{\epsilon} \log M \leq I(W; \hat{W}) \leq I(X^n; Y^n) \leq nC^{(I)}.$$

With feedback the probabilistic picture becomes more complicated as the following figure shows for $n = 3$ (dependence introduced by the extra squiggly arrows):



So, while the Markov chain relation in (23.2) is still true, the input-output relation is no longer memoryless²

$$P_{Y^n|X^n}(y^n|x^n) \neq \prod_{j=1}^n P_{Y|X}(y_j|x_j) \quad (!)$$

There is still a large degree of independence in the channel, though. Namely, we have

$$(Y^{t-1}, W) \rightarrow X_t \rightarrow Y_t, \quad t = 1, \dots, n \quad (23.3)$$

$$W \rightarrow Y^n \rightarrow \hat{W} \quad (23.4)$$

Then

$$\begin{aligned} -h(\epsilon) + \bar{\epsilon} \log M &\leq I(W; \hat{W}) && \text{(Fano)} \\ &\leq I(W; Y^n) && \text{(Data processing applied to (23.4))} \\ &= \sum_{t=1}^n I(W; Y_t | Y^{t-1}) && \text{(Chain rule)} \\ &\leq \sum_{t=1}^n I(W, Y^{t-1}; Y_t) && (I(W; Y_t | Y^{t-1}) = I(W, Y^{t-1}; Y_t) - I(Y^{t-1}; Y_t)) \\ &\leq \sum_{t=1}^n I(X_t; Y_t) && \text{(Data processing applied to (23.3))} \end{aligned}$$

² To see this, consider the example where $X_2 = Y_1$ and thus $P_{Y_1|X_1X_2} = \delta_{X_1}$ is a point mass.

$$\leq nC_t$$

□

In comparison with Theorem 22.2, the following result shows that, with fixed-length block coding, feedback does not even improve the speed of approaching capacity and can at most improve the third-order $\log n$ terms.

Theorem 23.4 (Dispersion with feedback). *For weakly input-symmetric DMC (e.g. additive noise, BSC, BEC) we have:*

$$\log M_{fb}^*(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)$$

23.2* Alternative proof of Theorem 23.3 and Massey's directed information

In this section we show an alternative proof of Theorem 23.3, which is more in the spirit of “channel substitution” ideas that we continue to emphasize in this book, see Sections 6.3, 17.4 and 22.3. In addition, it will also lead us to defining the concept of directed information $\vec{I}(X^n; Y^n)$ due to Massey [206]. The latter is deeply related to various causality studies, but we will not go into this side of \vec{I} here.

Proof. It is obvious that $C_{fb} \geq C$, we are left to show that $C_{fb} \leq C^{(I)}$.

1 Recap of the steps of showing the strong converse of $C \leq C^{(I)}$ previously in Section 22.1: take any (n, M, ϵ) code, compare the two distributions:

$$P : W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W} \tag{23.5}$$

$$Q : W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W} \tag{23.6}$$

two key observations:

- a Under Q , $W \perp\!\!\!\perp W$, so that $\mathbb{Q}[W = \hat{W}] = \frac{1}{M}$ while $\mathbb{P}[W = \hat{W}] \geq 1 - \epsilon$.
- b The two graphical models give the factorization:

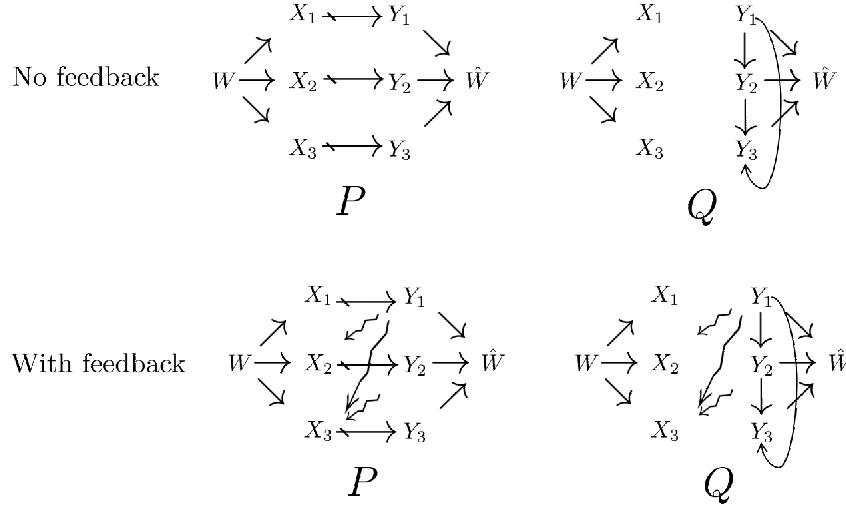
$$P_{W, X^n, Y^n, \hat{W}} = P_{W, X^n} P_{Y^n | X^n} P_{\hat{W} | Y^n}, \quad Q_{W, X^n, Y^n, \hat{W}} = P_{W, X^n} P_{Y^n} P_{\hat{W} | Y^n}$$

thus $D(P \| Q) = I(X^n; Y^n)$ measures the information flow through the links $X^n \rightarrow Y^n$.

$$-h(\epsilon) + \bar{\epsilon} \log M = d(1 - \epsilon \| \frac{1}{M}) \stackrel{\text{DPI}}{\leq} D(P \| Q) = I(X^n; Y^n) \stackrel{\text{mem-less, stat}}{=} \sum_{i=1}^n I(X_i; Y_i) \leq nC^{(I)} \tag{23.7}$$

2 Notice that when feedback is present, $X^n \rightarrow Y^n$ is not memoryless due to the transmission protocol. So let us unfold the probability space over time to see the dependence explicitly. As an example, the graphical model for $n = 3$ is given below:

23.2* Alternative proof of Theorem 23.3 and Massey's directed information 387



If we define Q similarly as in the case without feedback, we will encounter a problem at the second last inequality in (23.7), as with feedback $I(X^n; Y^n)$ can be significantly larger than $\sum_{i=1}^n I(X_i; Y_i)$. Consider the example where $X_2 = Y_1$, we have $I(X^n; Y^n) = +\infty$ independent of $I(X; Y)$.

We also make the observation that if Q is defined in (23.6), $D(P||Q) = I(X^n; Y^n)$ measures the information flow through all the $\not\rightarrow$ and \rightsquigarrow links. This motivates us to find a proper Q such that $D(P||Q)$ only captures the information flow through all the $\not\rightarrow$ links $\{X_i \rightarrow Y_i : i = 1, \dots, n\}$, so that $D(P||Q)$ closely relates to $nC^{(I)}$, while still guarantees that $W \perp\!\!\!\perp W$, so that $Q[W \neq \hat{W}] = \frac{1}{M}$.

- 3 Formally, we shall restrict $Q_{W, X^n, Y^n, \hat{W}} \in \mathcal{Q}$, where \mathcal{Q} is the set of distributions that can be factorized as follows:

$$Q_{W, X^n, Y^n, \hat{W}} = Q_W Q_{X_1|W} Q_{Y_1} Q_{X_2|W, Y_1} Q_{Y_2|Y_1} \cdots Q_{X_n|W, Y^{n-1}} Q_{Y_n|Y^{n-1}} Q_{\hat{W}|Y^n} \quad (23.8)$$

$$P_{W, X^n, Y^n, \hat{W}} = P_W P_{X_1|W} P_{Y_1|X_1} P_{X_2|W, Y_1} P_{Y_2|X_2} \cdots P_{X_n|W, Y^{n-1}} P_{Y_n|X_n} P_{\hat{W}|Y^n} \quad (23.9)$$

Verify that $W \perp\!\!\!\perp W$ under Q : W and \hat{W} are d-separated by X^n .

Notice that in the graphical models, when removing $\not\rightarrow$ we also added the directional links between the Y_i 's, these links serve to maximally preserve the dependence relationships between variables when $\not\rightarrow$ are removed, so that Q is the “closest” to P while $W \perp\!\!\!\perp W$ is satisfied.

Now we have that for $Q \in \mathcal{Q}$, $d(1 - \epsilon) \frac{1}{M} \leq D(P||Q)$, in order to obtain the least upper bound, in Lemma 23.5 we shall show that:

$$\begin{aligned} \inf_{Q \in \mathcal{Q}} D(P_{W, X^n, Y^n, \hat{W}} || Q_{W, X^n, Y^n, \hat{W}}) &= \sum_{t=1}^n I(X_t; Y_t | Y^{t-1}) \\ &= \sum_{t=1}^n \mathbb{E}_{Y^{t-1}} [I(P_{X_t|Y^{t-1}}, P_{Y_t|X_t})] \\ &\leq \sum_{t=1}^n I(\mathbb{E}_{Y^{t-1}} [P_{X_t|Y^{t-1}}], P_{Y_t|X_t}) \quad (\text{concavity of } I(\cdot, P_{Y_t|X_t})) \end{aligned}$$

$$= \sum_{t=1}^n I(P_{X_t}, P_{Y|X}) \\ \leq nC^{(I)}.$$

Following the same procedure as in (a) we have

$$-h(\epsilon) + \bar{\epsilon} \log M \leq nC^{(I)} \Rightarrow \log M \leq \frac{nC + h(\epsilon)}{1 - \epsilon} \Rightarrow C_{fb,\epsilon} \leq \frac{C}{1 - \epsilon} \Rightarrow C_{fb} \leq C.$$

4 Notice that the above proof is also valid even when cost constraint is present.

□

Lemma 23.5.

$$\inf_{Q \in \mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}} \| Q_{W,X^n,Y^n,\hat{W}}) = \sum_{t=1}^n I(X_t; Y_t | Y^{t-1}) \quad (23.10)$$

Remark 23.1 (Directed information). The quantity $\vec{I}(X^n; Y^n) \triangleq \sum_{t=1}^n I(X_t; Y_t | Y^{t-1})$ was defined by Massey and is known as directed information. In some sense, see [206] it quantifies the amount of *causal* information transfer from X -process to Y -process.

Proof. By chain rule, we can show that the minimizer $Q \in \mathcal{Q}$ must satisfy the following equalities:

$$\begin{aligned} Q_{X,W} &= P_{X,W}, \\ Q_{X_t|W,Y^{t-1}} &= P_{X_t|W,Y^{t-1}}, \quad (\text{exercise}) \\ Q_{\hat{W}|Y^n} &= P_{W|Y^n}. \end{aligned}$$

and therefore

$$\begin{aligned} &\inf_{Q \in \mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}} \| Q_{W,X^n,Y^n,\hat{W}}) \\ &= D(P_{Y_1|X_1} \| Q_{Y_1|X_1}) + D(P_{Y_2|X_2,Y_1} \| Q_{Y_2|Y_1|X_2,Y_1}) + \cdots + D(P_{Y_n|X_n,Y^{n-1}} \| Q_{Y_n|Y^{n-1}|X_n,Y^{n-1}}) \\ &= I(X_1; Y_1) + I(X_2; Y_2 | Y_1) + \cdots + I(X_n; Y_n | Y^{n-1}) \end{aligned}$$

□

23.3 When is feedback really useful?

Theorems 23.3 and 23.4 state that feedback does not improve communication rate neither asymptotically nor for moderate blocklengths. In this section, we shall examine three cases where feedback turns out to be very useful.

23.3.1 Code with very small (e.g. zero) error probability

Theorem 23.6 (Shannon [?]). *For any DMC $P_{Y|X}$,*

$$C_{fb,0} = \max_{P_X} \min_{y \in \mathcal{B}} \log \frac{1}{P_X(S_y)} \quad (23.11)$$

where

$$S_y = \{x \in \mathcal{A} : P_{Y|X}(y|x) > 0\}$$

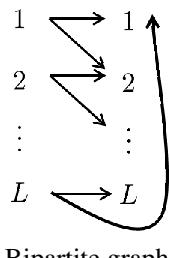
denotes the set of input symbols that can lead to the output symbol y .

Remark 23.2. For stationary memoryless channel, we have

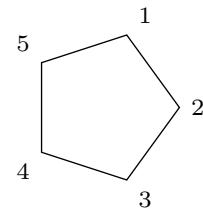
$$C_0 \stackrel{(a)}{\leq} C_{fb,0} \stackrel{(b)}{\leq} C_{fb} = \lim_{\epsilon \rightarrow 0} C_{fb,\epsilon} \stackrel{(c)}{=} C = \lim_{\epsilon \rightarrow 0} C_\epsilon \stackrel{(d)}{=} C^{(I)} = \sup_{P_X} I(X; Y)$$

where (a) and (b) are by definitions, (c) follows from Theorem 23.3, and (d) is due to Theorem 19.8. All capacity quantities above are defined with (fixed-length) block codes.

Remark 23.3. 1 In DMC for both zero-error capacities (C_0 and $C_{fb,0}$) only the support of the transition matrix $P_{Y|X}$, i.e., whether $P_{Y|X}(b|a) > 0$ or not, matters. The value of $P_{Y|X}(b|a) > 0$ is irrelevant. That is, C_0 and $C_{fb,0}$ are functions of a bipartite graph between input and output alphabets. Furthermore, the C_0 (but not $C_{fb,0}$!) is a function of the *confusability graph* – a simple undirected graph on \mathcal{A} with $a \neq a'$ connected by an edge iff $\exists b \in \mathcal{B}$ s.t. $P_{Y|X}(b|a)P_{Y|X}(b|a') > 0$.
2 That $C_{fb,0}$ is not a function of the confusability graph alone is easily seen from comparing the polygon channel (next remark) with $L = 3$ (for which $C_{fb,0} = \log \frac{3}{2}$) and the useless channel with $\mathcal{A} = \{1, 2, 3\}$ and $\mathcal{B} = \{1\}$ (for which $C_{fb,0} = 0$). Clearly in both cases confusability graph is the same – a triangle.
3 Oftentimes C_0 is very hard to compute, but $C_{fb,0}$ can be obtained in closed form as in (23.11). As an example, consider the following *polygon channel*:



Bipartite graph



Confusability graph

The following are known:

- Zero-error capacity C_0 :
 - $L = 3$: $C_0 = 0$

390

- $L = 5$: $C_0 = \frac{1}{2} \log 5$. For achievability, with blocklength one, one can use $\{1, 3\}$ to achieve rate 1 bit; with blocklength two, the codebook $\{(1, 1), (2, 3), (3, 5), (4, 2), (5, 4)\}$ achieves rate $\frac{1}{2} \log 5$ bits, as pointed out by Shannon in his original 1956 paper [?]. More than two decades later this was shown optimal by Lovász using a technique now known as semidefinite programming relaxation [?].
- Even L : $C_0 = \log \frac{L}{2}$ (Exercise IV.16).
- Odd L : The exact value of C_0 is unknown in general. For large L , $C_0 = \log \frac{L}{2} + o(1)$ [?].
- Zero-error capacity with feedback (Exercise IV.16)

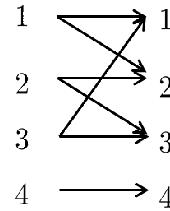
$$C_{fb,0} = \log \frac{L}{2}, \quad \forall L,$$

which can strictly exceed C_0 .

4 Notice that $C_{fb,0}$ is not necessarily equal to $C_{fb} = \lim_{\epsilon \rightarrow 0} C_{fb,\epsilon} = C$. Here is an example with

$$C_0 < C_{fb,0} < C_{fb} = C.$$

Consider the channel:



Then

$$\begin{aligned} C_0 &= \log 2 \\ C_{fb,0} &= \max_{\delta} -\log \max\left(\frac{2}{3}\delta, 1-\delta\right) & (P_X^* = (\delta/3, \delta/3, \delta/3, \bar{\delta})) \\ &= \log \frac{5}{2} > C_0 & (\delta^* = \frac{3}{5}) \end{aligned}$$

On the other hand, the Shannon capacity $C = C_{fb}$ can be made arbitrarily close to $\log 4$ by picking the cross-over probabilities arbitrarily close to zero, while the confusability graph stays the same.

Proof of Theorem 23.7. 1 Fix any $(n, M, 0)$ -code. Denote the confusability set of all possible messages that could have produced the received signal $y^t = (y_1, \dots, y_t)$ for all $t = 0, 1, \dots, n$ by:

$$E_t(y^t) \triangleq \{m \in [M] : f_1(m) \in S_{y_1}, f_2(m, y_1) \in S_{y_2}, \dots, f_n(m, y^{t-1}) \in S_{y_t}\}$$

Notice that zero-error means no ambiguity:

$$\epsilon = 0 \Leftrightarrow \forall y^n \in \mathcal{B}^n, |E_n(y^n)| = 1 \text{ or } 0. \quad (23.12)$$

23.3 When is feedback really useful? 391

2 The key quantities in the proof are defined as follows:

$$\theta_{fb} \triangleq \min_{P_X} \max_{y \in \mathcal{B}} P_X(S_y), P_X^* \triangleq \operatorname{argmin}_{P_X} \max_{y \in \mathcal{B}} P_X(S_y)$$

The goal is to show

$$C_{fb,0} = \log \frac{1}{\theta_{fb}}.$$

By definition, we have

$$\forall P_X, \exists y \in \mathcal{B}, \text{ such that } P_X(S_y) \geq \theta_{fb} \quad (23.13)$$

Notice the minimizer distribution P_X^* is usually not the capacity-achieving input distribution in the usual sense. This definition also sheds light on how the encoding and decoding should be proceeded and serves to lower bound the uncertainty reduction at each stage of the decoding scheme.

3 “ \leq ” (converse): Let P_{X^n} be the joint distribution of the codewords. Denote $E_0 = [M]$ – original message set.

$t = 1$: For P_{X_1} , by (23.13), $\exists y_1^*$ such that:

$$P_{X_1}(S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}\}|}{|\{m \in [M]\}|} = \frac{|E_1(y_1^*)|}{|E_0|} \geq \theta_{fb}.$$

$t = 2$: For $P_{X_2|X_1 \in S_{y_1^*}}$, by (23.13), $\exists y_2^*$ such that:

$$P_{X_2}(S_{y_2^*}|X_1 \in S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}, f_2(m, y_1^*) \in S_{y_2^*}\}|}{|\{m : f_1(m) \in S_{y_1^*}\}|} = \frac{|E_2(y_1^*, y_2^*)|}{|E_1(y_1^*)|} \geq \theta_{fb},$$

$t = n$: Continue the selection process up to y_n^* which satisfies that:

$$P_{X_n}(S_{y_n^*}|X_t \in S_{y_t^*} \text{ for } t = 1, \dots, n-1) = \frac{|E_n(y_1^*, \dots, y_n^*)|}{|E_{n-1}(y_1^*, \dots, y_{n-1}^*)|} \geq \theta_{fb}.$$

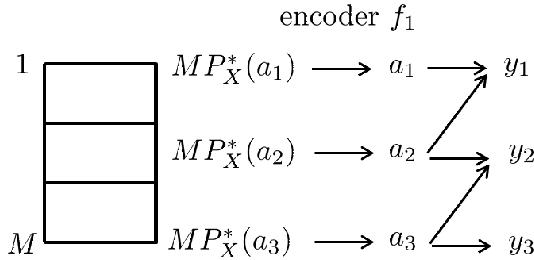
Finally, by (23.12) and the above selection procedure, we have

$$\begin{aligned} \frac{1}{M} &\geq \frac{|E_n(y_1^*, \dots, y_n^*)|}{|E_0|} \geq \theta_{fb}^n \\ \Rightarrow M &\leq -n \log \theta_{fb} \\ \Rightarrow C_{fb,0} &\leq -\log \theta_{fb} \end{aligned}$$

4 “ \geq ” (achievability)

Let's construct a code that achieves $(M, n, 0)$.

392



The above example with $|\mathcal{A}| = 3$ illustrates that the encoder f_1 partitions the space of all messages to 3 groups. The encoder f_1 at the first stage encodes the groups of messages into a_1, a_2, a_3 correspondingly. When channel outputs y_1 and assume that $S_{y_1} = \{a_1, a_2\}$, then the decoder can eliminate a total number of $MP_X^*(a_3)$ candidate messages in this round. The “confusability set” only contains the remaining $MP_X^*(S_{y_1})$ messages. By definition of P_X^* we know that $MP_X^*(S_{y_1}) \leq M\theta_{fb}$. In the second round, f_2 partitions the remaining messages into three groups, send the group index and repeat.

By similar arguments, each interaction reduces the uncertainty by a factor of *at least* θ_{fb} . After n iterations, the size of “confusability set” is upper bounded by $M\theta_{fb}^n$, if $M\theta_{fb}^n \leq 1$,³ then zero error probability is achieved. This is guaranteed by choosing $\log M = -n \log \theta_{fb}$. Therefore we have shown that $-n \log \theta_{fb}$ bits can be reliably delivered with $n + O(1)$ channel uses with feedback, thus

$$C_{fb,0} \geq -\log \theta_{fb}$$

□

23.3.2 Code with variable length

Consider the example of BEC_δ with feedback, send k bits in the following way: repeat sending each bit until it gets through the channel correctly. The expected number of channel uses for sending k bits is given by

$$l = \mathbb{E}[n] = \frac{k}{1 - \delta}$$

We state the result for *variable-length feedback* (VLF) code without proof:

$$\log M_{\text{VLF}}^*(l, 0) \geq lC$$

Notice that compared to the scheme without feedback, there is the improvement of $\sqrt{nVQ^{-1}(\epsilon)}$ in the order of $O(\sqrt{n})$, which is stronger than the result in Theorem 23.4.

This is also true in general [233]:

$$\log M_{\text{VLF}}^*(l, \epsilon) = \frac{lC}{1 - \epsilon} + O(\log l)$$

³ Some rounding-off errors need to be corrected in a few final steps (because P_X^* may not be closely approximable when very few messages are remaining). This does not change the asymptotics though.

23.3 When is feedback really useful? 393

Example 23.1. For the channel $\text{BSC}_{0.11}$ without feedback the minimal is $n = 3000$ needed to achieve 90% of the capacity C , while there exists a VLF code with $l = \mathbb{E}[n] = 200$ achieving that. This showcases how much feedback can improve the latency and decoding complexity.

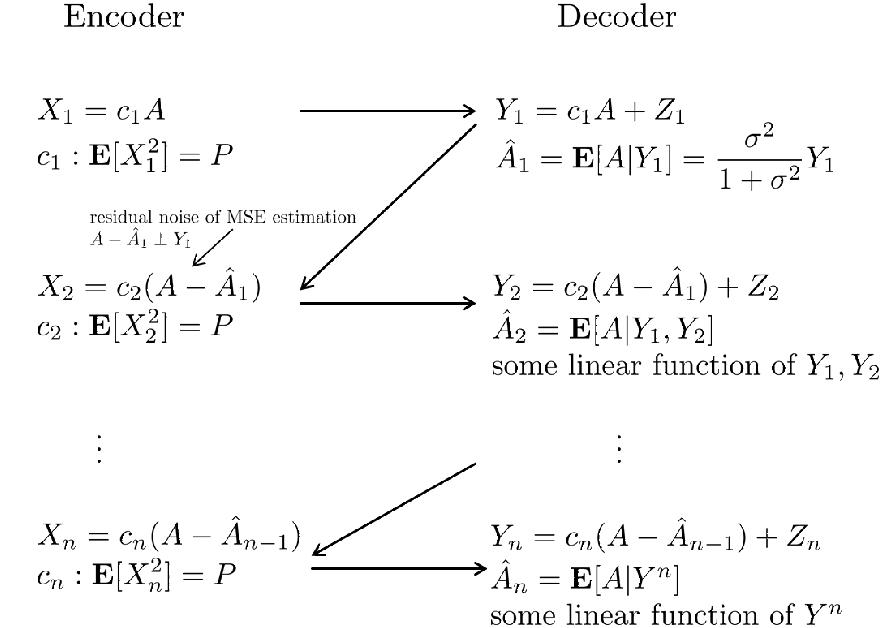
23.3.3 Code with variable power

Elias' scheme To send a number A drawn from a Gaussian distribution $\mathcal{N}(0, \text{Var}A)$, Elias' scheme uses *linear processing*. Consider the following set of AWGN channel:

$$Y_k = X_k + Z_k, \quad Z_k \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

$$\mathbb{E}[X_k^2] \leq P, \quad \text{power constraint in expectation}$$

Elias' scheme proceeds according to the figure below.⁴



According to the *orthogonality principle* of the minimum mean-square estimation (MMSE) of A at receiver side in every step:

$$A = \hat{A}_n + N_n, \quad N_n \perp\!\!\!\perp Y^n.$$

Moreover, since all operations are linear and everything is jointly Gaussian, $N_n \perp\!\!\!\perp Y^n$. Since $X_n \propto N_{n-1} \perp\!\!\!\perp Y^{n-1}$, the codeword we are sending at each time slot is independent of the history of the channel output ("innovation"), in order to maximize information transfer.

⁴ Note that if we insist each codeword satisfies power constraint almost surely instead on average, i.e., $\sum_{k=1}^n X_k^2 \leq nP$ a.s., then this scheme does not work!

Note that $Y^n \rightarrow \hat{A}_n \rightarrow A$, and the optimal estimator \hat{A}_n (a linear combination of Y^n) is a sufficient statistic of Y^n for A under Gaussianity. Then

$$\begin{aligned} I(A; Y^n) &= I(A; \hat{A}_n, Y^n) \\ &= I(A; \hat{A}_n) + I(A; Y^n | \hat{A}_n) \\ &= I(A; \hat{A}_n) \\ &= \frac{1}{2} \log \frac{\text{Var}(A)}{\text{Var}(N_n)}. \end{aligned}$$

where the last equality uses the fact that N follows a normal distribution. $\text{Var}(N_n)$ can be computed directly using standard linear MMSE results. Instead, we determine it information theoretically: Notice that we also have

$$\begin{aligned} I(A; Y^n) &= I(A; Y_1) + I(A; Y_2 | Y_1) + \cdots + I(A; Y_n | Y^{n-1}) \\ &= I(X_1; Y_1) + I(X_2; Y_2 | Y_1) + \cdots + I(X_n; Y_n | Y^{n-1}) \\ &\stackrel{\text{key}}{=} I(X_1; Y_1) + I(X_2; Y_2) + \cdots + I(X_n; Y_n) \\ &= n \frac{1}{2} \log(1 + P) = nC \end{aligned}$$

Therefore, with Elias' scheme of sending $A \sim \mathcal{N}(0, \text{Var } A)$, after the n -th use of the AWGN(P) channel with feedback,

$$\text{Var } N_n = \text{Var}(\hat{A}_n - A) = 2^{-2nC} \text{Var } A = \left(\frac{P}{P + \sigma^2} \right)^n \text{Var } A,$$

which says that the reduction of uncertainty in the estimation is exponential fast in n .

Schalkwijk-Kailath Elias' scheme can also be used to send digital data. Let $W \sim$ uniform on M -PAM constellation in $\in [-1, 1]$, i.e., $\{-1, -1 + \frac{2}{M}, \dots, -1 + \frac{2k}{M}, \dots, 1\}$. In the very first step W is sent (after scaling to satisfy the power constraint):

$$X_0 = \sqrt{P}W, \quad Y_0 = X_0 + Z_0$$

Since Y_0 and X_0 are both known at the encoder, it can compute Z_0 . Hence, to describe W it is sufficient for the encoder to describe the noise realization Z_0 . This is done by employing the Elias' scheme ($n - 1$ times). After $n - 1$ channel uses, and the MSE estimation, the equivalent channel output:

$$\tilde{Y}_0 = X_0 + \tilde{Z}_0, \quad \text{Var}(\tilde{Z}_0) = 2^{-2(n-1)C}$$

Finally, the decoder quantizes \tilde{Y}_0 to the nearest PAM point. Notice that

$$\epsilon \leq \mathbb{P} \left[|\tilde{Z}_0| > \frac{1}{2M} \right] = \mathbb{P} \left[2^{-(n-1)C} |Z| > \frac{\sqrt{P}}{2M} \right] = 2Q \left(\frac{2^{(n-1)C} \sqrt{P}}{2M} \right)$$

so that

$$\log M \geq (n - 1)C + \log \frac{\sqrt{P}}{2} - \log Q^{-1} \left(\frac{\epsilon}{2} \right) = nC + O(1).$$

23.3 When is feedback really useful? 395

Hence if the rate is strictly less than capacity, the error probability decays *doubly exponentially* as n increases. More importantly, we gained an \sqrt{n} term in terms of $\log M$, since for the case without feedback we have (by Theorem 22.2)

$$\log M^*(n, \epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n).$$

As an example, consider $P = 1$ and then channel capacity is $C = 0.5$ bit per channel use. To achieve error probability 10^{-3} , $2Q\left(\frac{2^{(n-1)C}}{2M}\right) \approx 10^{-3}$, so $\frac{e^{(n-1)C}}{2M} \approx 3$, and $\frac{\log M}{n} \approx \frac{n-1}{n}C - \frac{\log 8}{n}$. Notice that the capacity is achieved to within 99% in as few as $n = 50$ channel uses, whereas the best possible block codes without feedback require $n \approx 2800$ to achieve 90% of capacity.

The *take-away message* of this chapter is as follows: Feedback is best harnessed with *adaptive* strategies. Although it does not increase capacity under block coding, feedback greatly boosts reliability as well as reduces coding complexity.

Exercises for Part IV

IV.1 A code with $M = 2^k$, average probability of error $\epsilon < \frac{1}{2}$ and bit-error probability $p_b < \frac{1}{2}$ must satisfy both

$$\log M \leq \frac{C + h(\epsilon)}{1 - \epsilon} \quad (\text{IV.1})$$

and

$$\log M \leq \frac{C}{\log 2 - h(p_b)}, \quad (\text{IV.2})$$

where $C = \sup_{P_X} I(X; Y)$. Since $p_b \leq \epsilon$, in the bound (IV.2) we may replace $h(p_b)$ with $h(\epsilon)$ to obtain a new bound. Suppose that a value of k is fixed and the bounds are used to prove a lower bound on ϵ . When will the new bound be better than (IV.1)?

IV.2 A magician is performing card tricks on stage. In each round he takes a shuffled deck of 52 cards and asks someone to pick a random card N from the deck, which is then revealed to the audience. Assume the magician can prepare an arbitrary ordering of cards in the deck (before each round) and that N is distributed binomially on $\{0, \dots, 51\}$ with mean $\frac{51}{2}$.

- (a) What is the maximal number of *bits per round* that he can send over to his companion in the room? (in the limit of infinitely many rounds)
- (b) Is communication possible if N were uniform on $\{0, \dots, 51\}$? (In practice, however, nobody ever picks the top or the bottom ones)

IV.3 Find the capacity of the erasure-error channel (Fig. 23.2) with channel matrix

$$W = \begin{bmatrix} 1 - 2\delta & \delta & \delta \\ \delta & \delta & 1 - 2\delta \end{bmatrix}$$

where $0 \leq \delta \leq 1/2$.

FIXME

Figure 23.2 Binary erasure-error channel.

IV.4 Consider a binary symmetric channel with crossover probability $\delta \in (0, 1)$:

$$Y = X + Z \pmod{2}, Z \sim \text{Bern}(\delta).$$

Suppose that in addition to Y the receiver also gets to observe noise Z through a binary erasure channel with erasure probability $\delta_e \in (0, 1)$. Compute:

- (a) Capacity C of the channel.
- (b) Zero-error capacity C_0 of the channel.

- (c) Zero-error capacity in the presence of feedback $C_{fb,0}$.
 (d) (Bonus) Now consider the setup when in addition to feedback also the variable-length communication with feedback and termination (VLFT) is allowed. What is the zero-error capacity (in bits per average number of channel uses) in this case? (In VLFT model, transmitter can send a special symbol T that is received without error, but the channel dies after T has been sent.)

IV.5 (Time varying channel, Problem 9.12 [75]) A train pulls out of the station at constant velocity. The received signal energy thus falls off with time as $1/t^2$. The total received signal at time i is

$$Y_i = \left(\frac{1}{i} \right) X_i + Z_i,$$

where $Z_1, Z_2, \dots \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. The transmitter constraint for block length n is

$$\frac{1}{n} \sum_{i=1}^n x_i^2(w) \leq P, \quad w \in \{1, 2, \dots, 2^{nR}\}.$$

Using Fano's inequality, show that the capacity C is equal to zero for this channel.

IV.6 Randomized encoders and decoders may help for maximal probability of error:

- (a) Consider a binary asymmetric channel $P_{Y|X} : \{0, 1\} \rightarrow \{0, 1\}$ specified by $P_{Y|X=0} = \text{Ber}(1/2)$ and $P_{Y|X=1} = \text{Ber}(1/3)$. The encoder $f : [M] \rightarrow \{0, 1\}$ tries to transmit 1 bit of information, i.e., $M = 2$, with $f(1) = 0, f(2) = 1$. Show that the optimal decoder which minimizes the maximal probability of error is necessarily randomized. Find the optimal decoder and the optimal $P_{e,\max}$. (Hint: Recall binary hypothesis testing.)
 (b) Give an example of $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}, M > 1$ and $\epsilon > 0$ such that there is an $(M, \epsilon)_{\max}$ -code with a randomized encoder-decoder, but no such code with a deterministic encoder-decoder.

IV.7 Routers A and B are setting up a covert communication channel in which the data is encoded in the ordering of packets. Formally: router A receives n packets, each of type A or D (for Ack/Data), where type is i.i.d. Bernoulli(p) with $p \approx 0.9$. It encodes k bits of secret data by reordering these packets. The network between A and B delivers packets in-order with loss rate $\delta \approx 5\%$ (Note: packets have sequence numbers, so each loss is detected by B).

What is the maximum rate $\frac{k}{n}$ of reliable communication achievable for large n ? Justify your answer!

IV.8 (Strong converse for BSC) In this exercise we give a combinatorial proof of the strong converse for the binary symmetric channel. For BSC_δ with $0 < \delta < \frac{1}{2}$,

- (a) Given any $(n, M, \epsilon)_{\max}$ -code with deterministic encoder f and decoder g , recall that the decoding regions $\{D_i = g^{-1}(i)\}_{i=1}^M$ form a partition of the output space. Prove that for all $i \in [M]$,

$$|D_i| \geq \sum_{j=0}^L \binom{n}{j}$$

where L is the largest integer such that $\mathbb{P}[\text{Binomial}(n, \delta) \leq L] \leq 1 - \epsilon$.

398 Exercises for Part IV

(b) Conclude that

$$M \leq 2^{n(1-h(\delta))+o(n)}. \quad (\text{IV.3})$$

- (c) Show that (IV.3) holds for average probability of error. (Hint: how to go from maximal to average probability of error?)
- (d) Conclude that strong converse holds for BSC. (Hint: argue that requiring deterministic encoder/decoder does not change the asymptotics.)

IV.9 Recall that the AWGN channel is specified by

$$Y^n = X^n + Z^n, \quad Z^n \sim \mathcal{N}(0, I_n), \quad c(x^n) = \frac{1}{n} \|x^n\|^2$$

Prove the strong converse for the AWGN via the following steps:

- (a) Let $c_i = f(i)$ and $D_i = g^{-1}(i)$, $i = 1, \dots, M$ be the codewords and the decoding regions of an $(n, M, P, \epsilon)_{\max}$ code. Let

$$Q_{Y^n} = \mathcal{N}(0, (1+P)I_n).$$

Show that there must exist a codeword c and a decoding region D such that

$$P_{Y^n|X^n=c}[D] \geq 1 - \epsilon \quad (\text{IV.4})$$

$$Q_{Y^n}[D] \leq \frac{1}{M}. \quad (\text{IV.5})$$

(b) Show that then

$$\beta_{1-\epsilon}(P_{Y^n|X^n=c}, Q_{Y^n}) \leq \frac{1}{M}. \quad (\text{IV.6})$$

(c) Show that hypothesis testing problem

$$P_{Y^n|X^n=c} \quad \text{vs.} \quad Q_{Y^n}$$

is equivalent to

$$P_{Y^n|X^n=Uc} \quad \text{vs.} \quad Q_{Y^n}$$

where $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix. (Hint: use spherical symmetry of white Gaussian distributions.)(d) Choose U such that

$$P_{Y^n|X^n=Uc} = P^n,$$

where P^n is an iid Gaussian distribution of mean that depends on $\|c\|^2$.

(e) Apply Stein's lemma (Theorem 14.15) to show:

$$\beta_{1-\epsilon}(P^n, Q_{Y^n}) = \exp\{-nE + o(n)\}$$

(f) Conclude via (IV.6) that

$$\log M \leq nE + o(n) \implies C_\epsilon \leq \frac{1}{2} \log(1+P).$$

IV.10 (Capacity-cost at $P = P_0$.) Recall that we have shown that for stationary memoryless channels and $P > P_0$ capacity equals $f(P)$:

$$C(P) = f(P), \quad (\text{IV.7})$$

where

$$P_0 \triangleq \inf_{x \in \mathcal{A}} c(x) \quad (\text{IV.8})$$

$$f(P) \triangleq \sup_{X: \mathbb{E}[c(X)] \leq P} I(X; Y). \quad (\text{IV.9})$$

Show:

- (a) If P_0 is not admissible, i.e., $c(x) > P_0$ for all $x \in \mathcal{A}$, then $C(P_0)$ is undefined (even $M = 1$ is not possible)
- (b) If there exists a unique x_0 such that $c(x_0) = P_0$ then

$$C(P_0) = f(P_0) = 0.$$

- (c) If there are more than one x with $c(x) = P_0$ then we still have

$$C(P_0) = f(P_0).$$

- (d) Give example of a channel with discontinuity of $C(P)$ at $P = P_0$. (Hint: select a suitable cost function for the channel $Y = (-1)^Z \cdot \text{sign}(X)$, where Z is Bernoulli and $\text{sign} : \mathbb{R} \rightarrow \{-1, 0, 1\}$)

IV.11 Consider a stationary memoryless additive non-Gaussian noise channel:

$$Y_i = X_i + Z_i, \quad \mathbb{E}[Z_i] = 0, \quad \text{Var}[Z_i] = 1$$

with the input constraint

$$\|x^n\|_2 \leq \sqrt{nP} \iff \sum_{i=1}^n x_i^2 \leq nP.$$

- (a) Prove that capacity $C(P)$ of this channel satisfies

$$\frac{1}{2} \log(1 + P) \leq C(P) \leq \frac{1}{2} \log(1 + P) + D(P_Z \parallel \mathcal{N}(0, 1)),$$

where P_Z is the distribution of the noise. (Hints: Gaussian saddle point and the golden formula $I(X; Y) \leq D(P_{Y|X} \parallel Q_Y | P_X)$.)

- (b) If $D(P_Z \parallel \mathcal{N}(0, 1)) = \infty$ (Z is very non-Gaussian), then it is possible that the capacity is infinite. Consider Z is ± 1 equiprobably. Show that the capacity is infinite by a) proving the maximal mutual information is infinite; b) giving an explicit scheme to achieve infinite capacity.

IV.12 In Section 18.6 we showed that for additive noise, random linear codes achieves the same performance as Shannon's ensemble (fully random coding). The total number of possible generator matrices is q^{nk} , which is significant smaller than double exponential, but still quite large. Now

400 Exercises for Part IV

we show that without degrading the performance, we can reduce this number to q^n by restricting to *Toeplitz* generator matrix G , i.e., $G_{ij} = G_{i-1,j-1}$ for all $i, j > 1$.

Prove the following strengthening of Theorem 18.18: Let $P_{Y|X}$ be additive noise over \mathbb{F}_q^n . For any $1 \leq k \leq n$, there exists a linear code $f: \mathbb{F}_q^k \rightarrow \mathbb{F}_q^n$ with *Toeplitz* generator matrix, such that

$$P_{e,\max} = P_e \leq \mathbb{E} \left[q^{-\left(n - k - \log_q \frac{1}{P_{Z^n}(Z^n)} \right)^+} \right]$$

How many *Toeplitz* generator matrices are there?

Hint: Analogous to the proof Theorem 15.2, first consider random linear codewords plus random dithering, then argue that dithering can be removed without changing the performance of the codes. Show that codewords are pairwise independent and uniform.

- IV.13** (Wozencraft ensemble) Let $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^2$, a vector space of dimension two over Galois field with q elements. A Wozencraft code of rate 1/2 is a map parameterized by $0 \neq u \in \mathbb{F}_q$ given as $a \mapsto (a, a \cdot u)$, where $a \in \mathbb{F}_q$ corresponds to the original message, multiplication is over \mathbb{F}_q and (\cdot, \cdot) denotes a 2-dimensional vector in \mathbb{F}_q^2 . We will show there exists u yielding a $(q, \epsilon)_{\text{avg}}$ code with

$$\epsilon \leq \mathbb{E} \left[\exp \left\{ - \left| i(X; Y) - \log \frac{q^2}{2(q-1)} \right|^+ \right\} \right] \quad (\text{IV.10})$$

for the channel $Y = X + Z$ where X is uniform on \mathbb{F}_q^2 , noise $Z \in \mathbb{F}_q^2$ has distribution P_Z and

$$i(a; b) \triangleq \log \frac{P_Z(b-a)}{q^{-2}}.$$

- (a) Show that probability of error of the code $a \mapsto (av, au) + h$ is the same as that of $a \mapsto (a, auv^{-1})$.
- (b) Let $\{X_a, a \in \mathbb{F}_q\}$ be a random codebook defined as

$$X_a = (aV, aU) + H,$$

with V, U – uniform over non-zero elements of \mathbb{F}_q and H – uniform over \mathbb{F}_q^2 , the three being jointly independent. Show that for $a \neq a'$ we have

$$P_{X_a, X_{a'}}(x_1^2, \tilde{x}_1^2) = \frac{1}{q^2(q-1)^2} \mathbf{1}\{x_1 \neq \tilde{x}_1, x_2 \neq \tilde{x}_2\}$$

- (c) Show that for $a \neq a'$

$$\begin{aligned} \mathbb{P}[i(X'_a; X_a + Z) > \log \beta] &\leq \frac{q^2}{(q-1)^2} \mathbb{P}[i(\bar{X}; Y) > \log \beta] - \frac{1}{(q-1)^2} \mathbb{P}[i(X; Y) > \log \beta] \\ &\leq \frac{q^2}{(q-1)^2} \mathbb{P}[i(\bar{X}; Y) > \log \beta], \end{aligned}$$

where $P_{\bar{X}XY}(\bar{a}, a, b) = \frac{1}{q^4} P_Z(b-a)$.

- (d) Conclude by following the proof of the DT bound with $M = q$ that the probability of error averaged over the random codebook $\{X_a\}$ satisfies (IV.10).

Exercises for Part IV 401

IV.14 (Information density and types.) Let $P_{Y|X} : \mathcal{A} \rightarrow \mathcal{B}$ be a DMC and let P_X be some input distribution. Take $P_{X^n Y^n} = P_{XY}^n$ and define $i(a^n; b^n)$ with respect to this $P_{X^n Y^n}$.

- (a) Show that $i(x^n; y^n)$ is a function of only the “joint-type” \hat{P}_{XY} of (x^n, y^n) , which is a distribution on $\mathcal{A} \times \mathcal{B}$ defined as

$$\hat{P}_{XY}(a, b) = \frac{1}{n} \# \{i : x_i = a, y_i = b\},$$

where $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Therefore $\{\frac{1}{n} i(x^n; y^n) \geq \gamma\}$ can be interpreted as a constraint on the joint type of (x^n, y^n) .

- (b) Assume also that the input x^n is such that $\hat{P}_X = P_X$. Show that

$$\frac{1}{n} i(x^n; y^n) \leq I(\hat{P}_X, \hat{P}_{Y|X}).$$

The quantity $I(\hat{P}_X, \hat{P}_{Y|X})$, sometimes written as $I(x^n \wedge y^n)$, is an *empirical mutual information*⁵. Hint:

$$\begin{aligned} \mathbb{E}_{Q_{XY}} \left[\log \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] &= \\ D(Q_{Y|X} \| Q_Y | Q_X) + D(Q_Y \| P_Y) - D(Q_{Y|X} \| P_{Y|X} | Q_X) \end{aligned} \quad (\text{IV.11})$$

IV.15 (Fitingof-Goppa universal codes) Consider a finite abelian group \mathcal{X} . Define *Fitingof norm* as

$$\|x^n\|_\Phi \triangleq nH(\hat{P}_{x^n}) = nH(x_T), \quad T \sim \text{Unif}([n]),$$

where \hat{P}_{x^n} is the empirical distribution of x^n .

- (a) Show that $\|x^n\|_\Phi = \| -x^n \|_\Phi$ and triangle inequality

$$\|x^n - y^n\|_\Phi \leq \|x^n\|_\Phi + \|y^n\|_\Phi$$

Conclude that $d_\Phi(x^n, y^n) \triangleq \|x^n - y^n\|_\Phi$ is a translation invariant (Fitingof) metric on the set of equivalence classes in \mathcal{X}^n , with equivalence $x^n \sim y^n \iff \|x^n - y^n\|_\Phi = 0$.

- (b) Define Fitingof ball $B_r(x^n) \triangleq \{y^n : d_\Phi(x^n, y^n) \leq r\}$. Show that

$$\log |B_{\lambda n}(x^n)| = \lambda n + O(\log n)$$

for all $0 \leq \lambda \leq \log |\mathcal{X}|$.

- (c) Show that for any i.i.d. measure $P_{Z^n} = P_Z^n$ on \mathcal{X}^n we have

$$\lim_{n \rightarrow \infty} P_{Z^n}[B_{\lambda n}(0^n)] = \begin{cases} 1, & H(Z) < \lambda \\ 0, & H(Z) > \lambda \end{cases}$$

- (d) Conclude that a code $\mathcal{C} \subset \mathcal{X}^n$ with Fitingof minimal distance $d_{min,\Phi}(\mathcal{C}) \triangleq \min_{c \neq c' \in \mathcal{C}} d_\Phi(c, c') \geq 2\lambda n$ is decodable with vanishing probability of error on any additive-noise channel $Y = X + Z$, as long as $H(Z) < \lambda$.

⁵ Invented by V. Goppa for his maximal mutual information (MMI) decoder: $\hat{W} = \text{argmax}_i I(c_i \wedge y^n)$.

402 Exercises for Part IV

Comment: By Feinstein-lemma like argument it can be shown that there exist codes of size $\mathcal{X}^{n(1-\lambda)}$, such that balls of radius λn centered at codewords are almost disjoint. Such codes are universally capacity achieving for all memoryless additive noise channels on \mathcal{X} . Extension to general (non-additive) channels is done via introducing $d_\Phi(x^n, y^n) = nH(x_T|y_T)$, while extension to channels with Markov memory is done by introducing Markov-type norm $\|x^n\|_{\Phi_1} = nH(x_T|x_{T-1})$. See [140, Chapter 3].

IV.16 Consider the *polygon channel* discussed in Remark 23.9, where the input and output alphabet are both $\{1, \dots, L\}$, and $P_{Y|X}(b|a) > 0$ if and only if $b = a$ or $b = (a \bmod L) + 1$. The confusability graph is a *cycle* of L vertices. *Rigorously* prove the following:

- For all L , The zero-error capacity with feedback is $C_{fb,0} = \log \frac{L}{2}$.
- For even L , the zero-error capacity without feedback $C_0 = \log \frac{L}{2}$.
- Now consider the following channel, where the input and output alphabet are both $\{1, \dots, L\}$, and $P_{Y|X}(b|a) > 0$ if and only if $b = a$ or $b = a+1$. In this case the confusability graph is a *path* of L vertices. Show that the zero-error capacity is given by

$$C_0 = \log \left\lceil \frac{L}{2} \right\rceil$$

What is $C_{fb,0}$?

IV.17 (Input-output cost) Let $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$ be a DMC and consider a cost function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ (note that $c(x, y) \leq L < \infty$ for some L). Consider a problem of channel coding, where the error-event is defined as

$$\{\text{error}\} \triangleq \{\hat{W} \neq W\} \cup \left\{ \sum_{k=1}^n c(X_k, Y_k) > nP \right\},$$

where P is a fixed parameter. Define operational capacity $C(P)$ and show it is given by

$$C^{(I)}(P) = \max_{P_X: \mathbb{E}[c(X, Y)] \leq P} I(X; Y)$$

for all $P > P_0 \triangleq \min_{x_0} \mathbb{E}[c(X, Y)|X = x_0]$. Give a counter-example for $P = P_0$. (Hint: do a converse directly, and for achievability reduce to an appropriately chosen cost-function $c'(x)$).

IV.18 (Expurgated random coding bound)

- For any code \mathcal{C} show the following bound on probability of error

$$P_e(\mathcal{C}) \leq \frac{1}{M} \sum_{c \neq c'} 2^{-d_B(c, c')},$$

where Bhattacharya distance $d_B(x^n, \tilde{x}^n) = \sum_{j=1}^n d_B(x_j, \tilde{x}_j)$ and a single-letter

$$d_B(x, \tilde{x}) = -\log_2 \sum_{y \in \mathcal{Y}} \sqrt{W(y|x)W(y|\tilde{x})}.$$

Exercises for Part IV 403

- (b) Fix P_X and let $E_{0,x}(\rho, P_X) \triangleq -\rho \log_2 \mathbb{E}[2^{-\frac{1}{\rho} d_B(X, X')}]$, where $X \perp\!\!\!\perp X' \sim P_X$. Show by random coding that there always exists a code \mathcal{C} of rate R with

$$P_e(\mathcal{C}) \leq 2^{n(E_{0,x}(1, P_X) - R)}.$$

- (c) We improve the previous bound as follows. We still generate \mathcal{C} by random coding. But this time we expurgate all codewords with $f(c, \mathcal{C}) > \text{med}(f(c, \mathcal{C}))$, where $f(c) = \sum_{c' \neq c} 2^{-d_B(c, c')}$. Using the bound

$$\text{med}(V) \leq 2^\rho \mathbb{E}[V^{1/\rho}]^\rho \quad \forall \rho \geq 1$$

show that

$$\text{med}(f(c, \mathcal{C})) \leq 2^{n(\rho R - E_{0,x}(\rho, P_X))}.$$

- (d) Conclude that there must exist a code with rate $R - O(1/n)$ and $P_e(\mathcal{C}) \leq 2^{-nE_{ex}(R)}$, where

$$E_{ex}(R) \triangleq \max_{\rho \geq 1} -\rho R + \max_{P_X} E_{0,x}(\rho, P_X).$$

IV.19 Give example of a channel with discontinuity of $C(P)$ at $P = P_0$. (Hint: select a suitable cost function for the channel $Y = (-1)^Z \cdot \text{sign}(X)$, where Z is Bernoulli and $\text{sign} : \mathbb{R} \rightarrow \{-1, 0, 1\}$)

IV.20 (Sum of channels) Let W_1 and W_2 denote the channel matrices of discrete memoryless channel (DMC) $P_{Y_1|X_1}$ and $P_{Y_2|X_2}$ with capacity C_1 and C_2 , respectively. The sum of the two channels is another DMC with channel matrix $\begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix}$. Show that the capacity of the sum channel is given by

$$C = \log(\exp(C_1) + \exp(C_2)).$$

IV.21 (Product of channels) For $i = 1, 2$, let $P_{Y_i|X_i}$ be a channel with input space \mathcal{A}_i , output space \mathcal{B}_i , and capacity C_i . Their product channel is a channel with input space $\mathcal{A}_1 \times \mathcal{A}_2$, output space $\mathcal{B}_1 \times \mathcal{B}_2$, and transition kernel $P_{Y_1 Y_2 | X_1 X_2} = P_{Y_1|X_1} P_{Y_2|X_2}$. Show that the capacity of the product channel is given by

$$C = C_1 + C_2.$$

IV.22 Mixtures of DMCs. Consider two DMCs $U_{Y|X}$ and $V_{Y|X}$ with a common capacity achieving input distribution and capacities $C_U < C_V$. Let $T = \{0, 1\}$ be uniform and consider a channel $P_{Y^n|X^n}$ that uses U if $T = 0$ and V if $T = 1$, or more formally:

$$P_{Y^n|X^n}(y^n|x^n) = \frac{1}{2} U_{Y|X}^n(y^n|x^n) + \frac{1}{2} V_{Y|X}^n(y^n|x^n). \quad (\text{IV.12})$$

Show:

- (a) Is this channel $\{P_{Y^n|X^n}\}_{n \geq 1}$ stationary? Memoryless?
- (b) Show that the Shannon capacity C of this channel is not greater than C_U .
- (c) The maximal mutual information rate is

$$C^{(I)} = \lim_{n \rightarrow \infty} \frac{1}{n} \sup_{X^n} I(X^n; Y^n) = \frac{C_U + C_V}{2}$$

404 Exercises for Part IV

(d) Conclude that

$$C < C^{(I)}.$$

IV.23 Compound DMC [37] Compound DMC is a family of DMC's with common input and output alphabets $P_{Y_s|X} : \mathcal{A} \rightarrow \mathcal{B}, s \in \mathcal{S}$. An (n, M, ϵ) code is an encoder-decoder pair whose probability of error $\leq \epsilon$ over any channel $P_{Y_s|X}$ in the family (note that the same encoder and the same decoder are used for each $s \in \mathcal{S}$). Show that capacity is given by

$$C = \sup_{P_X} \inf_s I(X; Y_s).$$

IV.24 Consider the following (memoryless) channel. It has a side switch U that can be in positions ON and OFF. If U is on then the channel from X to Y is BSC_δ and if U is off then Y is Bernoulli $(1/2)$ regardless of X . The receiving party sees Y but not U . A design constraint is that U should be in the ON position no more than the fraction s of all channel uses, $0 \leq s \leq 1$. Questions:

- (a) One strategy is to put U into ON over the first sn time units and ignore the rest of the $(1-s)n$ readings of Y . What is the maximal rate in bits per channel use achievable with this strategy?
- (b) Can we increase the communication rate if the encoder is allowed to modulate the U switch together with the input X (while still satisfying the s -constraint on U)?
- (c) Now assume nobody has access to U , which is random, independent of X , memoryless across different channel uses and

$$P[U = \text{ON}] = s.$$

Find capacity.

IV.25 Let $\{Z_j, j = 1, 2, \dots\}$ be a stationary Gaussian process with variance 1 such that Z_j form a Markov chain $Z_1 \rightarrow \dots \rightarrow Z_n \rightarrow \dots$ Consider an additive channel

$$Y^n = X^n + Z^n$$

with power constraint $\sum_{j=1}^n |x_j|^2 \leq nP$. Suppose that $I(Z_1; Z_2) = \epsilon \ll 1$, then capacity-cost function

$$C(P) = \frac{1}{2} \log(1 + P) + B\epsilon + o(\epsilon)$$

as $\epsilon \rightarrow 0$. Compute B and interpret your answer.

How does the frequency spectrum of optimal signal change with increasing ϵ ?

IV.26 A semiconductor company offers a random number generator that outputs a block of random n bits Y_1, \dots, Y_n . The company wants to secretly embed a signature in every chip. To that end, it decides to encode the k -bit signature in n real numbers $X_j \in [0, 1]$. To each individual signature a chip is manufactured that produces the outputs $Y_j \sim \text{Ber}(X_j)$. In order for the embedding to be inconspicuous the average bias P should be small:

$$\frac{1}{n} \sum_{j=1}^n \left| X_j - \frac{1}{2} \right| \leq P.$$

Exercises for Part IV 405

As a function of P how many signature bits per output (k/n) can be reliably embedded in this fashion? Is there a simple coding scheme achieving this performance?

- IV.27** Consider a DMC with two outputs $P_{Y,U|X}$. Suppose that receiver observes only Y , while U is (causally) fed back to the transmitter. We know that when $Y = U$ the capacity is not increased.

- (a) Show that capacity is not increased in general (even when $Y \neq U$).
- (b) Suppose now that there is a cost function c and $c(x_0) = 0$. Show that capacity per unit cost (with U being fed back) is still given by

$$C_V = \max_{x \neq x_0} \frac{D(P_{Y|X=x} \| P_{Y|X=x_0})}{c(x)}$$

- IV.28** (Capacity of sneezing) A sick student is sneezing periodically every minute, with each sneeze happening i.i.d. with probability p . He decides to send k bits to a friend by modulating the sneezes. For that, every time he realizes he is about to sneeze he chooses to suppress a sneeze or not. A friend listens for n minutes and then tries to decode k bits.

- (a) Find capacity in bits per minute. (Hint: Think how to define the channel so that channel input at time t were not dependent on the arrival of the sneeze at time t . To rule out strategies that depend on arrivals of past sneezes, you may invoke Exercise IV.27.)
- (b) Suppose sender can suppress at most E sneezes and listener can wait indefinitely ($n = \infty$). Show that sender can transmit $C_{puc}E + o(E)$ bits reliably as $E \rightarrow \infty$ and find C_{puc} . Curiously, $C_{puc} \geq 1.44$ bits/sneeze regardless of p . (Hint: This is similar to Exercise IV.17.)
- (c) (Bonus, hard) Redo 1 and 2 for the case of a clairvoyant student who knows exactly when sneezes will happen in the future.

- IV.29** An inmate has n oranges that he is using to communicate with his conspirators by putting oranges in trays. Assume that infinitely many trays are available, each can contain zero or more oranges, and each orange in each tray is eaten by guards independently with probability δ . In the limit of $n \rightarrow \infty$ show that an arbitrary high rate (in bits per orange) is achievable.

- IV.30** Recall that in the proof of the DT bound we used the decoder that outputs (for a given channel output y) the first c_m that satisfies

$$\{i(c_m; y) > \log \beta\}. \quad (\text{IV.13})$$

One may consider the following generalization. Fix $E \subset \mathcal{X} \times \mathcal{Y}$ and let the decoder output the first c_m which satisfies

$$(c_m, y) \in E$$

By repeating the random coding proof steps (as in the DT bound) show that the average probability of error satisfies

$$\mathbb{E}[P_e] \leq \mathbb{P}[(X, Y) \notin E] + \frac{M-1}{2} \mathbb{P}[(\bar{X}, Y) \in E],$$

where

$$P_{XY\bar{X}}(a, b, \bar{a}) = P_X(a)P_{Y|X}(b|a)P_X(\bar{a}).$$

Conclude that the optimal E is given by (IV.13) with $\beta = \frac{M-1}{2}$.

406 Exercises for Part IV

IV.31 Bounds for the binary erasure channel (BEC). Consider a code with $M = 2^k$ operating over the blocklength n BEC with erasure probability $\delta \in [0, 1]$.

- (a) Show that regardless of the encoder-decoder pair:

$$\mathbb{P}[\text{error} | \#\text{erasures} = z] \geq |1 - 2^{n-z-k}|^+$$

- (b) Conclude by averaging over the distribution of z that the probability of error ϵ must satisfy

$$\epsilon \geq \sum_{\ell=n-k+1}^n \binom{n}{\ell} \delta^\ell (1-\delta)^{n-\ell} (1 - 2^{n-\ell-k}) , \quad (\text{IV.14})$$

- (c) By applying the DT bound with uniform P_X show that there exist codes with

$$\epsilon \leq \sum_{t=0}^n \binom{n}{t} \delta^t (1-\delta)^{n-t} 2^{-|n-t-k+1|^+} . \quad (\text{IV.15})$$

- (d) Fix $n = 500$, $\delta = 1/2$. Compute the smallest k for which the right-hand side of (IV.14) is greater than 10^{-3} .
(e) Fix $n = 500$, $\delta = 1/2$. Find the largest k for which the right-hand side of (IV.15) is smaller than 10^{-3} .
(f) Express your results in terms of lower and upper bounds on $\log M^*(500, 10^{-3})$.

Part V

Rate-distortion theory and metric entropy



In Part II we studied lossless data compression (source coding), where the goal is to compress a random variable (source) X into a minimal number of bits on average (resp. exactly) so that X can be reconstructed exactly (resp. with high probability) using these bits. In both cases, the fundamental limit is given by the entropy of the source X . Clearly, this paradigm is confined to discrete random variables.

In this part we will tackle the next topic, *lossy data compression*: Given a random variable X , encode it into a minimal number of bits, such that the decoded version \hat{X} is a faithful reconstruction of X , in the sense that the “distance” between X and \hat{X} is at most some prescribed accuracy either on average or with high probability.

The motivations for study lossy compression are at least two-fold:

- 1 Many natural signals (e.g. audio, images, or video) are continuously valued. As such, there is a need to represent these real-valued random variables or processes using finitely many bits, which can be fed to downstream digital processing; see Fig. 23.3 for an illustration.

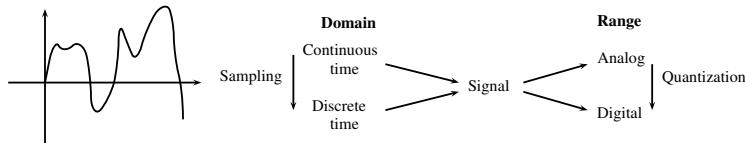


Figure 23.3 Sampling and quantization in engineering practice.

- 2 There is a lot to be gained in compression if we allow some reconstruction errors. This is especially important in applications where certain errors (such as high-frequency components in natural audio and visual signals) are imperceptible to humans. This observation is the basis of many important compression algorithms and standards that are widely deployed in practice, including JPEG for images, MPEG for videos, and MP3 for audios.

The operation of mapping (naturally occurring) continuous time/analog signals into (electronics-friendly) discrete/digital signals is known as *quantization*, which is an important subject in signal processing in its own right (cf. the encyclopedic survey [141]). In information theory, the study of optimal quantization is called *rate-distortion theory*, introduced by Shannon in 1959 [270]. To start, we will take a closer look at quantization next in Section 24.1, followed by the information-theoretic formulation in Section 24.2. A simple (and tight) converse bound is given in Section 24.3, with the matching achievability bound deferred to the next chapter.

Finally, in Chapter 27 we study Kolmogorov’s *metric entropy*, which is a non-probabilistic theory of quantization for sets in metric spaces. In addition to connections to the probabilistic theory of quantization in the preceding chapters, this concept has far-reaching consequences in both probability (e.g. empirical processes, small-ball probability) and statistical learning (e.g. entropic upper and lower bounds for estimation) that will be explored further in Part VI.

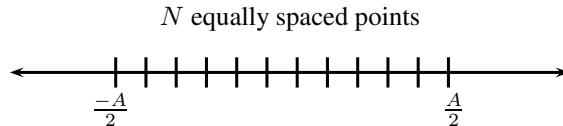
24 Rate-distortion theory

24.1 Scalar and vector quantization

24.1.1 Scalar Uniform Quantization

The idea of quantizing an inherently continuous-valued signal was most explicitly expounded in the patenting of Pulse-Coded Modulation (PCM) by A. Reeves; cf. [251] for some interesting historical notes. His argument was that unlike AM and FM modulation, quantized (digital) signals could be sent over long routes without the detrimental accumulation of noise. Some initial theoretical analysis of the PCM was undertaken in 1948 by Oliver, Pierce, and Shannon [222].

For a random variable $X \in [-A/2, A/2] \subset \mathbb{R}$, the scalar uniform quantizer $q_U(X)$ with N quantization points partitions the interval $[-A/2, A/2]$ uniformly

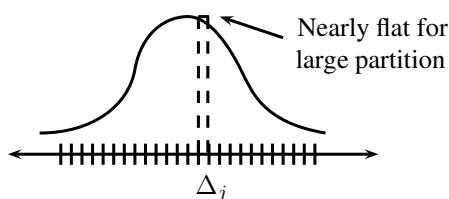


where the points are in $\{\frac{-A}{2} + \frac{kA}{N}, k = 0, \dots, N-1\}$.

What is the *quality* (or fidelity) of this quantization? Most of the time, mean squared error is used as the quality criterion:

$$D(N) = \mathbb{E}|X - q_U(X)|^2$$

where D denotes the average *distortion*. Often $R = \log_2 N$ is used instead of N , so that we think about the number of bits we can use for quantization instead of the number of points. To analyze this scalar uniform quantizer, we'll look at the high-rate regime ($R \gg 1$). The key idea in the high rate regime is that (assuming a smooth density P_X), each quantization interval Δ_j looks nearly flat, so conditioned on Δ_j , the distribution is accurately approximated by a uniform distribution.



24.1 Scalar and vector quantization 411

Let c_j be the j -th quantization point, and Δ_j be the j -th quantization interval. Here we have

$$D_U(R) = \mathbb{E}|X - q_U(X)|^2 = \sum_{j=1}^N \mathbb{E}[|X - c_j|^2 | X \in \Delta_j] \mathbb{P}[X \in \Delta_j] \quad (24.1)$$

$$\begin{aligned} \text{(high rate approximation)} &\approx \sum_{j=1}^N \frac{|\Delta_j|^2}{12} \mathbb{P}[X \in \Delta_j] \end{aligned} \quad (24.2)$$

$$= \frac{\left(\frac{A}{N}\right)^2}{12} = \frac{A^2}{12} 2^{-2R}, \quad (24.3)$$

where we used the fact that the variance of $\text{Unif}(-a, a)$ is $a^2/3$.

How much do we gain per bit?

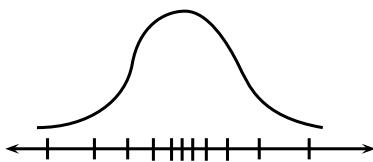
$$\begin{aligned} 10 \log_{10} \text{SNR} &= 10 \log_{10} \frac{\text{Var}(X)}{\mathbb{E}|X - q_U(X)|^2} \\ &= 10 \log_{10} \frac{12 \text{Var}(X)}{A^2} + (20 \log_{10} 2)R \\ &= \text{constant} + (6.02dB)R \end{aligned}$$

For example, when X is uniform on $[-\frac{A}{2}, \frac{A}{2}]$, the constant is 0. Every engineer knows the rule of thumb “6dB per bit”; adding one more quantization bit gets you 6 dB improvement in SNR. However, here we can see that this rule of thumb is valid only in the high rate regime. (Consequently, widely articulated claims such as “16-bit PCM (CD-quality) provides 96 dB of SNR” should be taken with a grain of salt.)

The above discussion deals with X with a bounded support. When X is unbounded, it is wise to allocate the quantization points to those values that are more likely and saturate the large values at the dynamic range of the quantizer, resulting in two types of contributions to the quantization error, known as the granular distortion and overload distortion. This leads us to the question: Perhaps uniform quantization is not optimal?

24.1.2 Scalar Non-uniform Quantization

Since our source has density p_X , a good idea might be to use more quantization points where p_X is larger, and less where p_X is smaller.



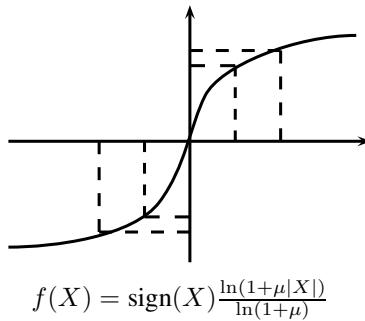
412

Often the way such quantizers are implemented is to take a monotone transformation of the source $f(X)$, perform uniform quantization, then take the inverse function:

$$\begin{array}{ccc} X & \xrightarrow{f} & U \\ \downarrow q & & \downarrow q_U \\ \hat{X} & \xleftarrow{f^{-1}} & q_U(U) \end{array} \quad (24.4)$$

i.e., $q(X) = f^{-1}(q_U(f(X)))$. The function f is usually called the *compander* (compressor+expander). One of the choice of f is the CDF of X , which maps X into uniform on $[0, 1]$. In fact, this compander architecture is optimal in the high-rate regime (fine quantization) but the optimal f is not the CDF (!). We defer this discussion till Section 24.1.4.

In terms of practical considerations, for example, the human ear can detect sounds with volume as small as 0 dB, and a painful, ear-damaging sound occurs around 140 dB. Achieving this is possible because the human ear inherently uses logarithmic companding function. Furthermore, many natural signals (such as *differences* of consecutive samples in speech or music (but not samples themselves!)) have an approximately Laplace distribution. Due to these two factors, a very popular and sensible choice for f is the μ -companding function



which compresses the dynamic range, uses more bits for smaller $|X|$'s, e.g. $|X|$'s in the range of human hearing, and less quantization bits outside this region. This results in the so-called μ -law which is used in the digital telecommunication systems in the US, while in Europe a slightly different compander called the A -law is used.

24.1.3 Optimal Scalar Quantizers

Now we look for the optimal scalar quantizer given R bits for reconstruction. Formally, this is

$$D_{\text{scalar}}(R) = \min_{q: |\text{Im } q| \leq 2^R} \mathbb{E}|X - q(X)|^2 \quad (24.5)$$

Intuitively, we would think that the optimal quantization regions should be contiguous; otherwise, given a point c_j , our reconstruction error will be larger. Therefore in one dimension quantizers are

24.1 Scalar and vector quantization 413

piecewise constant:

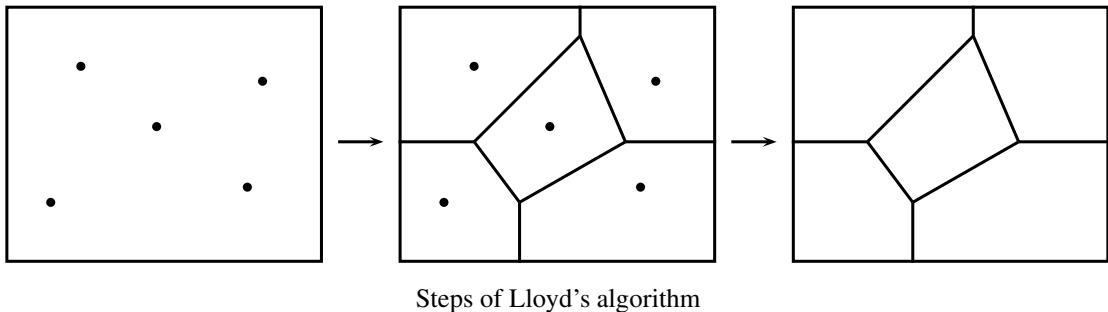
$$q(x) = c_j \mathbf{1}_{\{T_j \leq x \leq T_{j+1}\}}$$

for some $c_j \in [T_j, T_{j+1}]$.

Example 24.1. As a simple example, consider the one-bit quantization of $X \sim \mathcal{N}(0, \sigma^2)$. Then optimal quantization points are $c_1 = \mathbb{E}[X|X \geq 0] = \mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}\sigma$, $c_2 = \mathbb{E}[X|X \leq 0] = -\sqrt{\frac{2}{\pi}}\sigma$, with quantization error equal to $\text{Var}(|X|) = 1 - \frac{2}{\pi}$.

With ideas like this, in 1957 S. Lloyd developed an algorithm (called *Lloyd's algorithm* or *Lloyd's Method I*) for iteratively finding optimal quantization regions and points.¹ Suitable for both the scalar and vector cases, this method proceeds as follows: Initialized with some choice of $N = 2^k$ quantization points, the algorithm iterates between the following two steps:

- 1 Draw the Voronoi regions around the chosen quantization points (aka minimum distance tessellation, or set of points closest to c_j), which forms a partition of the space.
- 2 Update the quantization points by the centroids $\mathbb{E}[X|X \in D]$ of each Voronoi region D .



Lloyd's clever observation is that the centroid of each Voronoi region is (in general) different than the original quantization points. Therefore, iterating through this procedure gives the *Centroidal Voronoi Tessellation* (CVT - which are very beautiful objects in their own right), which can be viewed as the fixed point of this iterative mapping. The following theorem gives the results about Lloyd's algorithm

Theorem 24.1 (Lloyd).

- 1 *Lloyd's algorithm always converges to a Centroidal Voronoi Tessellation.*
- 2 *The optimal quantization strategy is always a CVT.*
- 3 *CVT's need not be unique, and the algorithm may converge to non-global optima.*

¹ This work at Bell Labs remained unpublished until 1982 [198].

Remark 24.1. The third point tells us that Lloyd's algorithm is not always guaranteed to give the optimal quantization strategy.² One sufficient condition for uniqueness of a CVT is the log-concavity of the density of X [127], e.g., Gaussians. On the other hand, even for Gaussian, if $N > 3$, optimal quantization points are not

Remark 24.2 (k-means). A popular clustering method called *k-means* is the following: Given n data points $x_1, \dots, x_n \in \mathbb{R}^d$, the goal is to find k centers $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ to minimize the objective function

$$\sum_{i=1}^n \min_{j \in [k]} \|x_i - \mu_j\|^2.$$

This is equivalent to solving the optimal vector quantization problem analogous to (24.5):

$$\min_{q: |\text{Im}(q)| \leq k} \mathbb{E}\|X - q(X)\|^2$$

where X is distributed according to the *empirical distribution* over the dataset, namely, $\frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Solving the *k-means* problem is NP-hard in the worst case, and Lloyd's algorithm is a commonly used heuristic.

24.1.4 Fine quantization

Following Panter-Dite [225], we now study the asymptotics of small quantization error. For this, introduce a probability density function $\lambda(x)$, which represents the density of quantization points in a given interval and allows us to approximate summations by integrals.³ Then the number of quantization points in any interval $[a, b]$ is $\approx N \int_a^b \lambda(x) dx$. For any point x , denote the size of the quantization interval that contains x by $\Delta(x)$. Then

$$N \int_x^{x+\Delta(x)} \lambda(t) dt \approx N\lambda(x)\Delta(x) \approx 1 \implies \Delta(x) \approx \frac{1}{N\lambda(x)}.$$

With this approximation, the quality of reconstruction is

$$\begin{aligned} \mathbb{E}|X - q(X)|^2 &= \sum_{j=1}^N \mathbb{E}[|X - c_j|^2 | X \in \Delta_j] \mathbb{P}[X \in \Delta_j] \\ &\approx \sum_{j=1}^N \mathbb{P}[X \in \Delta_j] \frac{|\Delta_j|^2}{12} \approx \int p(x) \frac{\Delta^2(x)}{12} dx \\ &= \frac{1}{12N^2} \int p(x) \lambda^{-2}(x) dx, \end{aligned}$$

² As a simple example one may consider $P_X = \frac{1}{3}\phi(x-1) + \frac{1}{3}f(x) + \frac{1}{3}\phi(x+1)$ where $f(\cdot)$ is a very narrow pdf, symmetric around 0. Here the CVT with centers $\pm \frac{2}{3}$ is not optimal among binary quantizers (just compare to any quantizer that quantizes two adjacent spikes to same value).

³ This argument is easy to make rigorous. We only need to define reconstruction points c_j as the solution of $\int_{-\infty}^{c_j} \lambda(x) dx = \frac{j}{N}$ (quantile).

24.1 Scalar and vector quantization 415

To find the optimal density λ that gives the best reconstruction (minimum MSE) when X has density p , we use Hölder's inequality: $\int p^{1/3} \leq (\int p\lambda^{-2})^{1/3}(\int \lambda)^{2/3}$. Therefore $\int p\lambda^{-2} \geq (\int p^{1/3})^3$, with equality iff $p\lambda^{-2} \propto \lambda$. Hence the optimizer is $\lambda^*(x) = \frac{p^{1/3}(x)}{\int p^{1/3}dx}$. Therefore when $N = 2^R$,⁴

$$D_{scalar}(R) \approx \frac{1}{12} 2^{-2R} \left(\int p^{1/3}(x) dx \right)^3$$

So our optimal quantizer density in the high rate regime is proportional to the cubic root of the density of our source. This approximation is called the *Panter-Dite approximation*. For example,

- When $X \in [-\frac{A}{2}, \frac{A}{2}]$, using Hölder's inequality again $\langle 1, p^{1/3} \rangle \leq \|1\|_{\frac{3}{2}} \|p^{1/3}\|_3 = A^{2/3}$, we have

$$D_{scalar}(R) \leq \frac{1}{12} 2^{-2R} A^2 = D_U(R)$$

where the RHS is the uniform quantization error given in (24.1). Therefore as long as the source distribution is not uniform, there is strict improvement. For uniform distribution, uniform quantization is, unsurprisingly, optimal.

- When $X \sim \mathcal{N}(0, \sigma^2)$, this gives

$$D_{scalar}(R) \approx \sigma^2 2^{-2R} \frac{\pi \sqrt{3}}{2} \quad (24.6)$$

Remark 24.3. In fact, in *scalar* case the optimal non-uniform quantizer can be realized using the compander architecture (24.4) that we discussed in Section 24.1.2: As an exercise, use Taylor expansion to analyze the quantization error of (24.4) when $N \rightarrow \infty$. The optimal compander $f: \mathbb{R} \rightarrow [0, 1]$ turns out to be $f(x) = \frac{\int_{-\infty}^x p^{1/3}(t) dt}{\int_{-\infty}^{\infty} p^{1/3}(t) dt}$ [28, 279].

24.1.5 Fine quantization and variable rate

So far we have been focusing on quantization with restriction on the cardinality of the image of $q(\cdot)$. If one, however, intends to further compress the values $q(X)$ losslessly, a more natural constraint is to bound $H(q(X))$.

Koshelev [179] discovered in 1963 that in the high rate regime uniform quantization is asymptotically optimal under the entropy constraint. Indeed, if q_Δ is a uniform quantizer with cell size Δ , then under appropriate assumptions we have (recall (2.21))

$$H(q_\Delta(X)) = h(X) - \log \Delta + o(1), \quad (24.7)$$

where $h(X) = -\int p_X(x) \log p_X(x) dx$ is the differential entropy of X . So a uniform quantizer with $H(q(X)) = R$ achieves

$$D = \frac{\Delta^2}{12} \approx 2^{-2R} \frac{2^{2h(X)}}{12}.$$

⁴ In fact when $R \rightarrow \infty$, “ \approx ” can be replaced by “ $= 1 + o(1)$ ” as shown by Zador [332, 333].

On the other hand, any quantizer with unnormalized point density function $\Lambda(x)$ (i.e. smooth function such that $\int_{-\infty}^{c_j} \Lambda(x) dx = j$) can be shown to achieve (assuming $\Lambda \rightarrow \infty$ pointwise)

$$D \approx \frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \quad (24.8)$$

$$H(q(X)) \approx \int p_X(x) \log \frac{\Lambda(x)}{p_X(x)} dx \quad (24.9)$$

Now, from Jensen's inequality we have

$$\frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \geq \frac{1}{12} \exp\{-2 \int p_X(x) \log \Lambda(x) dx\} \approx 2^{-2H(q(X))} \frac{2^{2h(X)}}{12},$$

concluding that uniform quantizer is asymptotically optimal.

Furthermore, it turns out that for any source, even the optimal vector quantizers (to be considered next) can not achieve distortion better than $2^{-2R} \frac{2^{2h(X)}}{2\pi e}$. That is, the maximal improvement they can gain for any i.i.d. source is 1.53 dB (or 0.255 bit/sample). This is one reason why scalar uniform quantizers followed by lossless compression is an overwhelmingly popular solution in practice.

24.2 Information-theoretic formulation

Before describing the mathematical formulation of optimal quantization, let us begin with two concrete examples.

Hamming Game. Given 100 unbiased bits, we are asked to inspect them and scribble something down on a piece of paper that can store 50 bits at most. Later we will be asked to guess the original 100 bits, with the goal of maximizing the number of correctly guessed bits. What is the best strategy? Intuitively, it seems the optimal strategy would be to store half of the bits then randomly guess on the rest, which gives 25% bit error rate (BER). However, as we will show in this chapter (Theorem 26.1), the optimal strategy amazingly achieves a BER of 11%. How is this possible? After all we are guessing independent bits and the loss function (BER) treats all bits equally.

Gaussian example. Given (X_1, \dots, X_n) drawn independently from $\mathcal{N}(0, \sigma^2)$, we are given a budget of one bit per symbol to compress, so that the decoded version $(\hat{X}_1, \dots, \hat{X}_n)$ has a small mean-squared error $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \hat{X}_i)^2]$.

To this end, a simple strategy is to quantize each coordinate into 1 bit. As worked out in Example 24.1, the optimal one-bit quantization error is $(1 - \frac{2}{\pi})\sigma^2 \approx 0.36\sigma^2$. In comparison, we will show later (Theorem 26.4) that there is a scheme that achieves an MSE of $\frac{\sigma^2}{4}$ per coordinate for large n ; furthermore, this is optimal. More generally, given R bits per symbol, by doing optimal vector quantization in high dimensions (namely, compressing (X_1, \dots, X_n) jointly to nR bits), rate-distortion theory will tell us that when n is large, we can achieve the per-coordinate MSE:

$$D_{vec}(R) = \sigma^2 2^{-2R}$$

24.2 Information-theoretic formulation 417

which, compared to (24.6), gains 4.35 dB (or 0.72 bit/sample).

The conclusions from both the Bernoulli and the Gaussian examples are rather surprising: Even when X_1, \dots, X_n are iid, there is something to be gained by quantizing these coordinates jointly. Some intuitive explanations for this high-dimensional phenomenon as follows:

- 1 Applying scalar quantization componentwise results in quantization region that are hypercubes, which may not suboptimal for covering in high dimensions.
- 2 Concentration of measures effectively removes many atypical source realizations. For example, when quantizing a single Gaussian X , we need to cover large portion of \mathbb{R} in order to deal with those significant deviations of X from 0. However, when we are quantizing many (X_1, \dots, X_n) together, the law of large numbers makes sure that many X_j 's cannot conspire together and all produce large values. Indeed, (X_1, \dots, X_n) concentrates near a sphere. As such, we may exclude large portions of the space \mathbb{R}^n from consideration.

Mathematical formulation A lossy compressor is an encoder/decoder pair (f, g) that induced the following Markov chain

$$X \xrightarrow{f} W \xrightarrow{g} \hat{X}$$

where $X \in \mathcal{X}$ is refereed to as the source, $W = f(X)$ is the compressed discrete data, and $\hat{X} = g(W)$ is the reconstruction which takes values in some alphabet $\hat{\mathcal{X}}$ that needs not be the same as \mathcal{X} .

A *distortion metric* (or loss function) is a measurable function $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R} \cup \{+\infty\}$. There are various formulations of the lossy compression problem:

- 1 Fixed length (fixed rate), average distortion: $W \in [M]$, minimize $\mathbb{E}[d(X, \hat{X})]$.
- 2 Fixed length, excess distortion: $W \in [M]$, minimize $\mathbb{P}[d(X, \hat{X}) > D]$.
- 3 Variable length, max distortion: $W \in \{0, 1\}^*$, $d(X, \hat{X}) \leq D$ a.s., minimize the average length $\mathbb{E}[l(W)]$ or entropy $H(W)$.

In this book we focus on lossy compression with fixed length and are chiefly concerned with average distortion (with the exception of joint source-channel coding in Section 26.3 where excess distortion will be needed). The difference between average and excess distortion is analogous to average and high-probability risk bound in statistics and machine learning. It turns out that under mild assumptions these two formulations lead to the same fundamental limit (cf. Remark 25.6).

As usual, of particular interest is when the source takes the form of a random vector $S^n = (S_1, \dots, S_n) \in \mathcal{S}^n$ and the reconstruction is $\hat{S}^n = (\hat{S}_1, \dots, \hat{S}_n) \in \hat{\mathcal{S}}^n$. We will be focusing on the so called *separable* distortion metric defined for n -letter vectors by averaging the single-letter distortions:

$$d(s^n, \hat{s}^n) \triangleq \frac{1}{n} \sum_{i=1}^n d(s_i, \hat{s}_i). \quad (24.10)$$

Definition 24.2. An (n, M, D) -code consists of an encoder $f : \mathcal{A}^n \rightarrow [M]$ and a decoder $g : [M] \rightarrow \hat{\mathcal{A}}^n$ such that the average distortion satisfies $\mathbb{E}[d(S^n, g(f(S^n)))] \leq D$. The nonasymptotic

and asymptotic fundamental limits are defined as follows:

$$M^*(n, D) = \min\{M : \exists(n, M, D)\text{-code}\} \quad (24.11)$$

$$R(D) = \limsup_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, D). \quad (24.12)$$

Note that, for stationary memoryless (iid) source, the large-blocklength limit in (24.12) in fact exists and coincides with the infimum over all blocklengths. This is a consequence of the average distortion criterion and the separability of the distortion metric – see Exercise V.2.

24.3 Converse bounds

Now that we have the definitions, we give a (surprisingly simple) general converse.

Theorem 24.3 (General Converse). *Suppose $X \rightarrow W \rightarrow \hat{X}$, where $W \in [M]$ and $\mathbb{E}[d(X, \hat{X})] \leq D$. Then*

$$\log M \geq \phi_X(D) \triangleq \inf_{P_{Y|X}: \mathbb{E}[d(X, Y)] \leq D} I(X; Y).$$

Proof.

$$\log M \geq H(W) \geq I(X; W) \geq I(X; \hat{X}) \geq \phi_X(D)$$

where the last inequality follows from the fact that $P_{\hat{X}|X}$ is a feasible solution (by assumption). \square

Theorem 24.4 (Properties of ϕ_X).

- (a) ϕ_X is convex, non-increasing.
- (b) ϕ_X continuous on (D_0, ∞) , where $D_0 = \inf\{D : \phi_X(D) < \infty\}$.
- (c) Suppose the distortion function satisfies $d(x, x) = D_0$ for all x and $d(x, y) > D_0$ for all $x \neq y$. Then $\phi_X(D_0) = I(X; X)$.
- (d) Let

$$D_{\max} = \inf_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E} d(X, \hat{x}).$$

Then $\phi_X(D) = 0$ for all $D > D_{\max}$. If $D_0 > D_{\max}$ then also $\phi_X(D_{\max}) = 0$.

Remark 24.4 (The role of D_0 and D_{\max}). By definition, D_{\max} is the distortion attainable without any information. Indeed, if $D_{\max} = \mathbb{E} d(X, \hat{x})$ for some fixed \hat{x} , then this \hat{x} is the “default” reconstruction of X , i.e., the best estimate when we have no information about X . Therefore $D \geq D_{\max}$ can be achieved for free. This is the reason for the notation D_{\max} despite that it is defined as an infimum. On the other hand, D_0 should be understood as the minimum distortion one can hope to attain. Indeed, suppose that $\hat{\mathcal{X}} = \mathcal{X}$ and d is a metric on \mathcal{X} . In this case, we have $D_0 = 0$, since we can choose Y to be a finitely-valued approximation of X .

24.3 Converse bounds 419

As an example, consider the Gaussian source with MSE distortion, namely, $X \sim \mathcal{N}(0, \sigma^2)$ and $d(x, \hat{x}) = (x - \hat{x})^2$. We will show later that $\phi_X(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}$. In this case $D_0 = 0$ which is however not attained; $D_{\max} = \sigma^2$ and if $D \geq \sigma^2$, we can simply output 0 as the reconstruction which requires zero bits.

Proof.

- (a) Convexity follows from the convexity of $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ (Theorem 5.5).
- (b) Continuity in the interior of the domain follows from convexity, since $D_0 = \inf_{P_{\hat{X}|X}} \mathbb{E}[d(X, \hat{X})] = \inf\{D : \phi_S(D) < \infty\}$.
- (c) The only way to satisfy the constraint is to take $X = Y$.
- (d) For any $D > D_{\max}$ we can set $\hat{X} = \hat{x}$ deterministically. Thus $I(X; \hat{x}) = 0$. The second claim follows from continuity. \square

In channel coding, the main result relates the Shannon capacity, an operational quantity, to the information capacity. Here we introduce the *information rate-distortion function* in an analogous way, which by itself is *not* an operational quantity.

Definition 24.5. The information rate-distortion function for a source $\{S_i\}$ is

$$R^{(I)}(D) = \limsup_{n \rightarrow \infty} \frac{1}{n} \phi_{S^n}(D), \quad \text{where } \phi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n} : \mathbb{E}[d(S^n, \hat{S}^n)] \leq D} I(S^n; \hat{S}^n).$$

The reason for defining $R^{(I)}(D)$ is because from Theorem 24.6 we immediately get:

Corollary 24.6. $\forall D, R(D) \geq R^{(I)}(D)$.

Naturally, the information rate-distortion function inherits the properties of ϕ from Theorem 24.7:

Theorem 24.7 (Properties of $R^{(I)}$).

- (a) $R^{(I)}(D)$ is convex, non-increasing
- (b) $R^{(I)}(D)$ is continuous on (D_0, ∞) , where $D_0 \triangleq \inf\{D : R^{(I)}(D) < \infty\}$.
- (c) Assume the same assumption on the distortion function as in Theorem 24.7(c). For stationary ergodic $\{S^n\}$, $R^{(I)}(D) = \mathcal{H}$ (entropy rate) or $+\infty$ if S_k is not discrete.
- (d) $R^{(I)}(D) = 0$ for all $D > D_{\max}$, where

$$D_{\max} \triangleq \limsup_{n \rightarrow \infty} \inf_{\hat{x}^n \in \hat{\mathcal{X}}} \mathbb{E} d(X^n, \hat{x}^n).$$

If $D_0 < D_{\max}$, then $R^{(I)}(D_{\max}) = 0$ too.

Proof. All properties follow directly from corresponding properties in Theorem 24.7 applied to ϕ_{S^n} . \square

Next we show that $R^{(I)}(D)$ can be easily calculated for stationary memoryless (iid) source without going through the multi-letter optimization problem. This parallels Corollary 20.1 for channel capacity (with separable cost function).

Theorem 24.8 (Single-letterization). *For stationary memoryless source $S_i \stackrel{i.i.d.}{\sim} P_S$ and separable distortion d in the sense of (24.10), we have for every n ,*

$$\phi_{S^n}(D) = n\phi_S(D).$$

Thus

$$R^{(I)}(D) = \phi_S(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} I(S; \hat{S}).$$

Proof. By definition we have that $\phi_{S^n}(D) \leq n\phi_S(D)$ by choosing a product channel: $P_{\hat{S}^n|S^n} = P_{\hat{S}|S}^{\otimes n}$. Thus $R^{(I)}(D) \leq \phi_S(D)$.

For the converse, for any $P_{\hat{S}^n|S^n}$ satisfying the constraint $\mathbb{E}[d(S^n, \hat{S}^n)] \leq D$, we have

$$\begin{aligned} I(S^n; \hat{S}^n) &\geq \sum_{j=1}^n I(S_j, \hat{S}_j) && (S^n \text{ independent}) \\ &\geq \sum_{j=1}^n \phi_S(\mathbb{E}[d(S_j, \hat{S}_j)]) \\ &\geq n\phi_S \left(\frac{1}{n} \sum_{j=1}^n \mathbb{E}[d(S_j, \hat{S}_j)] \right) && (\text{convexity of } \phi_S) \\ &\geq n\phi_S(D) && (\phi_S \text{ non-increasing}) \end{aligned}$$

□

24.4* Converting excess distortion to average

Finally, we discuss how to build a compressor for average distortion if we have one for excess distortion, the former of which is our focus.

Theorem 24.9 (Excess-to-Average). *Suppose that there exists (f, g) such that $W = f(X) \in [M]$ and $\mathbb{P}[d(X, g(W)) > D] \leq \epsilon$. Assume for some $p \geq 1$ and $\hat{x}_0 \in \hat{\mathcal{X}}$ that $(\mathbb{E}[d(X, \hat{x}_0)^p])^{1/p} = D_p < \infty$. Then there exists (f', g') such that $W' = f'(X) \in [M+1]$ and*

$$\mathbb{E}[d(X, g(W'))] \leq D(1 - \epsilon) + D_p \epsilon^{1-1/p}. \quad (24.13)$$

Remark 24.5. This result is only useful for $p > 1$, since for $p = 1$ the right-hand side of (24.13) does not converge to D as $\epsilon \rightarrow 0$. However, a different method (as we will see in the proof of

24.4* Converting excess distortion to average 421

Theorem 25.1) implies that under just $D_{\max} = D_1 < \infty$ the analog of the second term in (24.13) is vanishing as $\epsilon \rightarrow 0$, albeit at an unspecified rate.

Proof. We transform the first code into the second by adding one codeword:

$$f'(x) = \begin{cases} f(x) & d(x, g(f(x))) \leq D \\ M + 1 & \text{otherwise} \end{cases}$$

$$g'(j) = \begin{cases} g(j) & j \leq M \\ \hat{x}_0 & j = M + 1 \end{cases}$$

Then by Hölder's inequality,

$$\begin{aligned} \mathbb{E}[d(X, g'(W'))] &\leq \mathbb{E}[d(X, g(W)) | \hat{W} \neq M + 1](1 - \epsilon) + \mathbb{E}[d(X, \hat{x}_0) 1\{\hat{W} = M + 1\}] \\ &\leq D(1 - \epsilon) + D_p \epsilon^{1-1/p} \end{aligned}$$

□

25

Rate distortion: achievability bounds

Recall from the last chapter:

$$R(D) = \limsup_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, D), \quad (\text{rate-distortion function})$$

$$R^{(I)}(D) = \limsup_{n \rightarrow \infty} \frac{1}{n} \phi_{S^n}(D), \quad (\text{information rate-distortion function})$$

where

$$\phi_S(D) \triangleq \inf_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} I(S; \hat{S}) \quad (25.1)$$

$$\phi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}: \mathbb{E}[d(S^n, \hat{S}^n)] \leq D} I(S^n; \hat{S}^n) \quad (25.2)$$

and $d(S^n, \hat{S}^n) = \frac{1}{n} \sum_{i=1}^n d(S_i, \hat{S}_i)$ takes a separable form.

We have shown the following general converse in Theorem 24.6: For any $[M] \ni W \rightarrow X \rightarrow \hat{X}$ such that $\mathbb{E}[d(X, \hat{X})] \leq D$, we have $\log M \geq \phi_X(D)$, which implies in the special case of $X = S^n$, $\log M^*(n, D) \geq \phi_{S^n}(D)$ and hence, in the large- n limit, $R(D) \geq R^{(I)}(D)$. For a stationary memoryless source $S_i \stackrel{\text{i.i.d.}}{\sim} P_S$, Theorem 24.11 shows that ϕ_{S^n} single-letterizes as $\phi_{S^n}(D) = n\phi_S(D)$. As a result, we obtain the converse

$$R(D) \geq R^{(I)}(D) = \phi_S(D).$$

In this chapter, we will prove a matching achievability bound and establish the identity $R(D) = R^{(I)}(D)$ for stationary memoryless sources.

25.1 Shannon's rate-distortion theorem

The following result is proved by Shannon in his 1959 paper [270].

Theorem 25.1. Consider a stationary memoryless source $S^n \stackrel{\text{i.i.d.}}{\sim} P_S$. Suppose that the distortion metric d and the target distortion D satisfy:

- 1 $d(s^n, \hat{s}^n)$ is non-negative and separable.
- 2 $D > D_0$, where $D_0 = \inf\{D : \phi_S(D) < \infty\}$.

25.1 Shannon's rate-distortion theorem 423

3 D_{\max} is finite, i.e.

$$D_{\max} \triangleq \inf_{\hat{S}} \mathbb{E}[d(S, \hat{S})] < \infty. \quad (25.3)$$

Then

$$R(D) = R^{(I)}(D) = \phi_S(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} I(S; \hat{S}). \quad (25.4)$$

Remark 25.1.

- Note that $D_{\max} < \infty$ does not require that $d(\cdot, \cdot)$ only takes values in \mathbb{R} . That is, Theorem 25.1 permits $d(s, \hat{s}) = \infty$.
- When $D_{\max} = \infty$, typically we have $R(D) = \infty$ for all D . Indeed, suppose that $d(\cdot, \cdot)$ is a metric (i.e. real-valued and satisfies triangle inequality). Then, for any $x_0 \in \mathcal{A}^n$ we have

$$d(X, \hat{X}) \geq d(X, x_0) - d(x_0, \hat{X}).$$

Thus, for any finite codebook $\{c_1, \dots, c_M\}$ we have $\max_j d(x_0, c_j) < \infty$ and therefore

$$\mathbb{E}[d(X, \hat{X})] \geq \mathbb{E}[d(X, x_0)] - \max_j d(x_0, c_j) = \infty.$$

So that $R(D) = \infty$ for any finite D . This observation, however, should not be interpreted as the absolute impossibility of compressing such sources; it is just not possible with fixed-length codes. As an example, for quadratic distortion and Cauchy-distributed S , $D_{\max} = \infty$ since S has infinite second moment. But it is easy to see that¹ the *information* rate-distortion function $R^{(I)}(D) < \infty$ for any $D \in (0, \infty)$. In fact, in this case $R^{(I)}(D)$ is a hyperbola-like curve that never touches either axis. Using variable-length codes, S^n can be compressed non-trivially into W with bounded entropy (but unbounded cardinality) $H(W)$. An interesting question is: Is $H(W) = nR^{(I)}(D) + o(n)$ attainable?

- Techniques for proving (25.4) for memoryless sources can be extended to “stationary ergodic” sources with changes similar to those we have discussed in lossless compression (Chapter 12).

Before giving a formal proof, we give a heuristic derivation emphasizing the connection to large deviations estimates from Chapter 15.

25.1.1 Intuition

Let us throw M random points $\mathcal{C} = \{c_1, \dots, c_M\}$ into the space $\hat{\mathcal{A}}^n$ by generating them independently according to a product distribution $Q_{\hat{S}}^n$, where $Q_{\hat{S}}$ is some distribution on $\hat{\mathcal{A}}$ to be optimized. Consider the following simple coding strategy:

$$\text{Encoder : } f(s^n) = \operatorname{argmin}_{j \in [M]} d(s^n, c_j) \quad (25.5)$$

¹ Indeed, if we take W to be a quantized version of S with small quantization error D and notice that differential entropy of the Cauchy S is finite, we get from (24.7) that $R^{(I)}(D) \leq H(W) < \infty$.

$$\text{Decoder : } g(j) = c_j \quad (25.6)$$

The basic idea is the following: Since the codewords are generated independently of the source, the probability that a given codeword is close to the source realization is (exponentially) small, say, ϵ . However, since we have many codewords, the chance that there exists a good one can be of high probability. More precisely, the probability that no good codewords exist is approximately $(1-\epsilon)^M$, which can be made close to zero provided $M \gg \frac{1}{\epsilon}$.

To explain this intuition further, consider a discrete memoryless source $S^n \stackrel{\text{i.i.d.}}{\sim} P_S$ and let us evaluate the excess distortion of this random code: $\mathbb{P}[d(S^n, f(S^n)) > D]$, where the probability is over all random codewords c_1, \dots, c_M and the source S^n . Define

$$P_{\text{failure}} \triangleq \mathbb{P}[\forall c \in \mathcal{C}, d(S^n, c) > D] = \mathbb{E}_{S^n} [\mathbb{P}[d(S^n, c_1) > D | S^n]^M],$$

where the last equality follows from the assumption that c_1, \dots, c_M are iid and independent of S^n . To simplify notation, let $\hat{S}^n \stackrel{\text{i.i.d.}}{\sim} Q_{\hat{S}}^n$ independently of S^n , so that $P_{S^n, \hat{S}^n} = P_S^n Q_{\hat{S}}^n$. Then

$$\mathbb{P}[d(S^n, c_1) > D | S^n] = \mathbb{P}[d(S^n, \hat{S}^n) > D | S^n]. \quad (25.7)$$

To evaluate the failure probability, let us consider the special case of $P_S = \text{Ber}(\frac{1}{2})$ and also choose $Q_{\hat{S}} = \text{Ber}(\frac{1}{2})$ to generate the random codewords, aiming to achieve a normalized Hamming distortion at most $D < \frac{1}{2}$. Since $nd(S^n, \hat{S}^n) = \sum_{i:s_i=1} (1 - \hat{S}_i) + \sum_{i:s_i=0} \hat{S}_i \sim \text{Bin}(n, \frac{1}{2})$ for any s^n , the conditional probability (25.7) does not depend on S^n and is given by

$$\mathbb{P}[d(S^n, \hat{S}^n) > D | S^n] = \mathbb{P}\left[\text{Bin}\left(n, \frac{1}{2}\right) \geq nD\right] \approx 1 - 2^{-n(1-h(D))+o(n)}, \quad (25.8)$$

where in the last step we applied large-deviation estimates from Theorem 15.10 and Example 15.1. (Note that here we actually need lower estimates on these exponentially small probabilities.) Thus, $P_{\text{failure}} = (1 - 2^{-n(1-h(D))+o(n)})^M$, which vanishes if $M = 2^{n(1-h(D)+\delta)}$ for any $\delta > 0$.² As we will compute in Theorem 26.1, the rate-distortion function for $P_S = \text{Ber}(\frac{1}{2})$ is precisely $\phi_S(D) = 1 - h(D)$, so we have a rigorous proof of the optimal achievability in this special case.

For general distribution P_S (or even for $P_S = \text{Ber}(p)$ for which it is suboptimal to choose $Q_{\hat{S}}$ as $\text{Ber}(\frac{1}{2})$), the situation is more complicated as the conditional probability (25.7) depends on the source realization S^n through its empirical distribution (type). Let T_n be the set of typical realizations whose empirical distribution is close to P_S . We have

$$\begin{aligned} P_{\text{failure}} &\approx \mathbb{P}[d(S^n, \hat{S}^n) > D | S^n \in T_n]^M \\ &= (1 - \underbrace{\mathbb{P}[d(S^n, \hat{S}^n) \leq D | S^n \in T_n]}_{\approx 0, \text{ since } S^n \perp\!\!\!\perp \hat{S}^n})^M \\ &\approx (1 - 2^{-nE(Q_{\hat{S}})})^M, \end{aligned} \quad (25.9)$$

² In fact, this argument shows that $M = 2^{n(1-h(D))+o(n)}$ codewords suffice to cover the *entire* Hamming space within distance Dn . See (27.9) and Exercise V.11.

25.1 Shannon's rate-distortion theorem 425

where it can be shown (using large deviations analysis similar to information projection in Chapter 15) that

$$E(Q_{\hat{S}}) = \min_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \quad (25.10)$$

Thus we conclude that for any choice of $Q_{\hat{S}}$ (from which the random codewords were drawn) and any $\delta > 0$, the above code with $M = 2^{n(E(Q_{\hat{S}}) + \delta)}$ achieves vanishing excess distortion

$$P_{\text{failure}} = \mathbb{P}[\forall c \in \mathcal{C}, d(S^n, c) > D] \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Finally, we optimize $Q_{\hat{S}}$ to get the smallest possible M :

$$\begin{aligned} \min_{Q_{\hat{S}}} E(Q_{\hat{S}}) &= \min_{Q_{\hat{S}}} \min_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \\ &= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} \min_{Q_{\hat{S}}} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \\ &= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} I(S; \hat{S}) \\ &= \phi_S(D) \end{aligned}$$

where the third equality follows from the variational representation of mutual information (Corollary 4.1). This heuristic derivation explains how the constrained mutual information minimization arises. Below we make it rigorous using a different approach, again via random coding.

25.1.2 Proof of Theorem 25.1

Theorem 25.2 (Random coding bound of average distortion). *Fix P_X and suppose $d(x, \hat{x}) \geq 0$ for all x, \hat{x} . For any $P_{Y|X}$ and any $y_0 \in \hat{\mathcal{A}}$, there exists a code $X \rightarrow W \rightarrow \hat{X}$ with $W \in [M+1]$, such $d(X, \hat{X}) \leq d(X, y_0)$ almost surely and for any $\gamma > 0$,*

$$\mathbb{E}[d(X, \hat{X})] \leq \mathbb{E}[d(X, Y)] + \mathbb{E}[d(X, y_0)]e^{-M/\gamma} + \mathbb{E}[d(X, y_0)1_{\{i(X;Y) > \log \gamma\}}].$$

Here the first and the third expectations are over $(X, Y) \sim P_{X,Y} = P_X P_{Y|X}$ and the information density $i(\cdot; \cdot)$ is defined with respect to this joint distribution (cf. Definition 18.1).

Some remarks are in order:

- Theorem 25.3 says that from an arbitrary $P_{Y|X}$ such that $\mathbb{E}[d(X, Y)] \leq D$, we can extract a good code with average distortion D plus some extra terms which will vanish in the asymptotic regime for memoryless sources.
- The proof uses the random coding argument with codewords drawn independently from P_Y , the marginal distribution induced by the source distribution P_X and the auxiliary channel $P_{Y|X}$. As such, $P_{Y|X}$ plays no role in the code construction and is used only in analysis (by defining a coupling between P_X and P_Y).

- The role of the deterministic y_0 is a “fail-safe” codeword (think of y_0 as the default reconstruction with $D_{\max} = \mathbb{E}[d(X, y_0)]$). We add y_0 to the random codebook for “damage control”, to hedge against the (highly unlikely) event that we end up with a terrible codebook.

Proof. Similar to the intuitive argument sketched in Section 25.1.1, we apply random coding and generate the codewords randomly and independently of the source:

$$\mathcal{C} = \{c_1, \dots, c_M\} \stackrel{\text{i.i.d.}}{\sim} P_Y \perp\!\!\!\perp X$$

and add the “fail-safe” codeword $c_{M+1} = y_0$. We adopt the same encoder-decoder pair (25.5) – (25.6) and let $\hat{X} = g(f(X))$. Then by definition,

$$d(X, \hat{X}) = \min_{j \in [M+1]} d(X, c_j) \leq d(X, y_0).$$

To simplify notation, let \bar{Y} be an independent copy of Y (similar to the idea of introducing unsent codeword \bar{X} in channel coding – see Chapter 18):

$$P_{X,Y,\bar{Y}} = P_{X,Y}P_{\bar{Y}}$$

where $P_{\bar{Y}} = P_Y$. Recall the formula for computing the expectation of a random variable $U \in [0, a]$: $\mathbb{E}[U] = \int_0^a \mathbb{P}[U \geq u] du$. Then the average distortion is

$$\begin{aligned} \mathbb{E}d(X, \hat{X}) &= \mathbb{E} \min_{j \in [M+1]} d(X, c_j) \\ &= \mathbb{E}_X \mathbb{E} \left[\min_{j \in [M+1]} d(X, c_j) \middle| X \right] \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P} \left[\min_{j \in [M+1]} d(X, c_j) > u \middle| X \right] du \\ &\leq \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P} \left[\min_{j \in [M]} d(X, c_j) > u \middle| X \right] du \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P}[d(X, \bar{Y}) > u | X]^M du \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} (1 - \mathbb{P}[d(X, \bar{Y}) \leq u | X])^M du \\ &\leq \mathbb{E}_X \int_0^{d(X,y_0)} \underbrace{(1 - \mathbb{P}[d(X, \bar{Y}) \leq u, i(X, \bar{Y}) > -\infty | X])^M}_{\triangleq \delta(X,u)} du. \end{aligned} \quad (25.11)$$

Next we upper bound $(1 - \delta(X, u))^M$ as follows:

$$(1 - \delta(X, u))^M \leq e^{-M/\gamma} + |1 - \gamma\delta(X, u)|^+ \quad (25.12)$$

$$= e^{-M/\gamma} + |1 - \gamma\mathbb{E}[\exp\{-i(X; Y)\} 1_{\{d(X,Y) \leq u\}} | X]|^+ \quad (25.13)$$

$$\leq e^{-M/\gamma} + \mathbb{P}[i(X; Y) > \log \gamma | X] + \mathbb{P}[d(X, Y) > u | X] \quad (25.14)$$

where

25.1 Shannon's rate-distortion theorem 427

- (25.12) uses the following trick in dealing with $(1 - \delta)^M$ for $\delta \ll 1$ and $M \gg 1$. First, recall the standard rule of thumb:

$$(1 - \delta)^M \approx \begin{cases} 0, & \delta M \gg 1 \\ 1, & \delta M \ll 1 \end{cases}$$

In order to obtain firm bounds of a similar flavor, we apply, for any $\gamma > 0$,

$$(1 - \delta)^M \leq e^{-\delta M} \leq e^{-M/\gamma} + (1 - \gamma\delta)_+.$$

- (25.13) is simply a change of measure argument of Proposition 18.3. Namely we apply (18.5) with $f(x, y) = 1_{\{d(x, y) \leq u\}}$.
- For (25.14) consider the chain:

$$\begin{aligned} 1 - \gamma \mathbb{E}[\exp\{-i(X; Y)\} 1_{\{d(X, Y) \leq u\}} | X] &\leq 1 - \gamma \mathbb{E}[\exp\{-i(X; Y)\} 1_{\{d(X, Y) \leq u, i(X; Y) \leq \log \gamma\}} | X] \\ &\leq 1 - \mathbb{E}[1_{\{d(X, Y) \leq u, i(X; Y) \leq \log \gamma\}} | X] \\ &= \mathbb{P}[d(X, Y) > u \text{ or } i(X; Y) > \log \gamma | X] \\ &\leq \mathbb{P}[d(X, Y) > u | X] + \mathbb{P}[i(X; Y) > \log \gamma | X] \end{aligned}$$

Plugging (25.14) into (25.11), we have

$$\begin{aligned} \mathbb{E}[d(X, \hat{X})] &\leq \mathbb{E}_X \left[\int_0^{d(X, y_0)} (e^{-M/\gamma} + \mathbb{P}[i(X; Y) > \log \gamma | X] + \mathbb{P}[d(X, Y) > u | X]) du \right] \\ &\leq \mathbb{E}[d(X, y_0)] e^{-M/\gamma} + \mathbb{E}[d(X, y_0) \mathbb{P}[i(X; Y) > \log \gamma | X]] + \mathbb{E}_X \int_0^\infty \mathbb{P}[d(X, Y) > u | X] du \\ &= \mathbb{E}[d(X, y_0)] e^{-M/\gamma} + \mathbb{E}[d(X, y_0) 1_{\{i(X; Y) > \log \gamma\}}] + \mathbb{E}[d(X, Y)]. \end{aligned}$$

□

As a side product, we have the following achievability result for excess distortion.

Theorem 25.3 (Random coding bound of excess distortion). *For any $P_{Y|X}$, there exists a code $X \rightarrow W \rightarrow \hat{X}$ with $W \in [M]$, such that for any $\gamma > 0$,*

$$\mathbb{P}[d(X, \hat{X}) > D] \leq e^{-M/\gamma} + \mathbb{P}[\{d(X, Y) > D\} \cup \{i(X; Y) > \log \gamma\}]$$

Proof. Proceed exactly as in the proof of Theorem 25.3 (without using the extra codeword y_0), replace (25.11) by $\mathbb{P}[d(X, \hat{X}) > D] = \mathbb{P}[\forall j \in [M], d(X, c_j) > D] = \mathbb{E}_X[(1 - \mathbb{P}[d(X, \bar{Y}) \leq D | X])^M]$, and continue similarly. □

Finally, we give a rigorous proof of Theorem 25.1 by applying Theorem 25.3 to the iid source $X = S^n \stackrel{\text{i.i.d.}}{\sim} P_S$ and $n \rightarrow \infty$:

Proof of Theorem 25.1. Our goal is the achievability: $R(D) \leq R^{(I)}(D) = \phi_S(D)$.

WLOG we can assume that $D_{\max} = \mathbb{E}[d(S, \hat{s}_0)]$ is achieved at some fixed \hat{s}_0 – this is our default reconstruction; otherwise just take any other fixed symbol so that the expectation is finite. The

default reconstruction for S^n is $\hat{s}_0^n = (\hat{s}_0, \dots, \hat{s}_0)$ and $\mathbb{E}[d(S^n, \hat{s}_0^n)] = D_{\max} < \infty$ since the distortion is separable.

Fix some small $\delta > 0$. Take any $P_{\hat{S}|S}$ such that $\mathbb{E}[d(S, \hat{S})] \leq D - \delta$; such $P_{\hat{S}|S}$ since $D > D_0$ by assumption. Apply Theorem 25.3 to $(X, Y) = (S^n, \hat{S}^n)$ with

$$\begin{aligned} P_X &= P_{S^n} \\ P_{Y|X} &= P_{\hat{S}^n|S^n} = (P_{\hat{S}|S})^n \\ \log M &= n(I(S; \hat{S}) + 2\delta) \\ \log \gamma &= n(I(S; \hat{S}) + \delta) \\ d(X, Y) &= \frac{1}{n} \sum_{j=1}^n d(S_j, \hat{S}_j) \\ y_0 &= \hat{s}_0^n \end{aligned}$$

we conclude that there exists a compressor $f: \mathcal{A}^n \rightarrow [M+1]$ and $g: [M+1] \rightarrow \hat{\mathcal{A}}^n$, such that

$$\begin{aligned} \mathbb{E}[d(S^n, g(f(S^n)))] &\leq \mathbb{E}[d(S^n, \hat{S}^n)] + \mathbb{E}[d(S^n, \hat{s}_0^n)] e^{-M/\gamma} + \mathbb{E}[d(S^n, \hat{s}_0^n) \mathbf{1}_{\{i(S^n, \hat{S}^n) > \log \gamma\}}] \\ &\leq D - \delta + \underbrace{D_{\max} e^{-\exp(n\delta)}}_{\rightarrow 0} + \underbrace{\mathbb{E}[d(S^n, \hat{s}_0^n) \mathbf{1}_{E_n}]}_{\rightarrow 0 \text{ (later)}}, \end{aligned} \quad (25.15)$$

where

$$E_n = \{i(S^n; \hat{S}^n) > \log \gamma\} = \left\{ \frac{1}{n} \sum_{j=1}^n i(S_j; \hat{S}_j) > I(S; \hat{S}) + \delta \right\} \xrightarrow{\text{WLLN}} \mathbb{P}[E_n] \rightarrow 0$$

If we can show the expectation in (25.15) vanishes, then there exists an (n, M, \bar{D}) -code with:

$$M = 2^{n(I(S; \hat{S}) + 2\delta)}, \quad \bar{D} = D - \delta + o(1) \leq D.$$

To summarize, $\forall P_{\hat{S}|S}$ such that $\mathbb{E}[d(S, \hat{S})] \leq D - \delta$ we have shown that $R(D) \leq I(S; \hat{S})$. Sending $\delta \downarrow 0$, we have, by continuity of $\phi_S(D)$ in (D_0, ∞) (recall Theorem 24.7), $R(D) \leq \phi_S(D-) = \phi_S(D)$.

It remains to show the expectation in (25.15) vanishes. This is a simple consequence of the uniform integrability of the sequence $\{d(S^n, \hat{s}_0^n)\}$. We need the following lemma.

Lemma 25.4. *For any positive random variable U , define $g(\delta) = \sup_{H: \mathbb{P}[H] \leq \delta} \mathbb{E}[U \mathbf{1}_H]$, where the supremum is over all events measurable with respect to U . Then³ $\mathbb{E}U < \infty \Rightarrow g(\delta) \xrightarrow{\delta \rightarrow 0} 0$.*

Proof. For any $b > 0$, $\mathbb{E}[U \mathbf{1}_H] \leq \mathbb{E}[U \mathbf{1}_{\{U > b\}}] + b\delta$, where $\mathbb{E}[U \mathbf{1}_{\{U > b\}}] \xrightarrow{b \rightarrow \infty} 0$ by dominated convergence theorem. Then the proof is completed by setting $b = 1/\sqrt{\delta}$. \square

Now $d(S^n, \hat{s}_0^n) = \frac{1}{n} \sum_{j=1}^n U_j$, where U_j are iid copies of $U \triangleq d(S, \hat{s}_0)$. Since $\mathbb{E}[U] = D_{\max} < \infty$ by assumption, applying Lemma 25.5 yields $\mathbb{E}[d(S^n, \hat{s}_0^n) \mathbf{1}_{E_n}] = \frac{1}{n} \sum \mathbb{E}[U_j \mathbf{1}_{E_n}] \leq g(\mathbb{P}[E_n]) \rightarrow 0$, since $\mathbb{P}[E_n] \rightarrow 0$. This proves the theorem. \square

³ In fact, \Rightarrow is \Leftrightarrow .

25.2* Covering lemma 429

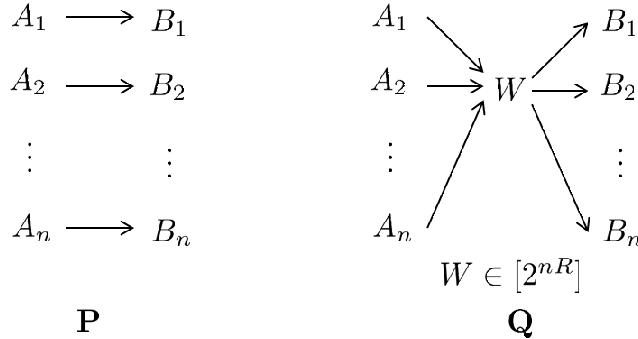


Figure 25.1 Description of channel simulation game. The distribution P (left) is to be simulated via the distribution Q (right) at minimal rate R . Depending on the exact formulation we either require $R = I(A; B)$ (covering lemma) or $R = C(A; B)$ (soft-covering lemma).

Remark 25.2 (Fundamental limit for excess distortion). Although Theorem 25.1 is stated for the average distortion, under certain mild extra conditions, it also holds for excess distortion where the goal is to achieve $d(S^n, \hat{S}^n) \leq D$ with probability arbitrarily close to one as opposed to in expectation. Indeed, the achievability proof of Theorem 25.1 is already stated in high probability. For converse, assume in addition to (25.3) that $D_p \triangleq \mathbb{E}[d(S, \hat{s})^p]^{1/p} < \infty$ for some $\hat{s} \in \hat{S}$ and $p > 1$. Applying Rosenthal's inequality [261, 166], we have $\mathbb{E}[d(S, \hat{s}^n)^p] = \mathbb{E}[(\sum_{i=1}^n d(S_i, \hat{s}_i))^p] \leq CD_p^p$ for some constant $C = C(p)$. Then we can apply Theorem 24.12 to convert a code for excess distortion to one for average distortion and invoke the converse for the latter.

To end this section, we note that in Section 25.1.1 and in Theorem 25.1 it seems we applied different proof techniques. How come they both turn out to yield the same *tight* asymptotic result? This is because the key to both proofs is to estimate the exponent (large deviations) of the underlined probabilities in (25.9) and (25.11), respectively. To obtain the right exponent, as we know, the key is to apply tilting (change of measure) to the distribution solving the information projection problem (25.10). When $P_{\bar{Y}} = (Q_{\hat{S}})^n = (P_{\hat{S}})^n$ with $P_{\hat{S}}$ chosen as the output distribution in the solution to rate-distortion optimization (25.1), the resulting exponent is precisely given by $2^{-i(X; Y)}$.

25.2* Covering lemma

In this section we consider the following curious problem, a version of *channel simulation/synthesis*. We want to simulate a sequence of iid correlated strings $(A_i, B_i) \stackrel{\text{i.i.d.}}{\sim} P_{A,B}$ following the following protocol. First, an sequence $A^n \stackrel{\text{i.i.d.}}{\sim} P_A$ is generated at one terminal. Then we can look at it, produces a short message $W \in [2^{nR}]$ which gets communicated to a remote destination (noiselessly). Upon receiver the message, remote decoder produces a string B^n out of it. The goal is to be able to fool the *tester* who inspects (A^n, B^n) and tries to check that it was indeed generated as $(A_i, B_i) \stackrel{\text{i.i.d.}}{\sim} P_{A,B}$. See figure Fig. 25.1 for an illustration.

How large a rate R is required depends on how we exactly understand the requirement to “fool the tester”. If the tester is fixed ahead of time (this just means that we know the set F such that $(A_i, B_i) \stackrel{i.i.d.}{\sim} P_{A,B}$ is declared whenever $(A^n, B^n) \in F$) then this is precisely the setting in which *covering lemma* operates. In the next section we show that a higher rate $R = C(A; B)$ is required if F is not known ahead of time. We leave out the celebrated theorem of Bennett and Shor [27] which shows that rate $R = I(A; B)$ is also attainable even if F is not known, but if encoder and decoder are given access to a source of common random bits (independent of A^n , of course).

Before proceeding, we note some simple corner cases:

- 1 If $R = H(A)$, we can compress A^n and send it to “B side”, who can reconstruct A^n perfectly and use that information to produce B^n through $P_{B^n|A^n}$.
- 2 If $R = H(B)$, “A side” can generate B^n according to $P_{A,B}^n$ and send that B^n sequence to the “B side”.
- 3 If $A \perp\!\!\!\perp B$, we know that $R = 0$, as “B side” can generate B^n independently.

Our previous argument for achieving the rate-distortion turns out to give a sharp answer (that $R = I(A; B)$ is sufficient) for the F -known case as follows.

Theorem 25.5 (Covering Lemma). *Fix $P_{A,B}$ and let $(A_j, B_j) \stackrel{i.i.d.}{\sim} P_{A,B}$, $R > I(A; B)$ and $\mathcal{C} = \{c_1, \dots, c_M\}$ where each codeword c_j is i.i.d. drawn from distribution P_B^n . $\forall \epsilon > 0$, for $M \geq 2^{n(I(A;B)+\epsilon)}$ we have that: $\forall F$*

$$\mathbb{P}[\exists c : (A^n, c) \in F] \geq \mathbb{P}[(A^n, B^n) \in F] + \underbrace{o(1)}_{\text{uniform in } F}$$

Remark 25.3. The origin of the name “covering” is due to the fact that sampling the \mathcal{A}^n space at rate slightly above $I(A; B)$ covers all of it, in the sense of reproducing the joint statistics of (A^n, B^n) .

Proof. Set $\gamma > M$ and following similar arguments of the proof for Theorem 25.3, we have

$$\begin{aligned} \mathbb{P}[\forall c \in \mathcal{C} : (A^n, c) \notin F] &\leq e^{-M/\gamma} + \mathbb{P}[\{(A^n, B^n) \notin F\} \cup \{i(A^n; B^n) > \log \gamma\}] \\ &= \mathbb{P}[(A^n, B^n) \notin F] + o(1) \\ \Rightarrow \mathbb{P}[\exists c \in \mathcal{C} : (A^n, c) \in F] &\geq \mathbb{P}[(A^n, B^n) \in F] + o(1) \end{aligned}$$

□

As we explained, the version of the covering lemma that we stated shows only that for one fixed test set F . However, if both A and B take values on finite alphabets then something stronger can be stated.

First, in this case $i(A^n; B^n)$ is a sum of bounded iid terms and thus the $o(1)$ is in fact $e^{-\Omega(n)}$. By applying the previous result to $F = \{(a^n, b^n) : \#\{i : a_i = \alpha, b_i = \beta\}\}$ with all possible $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$ we conclude that for every A^n there must exist a codeword c such that the empirical joint distribution (joint type) $\hat{P}_{A^n, c}$ satisfies

$$\mathbb{P}[\exists c \in \mathcal{C} \text{ such that } \text{TV}(\hat{P}_{A^n, c}, P_{A,B}) \leq \delta_n] \rightarrow 1,$$

25.3* Wyner's common information 431

where $\delta_n \rightarrow 0$. Thus, by communicating nR bits we are able to fully reproduce the correct empirical distribution as if the output were generated $\stackrel{\text{i.i.d.}}{\sim} P_{A,B}$.

That this is possible to do at rate $R \approx I(A;B)$ can be explained combinatorially: To generate B^n , there are around $2^{nH(B)}$ high probability sequences; for each A^n sequence, there are around $2^{nH(B|A)}$ B^n sequences that have the same joint distribution, therefore, it is sufficient to describe the class of B^n for each A^n sequence, and there are around $\frac{2^{nH(B)}}{2^{nH(B|A)}} = 2^{nI(A;B)}$ classes.

Let us now denote the selected codeword c by \hat{B}^n . From the previous discussion we have shown that

$$\frac{1}{n} \sum_{j=1}^n f(A_j, \hat{B}_j) \approx \mathbb{E}_{A,B \sim P_{A,B}} [f(A, B)],$$

for any bounded function f . A stronger requirement would be to demand that the joint distribution P_{A^n, \hat{B}^n} fools any permutation invariant tester, i.e.

$$\sup |P_{A^n, \hat{B}^n}(F) - P_{A,B}^n(F)| \rightarrow 0$$

where the supremum is taken over all permutation invariant subset $F \subset \mathcal{A}^n \times \mathcal{B}^n$. This is not guaranteed by the covering lemma. Indeed, a sufficient statistic for a permutation invariant tester is a joint type \hat{P}_{A^n, \hat{B}^n} . The construction above satisfies $\hat{P}_{A^n, \hat{B}^n} \approx P_{A,B}$, but it might happen that \hat{P}_{A^n, \hat{B}^n} although close to $P_{A,B}$ still takes highly unlikely values (for example, if we restrict all c to have the same composition P_0 , the tester can easily detect the problem since P_B^n -measure of all strings of composition P_0 cannot exceed $O(1/\sqrt{n})$). Formally, to fool permutation invariant tester we need to have small total variation between the distribution on the joint types under P and Q .

We conjecture, however, that nevertheless the rate $R = I(A;B)$ should be sufficient to achieve also this stronger requirement. In the next section we show that if one removes the permutation-invariance constraint, then a larger rate $R = C(A;B)$ is needed.

25.3* Wyner's common information

We continue discussing the channel simulation setting as in previous section. We now want to determine the minimal possible communication rate (i.e. cardinality of $W \in [2^{nR}]$) required to have small total variation:

$$\text{TV}(P_{A^n, \hat{B}^n}, P_{A,B}^n) \leq \epsilon \tag{25.16}$$

between the simulated and the true output (see Fig. 25.1).

Theorem 25.6 (Cuff [84]). *Let $P_{A,B}$ be an arbitrary distribution on the finite space $\mathcal{A} \times \mathcal{B}$. Consider a coding scheme where Alice observes $A^n \stackrel{\text{i.i.d.}}{\sim} P_A^n$, sends a message $W \in [2^{nR}]$ to Bob, who given W generates a (possibly random) sequence \hat{B}^n . If (25.16) is satisfied for all $\epsilon > 0$ and sufficiently large n , then we must have*

$$R \geq C(A;B) \triangleq \min_{A \rightarrow U \rightarrow B} I(A, B; U), \tag{25.17}$$

where $C(A; B)$ is known as the Wyner's common information [324]. Furthermore, for any $R > C(A; B)$ and $\epsilon > 0$ there exists $n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$ there exists a scheme satisfying (25.16).

Note that condition (25.16) guarantees that any tester (permutation invariant or not) is fooled to believe he sees the truly iid (A^n, B^n) with probability $\geq 1 - \epsilon$. However, compared to Theorem 25.7, this requires a higher communication rate since $C(A; B) \geq I(A; B)$, clearly.

Proof. Showing that Wyner's common information is a lower-bound is not hard. First, since $P_{A^n, \hat{B}^n} \approx P_{A, B}^n$ (in TV) we have

$$I(A_t, \hat{B}_t; A^{t-1}, \hat{B}^{t-1}) \approx I(A_t, B_t; A^{t-1}, B^{t-1}) = 0$$

(Here one needs to use finiteness of the alphabet of A and B and the bounds relating $H(P) - H(Q)$ with $\text{TV}(P, Q)$, cf. (7.18) and Corollary 6.2). Next, we have

$$nR = H(W) \geq I(A^n, \hat{B}^n; W) \quad (25.18)$$

$$\geq \sum_{t=1}^n I(A_t, \hat{B}_t; W) - I(A_t, \hat{B}_t; A^{t-1}, \hat{B}^{t-1}) \quad (25.19)$$

$$\approx \sum_{t=1}^n I(A_t, \hat{B}_t; W) \quad (25.20)$$

$$\gtrsim nC(A; B) \quad (25.21)$$

where in the last step we used the crucial observation that

$$A_t \rightarrow W \rightarrow \hat{B}_t$$

and that Wyner's common information $P_{A, B} \mapsto C(A; B)$ should be continuous in the total variation distance on $P_{A, B}$.

To show achievability, let us notice that the problem is equivalent to constructing three random variables $(\hat{A}^n, W, \hat{B}^n)$ such that a) $W \in [2^{nR}]$, b) the Markov relation

$$\hat{A}^n \leftarrow W \rightarrow \hat{B}^n \quad (25.22)$$

holds and c) $\text{TV}(P_{\hat{A}^n, \hat{B}^n}, P_{A, B}^n) \leq \epsilon/2$. Indeed, given such a triple we can use coupling characterization of TV (7.18) and the fact that $\text{TV}(P_{\hat{A}^n}, P_A^n) \leq \epsilon/2$ to extend the probability space to

$$A^n \rightarrow \hat{A}^n \rightarrow W \rightarrow \hat{B}^n$$

and $\mathbb{P}[A^n = \hat{A}^n] \geq 1 - \epsilon/2$. Again by (7.18) we conclude that $\text{TV}(P_{A^n, \hat{B}^n}, P_{\hat{A}^n, \hat{B}^n}) \leq \epsilon/2$ and by triangle inequality we conclude that (25.16) holds.

Finally, construction of the triple satisfying a)-c) follows from the soft-covering lemma (Corollary 25.1) applied with $V = (A, B)$ and W being uniform on the set of x_i 's there. \square

25.4* Approximation of output statistics and the soft-covering lemma

In this section we aim to prove the remaining ingredient (the soft-covering lemma) required for the proof of Theorem 25.9. To that end, recall that in Section 7.9 we have shown that generating iid samples $X_i \stackrel{\text{i.i.d.}}{\sim} P_X$ and passing their empirical distribution \hat{P}_n across the channel $P_{Y|X}$ results in a good approximation of $P_Y = P_{Y|X} \circ P_X$, i.e.

$$D(P_{Y|X} \circ \hat{P}_n \| P_Y) \rightarrow 0.$$

A natural question is how large n should be in order for the approximation $P_{Y|X} \circ \hat{P}_n \approx P_Y$ to hold. A remarkable fact that we establish in this section is that the answer is $n \approx 2^{I(X;Y)}$, assuming $I(X; Y) \gg 1$ and there is certain concentration properties of $i(X; Y)$ around $I(X; Y)$. This fact originated from Wyner [324] and was significantly strengthened in [150].

Here, we show a new variation of such results by strengthening our simple χ^2 -information bound of Proposition 7.23.

Theorem 25.7. Fix $P_{X,Y}$ and for any $\lambda \in \mathbb{R}$ let us define Rényi mutual information

$$I_\lambda(X; Y) = D_\lambda(P_{X,Y} \| P_X P_Y),$$

where D_λ is the Rényi-divergence, cf. Definition 7.34. We have for every $1 < \lambda \leq 2$

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \leq \frac{1}{\lambda - 1} \log(1 + \exp\{(\lambda - 1)(I_\lambda(X; Y) - \log n)\}). \quad (25.23)$$

Proof. Since $\lambda \rightarrow D_\lambda$ is non-decreasing, it is sufficient to prove an equivalent upper bound on $\mathbb{E}[D_\lambda(P_{Y|X} \circ \hat{P}_n \| P_Y)]$. From Jensen's inequality we see that

$$\begin{aligned} \mathbb{E}[D_\lambda(P_{Y|X} \circ \hat{P}_n \| P_Y)] &\triangleq \frac{1}{\lambda - 1} \mathbb{E}_{X^n} \log \mathbb{E}_{Y \sim P_Y} \left[\left\{ \frac{P_{Y|X} \circ \hat{P}_n}{P_Y} \right\}^\lambda \right] \\ &\leq \frac{1}{\lambda - 1} \log \mathbb{E}_{X^n} \mathbb{E}_{Y \sim P_Y} \left[\left\{ \frac{P_{Y|X} \circ \hat{P}_n}{P_Y} \right\}^\lambda \right] \triangleq I_\lambda(X^n; \bar{Y}), \end{aligned} \quad (25.24)$$

where similarly to (7.50) we introduced the channel $P_{\bar{Y}|X^n} = \frac{1}{n} \sum_{i=1}^n P_{Y|X=X_i}$. To analyze $I_\lambda(X^n; \bar{Y})$ we need to bound

$$\mathbb{E}_{(X^n, \bar{Y}) \sim P_X^n \times P_Y} \left[\left\{ \frac{1}{n} \sum_i \frac{P_{Y|X}(Y|X_i)}{P_Y(Y)} \right\}^\lambda \right]. \quad (25.25)$$

Note that conditioned on Y we get to analyze a λ -th moment of a sum of iid random variables. This puts us into a well-known setting of Rosenthal-type inequalities. In particular, we have that for any iid non-negative B_j we have, provided $1 \leq \lambda \leq 2$, that

$$\mathbb{E} \left[\left(\sum_{i=1}^n B_i \right)^\lambda \right] \leq n \mathbb{E}[B^\lambda] + (n \mathbb{E}[B])^\lambda. \quad (25.26)$$

This is known to be essentially tight [264]. It can be proven by applying $(a+b)^{\lambda-1} \leq a^{\lambda-1} + b^{\lambda-1}$ and Jensen's to get

$$\mathbb{E} B_i (B_i + \sum_{j \neq i} B_j)^{\lambda-1} \leq \mathbb{E}[B^\lambda] + \mathbb{E}[B]((n-1)\mathbb{E}[B])^{\lambda-1}.$$

Summing the latter over i and bounding $(n-1) \leq n$ we get (25.26).

Now using (25.26) we can overbound (25.25) as

$$\leq 1 + n^{1-\lambda} \mathbb{E}_{(X, \bar{Y}) \sim P_X \times P_Y} \left[\left\{ \frac{P_{Y|X}(Y|X_i)}{P_Y(Y)} \right\}^\lambda \right],$$

which implies

$$I_\lambda(X^n; \bar{Y}) \leq \frac{1}{\lambda-1} \log (1 + n^{1-\lambda} \exp\{(\lambda-1)I_\lambda(X; Y)\}),$$

which together with (25.24) recovers the main result (25.23). \square

Remark 25.4. Hayashi [154] upper bounds the LHS of (25.23) with

$$\frac{\lambda}{\lambda-1} \log(1 + \exp\{\frac{\lambda-1}{\lambda}(K_\lambda(X; Y) - \log n)\}),$$

where $K_\lambda(X; Y) = \inf_{Q_Y} D_\lambda(P_{X,Y} \| P_X Q_Y)$ is the so-called Sibson-Csiszár information, cf. [236]. This bound, however, does not have the right rate of convergence as $n \rightarrow \infty$, at least for $\lambda = 1$ as comparison with Proposition 7.23 reveals.

We note that [154, 150] also contain direct bounds on

$$\mathbb{E}[\text{TV}(P_{Y|X} \circ \hat{P}_n, P_Y)]$$

which do not assume existence of λ -th moment of $\frac{P_{Y|X}}{P_Y}$ for $\lambda > 1$ and instead rely on the distribution of $i(X; Y)$. We do not discuss these bounds here, however, since for the purpose of discussing finite alphabets the next corollary is sufficient.

Corollary 25.8 (Soft-covering lemma). *Suppose $X = (U_1, \dots, U_d)$ and $Y = (V_1, \dots, V_d)$ are vectors with $(U_i, V_i) \stackrel{i.i.d.}{\sim} P_{U,V}$ and $I_{\lambda_0}(U; V) < \infty$ for some $\lambda_0 > 1$ (e.g. if one of U or V is over a finite alphabet). Then for any $R > I(U; V)$ there exists $\epsilon > 0$, so that for all $d \geq 1$ there exists $x_1, \dots, x_n, n = \exp\{dR\}$, such that*

$$D\left(\frac{1}{n} \sum_{i=1}^n P_{Y|X=x_i} \| P_Y\right) \leq \exp\{-d\epsilon\}$$

as $d \rightarrow \infty$.

Remark 25.5. The origin of the name “soft-covering” is due to the fact that unlike the covering lemma (Theorem 25.7) which selects one x_i (trying to make $P_{Y|X=x_i}$ as close to P_Y as possible) here we mix over n choices uniformly.

25.4* Approximation of output statistics and the soft-covering lemma 435

Proof. By tensorization of Rényi divergence, cf. Section 7.12, we have

$$I_\lambda(X; Y) = dI_\lambda(U; V).$$

For every $1 < \lambda < \lambda_0$ we have that $\lambda \mapsto I_\lambda(U; V)$ is continuous and converging to $I(U; V)$ as $\lambda \rightarrow 1$. Thus, we can find λ sufficiently small so that $R > I_\lambda(U; V)$. Applying Theorem 25.10 with this λ completes the proof. \square

26

Evaluating rate-distortion function. Lossy Source-Channel separation.

In the previous chapters we have proved Shannon's main theorem for lossy data compression: For stationary memoryless (iid) sources and separable distortion, under the assumption that $D_{\max} < \infty$, the operational and information rate-distortion functions coincide, namely,

$$R(D) = R^{(I)}(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}d(S, \hat{S}) \leq D} I(S; \hat{S}).$$

In addition, we have shown various properties about the rate-distortion function (cf. Theorem 24.7). In this chapter we compute the rate-distortion function for several important source distributions by evaluating this constrained minimization of mutual information. Next we extending the paradigm of joint source-channel coding in Section 19.7 to the lossy setting; this reasoning will later be found useful in statistical applications in Part VI (cf. Chapter 30).

26.1 Evaluation of $R(D)$

26.1.1 Bernoulli Source

Let $S \sim \text{Ber}(p)$ with Hamming distortion $d(S, \hat{S}) = 1\{S \neq \hat{S}\}$ and alphabets $\mathcal{S} = \hat{\mathcal{S}} = \{0, 1\}$. Then $d(s^n, \hat{s}^n) = \frac{1}{n}d_H(s^n, \hat{s}^n)$ is the bit-error rate (fraction of erroneously decoded bits). By symmetry, we may assume that $p \leq 1/2$.

Theorem 26.1.

$$R(D) = (h(p) - h(D))_+. \quad (26.1)$$

For example, when $p = 1/2$, $D = .11$, we have $R(D) \approx 1/2$ bits. In the Hamming game described in Section 24.2 where we aim to compress 100 bits down to 50, we indeed can do this while achieving 11% average distortion, compared to the naive scheme of storing half the string and guessing on the other half, which achieves 25% average distortion.

Proof. Since $D_{\max} = p$, in the sequel we can assume $D < p$ for otherwise there is nothing to show.

For the converse, consider any $P_{\hat{S}|S}$ such that $P[S \neq \hat{S}] \leq D \leq p \leq \frac{1}{2}$. Then

$$\begin{aligned} I(S; \hat{S}) &= H(S) - H(S|\hat{S}) \\ &= H(S) - H(S + \hat{S}|\hat{S}) \end{aligned}$$

26.1 Evaluation of $R(D)$ 437

$$\begin{aligned} &\geq H(S) - H(S + \hat{S}) \\ &= h(p) - h(P[S \neq \hat{S}]) \\ &\geq h(p) - h(D). \end{aligned}$$

In order to achieve this bound, we need to saturate the above chain of inequalities, in particular, choose $P_{\hat{S}|S}$ so that the difference $S + \hat{S}$ is independent of \hat{S} . Let $S = \hat{S} + Z$, where $\hat{S} \sim \text{Ber}(p')$ $\perp\!\!\!\perp Z \sim \text{Ber}(D)$, and p' is such that the convolution gives exactly $\text{Ber}(p)$, namely,

$$p' * D = p'(1 - D) + (1 - p')D = p,$$

i.e., $p' = \frac{p-D}{1-2D}$. In other words, the backward channel $P_{S|\hat{S}}$ is exactly $\text{BSC}(D)$ and the resulting $P_{\hat{S}|S}$ is our choice of the forward channel $P_{\hat{S}|S}$. Then, $I(S; \hat{S}) = H(S) - H(S|\hat{S}) = H(S) - H(Z) = h(p) - h(D)$, yielding the upper bound $R(D) \leq h(p) - h(D)$. \square

Remark 26.1. Here is a more general strategy (which we will later implement in the Gaussian case.) Denote the optimal forward channel from the achievability proof by $P_{\hat{S}|S}^*$ and $P_{S|\hat{S}}^*$ the associated backward channel (which is $\text{BSC}(D)$). We need to show that there is no better $P_{\hat{S}|S}$ with $P[S \neq \hat{S}] \leq D$ and a smaller mutual information. Then

$$\begin{aligned} I(P_S, P_{\hat{S}|S}) &= D(P_{S|\hat{S}} \| P_S | P_{\hat{S}}) \\ &= D(P_{S|\hat{S}} \| P_{S|\hat{S}}^* | P_{\hat{S}}) + \mathbb{E}_P \left[\log \frac{P_{S|\hat{S}}^*}{P_S} \right] \\ &\geq H(S) + \mathbb{E}_P [\log D \mathbf{1}\{S \neq \hat{S}\} + \log \bar{D} \mathbf{1}\{S = \hat{S}\}] \\ &\geq h(p) - h(D) \end{aligned}$$

where the last inequality uses $P[S \neq \hat{S}] \leq D \leq \frac{1}{2}$.

Remark 26.2. By WLLN, the distribution $P_S^n = \text{Ber}(p)^n$ concentrates near the Hamming sphere of radius np as n grows large. Recall that in proving Shannon's rate distortion theorem, the optimal codebook are drawn independently from $P_{\hat{S}}^n = \text{Ber}(p')^n$ with $p' = \frac{p-D}{1-2D}$. Note that $p' = 1/2$ if $p = 1/2$ but $p' < p$ if $p < 1/2$. In the latter case, the reconstruction points concentrate on a smaller sphere of radius np' and *none* of them are typical source realizations, as illustrated in Fig. 26.1.

26.1.2 Gaussian Source

The following results compute the Gaussian rate-distortion function for quadratic distortion in both the scalar and vector case. (For general covariance, see Exercise V.5.)

Theorem 26.2. *Let $S \sim \mathcal{N}(0, \sigma^2)$ and $d(s, \hat{s}) = (s - \hat{s})^2$ for $s, \hat{s} \in \mathbb{R}$. Then*

$$R(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}. \quad (26.2)$$

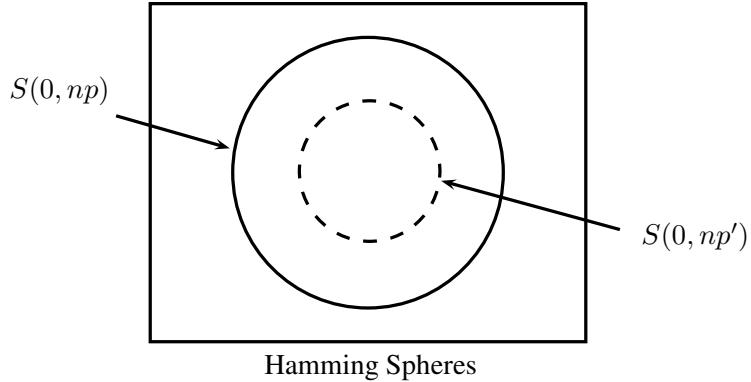


Figure 26.1 Source realizations (solid sphere) versus codewords (dashed sphere) in compressing Hamming sources.

In the vector case of $S \sim \mathcal{N}(0, \sigma^2 I_d)$ and $d(s, \hat{s}) = \|s - \hat{s}\|_2^2$,

$$R(D) = \frac{d}{2} \log^+ \frac{d\sigma^2}{D}. \quad (26.3)$$

Proof. Since $D_{\max} = \sigma^2$, in the sequel we can assume $D < \sigma^2$ for otherwise there is nothing to show.

(Achievability) Choose $S = \hat{S} + Z$, where $\hat{S} \sim \mathcal{N}(0, \sigma^2 - D) \perp\!\!\!\perp Z \sim \mathcal{N}(0, D)$. In other words, the backward channel $P_{S|\hat{S}}$ is AWGN with noise power D , and the forward channel can be easily found to be $P_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2 - D}{\sigma^2} S, \frac{\sigma^2 - D}{\sigma^2} D)$. Then

$$I(S; \hat{S}) = \frac{1}{2} \log \frac{\sigma^2}{D} \implies R(D) \leq \frac{1}{2} \log \frac{\sigma^2}{D}$$

(Converse) Formally, we can mimic the proof of Theorem 26.1 replacing Shannon entropy by the differential entropy and applying the maximal entropy result from Theorem 2.8; the caveat is that for \hat{S} (which may be discrete) the differential entropy may not be well-defined. As such, we follow the alternative proof given in Remark 26.2. Let $P_{\hat{S}|S}$ be any conditional distribution such that $\mathbb{E}_P[(S - \hat{S})^2] \leq D$. Denote the forward channel in the above achievability by $P_{\hat{S}|S}^*$. Then

$$\begin{aligned} I(P_S, P_{\hat{S}|S}) &= D(P_{S|\hat{S}} \| P_{S|\hat{S}}^* | P_{\hat{S}}) + \mathbb{E}_P \left[\log \frac{P_{S|\hat{S}}^*}{P_S} \right] \\ &\geq \mathbb{E}_P \left[\log \frac{P_{S|\hat{S}}^*}{P_S} \right] \\ &= \mathbb{E}_P \left[\log \frac{\frac{1}{\sqrt{2\pi D}} e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{S^2}{2\sigma^2}}} \right] \end{aligned}$$

26.1 Evaluation of $R(D)$ 439

$$\begin{aligned} &= \frac{1}{2} \log \frac{\sigma^2}{D} + \frac{\log e}{2} \mathbb{E}_P \left[\frac{S^2}{\sigma^2} - \frac{(S - \hat{S})^2}{D} \right] \\ &\geq \frac{1}{2} \log \frac{\sigma^2}{D}. \end{aligned}$$

Finally, for the vector case, (26.3) follows from (26.2) and the same single-letterization argument in Theorem 24.11 using the convexity of the rate-distortion function in Theorem 24.7(a). \square

The interpretation of the optimal reconstruction points in the Gaussian case is analogous to that of the Hamming source previously discussed in Remark 26.3: As n grows, the Gaussian random vector concentrates on $S(0, \sqrt{n\sigma^2})$ (n -sphere in Euclidean space rather than Hamming), but each reconstruction point drawn from $(P_{\hat{S}}^*)^n$ is close to $S(0, \sqrt{n(\sigma^2 - D)})$. So again the picture is similar to Fig. 26.1 of two nested spheres.

Note that the exact expression in Theorem 26.4 relies on the Gaussianity assumption of the source. How sensitive is the rate-distortion formula to this assumption? The following comparison result is a counterpart of Theorem 20.11 for channel capacity:

Theorem 26.3. *Assume that $\mathbb{E}S = 0$ and $\text{Var } S = \sigma^2$. Consider the MSE distortion. Then*

$$\frac{1}{2} \log^+ \frac{\sigma^2}{D} - D(P_S \| \mathcal{N}(0, \sigma^2)) \leq R(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}(\hat{S}-S)^2 \leq D} I(S; \hat{S}) \leq \frac{1}{2} \log^+ \frac{\sigma^2}{D}.$$

Remark 26.3. A simple consequence of Theorem 26.5 is that for source distributions with a density, the rate-distortion function grows according to $\frac{1}{2} \log \frac{1}{D}$ in the low-distortion regime as long as $D(P_S \| \mathcal{N}(0, \sigma^2))$ is finite. In fact, the first inequality, known as the *Shannon lower bound* (SLB), is asymptotically tight, in the sense that

$$R(D) = \frac{1}{2} \log \frac{\sigma^2}{D} - D(P_S \| \mathcal{N}(0, \sigma^2)) + o(1), \quad D \rightarrow 0 \quad (26.4)$$

under appropriate conditions on P_S [196, 173]. Therefore, by comparing (2.21) and (26.4), we see that, for small distortion, uniform scalar quantization (Section 24.1) is in fact asymptotically optimal within $\frac{1}{2} \log(2\pi e) \approx 2.05$ bits.

Later in Section 30.1 we will apply SLB to derive lower bounds for statistical estimation. For this we need the following general version of SLB (see Exercise V.6 for a proof): Let $\|\cdot\|$ be an arbitrary norm on \mathbb{R}^d and $r > 0$. Let X be a d -dimensional continuous random vector with finite differential entropy $h(X)$. Then

$$\inf_{P_{\hat{X}|X}: \mathbb{E}[\|\hat{X}-X\|^r] \leq D} I(X; \hat{X}) \geq h(X) + \frac{d}{r} \log \frac{d}{Dr} - \log \left(\Gamma \left(\frac{d}{r} + 1 \right) V \right), \quad (26.5)$$

where $V = \text{vol}(\{x \in \mathbb{R}^d : \|x\| \leq 1\})$ is the volume of the unit $\|\cdot\|$ -ball.

Proof. Again, assume $D < D_{\max} = \sigma^2$. Let $S_G \sim \mathcal{N}(0, \sigma^2)$.

“ \leq ”: Use the same $P_{\hat{S}|S}^* = \mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S, \frac{\sigma^2-D}{\sigma^2}D)$ in the achievability proof of Gaussian rate-distortion function:

$$R(D) \leq I(P_S, P_{\hat{S}|S}^*)$$

440

$$\begin{aligned}
&= I(S; \frac{\sigma^2 - D}{\sigma^2} S + W) & W \sim \mathcal{N}(0, \frac{\sigma^2 - D}{\sigma^2} D) \\
&\leq I(S_G; \frac{\sigma^2 - D}{\sigma^2} S_G + W) & \text{by Gaussian saddle point (Theorem 5.9)} \\
&= \frac{1}{2} \log \frac{\sigma^2}{D}.
\end{aligned}$$

“ \geq ”: For any $P_{\hat{S}|S}$ such that $\mathbb{E}(\hat{S} - S)^2 \leq D$. Let $P_{S|\hat{S}}^* = \mathcal{N}(\hat{S}, D)$ denote the AWGN channel with noise power D . Then

$$\begin{aligned}
I(S; \hat{S}) &= D(P_{S|\hat{S}} \| P_S | P_{\hat{S}}) \\
&= D(P_{S|\hat{S}} \| P_{S|\hat{S}}^* | P_{\hat{S}}) + \mathbb{E}_P \left[\log \frac{P_{S|\hat{S}}^*}{P_{S|\hat{S}}} \right] - D(P_S \| P_{S_G}) \\
&\geq \mathbb{E}_P \left[\log \frac{\frac{1}{\sqrt{2\pi D}} e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{S^2}{2\sigma^2}}} \right] - D(P_S \| P_{S_G}) \\
&\geq \frac{1}{2} \log \frac{\sigma^2}{D} - D(P_S \| P_{S_G}).
\end{aligned}$$

□

26.2* Analog of saddle-point property in rate-distortion

In the computation of $R(D)$ for the Hamming and Gaussian source, we guessed the correct form of the rate-distortion function. In both of their converse arguments, we used the same trick to establish that any other feasible $P_{\hat{S}|S}$ gave a larger value for $R(D)$. In this section, we formalize this trick, in an analogous manner to the saddle point property of the channel capacity. Note that typically we don't need any tricks to compute $R(D)$, since we can obtain a solution in a parametric form to the unconstrained convex optimization

$$\min_{P_{\hat{S}|S}} I(S; \hat{S}) + \lambda \mathbb{E}[d(S, \hat{S})]$$

In fact there are also iterative algorithms (Blahut-Arimoto) that computes $R(D)$. However, for the peace of mind it is good to know there are some general reasons why tricks like we used in Hamming/Gaussian actually are guaranteed to work.

Theorem 26.4.

I Suppose P_{Y^*} and $P_{X|Y^*} \ll P_X$ are such that $\mathbb{E}[d(X, Y^*)] \leq D$ and for any $P_{X,Y}$ with $\mathbb{E}[d(X, Y)] \leq D$ we have

$$\mathbb{E} \left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y) \right] \geq I(X; Y^*). \quad (26.6)$$

Then $R(D) = I(X; Y^*)$.

26.2* Analog of saddle-point property in rate-distortion 441

2 Suppose that $I(X; Y^*) = R(D)$. Then for any regular branch of conditional probability $P_{X|Y^*}$ and for any $P_{X,Y}$ satisfying

- $\mathbb{E}[d(X, Y)] \leq D$ and
- $P_Y \ll P_{Y^*}$ and
- $I(X; Y) < \infty$

the inequality (26.6) holds.

Remarks:

1 The first part is a sufficient condition for optimality of a given P_{XY^*} . The second part gives a necessary condition that is convenient to narrow down the search. Indeed, typically the set of $P_{X,Y}$ satisfying those conditions is rich enough to infer from (26.6):

$$\log \frac{dP_{X|Y^*}}{dP_X}(x|y) = R(D) - \theta[d(x, y) - D],$$

for a positive $\theta > 0$.

2 Note that the second part is not valid without assuming $P_Y \ll P_{Y^*}$. A counterexample to this and various other erroneous (but frequently encountered) generalizations is the following: $\mathcal{A} = \{0, 1\}$, $P_X = \text{Bern}(1/2)$, $\hat{\mathcal{A}} = \{0, 1, 0', 1'\}$ and

$$d(0, 0) = d(0, 0') = 1 - d(0, 1) = 1 - d(0, 1') = 0.$$

The $R(D) = |1 - h(D)|^+$, but there exist multiple non-equivalent optimal choices of $P_{Y|X}$, $P_{X|Y}$ and P_Y .

Proof. The first part is just a repetition of the proofs above for the Hamming and Gaussian case, so we focus on the second part. Suppose there exists a counterexample $P_{X,Y}$ achieving

$$I_1 = \mathbb{E} \left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y) \right] < I^* = R(D).$$

Notice that whenever $I(X; Y) < \infty$ we have

$$I_1 = I(X; Y) - D(P_{X|Y} \| P_{X|Y^*} | P_Y),$$

and thus

$$D(P_{X|Y} \| P_{X|Y^*} | P_Y) < \infty. \quad (26.7)$$

Before going to the actual proof, we describe the principal idea. For every λ we can define a joint distribution

$$P_{X,Y_\lambda} = \lambda P_{X,Y} + (1 - \lambda) P_{X,Y^*}.$$

Then, we can compute

$$I(X; Y_\lambda) = \mathbb{E} \left[\log \frac{P_{X|Y_\lambda}}{P_X}(X|Y_\lambda) \right] = \mathbb{E} \left[\log \frac{P_{X|Y_\lambda}}{P_{X|Y^*}} \frac{P_{X|Y^*}}{P_X} \right] \quad (26.8)$$

$$= D(P_{X|Y_\lambda} \| P_{X|Y^*} | P_{Y_\lambda}) + \mathbb{E} \left[\frac{P_{X|Y^*}(X|Y_\lambda)}{P_X} \right] \quad (26.9)$$

$$= D(P_{X|Y_\lambda} \| P_{X|Y^*} | P_{Y_\lambda}) + \lambda I_1 + (1 - \lambda) I_* . \quad (26.10)$$

From here we will conclude, similar to Proposition 2.17, that the first term is $o(\lambda)$ and thus for sufficiently small λ we should have $I(X; Y_\lambda) < R(D)$, contradicting optimality of coupling P_{X,Y^*} .

We proceed to details. For every $\lambda \in [0, 1]$ define

$$\rho_1(y) \triangleq \frac{dP_Y}{dP_{Y^*}}(y) \quad (26.11)$$

$$\lambda(y) \triangleq \frac{\lambda \rho_1(y)}{\lambda \rho_1(y) + \bar{\lambda}} \quad (26.12)$$

$$P_{X|Y=y}^{(\lambda)} = \lambda(y)P_{X|Y=y} + \bar{\lambda}(y)P_{X|Y^*=y} \quad (26.13)$$

$$dP_{Y_\lambda} = \lambda dP_Y + \bar{\lambda} dP_{Y^*} = (\lambda \rho_1(y) + \bar{\lambda}) dP_{Y^*} \quad (26.14)$$

$$D(y) = D(P_{X|Y=y} \| P_{X|Y^*=y}) \quad (26.15)$$

$$D_\lambda(y) = D(P_{X|Y=y}^{(\lambda)} \| P_{X|Y^*=y}) . \quad (26.16)$$

Notice:

$$\text{On } \{\rho_1 = 0\} : \quad \lambda(y) = D(y) = D_\lambda(y) = 0$$

and otherwise $\lambda(y) > 0$. By convexity of divergence

$$D_\lambda(y) \leq \lambda(y)D(y)$$

and therefore

$$\frac{1}{\lambda(y)} D_\lambda(y) \mathbf{1}_{\{\rho_1(y) > 0\}} \leq D(y) \mathbf{1}_{\{\rho_1(y) > 0\}} .$$

Notice that by (26.7) the function $\rho_1(y)D(y)$ is non-negative and P_{Y^*} -integrable. Then, applying dominated convergence theorem we get

$$\lim_{\lambda \rightarrow 0} \int_{\{\rho_1 > 0\}} dP_{Y^*} \frac{1}{\lambda(y)} D_\lambda(y) \rho_1(y) = \int_{\{\rho_1 > 0\}} dP_{Y^*} \rho_1(y) \lim_{\lambda \rightarrow 0} \frac{1}{\lambda(y)} D_\lambda(y) = 0 \quad (26.17)$$

where in the last step we applied the result from Chapter 5

$$D(P \| Q) < \infty \quad \Rightarrow \quad D(\lambda P + \bar{\lambda} Q \| Q) = o(\lambda)$$

since for each y on the set $\{\rho_1 > 0\}$ we have $\lambda(y) \rightarrow 0$ as $\lambda \rightarrow 0$.

On the other hand, notice that

$$\int_{\{\rho_1 > 0\}} dP_{Y^*} \frac{1}{\lambda(y)} D_\lambda(y) \rho_1(y) \mathbf{1}_{\{\rho_1(y) > 0\}} = \frac{1}{\lambda} \int_{\{\rho_1 > 0\}} dP_{Y^*} (\lambda \rho_1(y) + \bar{\lambda}) D_\lambda(y) \quad (26.18)$$

$$= \frac{1}{\lambda} \int_{\{\rho_1 > 0\}} dP_{Y_\lambda} D_\lambda(y) \quad (26.19)$$

26.3 Lossy joint source-channel coding 443

$$= \frac{1}{\lambda} \int_{\mathcal{Y}} dP_{Y_\lambda} D_\lambda(y) = \frac{1}{\lambda} D(P_{X|Y}^{(\lambda)} \| P_{X|Y^*} | P_{Y_\lambda}), \quad (26.20)$$

where in the penultimate step we used $D_\lambda(y) = 0$ on $\{\rho_1 = 0\}$. Hence, (26.17) shows

$$D(P_{X|Y}^{(\lambda)} \| P_{X|Y^*} | P_{Y_\lambda}) = o(\lambda), \quad \lambda \rightarrow 0.$$

Finally, since

$$P_{X|Y}^{(\lambda)} \circ P_{Y_\lambda} = P_X,$$

we have

$$I(X; Y_\lambda) = D(P_{X|Y}^{(\lambda)} \| P_{X|Y^*} | P_{Y_\lambda}) + \lambda \mathbb{E} \left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y) \right] + \bar{\lambda} \mathbb{E} \left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y^*) \right] \quad (26.21)$$

$$= I^* + \lambda(I_1 - I^*) + o(\lambda), \quad (26.22)$$

contradicting the assumption

$$I(X; Y_\lambda) \geq I^* = R(D).$$

□

26.3 Lossy joint source-channel coding

Extending the lossless joint source channel coding problem studied in Section 19.7, in this section we study the lossy version of this problem: How to transmit a source over a noisy channel such that the receiver can reconstruct the original source within a prescribed distortion.

The setup of the lossy joint source-channel coding problem is as follows. For each k and n , we are given a source $S^k = (S_1, \dots, S_k)$ taking values on \mathcal{S} , a distortion metric $d : \mathcal{S}^k \times \hat{\mathcal{S}}^k \rightarrow \mathbb{R}$, and a channel $P_{Y^n|X^n}$ acting from \mathcal{A}^n to \mathcal{B}^n . A lossy joint source-channel code (JSCL) consists of an encoder $f : \mathcal{S}^k \rightarrow \mathcal{A}^n$ and decoder $g : \mathcal{B}^n \rightarrow \hat{\mathcal{S}}^k$, such that the channel input is $X^n = f(S^k)$ and the reconstruction $\hat{S}^k = g(Y^n)$ satisfies $\mathbb{E}[d(S^k, \hat{S}^k)] \leq D$. By definition, we have the Markov chain

$$S^k \xrightarrow{f} X^n \xrightarrow{P_{Y^n|X^n}} Y^n \xrightarrow{g} \hat{S}^k$$

Such a pair (f, g) is called a (k, n, D) -JSCL, which transmits k symbols over n channel uses such that the end-to-end distortion is at most D in expectation. Our goal is to optimize the encoder/decoder pair so as to maximize the transmission rate (number of symbols per channel use) $R = \frac{k}{n}$.¹ As such, we define the asymptotic fundamental limit as

$$R_{\text{JSCL}}(D) \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n} \max \{k : \exists (k, n, D)\text{-JSCL}\}.$$

¹ Or equivalently, minimize the *bandwidth expansion factor* $\rho = \frac{n}{k}$.

To simplify the exposition, we will focus on JSCC for a stationary memoryless source $S^k \sim P_S^{\otimes k}$ transmitted over a stationary memoryless channel $P_{Y^n|X^n} = P_{Y|X}^{\otimes n}$ subject to a separable distortion function $d(s^k, \hat{s}^k) = \frac{1}{k} \sum_{i=1}^k d(s_i, \hat{s}_i)$.

26.3.1 Converse

The converse for the JSCC is quite simple, based on data processing inequality and following the weak converse of lossless JSCC using Fano's inequality.

Theorem 26.5 (Converse).

$$R_{\text{JSCC}}(D) \leq \frac{C}{R(D)},$$

where $C = \sup_{P_X} I(X; Y)$ is the capacity of the channel and $R(D) = \inf_{P_{\hat{S}|S}: \mathbb{E}[d(S, \hat{S})] \leq D} I(S; \hat{S})$ is the rate-distortion function of the source.

The interpretation of this result is clear: Since we need at least $R(D)$ bits per symbol to reconstruct the source up to a distortion D and we can transmit at most C bits per channel use, the overall transmission rate cannot exceed $C/R(D)$. Note that the above theorem clearly holds for channels with cost constraint with the corresponding capacity (Chapter 20).

Proof. Consider a (k, n, D) -code which induces the Markov chain $S^k \rightarrow X^n \rightarrow Y^n \rightarrow \hat{S}^k$ such that $\mathbb{E}[d(S^k, \hat{S}^k)] = \frac{1}{k} \sum_{i=1}^k \mathbb{E}[d(S_i, \hat{S}_i)] \leq D$. Then

$$kR(D) \stackrel{(a)}{=} \inf_{P_{\hat{S}^k|S^k}: \mathbb{E}[d(S^k, \hat{S}^k)] \leq D} I(S^k; \hat{S}^k) \stackrel{(b)}{\leq} I(S^k; \hat{S}^k) \leq I(X^n; Y^n) \leq \sup_{P_{X^n}} I(X^n; Y^n) \stackrel{(c)}{=} nC$$

where (b) applies data processing inequality for mutual information, (a) and (c) follow from the respective single-letterization result for lossy compression and channel coding (Theorem 24.11 and Proposition 19.9). \square

Remark 26.4. Consider the case where the source is $\text{Ber}(1/2)$ with Hamming distortion. Then Theorem 26.8 coincides with the converse for channel coding under bit error rate P_b in (19.33):

$$R = \frac{k}{n} \leq \frac{C}{1 - h(P_b)}$$

which was previously given in Theorem 19.21 and proved using ad hoc techniques. In the case of channel with cost constraints, e.g., the AWGN channel with $C(\text{SNR}) = \frac{1}{2} \log(1 + \text{SNR})$, we have

$$P_b \geq h^{-1} \left(1 - \frac{C(\text{SNR})}{R} \right)$$

This is often referred to as the Shannon limit in plots comparing the bit-error rate of practical codes. (See, e.g., Fig. 2 from [254] for BIAWGN (binary-input) channel.) *This is erroneous*, since the p_b above refers to the bit error rate of data bits (or systematic bits), not all of the codeword bits. The latter quantity is what typically called BER (see (19.33)) in the coding-theoretic literature.

26.3.2 Achievability via separation

The proof strategy is similar to lossless JSCC in Section 19.7 by *separately* constructing a channel coding scheme and a lossy compression scheme, as opposed to jointly optimizing the JSCC encoder/decoder pair. Specifically, first compress the data into bits then encode with a channel code; to decode, apply the channel decoder followed by the source decompressor. Under appropriate assumptions, this separately-designed scheme achieves the optimal rate in Theorem 26.8.

Theorem 26.6. *For any stationary memoryless source $(P_S, \mathcal{S}, \hat{\mathcal{S}}, d)$ with rate-distortion function $R(D)$ satisfying Assumption 26.1 (below), and for any stationary memoryless channel $P_{Y|X}$ with capacity C ,*

$$R_{\text{JSCC}}(D) = \frac{C}{R(D)}.$$

Assumption 26.1 on the source (which is rather technical and can be skipped in the first reading) is to control the distortion incurred by the channel decoder making an error. Despite this being a low-probability event, without any assumption on the distortion metric, we cannot say much about its contribution to the end-to-end average distortion. (Note that this issue does not arise in lossless JSCC). Assumption 26.1 is trivially satisfied by bounded distortion (e.g., Hamming), and can be shown to hold more generally such as for Gaussian sources and MSE distortion.

Proof. In view of Theorem 26.8, we only prove achievability. We constructed a separated compression/channel coding scheme as follows:

- Let (f_s, g_s) be a $(k, 2^{kR(D)+o(k)}, D)$ -code for compressing S^k such that $\mathbb{E}[d(S^k, g_s(f_s(S^k)))] \leq D$. By Lemma 26.12 (below), we may assume that all reconstruction points are not too far from some fixed string, namely,

$$d(s_0^k, g_s(i)) \leq L \quad (26.23)$$

for all i and some constant L , where $s_0^k = (s_0, \dots, s_0)$ is from Assumption 26.1 below.

- Let (f_c, g_c) be a $(n, 2^{nC+o(n)}, \epsilon_n)_{\max}$ -code for channel $P_{Y^n|X^n}$ such that $kR(D) + o(k) \leq nC + o(n)$ and the maximal probability of error $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$. Such a code exists thanks to Theorem 19.8 and Corollary 19.1.

Let the JSCC encoder and decoder be $f = f_c \circ f_s$ and $g = g_s \circ g_c$. So the overall system is

$$S^k \xrightarrow{f_s} W \xrightarrow{f_c} X^n \longrightarrow Y^n \xrightarrow{g_c} \hat{W} \xrightarrow{g_s} \hat{S}^k.$$

Note that here we need to control the *maximal* probability of error of the channel code since when we concatenate these two schemes, W at the input of the channel is the output of the source compressor, which need not be uniform.

To analyze the average distortion, we consider two cases depending on whether the channel decoding is successful or not:

$$\mathbb{E}[d(S^k, \hat{S}^k)] = \mathbb{E}[d(S^k, g_s(W))1\{W = \hat{W}\}] + \mathbb{E}[d(S^k, g_s(\hat{W}))1\{W \neq \hat{W}\}].$$

By assumption on our lossy code, the first term is at most D . For the second term, we have $\mathbb{P}[W \neq \hat{W}] \leq \epsilon_n = o(1)$ by assumption on our channel code. Then

$$\begin{aligned} \mathbb{E}[d(S^k, g_s(\hat{W}))1\{W \neq \hat{W}\}] &\stackrel{(a)}{\leq} \mathbb{E}[1\{W \neq \hat{W}\}\lambda(d(S^k, \hat{s}_0^k) + d(s_0^k, g_s(\hat{W})))] \\ &\stackrel{(b)}{\leq} \lambda \cdot \mathbb{E}[1\{W \neq \hat{W}\}d(S^k, \hat{s}_0^k)] + \lambda L \cdot \mathbb{P}[W \neq \hat{W}] \\ &\stackrel{(c)}{=} o(1), \end{aligned}$$

where (a) follows from the generalized triangle inequality from Assumption 26.1(a) below; (b) follows from (26.23); in (c) we apply Lemma 25.5 that were used to show the vanishing of the expectation in (25.15) before.

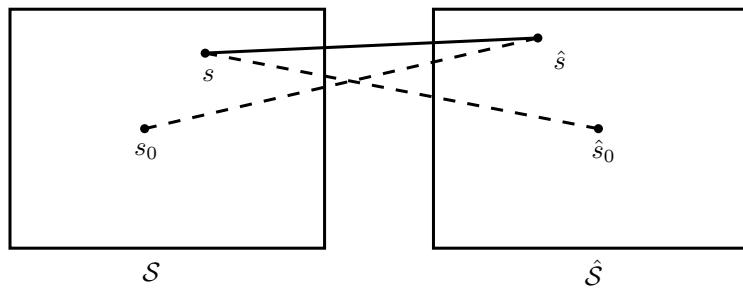
In all, our scheme meets the average distortion constraint. Hence we conclude that for all $R > C/R(D)$, there exists a sequence of $(k, n, D + o(1))$ -JSCC codes. \square

The following assumption is needed by the previous theorem:

Assumption 26.1. Fix D . For a source $(P_S, \mathcal{S}, \hat{\mathcal{S}}, d)$, there exists $\lambda \geq 0, s_0 \in \mathcal{S}, \hat{s}_0 \in \hat{\mathcal{S}}$ such that

- (a) Generalized triangle inequality: $d(s, \hat{s}) \leq \lambda(d(s, \hat{s}_0) + d(s_0, \hat{s})) \quad \forall s, \hat{s}$.
- (b) $\mathbb{E}[d(S, \hat{s}_0)] < \infty$ (so that $D_{\max} < \infty$ too).
- (c) $\mathbb{E}[d(s_0, \hat{\mathcal{S}})] < \infty$ for any output distribution $P_{\hat{\mathcal{S}}}$ achieving the rate-distortion function $R(D)$.
- (d) $d(s_0, \hat{s}_0) < \infty$.

The interpretation of this assumption is that the spaces \mathcal{S} and $\hat{\mathcal{S}}$ have ‘‘nice centers’’ s_0 and \hat{s}_0 , in the sense that the distance between any two points is upper bounded by a constant times the distance from the centers to each point (see figure below).



Note that Assumption 26.1 is not straightforward to verify. Next we give some more convenient sufficient conditions. First of all, Assumption 26.1 holds automatically for bounded distortion

26.3 Lossy joint source-channel coding 447

function. In other words, for a discrete source on a finite alphabet \mathcal{S} , a finite reconstruction alphabet $\hat{\mathcal{S}}$, and a finite distortion function $d(s, \hat{s}) < \infty$, Assumption 26.1 is fulfilled. More generally, we have the following criterion.

Theorem 26.7. *If $\mathcal{S} = \hat{\mathcal{S}}$ and $d(s, \hat{s}) = \rho(s, \hat{s})^q$ for some metric ρ and $q \geq 1$, and $D_{\max} \triangleq \inf_{\hat{s}_0} \mathbb{E}[d(S, \hat{s}_0)] < \infty$, then Assumption 26.1 holds.*

Proof. Take $s_0 = \hat{s}_0$ that achieves a finite $D_{\max} = \mathbb{E}[d(S, \hat{s}_0)]$. (In fact, any points can serve as centers in a metric space). Applying triangle inequality and Jensen's inequality, we have

$$\left(\frac{1}{2} \rho(s, \hat{s}) \right)^q \leq \left(\frac{1}{2} \rho(s, s_0) + \frac{1}{2} \rho(s_0, \hat{s}) \right)^q \leq \frac{1}{2} \rho^q(s, s_0) + \frac{1}{2} \rho^q(s_0, \hat{s}).$$

Thus $d(s, \hat{s}) \leq 2^{q-1}(d(s, s_0) + d(s_0, \hat{s}))$. Taking $\lambda = 2^{q-1}$ verifies (a) and (b) in Assumption 26.1. To verify (c), we can apply this generalized triangle inequality to get $d(s_0, \hat{S}) \leq 2^{q-1}(d(s_0, S) + d(S, \hat{S}))$. Then taking the expectation of both sides gives

$$\begin{aligned} \mathbb{E}[d(s_0, \hat{S})] &\leq 2^{q-1}(\mathbb{E}[d(s_0, S)] + \mathbb{E}[d(S, \hat{S})]) \\ &\leq 2^{q-1}(D_{\max} + D) < \infty. \end{aligned}$$

□

So we see that metrics raised to powers (e.g. squared norms) satisfy Assumption 26.1. Finally, we give the lemma used in the proof of Theorem 26.10.

Lemma 26.8. *Fix a source satisfying Assumption 26.1 and an arbitrary $P_{\hat{S}|S}$. Let $R > I(S; \hat{S})$, $L > \max\{\mathbb{E}[d(s_0, \hat{S})], d(s_0, \hat{s}_0)\}$ and $D > \mathbb{E}[d(S, \hat{S})]$. Then, there exists a $(k, 2^{kR}, D)$ -code such that $d(s_0^k, \hat{s}^k) \leq L$ for every reconstruction point \hat{s}^k , where $s_0^k = (s_0, \dots, s_0)$.*

Proof. Let $\mathcal{X} = \mathcal{S}^k$, $\hat{\mathcal{X}} = \hat{\mathcal{S}}^{\otimes k}$ and $P_X = P_S^k$, $P_{Y|X} = P_{\hat{S}|S}^{\otimes k}$. We apply the achievability bound for excess distortion from Theorem 25.4 with $\gamma = 2^{k(R+I(S; \hat{S}))/2}$ to the following *non-separable* distortion function

$$d_1(x, \hat{x}) = \begin{cases} d(x, \hat{x}) & d(s_0^k, \hat{x}) \leq L \\ +\infty & \text{otherwise.} \end{cases}$$

For any $D' \in (\mathbb{E}[d(S, \hat{S})], D)$, there exist $M = 2^{kR}$ reconstruction points (c_1, \dots, c_M) such that

$$\mathbb{P} \left[\min_{j \in [M]} d(s^k, c_j) > D' \right] \leq \mathbb{P}[d_1(s^k, \hat{S}^k) > D'] + o(1),$$

where on the right side $(S^k, \hat{S}^k) \sim P_{S, \hat{S}}^{\otimes k}$. Note that without any change in d_1 -distortion we can remove all (if any) reconstruction points c_j with $d(s_0^k, c_j) > L$. Furthermore, from the WLLN we have

$$\mathbb{P}[d_1(S, \hat{S}) > D'] \leq \mathbb{P}[d(S^k, \hat{S}^k) > D'] + \mathbb{P}[d(s_0^k, \hat{S}^k) > L] \rightarrow 0$$

as $k \rightarrow \infty$ (since $\mathbb{E}[d(S, \hat{S})] < D'$ and $\mathbb{E}[s_0, \hat{S}] < L$). Thus we have

$$\mathbb{P} \left[\min_{j \in [M]} d(S^k, c_j) > D' \right] \rightarrow 0$$

and $d(s_0^k, c_j) \leq L$. Finally, by adding another reconstruction point $c_{M+1} = \hat{s}_0^k = (\hat{s}_0, \dots, \hat{s}_0)$ we get

$$\mathbb{E} \left[\min_{j \in [M+1]} d(S^k, c_j) \right] \leq D' + \mathbb{E} \left[d(S^k, \hat{s}_0^k) \mathbf{1}_{\{\min_{j \in [M]} d(S^k, c_j) > D'\}} \right] = D' + o(1),$$

where the last estimate follows from the same argument that shows the vanishing of the expectation in (25.15). Thus, for sufficiently large n the expected distortion is at most D , as required. \square

26.4 What is lacking in classical lossy compression?

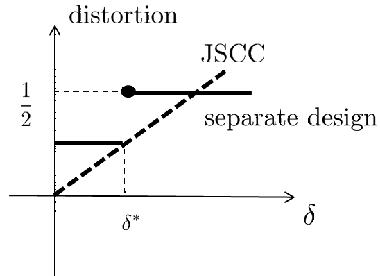
Let us discuss some issues and open problems in the classical compression theory. First, for the *compression* the standard results in lossless compression apply well for text files. The lossy compression theory, however, relies on the independence assumption and on separable distortion metrics. Because of that, while the scalar quantization theory has been widely used in practice (in the form of analog-to-digital converters, ADCs), the vector quantizers (rate-distortion) theory so far has not been employed. The assumptions of the rate-distortion theory can be seen to be especially problematic in the case of compressing digital images, which evidently have very strong spatial correlation compared to 1D signals. (For example, the first sentence and the last in Tolstoy's novel are pretty uncorrelated. But the regions in the upper-left and bottom-right corners of one image can be strongly correlated. At the same time, the uncompressed size of the novel and the image could be easily equal.) Thus, for practicing the lossy compression of videos and images the key problem is that of coming up with a good "whitening" bases, which is an art still being refined.

For the joint-source-channel coding, the separation principle has definitely been a guiding light for the entire development of digital information technology. But this now ubiquitous solution that Shannon's separation has professed led to a rather undesirable feature of dropped cellular calls (as opposed to slowly degraded quality of the old analog telephones) or "snow screen" on TV whenever the SNR drops below a certain limit. That is the separated systems can be very unstable, or lacks *graceful degradation*. To sketch this effect consider an example of JSCC, where the source distribution is $\text{Ber}(\frac{1}{2})$ and the channel is $= \text{BSC}_\delta$. For it we can consider two solutions:

- 1 a separate compressor and channel encoder designed for $\frac{R(D)}{C(\delta)} = 1$
- 2 a simple JSCC with $\rho = 1$ which transmits "uncoded" data, i.e. $X_i = S_i$.

26.4 What is lacking in classical lossy compression? 449

As a function of δ the resulting distortion (at large blocklength) will look like the solid and dashed lines in this graph:



no graceful degradation of separately designed source channel code

We can see that below $\delta < \delta^*$ the separated solution is much preferred since it achieves zero distortion. But at $\delta > \delta^*$ it undergoes a catastrophic failure and distortion becomes $1/2$ (that is, we observe pure noise). At the same time the simple “uncoded” JSCC has its distortion decreasing *gracefully*. It has been a long-standing problem since the early days of information theory to find schemes that would interpolate between these two extreme solutions.

Even theoretically the problem of JSCC still contains great many mysteries. For example, in Section 22.5 we described refined expansion of the channel coding rate as a function of block-length. However, similar expansions for the JSCC are not available. In fact, even showing that convergence of the $\frac{k}{n}$ to the ultimate limit of $\frac{C}{R(D)}$ happens at the speed of $\Theta(1/\sqrt{n})$ has only been demonstrated recently [174] and only for one special case (of a binary source and BSC_δ channel as in the example above).

27 Metric entropy

In the previous chapters of this part we discussed optimal quantization of random vectors in both fixed and high dimensions. Complementing this average-case perspective, the topic of this chapter is on the deterministic (worst-case) theory of quantization. The main object of interest is the *metric entropy* of a set, which allows us to answer two key questions (a) covering number: the minimum number of points to cover a set up to a given accuracy; (b) packing number: the maximal number of elements of a given set with a prescribed minimum pairwise distance.

The foundational theory of metric entropy were put forth by Kolmogorov, who, together with his students, also determined the behavior of metric entropy in a variety of problems for both finite and infinite dimensions. Kolmogorov's original interest in this subject stems from Hilbert's 13th problem, which concerns the possibility or impossibility of representing multi-variable functions as compositions of functions of fewer variables. It turns out that the theory of metric entropy can provide a surprisingly simple and powerful resolution to such problems. Over the years, metric entropy has found numerous connections to and applications in other fields such as approximation theory, empirical processes, small-ball probability, mathematical statistics, and machine learning. In particular, metric entropy will be featured prominently in Part VI of this book, wherein we discuss its applications to proving both lower and upper bounds for statistical estimation.

This chapter is organized as follows. Section 27.1 provides basic definitions and explains the fundamental connections between covering and packing numbers. In Section 27.2 we study metric entropy in finite-dimensional spaces and a popular approach for bounding the metric entropy known as the volume bound. To demonstrate the limitations of the volume method and the associated high-dimensional phenomenon, in Section 27.3 we discuss a few other approaches through concrete examples. Infinite-dimensional spaces are treated next for smooth functions in Section 27.4 (wherein we also discuss the application to Hilbert's 13th problem) and Hilbert spaces in Section 27.5 (wherein we also discuss the application to empirical processes). Section 27.6 gives an exposition of the connections between metric entropy and the small-ball problem in probability theory. Finally, in Section 27.7 we circle back to rate-distortion theory and discuss how it is related to metric entropy and how information-theoretic methods can be useful for the latter.

27.1 Covering and packing

Definition 27.1. Let (V, d) be a metric space and $\Theta \subset V$.

27.1 Covering and packing 451

- We say $\{v_1, \dots, v_N\} \subset V$ is an ϵ -covering of Θ if $\Theta \subset \cup_{i=1}^N B(v_i, \epsilon)$, where $B(v, \epsilon) \triangleq \{u \in V : d(u, v) \leq \epsilon\}$ is the (closed) ball of radius ϵ centered at v ; or equivalently, $\forall \theta \in \Theta, \exists i \in [N]$ such that $d(\theta, v_i) \leq \epsilon$.
- We say $\{\theta_1, \dots, \theta_M\} \subset \Theta$ is an ϵ -packing of Θ if $\min_{i \neq j} \|\theta_i - \theta_j\| > \epsilon$ ¹; or equivalently, the balls $\{B(\theta_i, \epsilon/2) : i \in [M]\}$ are disjoint.

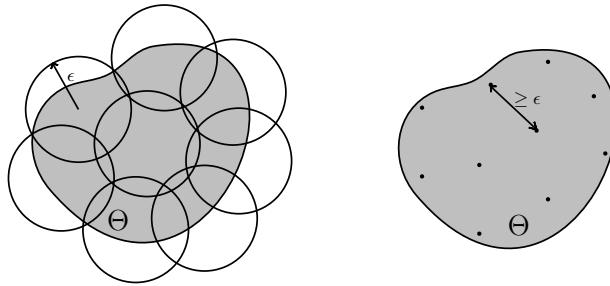


Figure 27.1 Illustration of ϵ -covering and ϵ -packing.

Upon defining ϵ -covering and ϵ -packing, a natural question concerns the size of the optimal covering and packing, leading to the definition of *covering* and *packing numbers*:

$$N(\Theta, d, \epsilon) \triangleq \min\{n : \exists \epsilon\text{-covering of } \Theta \text{ of size } n\} \quad (27.1)$$

$$M(\Theta, d, \epsilon) \triangleq \max\{m : \exists \epsilon\text{-packing of } \Theta \text{ of size } m\} \quad (27.2)$$

with $\min \emptyset$ understood as ∞ ; we will sometimes abbreviate these as $N(\epsilon)$ and $M(\epsilon)$ for brevity. Similar to volume and width, covering and packing numbers provide a meaningful measure for the “massiveness” of a set. The major focus of this chapter is to understanding their behavior in both finite and infinite-dimensional spaces as well as their statistical applications.

Some remarks are in order.

- Monotonicity: $N(\Theta, d, \epsilon)$ and $M(\Theta, d, \epsilon)$ are non-decreasing and right-continuous functions of ϵ . Furthermore, both are non-decreasing in Θ with respect to set inclusion.
- Finiteness: Θ is totally bounded (e.g. compact) if $N(\Theta, d, \epsilon) < \infty$ for all $\epsilon > 0$. For Euclidean spaces, this is equivalent to Θ being bounded, namely, $\text{diam}(\Theta) < \infty$ (cf. (5.4)).
- The logarithm of the covering and packing numbers are commonly referred to as *metric entropy*. In particular, $\log M(\epsilon)$ and $\log N(\epsilon)$ are called ϵ -entropy and ϵ -capacity in [176]. Quantitative connections between metric entropy and other information measures are explored in Section 27.7.
- Widely used in the literature of functional analysis [229, 199], the notion of *entropy numbers* essentially refers to the inverse of the metric entropy: The k th entropy number of Θ is $e_k(\Theta) \triangleq \inf\{\epsilon : N(\Theta, d, \epsilon) \leq 2^k\}$. In particular, $e_1(\Theta) = \text{rad}(\Theta)$, the radius of Θ defined in (5.5).

¹ Notice we imposed strict inequality for convenience.

Remark 27.1. Unlike the packing number $M(\Theta, d, \epsilon)$, the covering number $N(\Theta, d, \epsilon)$ defined in (27.1) depends implicitly on the ambient space $V \supset \Theta$, since, per Definition 27.1), an ϵ -covering is required to be a subset of V rather than Θ . Nevertheless, as the next Theorem 27.3 shows, this dependency on V has almost no effect on the behavior of the covering number.

As an alternative to (27.1), we can define $N'(\Theta, d, \epsilon)$ as the size of the minimal ϵ -covering of Θ that is also a subset of Θ , which is closely related to the original definition as

$$N(\Theta, d, \epsilon) \leq N'(\Theta, d, \epsilon) \leq N(\Theta, d, \epsilon/2) \quad (27.3)$$

Here, the left inequality is obvious. To see the right inequality,² let $\{\theta_1, \dots, \theta_N\}$ be an $\frac{\epsilon}{2}$ -covering of Θ . We can project each θ_i to Θ by defining $\theta'_i = \operatorname{argmin}_{u \in \Theta} d(\theta_i, u)$. Then $\{\theta'_1, \dots, \theta'_N\} \subset \Theta$ constitutes an ϵ -covering. Indeed, for any $\theta \in \Theta$, we have $d(\theta, \theta_i) \leq \epsilon/2$ for some θ_i . Then $d(\theta, \theta'_i) \leq d(\theta, \theta_i) + d(\theta_i, \theta'_i) \leq 2d(\theta, \theta_i) \leq \epsilon$. On the other hand, the N' covering numbers need not be monotone with respect to set inclusion.

The relation between the covering and packing numbers is described by the following fundamental result.

Theorem 27.2 (Kolomogrov-Tikhomirov [176]).

$$M(\Theta, d, 2\epsilon) \leq N(\Theta, d, \epsilon) \leq M(\Theta, d, \epsilon). \quad (27.4)$$

Proof. To prove the right inequality, fix a maximal packing $E = \{\theta_1, \dots, \theta_M\}$. Then $\forall \theta \in \Theta \setminus E$, $\exists i \in [M]$, such that $d(\theta, \theta_i) \leq \epsilon$ (for otherwise we can obtain a bigger packing by adding θ). Hence E must an ϵ -covering (which is also a subset of Θ). Since $N(\Theta, d, \epsilon)$ is the minimal size of all possible coverings, we have $M(\Theta, d, \epsilon) \geq N(\Theta, d, \epsilon)$.

We next prove the left inequality by contradiction. Suppose there exists a 2ϵ -packing $\{\theta_1, \dots, \theta_M\}$ and an ϵ -covering $\{x_1, \dots, x_N\}$ such that $M \geq N + 1$. Then by the pigeonhole principle, there exist distinct θ_i and θ_j belonging to the same ϵ -ball $B(x_k, \epsilon)$. By triangle inequality, $d(\theta_i, \theta_j) \leq 2\epsilon$, which is a contradiction since $d(\theta_i, \theta_j) > 2\epsilon$ for a 2ϵ -packing. Hence the size of any 2ϵ -packing is at most that of any ϵ -covering, that is, $M(\Theta, d, 2\epsilon) \leq N(\Theta, d, \epsilon)$. \square

The significance of (27.4) is that it shows that the small- ϵ behavior of the covering and packing numbers are essentially the same. In addition, the right inequality therein, namely, $N(\epsilon) \leq M(\epsilon)$, deserves some special mention. As we will see next, it is oftentimes easier to prove negative results (lower bound on the minimal covering or upper bound on the maximal packing) than positive results which require explicit construction. When used in conjunction with the inequality $N(\epsilon) \leq M(\epsilon)$, these converses turn into achievability statements,³ leading to many useful bounds on metric entropy (e.g. the volume bound in Theorem 27.4 and the Gilbert-Varshamov bound

² Another way to see this is from Theorem 27.3: Note that the right inequality in (27.4) yields a ϵ -covering that is included in Θ . Together with the left inequality, we get $N'(\epsilon) \leq M(\epsilon) \leq N(\epsilon/2)$.

³ This is reminiscent of duality-based argument in optimization: To bound a minimization problem from above, instead of constructing an explicit feasible solution, a fruitful approach is to equate it with the dual problem (maximization) and bound this maximum from above.

27.2 Finite-dimensional space and volume bound 453

Theorem 27.7 in the next section). Revisiting the proof of Theorem 27.3, we see that this logic actually corresponds to a *greedy construction* (greedily increase the packing until no points can be added).

27.2 Finite-dimensional space and volume bound

A commonly used method to bound metric entropy in finite dimensions is in terms of *volume ratio*. Consider the d -dimensional Euclidean space $V = \mathbb{R}^d$ with metric given by an arbitrary norm $d(x, y) = \|x - y\|$. We have the following result.

Theorem 27.3. *Let $\|\cdot\|$ be an arbitrary norm on \mathbb{R}^d and $B = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ the corresponding unit norm ball. Then for any $\Theta \subset \mathbb{R}^d$,*

$$\left(\frac{1}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)} \stackrel{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \stackrel{(b)}{\leq} \frac{\text{vol}(\Theta + \frac{\epsilon}{2}B)}{\text{vol}(\frac{\epsilon}{2}B)} \stackrel{(c)}{\leq} \left(\frac{3}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)}.$$

where (c) holds under the extra condition that Θ is convex and contains ϵB .

Proof. To prove (a), consider an ϵ -covering $\Theta \subset \cup_{i=1}^N B(\theta_i, \epsilon)$. Applying the union bound yields

$$\text{vol}(\Theta) \leq \text{vol}\left(\cup_{i=1}^N B(\theta_i, \epsilon)\right) \leq \sum_{i=1}^N \text{vol}(B(\theta_i, \epsilon)) = N\epsilon^d \text{vol}(B),$$

where the last step follows from the translation-invariance and scaling property of volume.

To prove (b), consider an ϵ -packing $\{\theta_1, \dots, \theta_M\} \subset \Theta$ such that the balls $B(\theta_i, \epsilon/2)$ are disjoint. Since $\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2) \subset \Theta + \frac{\epsilon}{2}B$, taking the volume on both sides yields

$$\text{vol}\left(\Theta + \frac{\epsilon}{2}B\right) \geq \text{vol}\left(\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2)\right) = M\text{vol}\left(\frac{\epsilon}{2}B\right).$$

This proves (b).

Finally, (c) follows from the following two statements: (1) if $\epsilon B \subset \Theta$, then $\Theta + \frac{\epsilon}{2}B \subset \Theta + \frac{1}{2}\Theta$; and (2) if Θ is convex, then $\Theta + \frac{1}{2}\Theta = \frac{3}{2}\Theta$. We only prove (2). First, $\forall \theta \in \frac{3}{2}\Theta$, we have $\theta = \frac{1}{3}\theta + \frac{2}{3}\theta$, where $\frac{1}{3}\theta \in \frac{1}{2}\Theta$ and $\frac{2}{3}\theta \in \Theta$. Thus $\frac{3}{2}\Theta \subset \Theta + \frac{1}{2}\Theta$. On the other hand, for any $x \in \Theta + \frac{1}{2}\Theta$, we have $x = y + \frac{1}{2}z$ with $y, z \in \Theta$. By the convexity of Θ , $\frac{2}{3}x = \frac{2}{3}y + \frac{1}{3}z \in \Theta$. Hence $x \in \frac{3}{2}\Theta$, implying $\Theta + \frac{1}{2}\Theta \subset \frac{3}{2}\Theta$. \square

Remark 27.2. Similar to the proof of (a) in Theorem 27.4, we can start from $\Theta + \frac{\epsilon}{2}B \subset \cup_{i=1}^N B(\theta_i, \frac{3\epsilon}{2})$ to conclude that

$$(2/3)^d \leq \frac{N(\Theta, \|\cdot\|, \epsilon)}{\text{vol}(\Theta + \frac{\epsilon}{2}B)/\text{vol}(\epsilon B)} \leq 2^d.$$

In other words, the volume of the fattened set $\Theta + \frac{\epsilon}{2}$ determines the metric entropy up to constants that only depend on the dimension. We will revisit this reasoning in Section 27.6 to adapt the volumetric estimates to infinite dimensions where this fattening step becomes necessary.

Next we discuss several applications of Theorem 27.4.

Corollary 27.4 (Metric entropy of balls and spheres). *Let $\|\cdot\|$ be an arbitrary norm on \mathbb{R}^d . Let $B \equiv B_{\|\cdot\|} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ and $S \equiv S_{\|\cdot\|} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be the corresponding unit ball and unit sphere. Then for $\epsilon < 1$,*

$$\left(\frac{1}{\epsilon}\right)^d \leq N(B, \|\cdot\|, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^d \quad (27.5)$$

$$\left(\frac{1}{2\epsilon}\right)^{d-1} \leq N(S, \|\cdot\|, \epsilon) \leq 2d \left(1 + \frac{1}{\epsilon}\right)^{d-1} \quad (27.6)$$

where the left inequality in (27.6) holds under the extra assumption that $\|\cdot\|$ is an absolute norm (invariant to sign changes of coordinates).

Proof. For balls, the estimate (27.5) directly follows from Theorem 27.4 since $B + \frac{\epsilon}{2}B = (1 + \frac{\epsilon}{2})B$. Next we consider the spheres. Applying (b) in Theorem 27.4 yields

$$\begin{aligned} N(S, \|\cdot\|, \epsilon) &\leq M(S, \|\cdot\|, \epsilon) \leq \frac{\text{vol}(S + \epsilon B)}{\text{vol}(\epsilon B)} \leq \frac{\text{vol}((1 + \epsilon)B) - \text{vol}((1 - \epsilon)B)}{\text{vol}(\epsilon B)} \\ &= \frac{(1 + \epsilon)^d - (1 - \epsilon)^d}{\epsilon^d} = \frac{d}{\epsilon^d} \int_{-\epsilon}^{\epsilon} (1 + x)^{d-1} dx \leq 2d \left(1 + \frac{1}{\epsilon}\right)^{d-1}. \end{aligned}$$

where the third inequality applies $S + \epsilon B \subset ((1 + \epsilon)B) \setminus ((1 - \epsilon)B)$ by triangle inequality.

Finally, we prove the lower bound in (27.6) for an absolute norm $\|\cdot\|$. To this end one cannot directly invoke the lower bound in Theorem 27.4 as the sphere has zero volume. Note that $\|\cdot\|' \triangleq \|\cdot, 0\|$ defines a norm on \mathbb{R}^{d-1} . We claim that every ϵ -packing in $\|\cdot\|'$ for the unit $\|\cdot\|'$ -ball induces an ϵ -packing in $\|\cdot\|$ for the unit $\|\cdot\|$ -sphere. Fix $x \in \mathbb{R}^{d-1}$ such that $\|(x, 0)\| \leq 1$ and define $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ by $f(y) = \|(x, y)\|$. Using the fact that $\|\cdot\|$ is an absolute norm, it is easy to verify that f is a continuous increasing function with $f(0) \leq 1$ and $f(\infty) = \infty$. By the mean value theorem, there exists y_x , such that $\|(x, y_x)\| = 1$. Finally, for any ϵ -packing $\{x'_1, \dots, x'_M\}$ of the unit ball $B_{\|\cdot\|'}$ with respect to $\|\cdot\|'$, setting $x'_i = (x_i, y_{x_i})$ we have $\|x'_i - x'_j\| \geq \|(x_i - x_j, 0)\| = \|x_i - x_j\|' \geq \epsilon$. This proves

$$M(S_{\|\cdot\|}, \|\cdot\|, \epsilon) \geq M(B_{\|\cdot\|'}, \|\cdot\|', \epsilon).$$

Then the left inequality of (27.6) follows from those of (27.4) and (27.5). \square

Remark 27.3. Several remarks on Corollary 27.1 are in order:

(a) Using (27.5), we see that for any compact Θ with nonempty interior, we have

$$N(\Theta, \|\cdot\|, \epsilon) \asymp M(\Theta, \|\cdot\|, \epsilon) \asymp \frac{1}{\epsilon^d} \quad (27.7)$$

for small ϵ , with proportionality constants depending on both Θ and the norm. In fact, the sharp constant is also known to exist. It is shown in [?, Theorem IX] that there exists a constant τ

27.2 Finite-dimensional space and volume bound 455

depending only on $\|\cdot\|$ and the dimension, such that

$$M(\Theta, \|\cdot\|, 2\epsilon) = (\tau + o(1)) \frac{\text{vol}(\Theta)}{\text{vol}(B)} \frac{1}{\epsilon^d}$$

holds for any Θ with positive volume. This constant τ is the *maximal sphere packing density* in \mathbb{R}^d (the proportion of the whole space covered by the balls in the packing – see [259, Chapter 1] for a formal definition); a similar result and interpretation hold for the covering number as well. Computing or bounding the value of τ is extremely difficult and remains open except for some special cases.⁴ For more on this subject see the monographs [259, 71].

- (b) The result (27.6) for spheres suggests that one may expect the metric entropy for a smooth manifold Θ to behave as $(\frac{1}{\epsilon})^{\dim}$, where \dim stands for the dimension of Θ as opposed to the ambient dimension. This is indeed true in many situations, for example, in the context of matrices, for the orthogonal group $O(d)$, unitary group $U(d)$, and Grassmannian manifolds [290, 291], in which case \dim corresponds to the “degrees of freedom” (for example, $\dim = d(d-1)/2$ for $O(d)$). More generally, for an arbitrary set Θ , one may define the limit $\lim_{\epsilon \rightarrow 0} \frac{\log N(\Theta, \|\cdot\|, \epsilon)}{\log \frac{1}{\epsilon}}$ as its dimension (known as the *Minkowski dimension* or *box-counting dimension*). For sets of a fractal nature, this dimension can be a non-integer (e.g. $\log_2 3$ for the Cantor set).
- (c) Since all norms on Euclidean space are equivalent (within multiplicative constant factors depending on dimension), the small- ϵ behavior in (27.7) holds for any norm as long as the dimension d is fixed. However, this result does not capture the full picture in *high dimensions* when ϵ is allowed to depend on d . Understanding these high-dimensional phenomena requires us to go beyond volume methods. See Section 27.3 for details.

Next we switch our attention to the discrete case of Hamming space. The following theorem bounds its packing number $M(\mathbb{F}_2^d, d_H, r) \equiv M(\mathbb{F}_2^d, r)$, namely, the maximal number of binary codewords of length d with a prescribed minimum distance $r+1$.⁵ This is a central question in coding theory, wherein the lower and upper bounds below are known as the *Gilbert-Varshamov bound* and the *Hamming bound*, respectively.

Theorem 27.5. *For any integer $1 \leq r \leq d-1$,*

$$\frac{2^d}{\sum_{i=0}^r \binom{d}{i}} \leq M(\mathbb{F}_2^d, r) \leq \frac{2^d}{\sum_{i=0}^{\lfloor r/2 \rfloor} \binom{d}{i}}. \quad (27.8)$$

Proof. Both inequalities in (27.8) follow from the same argument as that in Theorem 27.4, with \mathbb{R}^d replaced by \mathbb{F}_2^d and volume by the counting measure (which is translation invariant). \square

Of particular interest to coding theory is the asymptotic regime of $d \rightarrow \infty$ and $r = \rho d$ for some constant $\rho \in (0, 1)$. Using the asymptotics of the binomial coefficients (cf. Proposition 1.6), the

⁴ For example, it is easy to show that $\tau = 1$ for both ℓ_∞ and ℓ_1 balls in any dimension since cubes can be subdivided into smaller cubes; for ℓ_2 -ball in $d = 2$, $\tau = \frac{\pi}{\sqrt{12}}$ is the famous result of L. Fejes Tóth on the optimality of hexagonal arrangement for circle packing [259].

⁵ Recall that the packing number in Definition 27.1 is defined with a strict inequality.

Hamming and Gilbert-Varshamov bounds translate to

$$2^{d(1-h(\rho)+o(d)} \leq M(\mathbb{F}_2^d, \rho d) \leq 2^{d(1-h(\rho/2))+o(d)}.$$

Finding the exact exponent is one of the most significant open questions in coding theory. The best upper bound to date is due to McEliece, Rodemich, Rumsey and Welch [209] using the technique of linear programming relaxation.

In contrast, the corresponding covering problem in Hamming space is much simpler, as we have the following tight result

$$N(\mathbb{F}_2^d, \rho d) = 2^{dR(\rho)+o(d)}, \quad (27.9)$$

where $R(\rho) = (1 - h(\rho))_+$ is the rate-distortion function of $\text{Ber}(\frac{1}{2})$ from Theorem 26.1. Although this does not automatically follow from the rate-distortion theory, it can be shown using similar argument – see Exercise V.11.

Finally, we state a lower bound on the packing number of Hamming spheres, which is needed for subsequent application in sparse estimation (Exercise VI.11) and useful as basic building blocks for computing metric entropy in more complicated settings (Theorem 27.9).

Theorem 27.6 (Gilbert-Varshamov bound for Hamming spheres). *Denote by*

$$S_k^d = \{x \in \mathbb{F}_2^d : w_H(x) = k\} \quad (27.10)$$

the Hamming sphere of radius $0 \leq k \leq d$. Then

$$M(S_k^d, r) \geq \frac{\binom{d}{k}}{\sum_{i=0}^r \binom{d}{i}}. \quad (27.11)$$

In particular,

$$\log M(S_k^d, k/2) \geq \frac{k}{2} \log \frac{d}{2ek}. \quad (27.12)$$

Proof. Again (27.11) follows from the volume argument. To verify (27.12), note that for $r \leq d/2$, we have $\sum_{i=0}^r \binom{d}{i} \leq \exp(dh(\frac{r}{d}))$ (see Theorem 8.3 or (15.19) with $p = 1/2$). Using $h(x) \leq x \log \frac{e}{x}$ and $\binom{d}{k} \geq (\frac{d}{k})^k$, we conclude (27.12) from (27.11). \square

27.3 Beyond the volume bound

The volume bound in Theorem 27.4 provides a useful tool for studying metric entropy in Euclidean spaces. As a result, as $\epsilon \rightarrow 0$, the covering number of any set with non-empty interior always grows exponentially in d as $(\frac{1}{\epsilon})^d$ – cf. (27.7). This asymptotic result, however, has its limitations and does not apply if the dimension d is large and ϵ scales with d . In fact, one expects that there is some critical threshold of ϵ depending on the dimension d , below which the exponential asymptotics is tight, and above which the covering number can grow polynomially in d . This high-dimensional phenomenon is not fully captured by the volume method.

27.3 Beyond the volume bound 457

As a case in point, consider the maximum number of ℓ_2 -balls of radius ϵ packed into the unit ℓ_1 -ball, namely, $M(B_1, \|\cdot\|_2, \epsilon)$. (Recall that B_p denotes the unit ℓ_p -ball in \mathbb{R}^d with $1 \leq p \leq \infty$.) We have studied the metric entropy of arbitrary norm balls under the same norm in Corollary 27.1, where the specific value of the volume was canceled from the volume ratio. Here, although ℓ_1 and ℓ_2 norms are equivalent in the sense that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$, this relationship is too loose when d is large.

Let us start by applying the volume method in Theorem 27.4:

$$\frac{\text{vol}(B_1)}{\text{vol}(\epsilon B_2)} \leq N(B_1, \|\cdot\|_2, \epsilon) \leq M(B_1, \|\cdot\|_2, \epsilon) \leq \frac{\text{vol}(B_1 + \frac{\epsilon}{2}B_2)}{\text{vol}(\frac{\epsilon}{2}B_2)}.$$

Applying the formula for the volume of a unit ℓ_q -ball in \mathbb{R}^d :

$$\text{vol}(B_q) = \frac{\left[2\Gamma\left(1 + \frac{1}{q}\right)\right]^d}{\Gamma\left(1 + \frac{d}{q}\right)}, \quad (27.13)$$

we get⁶ $\text{vol}(B_1) = 2^d/d!$ and $\text{vol}(B_2) = \frac{\pi^d}{\Gamma(1+d/2)}$, which yield, by Stirling approximation,

$$\text{vol}(B_1)^{1/d} \asymp \frac{1}{d}, \quad \text{vol}(B_2)^{1/d} \asymp \frac{1}{\sqrt{d}}. \quad (27.14)$$

Then for some absolute constant C ,

$$M(B_1, \|\cdot\|_2, \epsilon) \leq \frac{\text{vol}(B_1 + \frac{\epsilon}{2}B_2)}{\text{vol}(\frac{\epsilon}{2}B_2)} \leq \frac{\text{vol}((1 + \frac{\epsilon\sqrt{d}}{2})B_1)}{\text{vol}(\frac{\epsilon}{2}B_2)} \leq \left(C\left(1 + \frac{1}{\epsilon\sqrt{d}}\right)\right)^d, \quad (27.15)$$

where the second inequality follows from $B_2 \subset \sqrt{d}B_1$ by Cauchy-Schwarz inequality. (This step is tight in the sense that $\text{vol}(B_1 + \frac{\epsilon}{2}B_2)^{1/d} \gtrsim \max\{\text{vol}(B_1)^{1/d}, \frac{\epsilon}{2}\text{vol}(B_2)^{1/d}\} \asymp \max\{\frac{1}{d}, \frac{\epsilon}{\sqrt{d}}\}$.) On the other hand, for some absolute constant c ,

$$M(B_1, \|\cdot\|_2, \epsilon) \geq \frac{\text{vol}(B_1)}{\text{vol}(\epsilon B_2)} = \left(\frac{1}{\epsilon}\right)^d \frac{\text{vol}(B_1)}{\text{vol}(B_2)} = \left(\frac{c}{\epsilon\sqrt{d}}\right)^d. \quad (27.16)$$

Overall, for $\epsilon \leq \frac{1}{\sqrt{d}}$, we have $M(B_1, \|\cdot\|_2, \epsilon)^{1/d} \asymp \frac{1}{\epsilon\sqrt{d}}$; however, the lower bound trivializes and the upper bound (which is exponential in d) is loose in the regime of $\epsilon \gg \frac{1}{\sqrt{d}}$, which requires different methods than volume calculation. The following result describes the complete behavior of this metric entropy. In view of Theorem 27.3, we will go back and forth between the covering and packing numbers in the argument.

Theorem 27.7. *For $0 < \epsilon < 1$ and $d \in \mathbb{N}$,*

$$\log M(B_1, \|\cdot\|_2, \epsilon) \asymp \begin{cases} d \log \frac{\epsilon}{\epsilon^2 d} & \epsilon \leq \frac{1}{\sqrt{d}} \\ \frac{1}{\epsilon^2} \log(e\epsilon^2 d) & \epsilon \geq \frac{1}{\sqrt{d}} \end{cases}.$$

⁶ For B_1 this can be proved directly by noting that B_1 consists 2^d disjoint “copies” of the simplex whose volume is $1/d!$ by induction on d .

Proof. The case of $\epsilon \leq \frac{1}{\sqrt{d}}$ follows from earlier volume calculation (27.15)–(27.16). Next we focus on $\frac{1}{\sqrt{d}} \leq \epsilon < 1$.

For the upper bound, we construct an ϵ -covering in ℓ_2 by quantizing each coordinate. Without loss of generality, assume that $\epsilon < 1/4$. Fix some $\delta < 1$. For each $\theta \in B_1$, there exists $x \in (\delta\mathbb{Z}^d) \cap B_1$ such that $\|x - \theta\|_\infty \leq \delta$. Then $\|x - \theta\|_2^2 \leq \|x - \theta\|_1 \|x - \theta\|_\infty \leq 2\delta$. Furthermore, x/δ belongs to the set

$$\mathcal{Z} = \left\{ z \in \mathbb{Z}^d : \sum_{i=1}^d |z_i| \leq k \right\} \quad (27.17)$$

with $k = \lfloor 1/\delta \rfloor$. Note that each $z \in \mathcal{Z}$ has at most k nonzeros. By enumerating the number of non-negative solutions (stars and bars calculation) and the sign pattern, we have⁷ $|\mathcal{Z}| \leq 2^{k \wedge d} \binom{d-1+k}{k}$. Finally, picking $\delta = \epsilon^2/2$, we conclude that $N(B_1, \|\cdot\|_2, \epsilon) \leq |\mathcal{Z}| \leq \left(\frac{2e(d+k)}{k}\right)^k$ as desired. (Note that this method also recovers the volume bound for $\epsilon \leq \frac{1}{\sqrt{d}}$, in which case $k \leq d$.)

For the lower bound, note that $M(B_1, \|\cdot\|_2, \sqrt{2}) \geq 2d$ by considering $\pm e_1, \dots, \pm e_d$. So it suffices to consider $d \geq 8$. We construct a packing of B_1 based on a packing of the Hamming sphere. Without loss of generality, assume that $\epsilon > \frac{1}{4\sqrt{d}}$. Fix some $1 \leq k \leq d$. Applying the Gilbert-Varshamov bound in Theorem 27.8, in particular, (27.12), there exists a $k/2$ -packing $\{x_1, \dots, x_M\} \subset S_k^d = \{x \in \{0, 1\}^d : \sum_{i=1}^d x_i = k\}$ and $\log M \geq \frac{k}{2} \log \frac{d}{2ek}$. Scale the Hamming sphere to fit the ℓ_1 -ball by setting $\theta_i = x_i/k$. Then $\theta_i \in B_1$ and $\|\theta_i - \theta_j\|_2^2 = \frac{1}{k^2} d_H(x_i, x_j) \geq \frac{1}{2k}$ for all $i \neq j$. Choosing $k = \lfloor \frac{1}{\epsilon^2} \rfloor$ which satisfies $k \leq d/8$, we conclude that $\{\theta_1, \dots, \theta_M\}$ is a $\frac{\epsilon}{2}$ -packing of B_1 in $\|\cdot\|_2$ as desired. \square

The above elementary proof can be adapted to give the following more general result (see Exercise V.12): Let $1 \leq p < q \leq \infty$. For all $0 < \epsilon < 1$ and $d \in \mathbb{N}$,

$$\log M(B_p, \|\cdot\|_q, \epsilon) \asymp_{p,q} \begin{cases} d \log \frac{\epsilon}{\epsilon^s d} & \epsilon \leq d^{-1/s} \\ \frac{1}{\epsilon^s} \log(e\epsilon^s d) & \epsilon \geq d^{-1/s} \end{cases}, \quad \frac{1}{s} \triangleq \frac{1}{p} - \frac{1}{q}. \quad (27.18)$$

In the remainder of this section, we discuss a few generic results in connection to Theorem 27.9, in particular, metric entropy upper bounds via the *Sudakov minorization* and *Maurey's empirical method*, as well as the duality of metric entropy in Euclidean spaces.

27.3.1 Sudakov minorization

Theorem 27.8 (Sudakov minoration). *Define the Gaussian width of $\Theta \subset \mathbb{R}^d$ as⁸*

$$w(\Theta) \triangleq \mathbb{E} \sup_{\theta \in \Theta} \langle \theta, Z \rangle, \quad Z \sim N(0, I_d). \quad (27.19)$$

⁷ By enumerating the support and counting positive solutions, it is easy to show that $|\mathcal{Z}| = \sum_{i=0}^d 2^{d-i} \binom{d}{i} \binom{k}{d-i}$.

⁸ To avoid measurability difficulty, $w(\Theta)$ should be understood as $\sup_{T \subset \Theta, |T| < \infty} \mathbb{E} \max_{\theta \in T} \langle \theta, Z \rangle$.

27.3 Beyond the volume bound 459

For any $\Theta \subset \mathbb{R}^d$,

$$w(\Theta) \gtrsim \sup_{\epsilon > 0} \epsilon \sqrt{\log M(\Theta, \|\cdot\|_2, \epsilon)}. \quad (27.20)$$

The preceding theorem relates the Gaussian width to the metric entropy, both of which are meaningful measure of the massiveness of a set. The following complementary result is due to R. Dudley. (See [229, Theorem 5.6] for both results.)

$$w(\Theta) \lesssim \int_0^\infty \sqrt{\log M(\Theta, \|\cdot\|_2, \epsilon)} d\epsilon. \quad (27.21)$$

Understanding the maximum of a Gaussian process is a field on its own; see the monograph [294]. In this section we focus on the upper bound (27.20) in order to develop upper bound for metric entropy using the Gaussian width.

The proof of Theorem 27.10 relies on the following Gaussian comparison lemma of Slepian (whom we have encountered earlier in Theorem 11.18). For a self-contained proof see [61]. See also [229, Lemma 5.7, p. 70] for a simpler proof of a weaker version $\mathbb{E} \max X_i \leq 2\mathbb{E} \max Y_i$, which suffices for our purposes.

Lemma 27.9 (Slepian's lemma). *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be Gaussian random vectors. If $\mathbb{E}(Y_i - Y_j)^2 \leq \mathbb{E}(X_i - X_j)^2$ for all i, j , then $\mathbb{E} \max Y_i \leq \mathbb{E} \max X_i$.*

We also need the result bounding the expectation of the maximum of n Gaussian random variables (see also Exercise I.45).

Lemma 27.10. *Let Z_1, \dots, Z_n be distributed as $\mathcal{N}(0, 1)$. Then*

$$\mathbb{E} \left[\max_{i \in [n]} Z_i \right] \leq \sqrt{2 \log n}. \quad (27.22)$$

In addition, if Z_1, \dots, Z_n are iid, then

$$\mathbb{E} \left[\max_{i \in [n]} Z_i \right] = \sqrt{2 \log n} (1 + o(1)). \quad (27.23)$$

Proof. By Jensen's inequality, for any $t > 0$,

$$e^{t \mathbb{E}[\max_i Z_i]} \leq \mathbb{E}[e^{t \max_i Z_i}] = \mathbb{E}[\max_i e^{t Z_i}] \leq \sum_i \mathbb{E}[e^{t Z_i}] = n e^{t^2/2}.$$

Therefore

$$\mathbb{E}[\max_i Z_i] \leq \frac{\log n}{t} + \frac{t}{2}.$$

Choosing $t = \sqrt{2 \log n}$ yields (27.22). Next, assume that Z_i are iid. For any $t > 0$,

$$\begin{aligned} \mathbb{E}[\max_i Z_i] &\geq t \mathbb{P}[\max_i Z_i \geq t] + \mathbb{E}[\max_i Z_i \mathbf{1}_{\{Z_1 < 0\}} \mathbf{1}_{\{Z_2 < 0\}} \dots \mathbf{1}_{\{Z_n < 0\}}] \\ &\geq t(1 - (1 - \Phi^c(t))^n) + \mathbb{E}[Z_1 \mathbf{1}_{\{Z_1 < 0\}} \mathbf{1}_{\{Z_2 < 0\}} \dots \mathbf{1}_{\{Z_n < 0\}}]. \end{aligned}$$

where $\Phi^c(t) = \mathbb{P}[Z_1 \geq t]$ is the normal tail probability. The second term equals $2^{-(n-1)}\mathbb{E}[Z_1 1_{\{Z_1 < 0\}}] = o(1)$. For the first term, recall that $\Phi^c(t) \geq \frac{t}{1+t^2}\varphi(t)$ (Exercise V.9). Choosing $t = \sqrt{(2-\epsilon)\log n}$ for small $\epsilon > 0$ so that $\Phi^c(t) = \omega(\frac{1}{n})$ and hence $\mathbb{E}[\max_i Z_i] \geq \sqrt{(2-\epsilon)\log n}(1 + o(1))$. By the arbitrariness of $\epsilon > 0$, the lower bound part of (27.23) follows. \square

Proof of Theorem 27.10. Let $\{\theta_1, \dots, \theta_M\}$ be an optimal ϵ -packing of Θ . Let $X_i = \langle \theta_i, Z \rangle$ for $i \in [M]$, where $Z \sim \mathcal{N}(0, I_d)$. Let $Y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \epsilon^2/2)$. Then

$$\mathbb{E}(X_i - X_j)^2 = (\theta_i - \theta_j)^\top \mathbb{E}[ZZ^\top](\theta_i - \theta_j) = \|\theta_i - \theta_j\|_2^2 \geq \epsilon^2 = \mathbb{E}(Y_i - Y_j)^2.$$

Then

$$\mathbb{E} \sup_{\theta \in \Theta} \langle \theta, Z \rangle \geq \mathbb{E} \max_{1 \leq i \leq M} X_i \geq \mathbb{E} \max_{1 \leq i \leq M} Y_i \asymp \epsilon \sqrt{\log M}$$

where the second and third step follows from Lemma 27.11 and Lemma 27.12 respectively. \square

Revisiting the packing number of the ℓ_1 -ball, we apply Sudakov minorization to $\Theta = B_1$. By duality and applying Lemma 27.12,

$$w(B_1) = \mathbb{E} \sup_{x: \|x\|_1 \leq 1} \langle x, Z \rangle = \mathbb{E} \|Z\|_\infty \leq \sqrt{2 \log d}.$$

Then Theorem 27.10 gives

$$\log M(B_1, \|\cdot\|_2, \epsilon) \lesssim \frac{\log d}{\epsilon^2}. \quad (27.24)$$

When $\epsilon \gtrsim 1/\sqrt{d}$, this is much tighter than the volume bound (27.15) and almost optimal (compared to $\frac{\log(de^2)}{\epsilon^2}$); however, when $\epsilon \asymp 1/\sqrt{d}$, (27.24) yields $d \log d$ but we know (even from the volume bound) that the correct behavior is d . Next we discuss another general approach that gives the optimal bound in this case.

27.3.2 Maurey's empirical method

In this section we discuss a powerful probabilistic method due to B. Maurey for constructing a good covering. It has found applications in approximation theory and especially that for neural nets [168, 24]. The following result gives a dimension-free bound on the cover number of convex hulls in Hilbert spaces:

Theorem 27.11. *Let H be an inner product space with the norm $\|x\| \triangleq \sqrt{\langle x, x \rangle}$. Let $T \subset H$ be a finite set, with radius $r = \text{rad}(T) = \inf_{y \in H} \sup_{x \in T} \|x - y\|$ (recall (5.3)). Denote the convex hull of T by $\text{co}(T)$. Then for any $0 < \epsilon \leq r$,*

$$\mathcal{N}(\text{co}(T), \|\cdot\|, \epsilon) \leq \binom{|T| + \lceil \frac{r^2}{\epsilon^2} \rceil - 2}{\lceil \frac{r^2}{\epsilon^2} \rceil - 1}. \quad (27.25)$$

27.3 Beyond the volume bound 461

Proof. Let $T = \{t_1, t_2, \dots, t_m\}$ and denote the Chebyshev center of T by $c \in H$, such that $r = \max_{i \in [m]} \|c - t_i\|$. For $n \in \mathbb{Z}_+$, let

$$\mathcal{Z} = \left\{ \frac{1}{n+1} \left(c + \sum_{i=1}^m n_i t_i \right) : n_i \in \mathbb{Z}_+, \sum_{i=1}^m n_i = n \right\}.$$

For any $x = \sum_{i=1}^m x_i t_i \in \text{co}(T)$ where $x_i \geq 0$ and $\sum x_i = 1$, let Z be a discrete random variable such that $Z = t_i$ with probability x_i . Then $\mathbb{E}[Z] = x$. Let $Z_0 = c$ and Z_1, \dots, Z_n be i.i.d. copies of Z . Let $\bar{Z} = \frac{1}{n+1} \sum_{i=0}^n Z_i$, which takes values in the set \mathcal{Z} . Since

$$\begin{aligned} \mathbb{E}\|\bar{Z} - x\|_2^2 &= \frac{1}{(n+1)^2} \mathbb{E} \left\| \sum_{i=0}^n (Z_i - x) \right\|^2 = \frac{1}{(n+1)^2} \left(\sum_{i=0}^n \mathbb{E} \|Z_i - x\|^2 + \sum_{i \neq j} \mathbb{E} \langle Z_i - x, Z_j - x \rangle \right) \\ &= \frac{1}{(n+1)^2} \sum_{i=0}^n \mathbb{E} \|Z_i - x\|^2 = \frac{1}{(n+1)^2} (\|c - x\|^2 + n\mathbb{E}[\|Z - x\|^2]) \leq \frac{r^2}{n+1}, \end{aligned}$$

where the last inequality follows from that $\|c - x\| \leq \sum_{i=1}^m x_i \|c - t_i\| \leq r$ (in other words, $\text{rad}(T) = \text{rad}(\text{co}(T))$ and $\mathbb{E}[\|Z - x\|^2] \leq \mathbb{E}[\|Z - c\|^2] \leq r^2$). Set $n = \lceil r^2/\epsilon^2 \rceil - 1$ so that $r^2/(n+1) \leq \epsilon^2$. There exists some $z \in N$ such that $\|z - x\| \leq \epsilon$. Therefore \mathcal{Z} is an ϵ -covering of $\text{co}(T)$. Similar to (27.17), we have

$$|\mathcal{Z}| \leq \binom{n+m-1}{n} = \binom{m + \lceil r^2/\epsilon^2 \rceil - 2}{\lceil r^2/\epsilon^2 \rceil - 1}.$$

□

We now apply Theorem 27.13 to recover the result for the unit ℓ_1 -ball B_1 in \mathbb{R}^d in Theorem 27.9: Note that $B_1 = \text{co}(T)$, where $T = \{\pm e_1, \dots, \pm e_d, 0\}$ satisfies $\text{rad}(T) = 1$. Then

$$N(B_1, \|\cdot\|_2, \epsilon) \leq \binom{2d + \lceil \frac{1}{\epsilon^2} \rceil - 1}{\lceil \frac{1}{\epsilon^2} \rceil - 1}, \quad (27.26)$$

which recovers the optimal upper bound in Theorem 27.9 at both small and big scale.

27.3.3 Duality of metric entropy

First we define a more general notion of covering number. For $K, T \subset \mathbb{R}^d$, define the covering number of K using translates of T as

$$N(K, T) = \min\{N : \exists x_1, \dots, x_N \in \mathbb{R}^d \text{ such that } K \subset \bigcup_{i=1}^N T + x_i\}.$$

Then the usual covering number in Definition 27.1 satisfies $N(K, \|\cdot\|, \epsilon) = N(K, \epsilon B)$, where B is the corresponding unit norm ball.

A deep result of Artstein, Milman, and Szarek [18] establishes the following duality for metric entropy: There exist absolute constants α and β such that for any symmetric convex body K ,⁹

$$\frac{1}{\beta} \log N\left(B_2, \frac{\epsilon}{\alpha} K^\circ\right) \leq \log N(K, \epsilon B_2) \leq \log N(B_2, \alpha \epsilon K^\circ), \quad (27.27)$$

where B_2 is the usual unit ℓ_2 -ball, and $K^\circ = \{y : \sup_{x \in K} \langle x, y \rangle \leq 1\}$ is the polar body of K .

As an example, consider $p < 2 < q$ and $\frac{1}{p} + \frac{1}{q} = 1$. By duality, $B_p^\circ = B_q$. Then (27.27) shows that $N(B_p, \|\cdot\|_2, \epsilon)$ and $N(B_2, \|\cdot\|_q, \epsilon)$ have essentially the same behavior, as verified by (27.18).

27.4 Infinite-dimensional space: smooth functions

Unlike Euclidean spaces, in infinite-dimensional spaces, the metric entropy can grow arbitrarily fast [?, Theorem XI]. Studying of metric entropy in functional space (for example, under shape or smoothness constraints) is an area of interest in functional analysis (cf. [315]), and has important applications in nonparametric statistics, empirical processes, and learning theory [103]. To gain some insight on the fundamental distinction between finite- and infinite-dimensional spaces, let us work out a concrete example, which will later be used in the application of density estimation in Section 32.4. For more general and more precise results (including some cases of equality), see [176, Sec. 4 and 7]. Consider the class $\mathcal{F}(A, L)$ of all L -Lipschitz probability densities on the compact interval $[0, A]$.

Theorem 27.12. *Assume that $L, A > 0$ and $p \in [1, \infty]$ are constants. Then*

$$\log N(\mathcal{F}(A, L), \|\cdot\|_p, \epsilon) = \Theta\left(\frac{1}{\epsilon}\right). \quad (27.28)$$

Furthermore, for the sup-norm we have the sharp asymptotics:

$$\log_2 N(\mathcal{F}(A, L), \|\cdot\|_\infty, \epsilon) = \frac{LA}{\epsilon}(1 + o(1)), \quad \epsilon \rightarrow 0. \quad (27.29)$$

Proof. By replacing $f(x)$ by $\frac{1}{\sqrt{L}}f(\frac{x}{\sqrt{L}})$, we have

$$N(\mathcal{F}(A, L), \|\cdot\|_p, \epsilon) = N(\mathcal{F}(\sqrt{L}A, 1), \|\cdot\|_p, L^{\frac{1-p}{2p}}\epsilon). \quad (27.30)$$

Thus, it is sufficient to consider $\mathcal{F}(A, 1) \triangleq \mathcal{F}(A)$, the collection of 1-Lipschitz densities on $[0, A]$. Next, observe that any such density function f is bounded from above. Indeed, since $f(x) \geq (f(0) - x)_+$ and $\int_0^A f = 1$, we conclude that $f(0) \leq \max\{A, \frac{A}{2} + \frac{1}{A}\} \triangleq m$.

To show (27.28), it suffices to prove the upper bound for $p = \infty$ and the lower bound for $p = 1$. Specifically, we aim to show, by explicit construction,

$$N(\mathcal{F}(A), \|\cdot\|_\infty, \epsilon) \leq \frac{C}{\epsilon} 2^{\frac{A}{\epsilon}} \quad (27.31)$$

⁹ A convex body K is a compact convex set with non-empty interior. We say K is symmetric if $K = -K$.

27.4 Infinite-dimensional space: smooth functions 463

$$M(\mathcal{F}(A), \|\cdot\|_1, \epsilon) \geq 2^{\frac{\epsilon}{\epsilon}} \quad (27.32)$$

which imply the desired (27.28) in view of Theorem 27.3. Here and below, c, C are constants depending on A . We start with the easier (27.32). We construct a packing by perturbing the uniform density. Define a function T by $T(x) = x1_{\{x \leq \epsilon\}} + (2\epsilon - x)1_{\{x \geq \epsilon\}} + \frac{1}{A}$ on $[0, 2\epsilon]$ and zero elsewhere. Let $n = \lceil \frac{A}{4\epsilon} \rceil$ and $a = 2n\epsilon$. For each $y \in \{0, 1\}^n$, define a density f_y on $[0, A]$ such that

$$f_y(x) = \sum_{i=1}^n y_i T(x - 2(i-1)\epsilon), \quad x \in [0, a],$$

and we linearly extend f_y to $[a, A]$ so that $\int_0^A f_y = 1$; see Fig. 27.2. For sufficiently small ϵ , the resulting f_y is 1-Lipschitz since $\int_0^a f_y = \frac{1}{2} + O(\epsilon)$ so that the slope of the linear extension is $O(\epsilon)$.

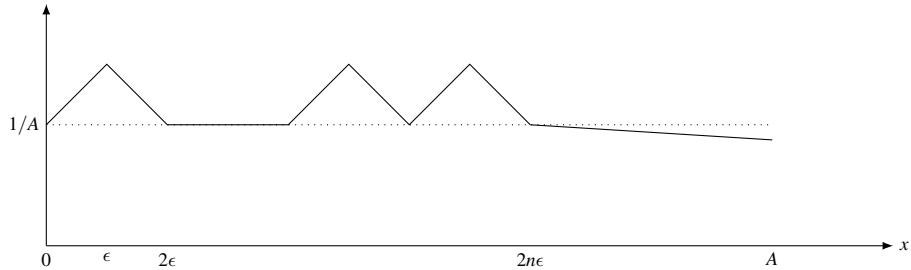


Figure 27.2 Packing that achieves (27.32). The solid line represent one such density $f_y(x)$ with $y = (1, 0, 1, 1)$. The dotted line is the density of $\text{Unif}(0, A)$.

Thus we conclude that each f_y is a valid member of $\mathcal{F}(A)$. Furthermore, for $y, z \in \{0, 1\}^n$, we have $\|f_y - f_z\|_1 = d_H(y, z)\|T\|_1 = \epsilon^2 d_H(y, z)$. Invoking the Gilbert-Varshamov bound Theorem 27.7, we obtain an $\frac{n}{2}$ -packing \mathcal{Y} of the Hamming space $\{0, 1\}^n$ with $|\mathcal{Y}| \geq 2^{cn}$ for some absolute constant c . Thus $\{f_y : y \in \mathcal{Y}\}$ constitutes an $\frac{n\epsilon^2}{2}$ -packing of $\mathcal{F}(A)$ with respect to the L_1 -norm. This is the desired (27.32) since $\frac{n\epsilon^2}{2} = \Theta(\epsilon)$.

To construct a covering, set $J = \lceil \frac{m}{\epsilon} \rceil$, $n = \lceil \frac{A}{\epsilon} \rceil$, and $x_k = k\epsilon$ for $k = 0, \dots, n$. Let \mathcal{G} be the collection of all lattice paths (with grid size ϵ) of n steps starting from the coordinate $(0, j\epsilon)$ for some $j \in \{0, \dots, J\}$. In other words, each element g of \mathcal{G} is a continuous piecewise linear function on each subinterval $I_k = [x_k, x_{k+1})$ with slope being either $+1$ or -1 . Evidently, the number of such paths is at most $(J+1)2^n = O(\frac{1}{\epsilon}2^{A/\epsilon})$. To show that \mathcal{G} is an ϵ -covering, for each $f \in \mathcal{F}(A)$, we show that there exists $g \in \mathcal{G}$ such that $|f(x) - g(x)| \leq \epsilon$ for all $x \in [0, A]$. This can be shown by a simple induction. Suppose that there exists g such that $|f(x) - g(x)| \leq \epsilon$ for all $x \in [0, x_k]$, which clearly holds for the base case of $k = 0$. We show that g can be extended to I_k so that this holds for $k + 1$. Since $|f(x_k) - g(x_k)| \leq \epsilon$ and f is 1-Lipschitz, either $f(x_{k+1}) \in [g(x_k), g(x_k) + 2\epsilon]$ or $[g(x_k) - 2\epsilon, g(x_k)]$, in which case we extend g upward or downward, respectively. The resulting g satisfies $|f(x) - g(x)| \leq \epsilon$ on I_k , completing the induction.

Finally, we prove the sharp bound (27.29) for $p = \infty$. The upper bound readily follows from (27.31) plus the scaling relation (27.30). For the lower bound, we apply Theorem 27.3 converting

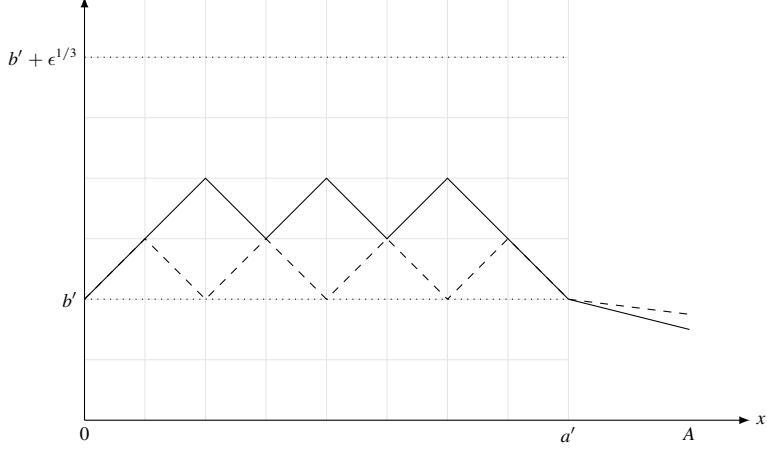


Figure 27.3 Improved packing for (27.33). Here the solid and dashed lines are two lattice paths on a grid of size ϵ starting from $(0, b')$ and staying in the range of $[b', b' + \epsilon^{1/3}]$, followed by their respective linear extensions.

the problem to the construction of 2ϵ -packing. Following the same idea of lattice paths, next we give an improved packing construction such that

$$M(\mathcal{F}(A), \|\cdot\|_\infty, 2\epsilon) \geq \Omega(\epsilon^{3/2} 2^{\frac{a}{\epsilon}}). \quad (27.33)$$

for any $a < A$. Choose any b such that $\frac{1}{A} < b < \frac{1}{A} + \frac{(A-a)^2}{2A}$. Let $a' = \epsilon \lfloor \frac{a}{\epsilon} \rfloor$ and $b' = \epsilon \lfloor \frac{b}{\epsilon} \rfloor$. Consider a density f on $[0, A]$ of the following form (cf. Fig. 27.3): on $[0, a']$, f is a lattice path from $(0, b')$ to (a', b') that stays in the vertical range of $[b', b' + \epsilon^{1/3}]$; on $[a', A]$, f is a linear extension chosen so that $\int_0^A f = 1$. This is possible because by the 1-Lipschitz constraint we can linearly extend f so that $\int_{a'}^A f$ takes any value in the interval $[b'(A-a') - \frac{(A-a')^2}{2}, b'(A-a') + \frac{(A-a')^2}{2}]$. Since $\int_0^{a'} f = ab + o(1)$, we need $\int_{a'}^A f = 1 - \int_0^{a'} f = 1 - ab + o(1)$, which is feasible due to the choice of b . The collection \mathcal{G} of all such functions constitute a 2ϵ -packing in the sup norm (for two distinct paths consider the first subinterval where they differ). Finally, we bound the cardinality of this packing by counting the number of such paths. This can be accomplished by standard estimates on random walks (see e.g. [120, Chap. III]). For any constant $c > 0$, the probability that a symmetric random walk on \mathbb{Z} returns to zero in n (even) steps and stays in the range of $[0, n^{1+c}]$ is $\Theta(n^{-3/2})$; this implies the desired (27.33). Finally, since $a < A$ is arbitrary, the lower bound part of (27.29) follows in view of Theorem 27.3. \square

The following result, due to Birman and Solomjak [36] (cf. [199, Sec. 15.6] for an exposition), is an extension of Theorem 27.14 to the more general Hölder class.

Theorem 27.13. Fix positive constants A, L and $d \in \mathbb{N}$. Let $\beta > 0$ and write $\beta = \ell + \alpha$, where $\ell = \lfloor \beta \rfloor$ and $\alpha \in [0, 1)$. Let $\mathcal{F}_\beta(A, L)$ denote the collection of ℓ -times continuously differentiable densities f on $[0, A]^d$ whose ℓ th derivative is (L, α) -Hölder continuous, namely,

27.5 Hilbert ball has metric entropy $\frac{1}{\epsilon^2}$ 465

$\|D^{(\ell)}f(x) - D^{(\ell)}f(y)\|_\infty \leq L\|x - y\|_\infty^\alpha$ for all $x, y \in [0, A]^d$. Then for any $1 \leq p \leq \infty$,

$$\log N(\mathcal{F}_\beta(A, L), \|\cdot\|_p, \epsilon) \asymp \left(\frac{1}{\epsilon}\right)^{\frac{d}{\beta}}. \quad (27.34)$$

The main message of the preceding theorem is that the entropy of the function class grows more slowly if the dimension decreases or the smoothness increases. As such, the metric entropy for very smooth functions can grow subpolynomially in $\frac{1}{\epsilon}$. For example, Vitushkin (cf. [176, Eq. (129)]) showed that for the class of analytic functions on the unit complex disk D having analytic extension to a bigger disk rD for $r > 1$, the metric entropy (with respect to the sup-norm on D) is $\Theta((\log \frac{1}{\epsilon})^2)$; see [176, Sec. 7 and 8] for more such results.

As mentioned at the beginning of this chapter, the conception and development of the subject on metric entropy, in particular, Theorem 27.15, are motivated by and plays an important role in the study of Hilbert's 13th problem. In 1900, Hilbert conjectured that there exist functions of several variables which cannot be represented as a superposition (composition) of finitely many functions of fewer variables. This was disproved by Kolmogorov and Arnold in 1950s who showed that every continuous function of d variables can be represented by sums and superpositions of single-variable functions; however, their construction does not work if one requires the constituent functions to have specific smoothness. Subsequently, Hilbert's conjecture for smooth functions was positively resolved by Vitushkin [314], who showed that there exist functions of d variables in the β -Hölder class (in the sense of Theorem 27.15) that cannot be expressed as finitely many superpositions of functions of d' variables in the β' -Hölder class, provided $d/\beta > d'/\beta'$. The original proof of Vitushkin is highly involved. Later, Kolmogorov gave a much simplified proof by proving and applying the $\|\cdot\|_\infty$ -version of Theorem 27.15. As evident in (27.34), the index d/β provides a complexity measure for the function class; this allows an proof of impossibility of superposition by an entropy comparison argument. For concreteness, let us prove the following simpler version: There exists a 1-Lipschitz function $f(x, y, z)$ of three variables on $[0, 1]^3$ that cannot be written as $g(h_1(x, y), h_2(y, z))$ where g, h_1, h_2 are 1-Lipschitz functions of two variables on $[0, 1]^2$. Suppose, for the sake of contradiction, that this is possible. Fixing an ϵ -covering of cardinality $\exp(O(\frac{1}{\epsilon^2}))$ for 1-Lipschitz functions on $[0, 1]^2$ and using it to approximate the functions g, h_1, h_2 , we obtain by superposition $g(h_1, h_2)$ an $O(\epsilon)$ -covering of cardinality $\exp(O(\frac{1}{\epsilon^2}))$ of 1-Lipschitz functions on $[0, 1]^3$; however, this is a contradiction as any such covering must be of size $\exp(\Omega(\frac{1}{\epsilon^3}))$. For stronger and more general results along this line, see [176, Appendix I].

27.5 Hilbert ball has metric entropy $\frac{1}{\epsilon^2}$

Consider the following set of linear functions $f_\theta(x) = (\theta, x)$ with $\theta, x \in B$ – a unit ball in infinite dimensional Hilbert space with inner product (\cdot, \cdot) .

Theorem 27.14. Consider any measure P on B and let $d_P(\theta, \theta') = \sqrt{\mathbb{E}_{X \sim P}[|f_\theta(X) - f_{\theta'}(X)|^2]}$. Then we have

$$\log N(\epsilon, d_P) \leq \frac{1}{e\epsilon^2}.$$

Proof. We have $\log N(\epsilon) \leq \log M(\epsilon)$. By some continuity argument, let's consider only empirical measures $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. First consider the special case when x_i 's are orthogonal basis. Then the ϵ -packing in d_P is simply an $\sqrt{n}\epsilon$ -packing of n -dimensional Euclidean unit ball. From Varshamov's argument we have

$$\log M(\epsilon) \leq -n \log \sqrt{n}\epsilon. \quad (27.35)$$

Thus, we have

$$\log N(\epsilon, d_{P_n}) \leq \max_n n \log \frac{1}{\sqrt{n}\epsilon} = \frac{1}{e\epsilon^2}.$$

Now, for a general case, after some linear algebra we get that the goal is to do $\sqrt{n}\epsilon$ -packing in Euclidean metric of an ellipsoid:

$$\{y^n : \sum_{j=1}^n y_j^2 / \lambda_j \leq 1\},$$

where λ_j are eigenvalues of the Gram matrix of $\{x_i, i \in [n]\}$. By calculating the volume of this ellipsoid the bound (27.35) is then replaced by

$$\log M(\epsilon) \leq \sum_{j=1}^n \log \lambda_j - n \log \sqrt{n}\epsilon.$$

Since $\sum_j \lambda_j \leq n$ (x_i 's are unit norm!) we get from Jensen's that the first sum above is ≤ 0 and we reduced to the previous case. \square

To see one simple implication of the result, recall the standard bound on empirical processes

$$\mathbb{E} \left[\sup_{\theta} \mathbb{E}[f_\theta(X)] - \hat{\mathbb{E}}_n[f_\theta(X)] \right] \lesssim \mathbb{E} \left[\inf_{\delta > 0} \delta + \int_{\delta}^{\infty} \sqrt{\frac{\log N(\Theta, L_2(\hat{P}_n), \epsilon)}{n}} d\epsilon \right].$$

It can be seen that when entropy behaves as ϵ^{-p} we get rate $n^{-\min(1/p, 1/2)}$ except for $p = 2$ for which the upper bound yields $n^{-\frac{1}{2}} \log n$. The significance of the previous theorem is that the Hilbert ball is precisely “at the phase transition” from parametric to nonparametric rate.

As a sanity check, let us take $\Theta = B$ and any P_X over the unit (possibly infinite dimensional) ball with $\mathbb{E}[X] = 0$ we have

$$\mathbb{E}[\|\bar{X}_n\|] = \mathbb{E} \left[\sup_{\theta} \frac{1}{n} \sum_{i=1}^n (\theta, X_i) \right] \lesssim \sqrt{\frac{\log n}{n}},$$

where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is the empirical mean vector. In this special case it is easy to bound $\mathbb{E}[\|\bar{X}_n\|] \leq \sqrt{\mathbb{E}[\|\bar{X}_n\|^2]} \leq \frac{1}{\sqrt{n}}$ by an explicit calculation.

27.6 Metric entropy and small-ball probability

The small ball problem in probability theory concerns the behavior of the function

$$\phi(\epsilon) \triangleq \log \frac{1}{\mathbb{P}[\|X\| \leq \epsilon]}$$

as $\epsilon \rightarrow 0$, where X is a random variable taking values on some real separable Banach space $(V, \|\cdot\|)$. For example, for standard normal $X \sim \mathcal{N}(0, I_d)$ and the ℓ_2 -ball, a simple large-deviation calculation (Exercise III.10) shows that

$$\phi(\epsilon) \asymp d \log \frac{1}{\epsilon}.$$

Of more interest is the infinite-dimensional case of Gaussian processes. For example, for the standard Brownian motion on the unit interval and the sup norm, it is elementary to show (Exercise V.10) that

$$\phi(\epsilon) \asymp \frac{1}{\epsilon^2}. \quad (27.36)$$

We refer the reader to the excellent survey [195] for this field.

There is a deep connection between the small-ball probability and metric entropy, which allows one to translate results from one area to the other in fruitful ways. To identify this link, the starting point is the volume argument in Theorem 27.4. On the one hand, it is well-known that there exists no analog of Lebesgue measure (translation-invariant) in infinite-dimensional spaces. As such, for functional spaces, one frequently uses a Gaussian measure. On the other hand, the “volume” argument in Theorem 27.4 and Remark 27.5 can be adapted to a measure γ that need not be translation-invariant, leading to

$$\frac{\gamma(\Theta + B(0, \epsilon))}{\max_{\theta \in V} \gamma(B(\theta, 2\epsilon))} \leq N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \leq \frac{\gamma(\Theta + B(0, \epsilon/2))}{\min_{\theta \in \Theta} \gamma(B(\theta, \epsilon/2))}, \quad (27.37)$$

where we recall that $B(\theta, \epsilon)$ denotes the norm ball centered at θ of radius ϵ . From here we have already seen the natural appearance of small-ball probabilities. Using properties native to the Gaussian measure, this can be further analyzed and reduced to balls centered at zero.

To be precise, let γ be a zero-mean Gaussian measure on V such that $\mathbb{E}_{X \sim \gamma}[\|X\|^2] < \infty$. Let $H \subset V$ be the reproducing kernel Hilbert space (RKHS) generated by γ and K the unit ball in H . We refer the reader to, e.g., [182, Sec. 2] and [219, III.3.2], for the precise definition of this object.¹⁰ For the purpose of this section, it is enough to consider the following examples (for more see [182]):

- Finite dimensions. Let $\gamma = \mathcal{N}(0, \Sigma)$. Then

$$K = \{\Sigma^{1/2}x : \|x\|_2 \leq 1\} \quad (27.38)$$

¹⁰ In particular, if γ is the law of a Gaussian process X on $C([0, 1])$ with $\mathbb{E}[\|X\|_2^2] < \infty$, the kernel $K(s, t) = \mathbb{E}[X(s)X(t)]$ admits the eigendecomposition $K(s, t) = \sum \lambda_k \psi_k(s)\psi_k(t)$ (Mercer's theorem), where $\{\phi_k\}$ is an orthonormal basis for $L_2([0, 1])$ and $\lambda_k > 0$. Then H is the closure of the span of $\{\phi_k\}$ with the inner product $\langle x, y \rangle_H = \sum_k \langle x, \psi_k \rangle \langle y, \psi_k \rangle / \lambda_k$.

is a rescaled Euclidean ball, with inner product $\langle x, y \rangle_H = x^\top \Sigma^{-1} y$.

- Brownian motion: Let γ be the law of the standard Brownian motion on the unit interval $[0, 1]$. Then

$$K = \left\{ f(t) = \int_0^t f'(s) ds : \|f'\|_2 \leq 1 \right\} \quad (27.39)$$

with inner product $\langle f, g \rangle_H = \langle f', g' \rangle \equiv \int_0^1 f'(t) g'(t) dt$.

The following fundamental result due to Kuelbs and Li [183] (see also the earlier work of Goodman [139]) describes a precise connection between the small-ball probability function $\phi(\epsilon)$ and the metric entropy of the unit Hilbert ball $N(K, \|\cdot\|, \epsilon) \equiv N(\epsilon)$.

Theorem 27.15. *For all $\epsilon > 0$,*

$$\phi(2\epsilon) - \log 2 \leq \log N\left(\frac{\epsilon}{\sqrt{2\phi(\epsilon/2)}}\right) \leq 2\phi(\epsilon/2) \quad (27.40)$$

Proof. We show that for any $\lambda > 0$,

$$\phi(2\epsilon) + \log \Phi(\lambda + \Phi^{-1}(e^{-\phi(\epsilon)})) \leq \log N(\lambda K, \epsilon) \leq \log M(\lambda K, \epsilon) \leq \frac{\lambda^2}{2} + \phi(\epsilon/2) \quad (27.41)$$

To deduce (27.40), choose $\lambda = \sqrt{2\phi(\epsilon/2)}$ and note that by scaling $N(\lambda K, \epsilon) = N(K, \epsilon/\lambda)$. Applying the normal tail bound $\Phi(-t) = \Phi^c(t) \leq e^{-t^2/2}$ (Exercise V.9) yields $\Phi^{-1}(e^{-\phi(\epsilon)}) \geq -\sqrt{2\phi(\epsilon)} \geq -\lambda$ so that $\Phi(\Phi^{-1}(e^{-\phi(\epsilon)}) + \lambda) \geq \Phi(0) = 1/2$.

We only give the proof in finite dimensions as the results are dimension-free and extend naturally to infinite-dimensional spaces. Let $Z \sim \gamma = N(0, \Sigma)$ on \mathbb{R}^d so that $K = \Sigma^{1/2} B_2$ is given in (27.38). Applying (27.37) to λK and noting that γ is a probability measure, we have

$$\frac{\gamma(\lambda K + B(0, \epsilon))}{\max_{\theta \in \mathbb{R}^d} \gamma(B(\theta, 2\epsilon))} \leq N(\lambda K, \epsilon) \leq M(\lambda K, \epsilon) \leq \frac{1}{\min_{\theta \in \lambda K} \gamma(B(\theta, \epsilon/2))}. \quad (27.42)$$

Next we further bound (27.42) using properties native to the Gaussian measure.

- For the upper bound, for any *symmetric* set $A = -A$ and any $\theta \in \lambda K$, by a change of measure

$$\begin{aligned} \gamma(\theta + A) &= \mathbb{P}[Z - \theta \in A] \\ &= e^{-\frac{1}{2}\theta^\top \Sigma^{-1}\theta} \mathbb{E}\left[e^{\langle \Sigma^{-1}\theta, Z \rangle} \mathbf{1}_{\{Z \in A\}}\right] \\ &\geq e^{-\lambda^2/2} \mathbb{P}[Z \in A], \end{aligned}$$

where the last step follows from $\theta^\top \Sigma^{-1}\theta \leq \lambda^2$ and by Jensen's inequality $\mathbb{E}\left[e^{\langle \Sigma^{-1}\theta, Z \rangle} | Z \in A\right] \geq e^{\langle \Sigma^{-1}\theta, \mathbb{E}[Z | Z \in A] \rangle} = 1$, using crucially that $\mathbb{E}[Z | Z \in A] = 0$ by symmetry. Applying the above to $A = B(0, \epsilon/2)$ yields the right inequality in (27.41).

27.7 Metric entropy and rate-distortion theory 469

- For the lower bound, recall Anderson's lemma (Lemma 28.15) stating that the Gaussian measure of a ball is maximized when centered at zero, so $\gamma(B(\theta, 2\epsilon)) \leq \gamma(B(0, 2\epsilon))$ for all θ . To bound the numerator, recall the Gaussian isoperimetric inequality (see e.g. [45, Theorem 10.15]):¹¹

$$\gamma(A + \lambda K) \geq \Phi(\Phi^{-1}(\gamma(A)) + \lambda). \quad (27.43)$$

Applying this with $A = B(0, \epsilon)$ proves the left inequality in (27.41) and the theorem. \square

The implication of Theorem 27.17 is the following. Provided that $\phi(\epsilon) \asymp \phi(\epsilon/2)$, then we should expect that approximately

$$\log N\left(\frac{\epsilon}{\sqrt{\phi(\epsilon)}}\right) \asymp \phi(\epsilon)$$

With more effort this can be made precise unconditionally (see e.g. [195, Theorem 3.3], incorporating the later improvement by [194]), leading to *very precise* connections between metric entropy and small-ball probability, for example: for fixed $\alpha > 0, \beta \in \mathbb{R}$,

$$\phi(\epsilon) \asymp \epsilon^{-\alpha} \left(\log \frac{1}{\epsilon}\right)^\beta \iff \log N(\epsilon) \asymp \epsilon^{-\frac{2\alpha}{2+\alpha}} \left(\log \frac{1}{\epsilon}\right)^{\frac{2\beta}{2+\alpha}} \quad (27.44)$$

As a concrete example, consider the unit ball (27.39) in the RKHS generated by the standard Brownian motion, which is similar to a Sobolev ball.¹² Using (27.36) and (27.44), we conclude that $\log N(\epsilon) \asymp \frac{1}{\epsilon}$, recovering the metric entropy of Sobolev ball determined in [301]. This result also coincides with the metric entropy of Lipschitz ball in Theorem 27.15 which requires the derivative to be bounded everywhere as opposed to on average in L_2 . For more applications of small-ball probability on metric entropy (and vice versa), see [183, 194].

27.7 Metric entropy and rate-distortion theory

In this section we discuss a connection between metric entropy and rate-distortion function. Note that the former is a non-probabilistic quantity whereas the latter is an information measure depending on the source distribution; nevertheless, if we consider the rate-distortion function induced by the “least favorable” source distribution, it turns out to behave similarly to the metric entropy.

To make this precise, consider a metric space (\mathcal{X}, d) . For an \mathcal{X} -valued random variable X , denote by

$$\phi_X(\epsilon) = \inf_{P_{\hat{X}|X}: \mathbb{E}[d(X, \hat{X})] \leq \epsilon} I(X; \hat{X}) \quad (27.45)$$

¹¹ The connection between (27.43) and isoperimetry is that if we interpret $\lim_{\lambda \rightarrow 0} (\gamma(A + \lambda K) - \gamma(A))/\lambda$ as the surface measure of A , then among all sets with the same Gaussian measure, the half space has maximal surface measure.

¹² The Sobolev norm is $\|f\|_{W^{1,2}} \triangleq \|f\|_2 + \|f'\|_2$. Nevertheless, it is simple to verify a priori that the metric entropy of (27.39) and that of the Sobolev ball share the same behavior (see [183, p. 152]).

its rate-distortion function (recall Section 24.3). Denote the *worst-case* rate-distortion function on \mathcal{X} by

$$\phi_{\mathcal{X}}(\epsilon) = \sup_{P_X \in \mathcal{P}(\mathcal{X})} \phi_X(\epsilon). \quad (27.46)$$

The next theorem relates $\phi_{\mathcal{X}}$ to the covering and packing number of \mathcal{X} . The lower bound simply follows from a “Bayesian” argument, which bounds the worst case from below by the average case, akin to the relationship between minimax and Bayes risk (see Section 28.3). The upper bound was shown in [170] using the dual representation of rate-distortion functions; here we give a simpler proof via Fano’s inequality.

Theorem 27.16. *For any $0 < c < 1/2$,*

$$\phi_{\mathcal{X}}(\epsilon) \leq \log N(\mathcal{X}, d, \epsilon) \leq \log M(\mathcal{X}, d, \epsilon) \leq \frac{\phi_{\mathcal{X}}(c\epsilon) + \log 2}{1 - 2c}. \quad (27.47)$$

Proof. Fix an ϵ -covering of \mathcal{X} in d of size N . Let \hat{X} denote the closest element in the covering to X . Then $d(X, \hat{X}) \leq \epsilon$ almost surely. Thus $\phi_X(\epsilon) \leq I(X; \hat{X}) \leq \log N$. Optimizing over P_X proves the left inequality.

For the right inequality, let X be uniformly distributed over a maximal ϵ -packing of \mathcal{X} . For any $P_{\tilde{X}|X}$ such that $\mathbb{E}[d(X, \tilde{X})] \leq c\epsilon$. Let \tilde{X} denote the closest point in the packing to \hat{X} . Then we have the Markov chain $X \rightarrow \hat{X} \rightarrow \tilde{X}$. By definition, $d(X, \tilde{X}) \leq d(\hat{X}, \tilde{X}) + d(\hat{X}, X) \leq 2d(\hat{X}, X)$ so $\mathbb{E}[d(X, \tilde{X})] \leq 2c\epsilon$. Since either $X = \tilde{X}$ or $d(X, \tilde{X}) > \epsilon$, we have $\mathbb{P}[X \neq \tilde{X}] \leq 2c$. On the other hand, Fano’s inequality (Corollary 6.1) yields $\mathbb{P}[X \neq \tilde{X}] \geq 1 - \frac{I(X; \tilde{X}) + \log 2}{\log M}$. In all, $I(X; \hat{X}) \geq (1 - 2c) \log M - \log 2$, proving the upper bound. \square

Remark 27.4. (a) Clearly, Theorem 27.18 can be extended to the case where the distortion function equals a power of the metric, namely, replacing (27.45) with

$$\phi_{X,r}(\epsilon) \triangleq \inf_{P_{\tilde{X}|X}: \mathbb{E}[d(X, \tilde{X})] \leq \epsilon^r} I(X; \tilde{X}).$$

Then (27.47) continues to hold with $1 - 2c$ replaced by $1 - (2c)^r$. This will be useful, for example, in the forthcoming applications where second moment constraint is easier to work with.

(b) In the earlier literature a variant of the rate-distortion function is also considered, known as the ϵ -entropy of X , where the constraint is $d(X, \hat{X}) \leq \epsilon$ with probability one as opposed to in expectation (cf. e.g. [176, Appendix II] and [245]). With this definition, it is natural to conjecture that the maximal ϵ -entropy over all distributions on \mathcal{X} coincides with the metric entropy $\log N(\mathcal{X}, \epsilon)$; nevertheless, this need not be true (see [210, Remark, p. 1708] for a counterexample).

27.7 Metric entropy and rate-distortion theory 471

Theorem 27.18 points out an information-theoretic route to bound the metric entropy by the worst-case rate-distortion function (27.46).¹³ Solving this maximization, however, is not easy as $P_X \mapsto \phi_X(D)$ is in general neither convex nor concave [3].¹⁴ Fortunately, for certain spaces, one can show via a symmetry argument that the “uniform” distribution maximizes the rate-distortion function at every distortion level; see Exercise V.8 for a formal statement. As a consequence, we have:

- For Hamming space $\mathcal{X} = \{0, 1\}^d$ and Hamming distortion, $\phi_{\mathcal{X}}(D)$ is attained by $\text{Ber}(\frac{1}{2})^d$. (We already knew this from Theorem 26.1 and Theorem 24.11.)
- For the unit sphere $\mathcal{X} = S^{d-1}$ and distortion function defined by the Euclidean distance, $\phi_{\mathcal{X}}(D)$ is attained by $\text{Unif}(S^{d-1})$.
- For the orthogonal group $\mathcal{X} = O(d)$ or unitary group $U(d)$ and distortion function defined by the Frobenius norm, $\phi_{\mathcal{X}}(D)$ is attained by the Haar measure. Similar statements also hold for the Grassmann manifold (collection of linear subspaces).

Next we give a concrete example by computing the rate-distortion function of $\theta \sim \text{Unif}(S^{d-1})$:

Theorem 27.17. *Let θ be uniformly distributed over the unit sphere S^{d-1} . Then for all $0 < \epsilon < 1$,*

$$(d-1) \log \frac{1}{\epsilon} - C \leq \inf_{P_{\hat{\theta}|\theta}: \mathbb{E}[\|\hat{\theta} - \theta\|_2^2] \leq \epsilon^2} I(\theta; \hat{\theta}) \leq (d-1) \log \left(1 + \frac{1}{\epsilon}\right) + \log(2d)$$

for some universal constant C .

Note that the random vector θ have dependent entries so we cannot invoke the single-letterization technique in Theorem 24.11. Nevertheless, we have the representation $\theta \stackrel{d}{=} Z / \|Z\|_2$ for $Z \sim \mathcal{N}(0, I_d)$, which allows us to relate the rate-distortion function of θ to that of the Gaussian found in Theorem 26.4. The resulting lower bound agree with the metric entropy for spheres in Corollary 27.1, which scales as $(d-1) \log \frac{1}{\epsilon}$. Using similar reduction arguments (see [191, Theorem VIII.18]), one can obtain tight lower bound for the metric entropy of the orthogonal group $O(d)$ and the unitary group $U(d)$, which scales as $\frac{d(d-1)}{2} \log \frac{1}{\epsilon}$ and $d^2 \log \frac{1}{\epsilon}$, with pre-log factors commensurate with their respective degrees of freedoms. As mentioned in Remark 27.6(b), these results were obtained by Szarek in [290] using a volume argument with Haar measures; in comparison, the information-theoretic approach is more elementary as we can again reduce to Gaussian rate-distortion computation.

Proof. The upper bound follows from Theorem 27.18 and Remark 27.19(a), applying the metric entropy bound for spheres in Corollary 27.1.

¹³ A striking parallelism between the metric entropy of Sobolev balls and the rate-distortion function of smooth Gaussian processes has been observed by Donoho in [98]. However, we cannot apply Theorem 27.18 to formally relate one to the other since it is unclear whether the Gaussian rate-distortion function is maximal.

¹⁴ As a counterexample, consider Theorem 26.1 for the binary source.

472

To prove the lower bound, let $Z \sim \mathcal{N}(0, I_d)$. Define $\theta = \frac{Z}{\|Z\|}$ and $A = \|Z\|$, where $\|\cdot\| \equiv \|\cdot\|_2$ henceforth. Then $\theta \sim \text{Unif}(S^{d-1})$ and $A \sim \chi_d$ are independent. Fix $P_{\hat{\theta}|\theta}$ such that $\mathbb{E}[\|\hat{\theta} - \theta\|^2] \leq \epsilon^2$. Since $\text{Var}(A) \leq 1$, the Shannon lower bound (Theorem 26.5) shows that the rate-distortion function of A is majorized by that of the standard Gaussian. So for each $\delta \in (0, 1)$, there exists $P_{\hat{A}|A}$ such that $\mathbb{E}[(\hat{A} - A)^2] \leq \delta^2$, $I(A, \hat{A}) \leq \log \frac{1}{\delta}$, and $\mathbb{E}[A] = \mathbb{E}[\hat{A}]$. Set $\hat{Z} = \hat{A}\hat{\theta}$. Then

$$I(Z; \hat{Z}) = I(\theta, A; \hat{Z}) \leq I(\theta, A; \hat{\theta}, \hat{A}) = I(\theta; \hat{\theta}) + I(A, \hat{A}).$$

Furthermore, $\mathbb{E}[\hat{A}^2] = \mathbb{E}[(\hat{A} - A)^2] + \mathbb{E}[A^2] + 2\mathbb{E}[(\hat{A} - A)(A - \mathbb{E}[A])] \leq d + \delta^2 + 2\delta \leq d + 3\delta$. Similarly, $|\mathbb{E}[\hat{A}(\hat{A} - A)]| \leq 2\delta$ and $\mathbb{E}[\|Z - \hat{Z}\|^2] \leq d\epsilon^2 + 7\delta\epsilon + \delta$. Choosing $\delta = \epsilon$, we have $\mathbb{E}[\|Z - \hat{Z}\|^2] \leq (d+8)\epsilon^2$. Combining Theorem 24.11 with the Gaussian rate-distortion function in Theorem 26.4, we have $I(Z; \hat{Z}) \geq \frac{d}{2} \log \frac{1}{(d+8)\epsilon^2}$, so applying $\log(1+x) \leq x$ yields

$$I(\theta; \hat{\theta}) \geq (d-1) \log \frac{1}{\epsilon^2} - 4 \log e.$$

□

Exercises for Part V

- V.1** Let $\mathcal{S} = \hat{\mathcal{S}} = \{0, 1\}$ and let the source X^{10} be fair coin flips. Denote the output of the decompressor by \hat{X}^{10} . Show that it is possible to achieve average Hamming distortion $\frac{1}{20}$ with 512 codewords.
- V.2** Assume the distortion function is separable. Show that the minimal number of codewords $M^*(n, D)$ required to represent memoryless source X^n with average distortion D satisfies

$$\log M^*(n_1 + n_2, D) \leq \log M^*(n_1, D) + \log M^*(n_2, D).$$

Conclude that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M^*(n, D) = \inf_n \frac{1}{n} \log M^*(n, D). \quad (\text{V.1})$$

(i.e. one can always achieve a better compression rate by using a longer blocklength). Neither claim holds for $\log M^*(n, \epsilon)$ in channel coding (with inf replaced by sup in (V.1) of course). Explain why this different behavior arises.

- V.3** Consider a source $S^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. Answer the following questions *when n is large*.
- Suppose the goal is to compress S^n into k bits so that one can reconstruct S^n with at most one bit of error. That is, the decoded version \hat{S}^n satisfies $\mathbb{E}[d_H(\hat{S}^n, S^n)] \leq 1$. Show that this can be done (if possible, with an explicit algorithm) with $k = n - C \log n$ bits for some constant C . Is it optimal?
 - Suppose we are required to compress S^n into only 1 bit. Show that one can achieve (if possible, with an explicit algorithm) a reconstruction error $\mathbb{E}[d_H(\hat{S}^n, S^n)] \leq \frac{n}{2} - C\sqrt{n}$ for some constant C . Is it optimal?

Warning: We cannot blindly apply asymptotic the rate-distortion theory to show achievability since here the distortion changes with n . The converse, however, directly applies.

- V.4** (Noisy source coding [93]) Let $Z^n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. Let X^n be the output of a stationary memoryless binary erasure channel with erasure probability δ when the input is Z^n .
- Find the best compression rate for X^n so that the decompressor can reconstruct Z^n with bit error rate D .
 - What if the input is a $\text{Ber}(p)$ sequence?

- V.5** (a) Let $0 \prec \Delta \preceq \Sigma$ be positive definite matrices. For $S \sim \mathcal{N}(0, \Sigma)$, show that

$$\inf_{P_{\hat{S}|S}: \mathbb{E}[(S - \hat{S})(S - \hat{S})^\top] \preceq \Delta} I(S; \hat{S}) = \frac{1}{2} \log \frac{\det \Sigma}{\det \Delta}.$$

(Hint: for achievability, consider $S = \hat{S} + Z$ with $\hat{S} \sim \mathcal{N}(0, \Sigma - \Delta) \perp\!\!\!\perp Z \sim \mathcal{N}(0, \Delta)$ and apply Example 3.4; for converse, follow the proof of Theorem 26.4.)

474 Exercises for Part V

(b) Prove the following extension of (26.3): Let $\sigma_1^2, \dots, \sigma_d^2$ be the eigenvalues of Σ . Then

$$\inf_{P_{\hat{S}|S}: \mathbb{E}[\|S - \hat{S}\|_2^2] \leq D} I(S; \hat{S}) = \frac{1}{2} \sum_{i=1}^d \log^+ \frac{\sigma_i^2}{\lambda}$$

where $\lambda > 0$ is such that $\sum_{i=1}^d \min\{\sigma_i^2, \lambda\} = D$. This is the counterpart of the waterfilling solution in Theorem 20.14.

(Hint: First, using the orthogonal invariance of distortion metric we can assume that Σ is diagonal. Next, apply the same single-letterization argument for (26.3) and solve $\min_{\sum D_i=D} \frac{1}{2} \sum_{i=1}^d \log^+ \frac{\sigma_i^2}{D_i}$.)

V.6 (Shannon lower bound) Let $\|\cdot\|$ be an arbitrary norm on \mathbb{R}^d and $r > 0$. Let X be a \mathbb{R}^d -valued random vector with a probability density function p_X . Denote the rate-distortion function

$$\phi_X(D) \triangleq \inf_{P_{\hat{X}|X}: \mathbb{E}[\|\hat{X} - X\|^r] \leq D} I(X; \hat{X})$$

Prove the Shannon lower bound (26.5), namely

$$\phi_X(D) \geq h(X) + \frac{d}{r} \log \frac{d}{Dr} - \log \left(\Gamma \left(\frac{d}{r} + 1 \right) V \right), \quad (\text{V.2})$$

where the differential entropy $h(X) = \int_{\mathbb{R}^d} p_X(x) \frac{1}{p_X(x)} dx$ is assumed to be finite and $V = \text{vol}(\{x \in \mathbb{R}^d : \|x\| \leq 1\})$.

(a) Show that $0 < V < \infty$.

(b) Show that for any $s > 0$,

$$Z(s) \triangleq \int_{\mathbb{R}^d} \exp(-s\|w\|^r) dw = \Gamma \left(\frac{d}{r} + 1 \right) Vs^{-\frac{d}{r}}.$$

(Hint: Apply Fubini's theorem to $\int_{\mathbb{R}^d} \exp(-s\|w\|^r) dw = \int_{\mathbb{R}^d} \int_{\|w\|^r}^{\infty} s \exp(-sx) dx dw$ and use $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$.)

(c) Show that for any feasible $P_{X|\hat{X}}$ such that $\mathbb{E}[\|X - \hat{X}\|^r] \leq D$,

$$I(X; \hat{X}) \geq h(X) - \log Z(s) - sD.$$

(Hint: Define an auxiliary backward channel $Q_{X|\hat{X}}(dx|\hat{x}) = q_s(x - \hat{x})dx$, where $q_s(w) = \frac{1}{Z(s)} \exp(-s\|w\|^r)$. Then $I(X; \hat{X}) = \mathbb{E}_P[\log \frac{Q_{X|\hat{X}}}{P_X}] + D(P_{X|\hat{X}}||Q_{X|\hat{X}}||P_{\hat{X}})$.)

(d) Optimize over $s > 0$ to conclude (V.2).

(e) Verify that the lower bound of Theorem 26.5 is a special case of (V.2).

Note: Alternatively, the SLB can be written in the following form:

$$\phi_X(D) \geq h(X) - \sup_{P_W: \mathbb{E}[\|W\|^r] \leq D} h(W)$$

and this entropy maximization can be solved following the argument in Example 5.2.

V.7 (Uniform distribution minimizes convex symmetric functional.) Let G be a group acting on a set \mathcal{X} such that each $g \in G$ sends $x \in \mathcal{X}$ to $gx \in \mathcal{X}$. Suppose G acts transitively, i.e., for each $x, x' \in \mathcal{X}$ there exists $g \in G$ such that $gx = x'$. Let g be a random element of G with an invariant

Exercises for Part V 475

distribution, namely $hg \stackrel{d}{=} g$ for any $h \in G$. (Such a distribution, known as the Haar measure, exists for compact topological groups.)

- (a) Show that for any $x \in \mathcal{X}$, gx has the same law, denoted by $\text{Unif}(\mathcal{X})$, the uniform distribution on \mathcal{X} .
- (b) Let $f : \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$ be convex and G -invariant, i.e., $f(P_{gx}) = f(P_X)$ for any \mathcal{X} -valued random variable X and any $g \in G$. Show that $\min_{P_X \in \mathcal{P}(\mathcal{X})} f(P_X) = f(\text{Unif}(\mathcal{X}))$.

V.8 (Uniform distribution maximizes rate-distortion function.) Under the setup of Exercise V.7, let $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a G -invariant distortion function, i.e., $d(gx, gx') = d(x, x')$ for any $g \in G$. Denote the rate-distortion function of an \mathcal{X} -valued X by $\phi_X(D) = \inf_{P_{\hat{X}|X}: \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X})$. Suppose that $\phi_X(D) < \infty$ for all X and all $D > 0$.

- (a) Let $\phi_X^*(\lambda) = \sup_D \{\lambda D - \phi_X(D)\}$ denote the conjugate of ϕ_X . Applying Theorem 24.7 and Fenchel-Moreau's biconjugation theorem to conclude that $\phi_X(D) = \sup_\lambda \{\lambda D - \phi_X^*(\lambda)\}$.
- (b) Show that

$$\phi_X^*(\lambda) = \sup_{P_{\hat{X}|X}} \{\lambda \mathbb{E}[d(X, \hat{X})] - I(X; \hat{X})\}.$$

As such, for each λ , $P_X \mapsto \phi_X^*(\lambda)$ is convex and G -invariant. (Hint: Theorem 5.5.)

- (c) Applying Exercise V.7 to conclude that $\phi_U^*(\lambda) \leq \phi_X^*(\lambda)$ for $U \sim \text{Unif}(\mathcal{X})$ and that

$$\phi_X(D) \leq \phi_U(D), \quad \forall D > 0.$$

V.9 (Normal tail bound.) Denote the standard normal density and tail probability by $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi^c(t) = \int_t^\infty \varphi(x) dx$. Show that for all $t > 0$,

$$\frac{t}{1+t^2} \varphi(t) \leq \Phi^c(t) \leq \min \left\{ \frac{\varphi(t)}{t}, e^{-t^2/2} \right\}. \quad (\text{V.3})$$

(Hint: For $\Phi^c(t) \leq e^{-t^2/2}$ apply the Chernoff bound (15.2); for the rest, note that by integration by parts $\Phi^c(t) = \frac{\varphi(t)}{t} - \int_t^\infty \frac{\varphi(x)}{x^2} dx$.)

V.10 (Small-ball probability II.) In this exercise we prove (27.36). Let $\{W_t : t \geq 0\}$ be a standard Brownian motion. Show that for small ϵ ,¹⁵

$$\phi(\epsilon) = \log \frac{1}{\mathbb{P} \left[\sup_{t \in [0,1]} |W_t| \leq \epsilon \right]} \asymp \frac{1}{\epsilon^2}$$

- (a) By rescaling space and time, show that $\mathbb{P} \left[\sup_{t \in [0,1]} |W_t| \leq \epsilon \right] = \mathbb{P} \left[\sup_{t \in [0,T]} |W_t| \leq 1 \right] \triangleq p_T$, where $T = 1/\epsilon^2$. To show $p_T = e^{-\Theta(T)}$, there is no loss of generality to assume that T is an integer.
- (b) (Upper bound) Using the independent increment property, show that $p_{T+1} \leq ap_T$, where $a = \mathbb{P} [|Z| \leq 1]$ with $Z \sim N(0, 1)$. (Hint: $g(z) \triangleq \mathbb{P} [|Z - z| \leq 1]$ for $z \in [-1, 1]$ is maximized at $z = 0$ and minimized at $z = \pm 1$.)

¹⁵ Using the large-deviations theory developed by Donsker-Varadhan, the sharp constant can be found to be $\lim_{\epsilon \rightarrow 0} \epsilon^2 \phi(\epsilon) = \frac{\pi^2}{8}$. see for example [195, Sec. 6.2].

476 Exercises for Part V

(c) (Lower bound) Again by scaling, it is equivalent to show $\mathbb{P} \left[\sup_{t \in [0, T]} |W_t| \leq C \right] \geq C^{-T}$ for some constant C . Let $q_T \triangleq \mathbb{P} \left[\sup_{t \in [0, T]} |W_t| \leq 2, \max_{t=1, \dots, T} |W_t| \leq 1 \right]$. Show that $q_{T+1} \geq b q_T$, where $b = \mathbb{P} [|Z - 1| \leq 1] \mathbb{P} [\sup_{t \in [0, 1]} |B_t| \leq 1]$, and $B_t = B_t - tB_1$ is a Brownian bridge. (Hint: $\{W_t : t \in [0, T]\}$, $W_{T+1} - W_T$, and $\{W_{T+t} - (1-t)W_T - tW_{T+1} : t \in [0, 1]\}$ are mutually independent, with the latter distributed as a Brownian bridge.)

V.11 (Covering radius in Hamming space) In this exercise we prove (27.9), namely, for any fixed $0 \leq D \leq 1$, as $n \rightarrow \infty$,

$$N(\mathbb{F}_2^n, d_H, Dn) = 2^{n(1-h(D))_+ + o(n)},$$

where $h(\cdot)$ is the binary entropy function.

- (a) Prove the lower bound by invoking the volume bound in Theorem 27.4 and the large-deviations estimate in Example 15.1.
- (b) Prove the upper bound using probabilistic construction and a similar argument to (25.8).
- (c) Show that for $D \geq \frac{1}{2}$, $N(\mathbb{F}_2^n, d_H, Dn) \leq 2$ – cf. Ex. V.3a.

V.12 (Covering ℓ_p -ball with ℓ_q -balls)

- (a) For $1 \leq p < q \leq \infty$, prove the bound (27.18) on the metric entropy of the unit ℓ_p -ball with respect to the ℓ_q -norm (Hint: for small ϵ , apply the volume calculation in (27.15)–(27.16) and the formula in (27.13); for large ϵ , proceed as in the proof of Theorem 27.9 by applying the quantization argument and the Gilbert-Varshamov bound of Hamming spheres.)
- (b) What happens when $p > q$?

V.13 (Random matrix) Let A be an $m \times n$ matrix of iid $\mathcal{N}(0, 1)$ entries. Denote its operator norm by $\|A\|_{\text{op}} = \max_{v \in S^{n-1}} \|Av\|$, which is also the largest singular value of A .

- (a) Show that

$$\|A\|_{\text{op}} = \max_{u \in S^{m-1}, v \in S^{n-1}} \langle A, uv' \rangle. \quad (\text{V.4})$$

- (b) Let $\mathcal{U} = \{u_1, \dots, u_M\}$ and $\mathcal{V} = \{v_1, \dots, v_M\}$ be an ϵ -net for the spheres S^{m-1} and S^{n-1} respectively. Show that

$$\|A\|_{\text{op}} \leq \frac{1}{(1-\epsilon)^2} \max_{u \in \mathcal{U}, v \in \mathcal{V}} \langle A, uv' \rangle.$$

- (c) Apply Corollary 27.1 and Lemma 27.12 to conclude that

$$\mathbb{E}[\|A\|] \lesssim \sqrt{n} + \sqrt{m} \quad (\text{V.5})$$

- (d) By choosing u and v in (V.4) smartly, show a matching lower bound and conclude that

$$\mathbb{E}[\|A\|] \asymp \sqrt{n} + \sqrt{m} \quad (\text{V.6})$$

- (e) Use Sudakov minorization (Theorem 27.10) to prove a matching lower bound. (Hint: use (27.6)).

Part VI

Statistical applications



This part gives an exposition on the application of information-theoretic principles and methods in mathematical statistics; we do so by discussing a selection of topics. To start, Chapter 28 introduces the basic decision-theoretic framework of statistical estimation and the *Bayes risk* and the *minimax risk* as the fundamental limits. Chapter 29 gives an exposition of the classical large-sample asymptotics for smooth parametric models in fixed dimensions, highlighting the role of Fisher information introduced in Chapter 2. Notably, we discuss how to deduce classical lower bounds (Hammersley-Chapman-Robbins, Cramér-Rao, van Trees) from the variational characterization and the data processing inequality (DPI) of χ^2 -divergence in Chapter 7.

Moving into high dimensions, Chapter 30 introduces the *mutual information method* for statistical lower bound, based on the DPI for mutual information as well as the theory of capacity and rate-distortion function from Parts IV and V. This principled approach includes three popular methods for proving minimax lower bounds (Le Cam, Assouad, and Fano) as special cases, which are discussed at length in Chapter 31 drawing results from metric entropy in Chapter 27 also.

Complementing the exposition on lower bounds in Chapters 30 and 31, in Chapter 32 we present three upper bounds on statistical estimation based on metric entropy. These bounds appear strikingly similar but follow from completely different methodologies.

Chapter 33 introduces *strong data processing inequalities* (SDPI), which are quantitative strengthenings of DPIS in Part I. As applications we show how to apply SDPI to deduce lower bounds for various estimation problems on graphs or in distributed settings.

28

Basics of statistical decision theory

28.1 Basic setting

We start by presenting the basic elements of statistical decision theory. We refer to the classics [122, 189, 289] for a systematic treatment.

A *statistical experiment* or *statistical model* refers to a collection \mathcal{P} of probability distributions (over a common measurable space $(\mathcal{X}, \mathcal{F})$). Specifically, let us consider

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \quad (28.1)$$

where each distribution is indexed by a *parameter* θ taking values in the *parameter space* Θ .

In the decision-theoretic framework, we play the following game: Nature picks some parameter $\theta \in \Theta$ and generates a random variable $X \sim P_\theta$. A statistician observes the data X and wants to infer the parameter θ or its certain attributes. Specifically, consider some functional $T : \Theta \rightarrow \mathcal{Y}$ and the goal is to estimate $T(\theta)$ on the basis of the observation X . Here the *estimand* $T(\theta)$ may be the parameter θ itself, or some function thereof (e.g. $T(\theta) = 1_{\{\theta > 0\}}$ or $\|\theta\|$).

An *estimator* (decision rule) is a function $\hat{T} : \mathcal{X} \rightarrow \hat{\mathcal{Y}}$. Note that the action space $\hat{\mathcal{Y}}$ need not be the same as \mathcal{Y} (e.g. \hat{T} may be a confidence interval). Here \hat{T} can be either *deterministic*, i.e. $\hat{T} = \hat{T}(X)$, or *randomized*, i.e., \hat{T} obtained by passing X through a conditional probability distribution (Markov transition kernel) $P_{\hat{T}|X}$ or a channel in the language of Part I. For all practical purposes, we can write $\hat{T} = \hat{T}(X, U)$, where U denotes external randomness uniform on $[0, 1]$ and independent of X .

To measure the quality of an estimator \hat{T} , we introduce a *loss function* $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \rightarrow \mathbb{R}$ such that $\ell(T, \hat{T})$ is the risk of \hat{T} for estimating T . Since we are dealing with loss (as opposed to reward), all the negative (converse) results are lower bounds and all the positive (achievable) results are upper bounds. Note that X is a random variable, so are \hat{T} and $\ell(T, \hat{T})$. Therefore, to make sense of “minimizing the loss”, we consider the average risk:

$$R_\theta(\hat{T}) = \mathbb{E}_\theta[\ell(T, \hat{T})] = \int P_\theta(dx) P_{\hat{T}|X}(d\hat{T}|x) \ell(T(\theta), \hat{T}), \quad (28.2)$$

which we refer to as the risk of \hat{T} at θ . The subscript in \mathbb{E}_θ indicates the distribution with respect to which the expectation is taken. Note that the expected risk depends on the estimator as well as the ground truth.

28.1 Basic setting 481

Remark 28.1. We note that the problem of hypothesis testing and inference can be encompassed as special cases of the estimation paradigm. As previously discussed in Section 16.4, there are three formulations for testing:

- Simple vs. simple hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1, \quad \theta_0 \neq \theta_1$$

- Simple vs. composite hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1, \quad \theta_0 \notin \Theta_1$$

- Composite vs. composite hypotheses

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

For each case one can introduce the appropriate parameter space and loss function. For example, in the last (most general) case, we may take

$$\Theta = \Theta_0 \cup \Theta_1, \quad T(\theta) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_1 \end{cases}, \quad \hat{T} \in \{0, 1\}$$

and use the zero-one loss $\ell(T, \hat{T}) = 1_{\{T \neq \hat{T}\}}$ so that the expected risk $R_\theta(\hat{T}) = P_\theta\{\theta \notin \Theta_{\hat{T}}\}$ is the probability of error.

For the problem of inference, the goal is to output a confidence interval (or region) which covers the true parameter with high probability. In this case \hat{T} is a subset of Θ and we may choose the loss function $\ell(\theta, \hat{T}) = 1_{\{\theta \notin \hat{T}\}} + \lambda \text{length}(\hat{T})$ for some $\lambda > 0$, in order to balance the coverage and the size of the confidence interval.

Remark 28.2 (Randomized versus deterministic estimators). Although most of the estimators used in practice are deterministic, there are a number of reasons to consider randomized estimators:

- For certain formulations, such as the minimizing worst-case risk (minimax approach), deterministic estimators are suboptimal and it is necessary to randomize. On the other hand, if the objective is to minimize the average risk (Bayes approach), then it does not lose generality to restrict to deterministic estimators.
- The space of randomized estimators (viewed as Markov kernels) is convex which is the convex hull of deterministic estimators. This convexification is needed for example for the treatment of minimax theorems.

See Section 28.3 for a detailed discussion and examples.

A well-known fact is that for convex loss function (i.e., $\hat{T} \mapsto \ell(T, \hat{T})$ is convex), randomization does not help. Indeed, for any randomized estimator \hat{T} , we can derandomize it by considering its conditional expectation $\mathbb{E}[\hat{T}|X]$, which is a deterministic estimator and whose risk dominates that of the original \hat{T} at every θ , namely, $R_\theta(\hat{T}) = \mathbb{E}_\theta \ell(T, \hat{T}) \geq \mathbb{E}_\theta \ell(T, \mathbb{E}[\hat{T}|X])$, by Jensen's inequality.

28.2 Gaussian Location Model (GLM)

Note that, without loss of generality, all statistical models can be expressed in the parametric form of (28.1) (since we can take θ to be the distribution itself). In the statistics literature, it is customary to refer to a model as *parametric* if θ takes values in a finite-dimensional Euclidean space (so that each distribution is specified by finitely many parameters), and *nonparametric* if θ takes values in some infinite-dimensional space (e.g. density estimation or sequence model).

Perhaps the most basic parametric model is the Gaussian Location Model (GLM), also known as the Normal Mean Model, which corresponds to our familiar Gaussian channel in Example 3.3. This will be our running example in this part of the book. In this model, we have

$$\mathcal{P} = \{N(\theta, \sigma^2 I_d) : \theta \in \Theta\}$$

where I_d is the d -dimensional identity matrix and the parameter space $\Theta \subset \mathbb{R}^d$. Equivalently, we can express the data as a noisy observation of the unknown vector θ as:

$$X = \theta + Z, \quad Z \sim N(0, \sigma^2 I_d).$$

The case of $d = 1$ and $d > 1$ refers to the univariate (scalar) and multivariate (vector) case, respectively. (Also of interest is the case where θ is a $d_1 \times d_2$ matrix, which can be vectorized into a $d = d_1 d_2$ -dimensional vector.)

The choice of the parameter space Θ represents our prior knowledges of the unknown parameter θ , for example,

- $\Theta = \mathbb{R}^d$, in which case there is no assumption on θ .
- $\Theta = \ell_p$ -norm balls.
- $\Theta = \{\text{all } k\text{-sparse vectors}\} = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}$, where $\|\theta\|_0 \triangleq |\{i : \theta_i \neq 0\}|$ denotes the size of the support, informally referred to as the ℓ_0 -“norm”.
- $\Theta = \{\theta \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(\theta) \leq r\}$, the set of low-rank matrices.

By definition, more structure (smaller parameter space) always makes the estimation task easier (smaller worst-case risk), but not necessarily so in terms of computation.

For estimating θ itself (denoising), it is customary to use a loss function defined by certain norms, e.g., $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_p^\alpha$ for some $1 \leq p \leq \infty$ and $\alpha > 0$, where $\|\theta\|_p \triangleq (\sum |\theta_i|^p)^{\frac{1}{p}}$, with $p = \alpha = 2$ corresponding to the commonly used *quadratic loss* (squared error). Some well-known estimators include the Maximum Likelihood Estimator (MLE)

$$\hat{\theta}_{\text{ML}} = X \tag{28.3}$$

and the James-Stein estimator based on shrinkage

$$\hat{\theta}_{\text{JS}} = \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right) X \tag{28.4}$$

The choice of the estimator depends on both the objective and the parameter space. For instance, if θ is known to be sparse, it makes sense to set the smaller entries in the observed X to zero (thresholding) in order to better denoise θ (cf. Section 30.2).

28.3 Bayes risk, minimax risk, and the minimax theorem 483

In addition to estimating the vector θ itself, it is also of interest to estimate certain functionals $T(\theta)$ thereof, e.g., $T(\theta) = \|\theta\|_p$, $\max\{\theta_1, \dots, \theta_d\}$, or eigenvalues in the matrix case. In addition, the hypothesis testing problem in the GLM has been well-studied. For example, one can consider detecting the presence of a signal by testing $H_0 : \theta = 0$ against $H_1 : \|\theta\| \geq \epsilon$, or testing weak signal $H_0 : \|\theta\| \leq \epsilon_0$ versus strong signal $H_1 : \|\theta\| \geq \epsilon_1$, with or without further structural assumptions on θ . We refer the reader to the monograph [161] devoted to these problems.

28.3 Bayes risk, minimax risk, and the minimax theorem

One of our main objectives in this part of the book is to understand the fundamental limit of statistical estimation, that is, to determine the performance of the best estimator. As in (28.2), the risk $R_\theta(\hat{T})$ of an estimator \hat{T} for $T(\theta)$ depends on the ground truth θ . To compare the risk profiles of different estimators meaningfully requires some thought. As a toy example, Fig. 28.1 depicts the risk functions of three estimators. It is clear that $\hat{\theta}_1$ is superior to $\hat{\theta}_2$ in the sense that the risk of the former is pointwise lower than that of the latter. (In statistical literature we say $\hat{\theta}_2$ is inadmissible.) However, the comparison of $\hat{\theta}_1$ and $\hat{\theta}_3$ is less clear. Although the peak risk value of $\hat{\theta}_3$ is bigger than that of $\hat{\theta}_1$, on average its risk (area under the curve) is smaller. In fact, both views are valid and meaningful, and they correspond to the worst-case (minimax) and average-case (Bayesian) approach, respectively. In the minimax formulation, we summarize the risk function into a scalar quantity, namely, the worst-case risk, and seek the estimator that minimize this objective. In the Bayesian formulation, the objective is the average risk. Below we discuss these two approaches and their connections. For notational simplicity, we consider the task of estimating $T(\theta) = \theta$.

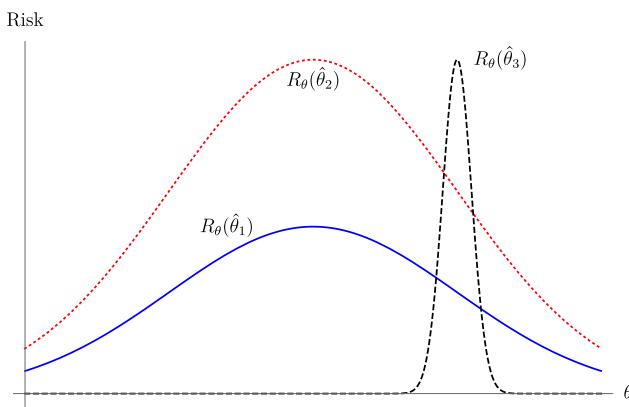


Figure 28.1 Risk profiles of three estimators.

28.3.1 Bayes risk

The Bayesian approach is an average-case formulation in which the statistician acts as if the parameter θ is random with a known distribution. Concretely, let π be a probability distribution (prior) on Θ . Then the *average risk* (w.r.t. π) of an estimator $\hat{\theta}$ is defined as

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta \sim \pi}[R_\theta(\hat{\theta})] = \mathbb{E}_{\theta, X}[\ell(\theta, \hat{\theta})]. \quad (28.5)$$

Given a prior π , its *Bayes risk* is the minimal average risk, namely

$$R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta}).$$

An estimator $\hat{\theta}^*$ is called a *Bayes estimator* if it attains the Bayes risk, namely, $R_\pi^* = \mathbb{E}_{\theta \sim \pi}[R_\theta(\hat{\theta}^*)]$.

Remark 28.3. Bayes estimator is always deterministic – this fact holds for any loss function. To see this, note that for any randomized estimator, say $\hat{\theta} = \hat{\theta}(X, U)$, where U is some external randomness independent of X and θ , its risk is lower bounded by

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta, X, U} \ell(\theta, \hat{\theta}(X, U)) = \mathbb{E}_U R_\pi(\hat{\theta}(\cdot, U)) \geq \inf_u R_\pi(\hat{\theta}(\cdot, u)).$$

Note that for any u , $\hat{\theta}(\cdot, u)$ is a deterministic estimator. This shows that we can find a deterministic estimator whose average risk is no worse than that of the randomized estimator.

An alternative way to under this fact is the following: Note that the average risk $R_\pi(\hat{\theta})$ defined in (28.5) is an affine function of the randomized estimator (understood as a Markov kernel $P_{\hat{\theta}|X}$) is affine, whose minimum is achieved at the extremal points. In this case the extremal points of Markov kernels are simply delta measures, which corresponds to deterministic estimators.

In certain settings the Bayes estimator can be found explicitly. Consider the problem of estimating $\theta \in \mathbb{R}^d$ drawn from a prior π . Under the quadratic loss $\ell(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$, the Bayes estimator is the conditional mean $\hat{\theta}(X) = \mathbb{E}[\theta|X]$ and the Bayes risk is the *minimum mean-square error (MMSE)*

$$R_\pi^* = \mathbb{E}\|\theta - \mathbb{E}[\theta|X]\|_2^2 = \text{Tr}(\text{Cov}(\theta|X)),$$

where $\text{Cov}(\theta|X = x)$ is the conditional covariance of θ given $X = x$.

As a concrete example, let us consider the Gaussian Location Model in Section 28.2 with a Gaussian prior.

Example 28.1 (Bayes risk in GLM). Consider the scalar case, where $X = \theta + Z$ and $Z \sim N(0, \sigma^2)$ is independent of θ . Consider a Gaussian prior $\theta \sim \pi = N(0, s)$. One can verify that the posterior distribution $P_{\theta|X=x}$ is $N(\frac{s}{s+\sigma^2}x, \frac{\sigma^2}{s+\sigma^2})$. As such, the Bayes estimator is $\mathbb{E}[\theta|X] = \frac{s}{s+\sigma^2}X$ and the Bayes risk is

$$R_\pi^* = \frac{s\sigma^2}{s + \sigma^2}. \quad (28.6)$$

Similarly, for multivariate GLM: $X = \theta + Z$, $Z \sim N(0, I_d)$, if $\theta \sim \pi = N(0, sI_d)$, then we have

$$R_\pi^* = \frac{s\sigma^2}{s + \sigma^2}d. \quad (28.7)$$

28.3 Bayes risk, minimax risk, and the minimax theorem 485

28.3.2 Minimax risk

A common criticism of the Bayesian approach is the arbitrariness of the selected prior. A framework related to this but not discussed in this case is the empirical Bayes approach [?], where one “estimates” the prior from the data instead of choosing a prior a priori. Instead, we take a frequentist viewpoint by considering the worst-case situation. The *minimax risk* is defined as

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta}). \quad (28.8)$$

If there exists $\hat{\theta}$ s.t. $\sup_{\theta \in \Theta} R_\theta(\hat{\theta}) = R^*$, then the estimator $\hat{\theta}$ is minimax (minimax optimal).

Finding the value of the minimax risk R^* entails proving two things, namely,

- a minimax upper bound, by exhibiting an estimator $\hat{\theta}^*$ such that $R_\theta(\hat{\theta}^*) \leq R^* + \epsilon$ for all $\theta \in \Theta$;
- a minimax lower bound, by proving that for any estimator $\hat{\theta}$, there exists some $\theta \in \Theta$, such that $R_\theta \geq R^* - \epsilon$,

where $\epsilon > 0$ is arbitrary. This task is frequently difficult especially in high dimensions. Instead of the exact minimax risk, it is often useful to find a constant-factor approximation Ψ , which we call *minimax rate*, such that

$$R^* \asymp \Psi, \quad (28.9)$$

that is, $c\Psi \leq R^* \leq C\Psi$ for some universal constants $c, C \geq 0$. Establishing Ψ is the minimax rate still entails proving the minimax upper and lower bounds, albeit within multiplicative constant factors.

In practice, minimax lower bounds are rarely established according to the original definition. The next result shows that the Bayes risk is always lower than the minimax risk. Throughout this book, all lower bound techniques essentially boil down to evaluating the Bayes risk with a sagaciously chosen prior.

Theorem 28.1. *Let $\Delta(\Theta)$ denote the collection of probability distributions on Θ . Then*

$$R^* \geq R_{\text{Bayes}}^* \triangleq \sup_{\pi \in \Delta(\Theta)} R_\pi^*. \quad (28.10)$$

Proof. Two (equivalent) ways to prove this fact:

- 1 “max \geq mean”: For any $\hat{\theta}$, $R_\pi(\hat{\theta}) = \mathbb{E}_{\theta \sim \pi} R_\theta(\hat{\theta}) \leq \sup_{\theta \in \Theta} R_\theta(\hat{\theta})$. Taking the infimum over $\hat{\theta}$ completes the proof;
- 2 “min max \geq max min”:

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\pi \in \Delta(\Theta)} R_\pi(\hat{\theta}) \geq \sup_{\pi \in \Delta(\Theta)} \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = \sup_{\pi} R_\pi^*,$$

where the inequality follows from the generic fact that $\min_x \max_y f(x, y) \geq \max_y \min_x f(x, y)$. \square

Remark 28.4. Unlike Bayes estimators which, as shown in Remark 28.3, are always deterministic, to minimize the worst-case risk it is sometimes necessary to randomize for example in the context of hypotheses testing (Chapter 14). Specifically, consider a trivial experiment where $\theta \in \{0, 1\}$ and X is absent, so that we are forced to guess the value of θ under the zero-one loss $\ell(\theta, \hat{\theta}) = 1_{\{\theta \neq \hat{\theta}\}}$. It is clear that in this case the minimax risk is $\frac{1}{2}$, achieved by random guessing $\hat{\theta} \sim \text{Ber}(\frac{1}{2})$ but not by any deterministic $\hat{\theta}$.

As an application of Theorem 28.4, let us determine the minimax risk of the Gaussian location model under the quadratic loss function.

Example 28.2 (Minimax quadratic risk of GLM). Consider the Gaussian location model without structural assumptions, where $X \sim N(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$. We show that

$$R^* \equiv \inf_{\theta \in \mathbb{R}^d} \sup_{\hat{\theta} \in \mathbb{R}^d} \mathbb{E}_{\theta}[\|\hat{\theta}(X) - \theta\|_2^2] = d\sigma^2. \quad (28.11)$$

By scaling, it suffices to consider $\sigma = 1$. For the upper bound, we consider $\hat{\theta}_{\text{ML}} = X$ which achieves $R_{\theta}(\hat{\theta}_{\text{ML}}) = d$ for all θ . To get a matching minimax lower bound, we consider the prior $\theta \sim N(0, s)$. Using the Bayes risk previously computed in (28.6), we have $R^* \geq R_{\pi}^* = \frac{sd}{s+1}$. Sending $s \rightarrow \infty$ yields $R^* \geq d$.

Remark 28.5 (Non-uniqueness of minimax estimators). In general, estimators that achieve the minimax risk need not be unique. For instance, as shown in Example 28.2, the MLE $\hat{\theta}_{\text{ML}} = X$ is minimax for the unconstrained GLM in any dimension. On the other hand, it is known that whenever $d \geq 3$, the risk of the James-Stein estimator (28.4) is smaller than of the MLE everywhere (see Fig. 28.2) and thus is also minimax. In fact, there exist a continuum of estimators that are minimax for (28.11) [192, Theorem 5.5].

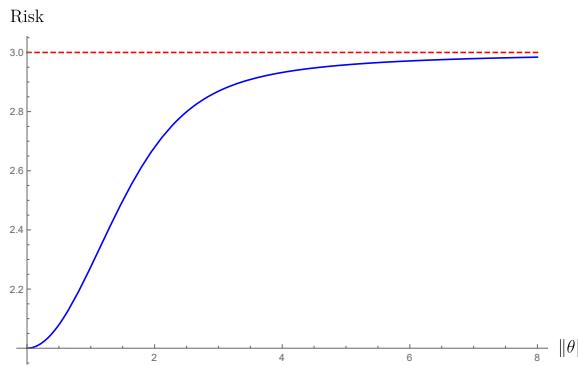


Figure 28.2 Risk of the James-Stein estimator (28.4) in dimension $d = 3$ and $\sigma = 1$ as a function of $\|\theta\|$.

For most of the statistical models, Theorem 28.4 in fact holds with equality; such a result is known as a *minimax theorem*. Before discussing this important topic, here is an example where minimax risk is strictly bigger than the worst-case Bayes risk.

28.3 Bayes risk, minimax risk, and the minimax theorem 487

Example 28.3. Let $\theta, \hat{\theta} \in \mathbb{N} \triangleq \{1, 2, \dots\}$ and $\ell(\theta, \hat{\theta}) = 1_{\{\hat{\theta} < \theta\}}$, i.e., the statistician loses one dollar if the Nature's choice exceeds the statistician's guess and loses nothing if otherwise. Consider the extreme case of blind guessing (i.e., no data is available, say, $X = 0$). Then for any $\hat{\theta}$ possibly randomized, we have $R_\theta(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$. Thus $R^* \geq \lim_{\theta \rightarrow \infty} \mathbb{P}(\hat{\theta} < \theta) = 1$, which is clearly achievable. On the other hand, for any prior π on \mathbb{N} , $R_\pi(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$, which vanishes as $\hat{\theta} \rightarrow \infty$. Therefore, we have $R_\pi^* = 0$. Therefore in this case $R^* = 1 > R_{\text{Bayes}}^* = 0$.

As an exercise, one can show that the minimax quadratic risk of the GLM $X \sim N(\theta, 1)$ with parameter space $\theta \geq 0$ is the same as the unconstrained case. (This might be a bit surprising because the thresholded estimator $X_+ = \max(X, 0)$ achieves a better risk pointwise at every $\theta \geq 0$; nevertheless, just like the James-Stein estimator (cf. Fig. 28.2), in the worst case the gain is asymptotically diminishing.)

28.3.3 Minimax and Bayes risk: a duality perspective

Recall from Theorem 28.4 the inequality

$$R^* \geq R_{\text{Bayes}}^*.$$

This result can be interpreted from an optimization perspective. More precisely, R^* is the value of a convex optimization problem (primal) and R_{Bayes}^* is precisely the value of its dual program. Thus the inequality (28.10) is simply *weak duality*. If *strong duality* holds, then (28.10) is in fact an equality, in which case the minimax theorem holds.

For simplicity, we consider the case where Θ is a finite set. Then

$$R^* = \min_{P_{\hat{\theta}|X}} \max_{\theta \in \Theta} \mathbb{E}_\theta[\ell(\theta, \hat{\theta})]. \quad (28.12)$$

This is a convex optimization problem. Indeed, $P_{\hat{\theta}|X} \mapsto \mathbb{E}_\theta[\ell(\theta, \hat{\theta})]$ is affine and the pointwise supremum of affine functions is convex. To write down its dual problem, first let us rewrite (28.12) in an augmented form

$$\begin{aligned} R^* &= \min_{P_{\hat{\theta}|X}, t} \quad t \\ \text{s.t. } &\mathbb{E}_\theta[\ell(\theta, \hat{\theta})] \leq t, \quad \forall \theta \in \Theta. \end{aligned} \quad (28.13)$$

Let $\pi_\theta \geq 0$ denote the Lagrange multiplier (dual variable) for each inequality constraint. The Lagrangian of (28.13) is

$$L(P_{\hat{\theta}|X}, t, \pi) = t + \sum_{\theta \in \Theta} \pi_\theta \left(\mathbb{E}_\theta[\ell(\theta, \hat{\theta})] - t \right) = \left(1 - \sum_{\theta \in \Theta} \pi_\theta \right) t + \sum_{\theta \in \Theta} \pi_\theta \mathbb{E}_\theta[\ell(\theta, \hat{\theta})].$$

By definition, we have $R^* \geq \min_{t, P_{\hat{\theta}|X}} L(\hat{\theta}, t, \pi)$. Note that unless $\sum_{\theta \in \Theta} \pi_\theta = 1$, $\min_{t \in \mathbb{R}} L(\hat{\theta}, t, \pi)$ is $-\infty$. Thus $\pi = (\pi_\theta : \theta \in \Theta)$ must be a probability measure and the dual problem is

$$\max_{\pi} \min_{P_{\hat{\theta}|X}, t} L(P_{\hat{\theta}|X}, t, \pi) = \max_{\pi \in \Delta(\Theta)} \min_{P_{\hat{\theta}|X}} R_\pi(\hat{\theta}) = \max_{\pi \in \Delta(\Theta)} R_\pi^*.$$

Hence, $R^* \geq R_{\text{Bayes}}^*$.

In summary, the minimax risk and the worst-case Bayes risk are related by convex duality, where the primal variables are (randomized) estimators and the dual variables are priors. This view can in fact be operationalized. For example, [169, 242] showed that for certain problems dualizing Le Cam's two-point lower bound (Theorem 31.1) leads to optimal minimax upper bound; see Exercise VI.16.

28.3.4 Minimax theorem

Next we state the minimax theorem which gives conditions that ensure (28.10) holds with equality, namely, the minimax risk R^* and the worst-case Bayes risk R_{Bayes}^* coincide. For simplicity, let us consider the case of estimating θ itself where the estimator $\hat{\theta}$ takes values in the action space $\hat{\Theta}$ with a loss function $\ell : \Theta \times \hat{\Theta} \rightarrow \mathbb{R}$. A very general result (cf. [289, Theorem 46.6]) asserts that $R^* = R_{\text{Bayes}}^*$, provided that the following condition hold:

- The experiment is dominated, i.e., $P_\theta \ll \nu$ holds for all $\theta \in \Theta$ for some ν on \mathcal{X} .
- The action space $\hat{\Theta}$ is a locally compact topological space with a countable base (e.g. the Euclidean space).
- The loss function is level-compact (i.e., for each $\theta \in \Theta$, $\ell(\theta, \cdot)$ is bounded from below and the sublevel set $\{\hat{\theta} : \ell(\theta, \hat{\theta}) \leq a\}$ is compact for each a).

This result shows that for virtually all problems encountered in practice, the minimax risk coincides with the least favorable Bayes risk. At the heart of any minimax theorem, there is an application of the separating hyperplane theorem. Below we give a proof of a special case illustrating this type of argument.

Theorem 28.2 (Minimax theorem).

$$R^* = R_{\text{Bayes}}^*$$

in either of the following cases:

- Θ is a finite set and the data X takes values in a finite set \mathcal{X} .
- Θ is a finite set and the loss function ℓ is bounded from below, i.e., $\inf_{\theta, \hat{\theta}} \ell(\theta, \hat{\theta}) > -\infty$.

Proof. The first case directly follows from the duality interpretation in Section 28.3.3 and the fact that strong duality holds for finite-dimensional linear programming (see for example [266, Sec. 7.4].

For the second case, we start by showing that if $R^* = \infty$, then $R_{\text{Bayes}}^* = \infty$. To see this, consider the uniform prior π on Θ . Then for any estimator $\hat{\theta}$, there exists $\theta \in \Theta$ such that $R(\theta, \hat{\theta}) = \infty$. Then $R_\pi(\hat{\theta}) \geq \frac{1}{|\Theta|} R(\theta, \hat{\theta}) = \infty$.

Next we assume that $R^* < \infty$. Then $R^* \in \mathbb{R}$ since ℓ is bounded from below (say, by a) by assumption. Given an estimator $\hat{\theta}$, denote its risk vector $R(\hat{\theta}) = (R_\theta(\hat{\theta}))_{\theta \in \Theta}$. Then its average risk

28.4 Multiple observations and sample complexity 489

with respect to a prior π is given by the inner product $\langle R(\hat{\theta}), \pi \rangle = \sum_{\theta \in \Theta} \pi_\theta R_\theta(\hat{\theta})$. Define

$$\begin{aligned} S &= \{R(\hat{\theta}) \in \mathbb{R}^\Theta : \hat{\theta} \text{ is a randomized estimator}\} = \text{set of all possible risk vectors}, \\ T &= \{t \in \mathbb{R}^\Theta : t_\theta < R^*, \theta \in \Theta\}. \end{aligned}$$

Note that both S and T are convex (why?) subsets of Euclidean space \mathbb{R}^Θ and $S \cap T = \emptyset$ by definition of R^* . By the separation hyperplane theorem, there exists a non-zero $\pi \in \mathbb{R}^\Theta$ and $c \in \mathbb{R}$, such that $\inf_{s \in S} \langle \pi, s \rangle \geq c \geq \sup_{t \in T} \langle \pi, t \rangle$. Obviously, π must be componentwise positive, for otherwise $\sup_{t \in T} \langle \pi, t \rangle = \infty$. Therefore by normalization we may assume that π is a probability vector, i.e., a prior on Θ . Then $R_{\text{Bayes}}^* \geq R_\pi^* = \inf_{s \in S} \langle \pi, s \rangle \geq \sup_{t \in T} \langle \pi, t \rangle \geq R^*$, completing the proof. \square

28.4 Multiple observations and sample complexity

Given a experiment $\{P_\theta : \theta \in \Theta\}$, consider the experiment

$$\mathcal{P}_n = \{P_\theta^{\otimes n} : \theta \in \Theta\}, \quad n \geq 1. \quad (28.14)$$

We refer to this as the *independent sampling model*, in which we observe a sample $X = (X_1, \dots, X_n)$ consisting of independent observations drawn from P_θ for some $\theta \in \Theta \subset \mathbb{R}^d$. Given a loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$, the minimax risk is denoted by

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta}). \quad (28.15)$$

Clearly, $n \mapsto R_n^*(\Theta)$ is non-increasing since we can always discard the extra observations. Typically, when Θ is a fixed subset of \mathbb{R}^d , $R_n^*(\Theta)$ vanishes as $n \rightarrow \infty$. Thus a natural question is at what rate R_n^* converges to zero. Equivalently, one can consider the *sample complexity*, namely, the minimum sample size to attain a prescribed error ϵ even in the worst case:

$$n^*(\epsilon) \triangleq \min \{n \in \mathbb{N} : R_n^*(\Theta) \leq \epsilon\}. \quad (28.16)$$

In the classical large-sample asymptotics (Chapter 29), the rate of convergence for the quadratic risk is usually $\Theta(\frac{1}{n})$, which is commonly referred to as the “parametric rate”. In comparison, in this book we focus on understanding the dependency on the dimension and other structural parameters nonasymptotically.

As a concrete example, let us revisit the GLM in Section 28.2 with sample size n , in which case we observe $X = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2 I_d)$, $\theta \in \mathbb{R}^d$. In this case, the minimax quadratic risk is¹

$$R_n^* = \frac{d\sigma^2}{n}. \quad (28.17)$$

To see this, note that in this case $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ is a sufficient statistic (cf. Section 3.5) of X for θ . Therefore the model reduces to $\bar{X} \sim N(\theta, \frac{\sigma^2}{n} I_d)$ and (28.17) follows from the minimax risk (28.11) for a single observation.

¹ See Exercise VI.10 for an extension of this result to nonparametric location models.

From (28.17), we conclude that the sample complexity is $n^*(\epsilon) = \lceil \frac{d\sigma^2}{\epsilon} \rceil$, which grows linearly with the dimension d . This is the common wisdom that ‘‘sample complexity scales proportionally to the number of parameters’’, also known as ‘‘counting the degrees of freedom’’. Indeed in high dimensions we typically expect the sample complexity to grow with the ambient dimension; however, the exact dependency need not be linear as it depends on the loss function and the objective of estimation. For example, consider the matrix case $\theta \in \mathbb{R}^{d \times d}$ with n independent observations in Gaussian noise. Let ϵ be a small constant. Then we have

- For quadratic loss, namely, $\|\theta - \hat{\theta}\|_{\text{F}}^2$, we have $R_n^* = \frac{d^2}{n}$ and hence $n^*(\epsilon) = \Theta(d^2)$;
- If the loss function is $\|\theta - \hat{\theta}\|_{\text{op}}^2$, then $R_n^* \asymp \frac{d}{n}$ and hence $n^*(\epsilon) = \Theta(d)$ (Example 28.4);
- As opposed to θ itself, suppose we are content with estimating only the scalar functional $\theta_{\max} = \max\{\theta_1, \dots, \theta_d\}$ up to accuracy ϵ , then $n^*(\epsilon) = \Theta(\sqrt{\log d})$ (Exercise VI.13).

In the last two examples, the sample complexity scales *sublinearly* with the dimension.

28.5 Tensor product of experiments

Tensor product is a way to define a high-dimensional model from low-dimensional models. Given statistical experiments $\mathcal{P}_i = \{P_{\theta_i} : \theta_i \in \Theta_i\}$ and the corresponding loss function ℓ_i , for $i \in [d]$, their tensor product refers to the following statistical experiment:

$$\begin{aligned} \mathcal{P} &= \left\{ P_{\theta} = \prod_{i=1}^d P_{\theta_i} : \theta = (\theta_1, \dots, \theta_d) \in \Theta \triangleq \prod_{i=1}^d \Theta_i \right\}, \\ \ell(\theta, \hat{\theta}) &\triangleq \sum_{i=1}^d \ell_i(\theta_i, \hat{\theta}_i), \forall \theta, \hat{\theta} \in \Theta. \end{aligned}$$

In this model, the observation $X = (X_1, \dots, X_d)$ consists of independent (not identically distributed) $X_i \stackrel{\text{ind}}{\sim} P_{\theta_i}$. This should be contrasted with the multiple-observation model in (28.14), in which n iid observations drawn from the same distribution are given.

The minimax risk of the tensorized experiment is related to the minimax risk $R^*(\mathcal{P}_i)$ and worst-case Bayes risks $R_{\text{Bayes}}^*(\mathcal{P}_i) \triangleq \sup_{\pi_i \in \Delta(\Theta_i)} R_{\pi_i}(\mathcal{P}_i)$ of each individual experiment as follows:

Theorem 28.3 (Minimax risk of tensor product).

$$\sum_{i=1}^d R_{\text{Bayes}}^*(\mathcal{P}_i) \leq R^*(\mathcal{P}) \leq \sum_{i=1}^d R^*(\mathcal{P}_i). \quad (28.18)$$

Consequently, if minimax theorem holds for each experiment, i.e., $R^*(\mathcal{P}_i) = R_{\text{Bayes}}^*(\mathcal{P}_i)$, then it also holds for the product experiment and, in particular,

$$R^*(\mathcal{P}) = \sum_{i=1}^d R^*(\mathcal{P}_i). \quad (28.19)$$

28.5 Tensor product of experiments 491

Proof. The right inequality of (28.18) simply follows by separately estimating θ_i on the basis of X_i , namely, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d)$, where $\hat{\theta}_i$ depends only on X_i . For the left inequality, consider a product prior $\pi = \prod_{i=1}^d \pi_i$, under which θ_i 's are independent and so are X_i 's. Consider any randomized estimator $\hat{\theta}_i = \hat{\theta}_i(X, U_i)$ of θ_i based on X , where U_i is some auxiliary randomness independent of X . We can rewrite it as $\hat{\theta}_i = \hat{\theta}_i(X_i, \tilde{U}_i)$, where $\tilde{U}_i = (X_{\setminus i}, U_i) \perp\!\!\!\perp X_i$. Thus $\hat{\theta}_i$ can be viewed as it a randomized estimator based on X_i alone and its the average risk must satisfy $R_{\pi_i}(\hat{\theta}_i) = \mathbb{E}[\ell(\theta_i, \hat{\theta}_i)] \geq R_{\pi_i}^*$. Summing over i and taking the suprema over priors π_i 's yields the left inequality of (28.18). \square

As an example, we note that the unstructured d -dimensional GLM $\{N(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$ with quadratic loss is simply the d -fold tensor product of the one-dimensional GLM. Since minimax theorem holds for the GLM (cf. Section 28.3.4), Theorem 28.8 shows the minimax risks sum up to $\sigma^2 d$, which agrees with Example 28.2. In general, however, it is possible that the minimax risk of the tensorized experiment is less than the sum of individual minimax risks and the right inequality of (28.19) can be strict. This might appear surprising since X_i only carries information about θ_i and it makes sense intuitively to estimate θ_i based solely on X_i . Nevertheless, the following is a counterexample:

Remark 28.6. Consider $X = \theta Z$, where $\theta \in \mathbb{N}$, $Z \sim \text{Ber}(\frac{1}{2})$. The estimator $\hat{\theta}$ takes values in \mathbb{N} as well and the loss function is $\ell(\theta, \hat{\theta}) = 1_{\{\hat{\theta} < \theta\}}$, i.e., whoever guesses the greater number wins. The minimax risk for this experiment is equal to $\mathbb{P}[Z = 0] = \frac{1}{2}$. To see this, note that if $Z = 0$, then all information about θ is erased. Therefore for any (randomized) estimator $P_{\hat{\theta}|X}$, the risk is lower bounded by $R_\theta(\hat{\theta}) = \mathbb{P}[\hat{\theta} < \theta] \geq \mathbb{P}[\hat{\theta} < \theta, Z = 0] = \frac{1}{2}\mathbb{P}[\hat{\theta} < \theta|X = 0]$. Therefore sending $\theta \rightarrow \infty$ yields $\sup_\theta R_\theta(\hat{\theta}) \geq \frac{1}{2}$. This is achievable by $\hat{\theta} = X$. Clearly, this is a case where minimax theorem does not hold, which is very similar to the previous Example 28.3.

Next consider the tensor product of two copies of this experiment with loss function $\ell(\theta, \hat{\theta}) = 1_{\{\hat{\theta}_1 < \theta_1\}} + 1_{\{\hat{\theta}_2 < \theta_2\}}$. We show that the minimax risk is strictly less than one. For $i = 1, 2$, let $X_i = \theta_i Z_i$, where $Z_1, Z_2 \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. Consider the following estimator

$$\hat{\theta}_1 = \hat{\theta}_2 = \begin{cases} X_1 \vee X_2 & X_1 > 0 \text{ or } X_2 > 0 \\ 1 & \text{otherwise.} \end{cases}$$

Then for any $\theta_1, \theta_2 \in \mathbb{N}$, averaging over Z_1, Z_2 , we get

$$\mathbb{E}[\ell(\theta, \hat{\theta})] \leq \frac{1}{4} (1_{\{\theta_1 < \theta_2\}} + 1_{\{\theta_2 < \theta_1\}} + 1) \leq \frac{3}{4}.$$

We end this section by consider the minimax risk of GLM with non-quadratic loss. The following result extends Example 28.2:

Theorem 28.4. *Consider the Gaussian location model $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\theta, I_d)$. Then for $1 \leq q < \infty$,*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta [\|\theta - \hat{\theta}\|_q^q] = \frac{\mathbb{E}[\|Z\|_q^q]}{n^{q/2}}, \quad Z \sim N(0, I_d).$$

Proof. Note that $N(\theta, I_d)$ is a product distribution and the loss function is separable: $\|\theta - \hat{\theta}\|_q^q = \sum_{i=1}^d |\theta_i - \hat{\theta}_i|^q$. Thus the experiment is a d -fold tensor product of the one-dimensional version. By Theorem 28.8, it suffices to consider $d = 1$. The upper bound is achieved by the sample mean $X = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\theta, \frac{1}{n})$, which is a sufficient statistic.

For the lower bound, following Example 28.2, consider a Gaussian prior $\theta \sim \pi = N(0, s)$. Then the posterior distribution is also Gaussian: $P_{\theta|X} = N(\mathbb{E}[\theta|X], \frac{s}{1+sn})$. The following lemma shows that the Bayes estimator is simply the conditional mean:

Lemma 28.5. *Let $Z \sim N(0, 1)$. Then $\min_{y \in \mathbb{R}} \mathbb{E}[|y + Z|^q] = \mathbb{E}[|Z|^q]$.*

Thus the Bayes risk is

$$R_\pi^* = \mathbb{E}[|\theta - \mathbb{E}[\theta|X]|^q] = \left(\frac{s}{1+sn} \right)^{q/2} \mathbb{E}|Z|^q.$$

Sending $s \rightarrow \infty$ proves the matching lower bound. \square

Proof of Lemma 28.11. Write

$$\mathbb{E}|y + Z|^q = \int_0^\infty \mathbb{P}[|y + Z|^q > c] dc \geq \int_0^\infty \mathbb{P}[|Z|^q > c] dc = \mathbb{E}|Z|^q,$$

where the inequality follows from the simple observation that for any $a > 0$, $\mathbb{P}[|y + Z| \leq a] \leq \mathbb{P}[|Z| \leq a]$, due to the symmetry and unimodality of the normal density. \square

28.6 Log-concavity, Anderson's lemma and exact minimax risk in GLM

As mentioned in Section 28.3.2, computing the exact minimax risk is frequently difficult especially in high dimensions. Nevertheless, for the special case of (unconstrained) GLM, the minimax risk is known exactly in arbitrary dimensions for a large collection of loss functions.² We have previously seen in Theorem 28.10 that this is possible for loss functions of the form $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_q^q$. Examining the proof of this result, we note that the major limitation is that it only applies to separable loss functions, so that tensorization allows us to reduce the problem to one dimension. This does not apply to (and actually fails) for inseparable loss, since Theorem 28.8, if applicable, dictates the risk to grow linearly with the dimension, which is not always the case. We next discuss a more general result that goes beyond separable losses.

Definition 28.6. A function $\rho : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is called *bowl-shaped* if its sublevel set $K_c \triangleq \{x : \rho(x) \leq c\}$ is convex and symmetric (i.e. $K_c = -K_c$) for all $c \in \mathbb{R}$.

² Another example is the multivariate model with the squared error; cf. Exercise VI.7.

28.6 Log-concavity, Anderson's lemma and exact minimax risk in GLM 493

Theorem 28.7. Consider the d -dimensional GLM where $X_1, \dots, X_n \sim N(0, I_d)$ are observed. Let the loss function be $\ell(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta})$, where $\rho : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is bowl-shaped and lower-semicontinuous. Then the minimax risk is given by

$$R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{\theta}[\rho(\theta - \hat{\theta})] = \mathbb{E}\rho\left(\frac{Z}{\sqrt{n}}\right), \quad Z \sim N(0, I_d).$$

Furthermore, the upper bound is attained by $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Corollary 28.8. Let $\rho(\cdot) = \|\cdot\|^q$ for some $q > 0$, where $\|\cdot\|$ is an arbitrary norm on \mathbb{R}^d . Then

$$R^* = \frac{\mathbb{E}\|Z\|^q}{n^{q/2}}. \quad (28.20)$$

Example 28.4. Some applications of Corollary 28.1:

- For $\rho = \|\cdot\|_2^2$, $R^* = \frac{1}{n} \mathbb{E}\|Z\|^2 = \frac{d}{n}$, which has been shown in (28.17).
- For $\rho = \|\cdot\|_\infty$, $\mathbb{E}\|Z\|_\infty \asymp \sqrt{\log d}$ (Lemma 27.12) and $R^* \asymp \sqrt{\frac{\log d}{n}}$.
- For a matrix $\theta \in \mathbb{R}^{d \times d}$, let $\rho(\theta) = \|\theta\|_{\text{op}}$ denote the operator norm (maximum singular value). It has been shown in Exercise V.13 that $\mathbb{E}\|Z\|_{\text{op}} \asymp \sqrt{d}$ and so $R^* \asymp \sqrt{\frac{d}{n}}$; for $\rho(\cdot) = \|\cdot\|_{\text{F}}$, $R^* \asymp \frac{d}{\sqrt{n}}$.

We can also phrase the result of Corollary 28.1 in terms of the sample complexity $n^*(\epsilon)$ as defined in (28.16). For example, for $q = 2$ we have $n^*(\epsilon) = \lceil \mathbb{E}[\|Z\|^2]/\epsilon \rceil$. The above examples show that the scaling of $n^*(\epsilon)$ with dimension depends on the loss function and the “rule of thumb” that the sampling complexity is proportional to the number of parameters need not always hold.

Finally, for the sake of high-probability (as opposed to average) risk bound, consider $\rho(\theta - \hat{\theta}) = 1\{\|\theta - \hat{\theta}\| > \epsilon\}$, which is lower semicontinuous and bowl-shaped. Then the exact expression $R^* = \mathbb{P}[\|Z\| \geq \epsilon\sqrt{n}]$. This result is stronger since the sample mean is optimal simultaneously for all ϵ , so that integrating over ϵ recovers (28.20).

Proof of Theorem 28.13. We only prove the lower bound. We bound the minimax risk R^* from below by the Bayes risk R_π^* with the prior $\pi = N(0, sI_d)$:

$$\begin{aligned} R^* &\geq R_\pi^* = \inf_{\hat{\theta}} \mathbb{E}_\pi[\rho(\theta - \hat{\theta})] \\ &= \mathbb{E} \left[\inf_{\hat{\theta}} \mathbb{E}[\rho(\theta - \hat{\theta})|X] \right] \\ &\stackrel{(a)}{=} \mathbb{E}[\mathbb{E}[\rho(\theta - \mathbb{E}[\theta|X])|X]] \\ &\stackrel{(b)}{=} \mathbb{E} \left[\rho \left(\sqrt{\frac{s}{1+sn}} Z \right) \right]. \end{aligned}$$

where (a) follows from the crucial Lemma 28.14 below; (b) uses the fact that $\theta - \mathbb{E}[\theta|X] \sim N(0, \frac{s}{1+sn} I_d)$ under the Gaussian prior. Since $\rho(\cdot)$ is Lower semicontinuous, sending $s \rightarrow \infty$ and

applying Fatou's lemma, we obtain the matching lower bound:

$$R^* \geq \lim_{s \rightarrow \infty} \mathbb{E} \left[\rho \left(\sqrt{\frac{s}{1+sn}} Z \right) \right] \geq \mathbb{E} \left[\rho \left(\frac{Z}{\sqrt{n}} \right) \right]. \quad \square$$

The following lemma establishes the conditional mean as the Bayes estimator under the Gaussian prior for all bowl-shaped losses, extending the previous Lemma 28.11 in one dimension:

Lemma 28.9 (Anderson [13]). *Let $X \sim N(0, \Sigma)$ for some $\Sigma \succ 0$ and $\rho : \mathbb{R}^d \rightarrow \mathbb{R}_+$ be a bowl-shaped loss function. Then*

$$\min_{y \in \mathbb{R}^d} \mathbb{E}[\rho(y + X)] = \mathbb{E}[\rho(X)].$$

In order to prove Lemma 28.14, it suffices to consider ρ being indicator functions. This is done in the next lemma, which we prove later.

Lemma 28.10. *Let $K \in \mathbb{R}^d$ be a symmetric convex set and $X \sim N(0, \Sigma)$. Then $\max_{y \in \mathbb{R}^d} \mathbb{P}(X + y \in K) = \mathbb{P}(X \in K)$.*

Proof of Lemma 28.14. Denote the sublevel set $K_c = \{x \in \mathbb{R}^d : \rho(x) \leq c\}$. Since ρ is bowl-shaped, K_c is convex and symmetric, which satisfies the conditions of Lemma 28.15. So,

$$\begin{aligned} \mathbb{E}[\rho(y + x)] &= \int_0^\infty \mathbb{P}(\rho(y + x) > c) dc, \\ &= \int_0^\infty (1 - \mathbb{P}(y + x \in K_c)) dc, \\ &\geq \int_0^\infty (1 - \mathbb{P}(x \in K_c)) dc, \\ &= \int_0^\infty \mathbb{P}(\rho(x) \geq c) dc, \\ &= \mathbb{E}[\rho(x)]. \end{aligned}$$

Hence, $\min_{y \in \mathbb{R}^d} \mathbb{E}[\rho(y + x)] = \mathbb{E}[\rho(x)]$. \square

Before going into the proof of Lemma 28.15, we need the following definition.

Definition 28.11. A measure μ on \mathbb{R}^d is said to be *log-concave* if

$$\mu(\lambda A + (1 - \lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}$$

for all measurable $A, B \subset \mathbb{R}^d$ and any $\lambda \in [0, 1]$.

The following result, due to Prékopa [246], characterizes the log-concavity of measures in terms of that of its density function; see also [257] (or [132, Theorem 4.2]) for a proof.

Theorem 28.12. *Suppose that μ has a density f with respect to the Lebesgue measure on \mathbb{R}^d . Then μ is log-concave if and only iff f is log-concave.*

28.6 Log-concavity, Anderson's lemma and exact minimax risk in GLM 495

Example 28.5. Examples of log-concave measures:

- Lebesgue measure: Let $\mu = \text{vol}$ be the Lebesgue measure on \mathbb{R}^d , which satisfies Theorem 28.17 ($f \equiv 1$). Then

$$\text{vol}(\lambda A + (1 - \lambda)B) \geq \text{vol}(A)^\lambda \text{vol}(B)^{1-\lambda}, \quad (28.21)$$

which implies³ the Brunn-Minkowski inequality:

$$\text{vol}(A + B)^{\frac{1}{d}} \geq \text{vol}(A)^{\frac{1}{d}} + \text{vol}(B)^{\frac{1}{d}}. \quad (28.22)$$

- Gaussian distribution: Let $\mu = N(0, \Sigma)$, with a log-concave density f since $\log f(x) = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log \det(\Sigma) - \frac{1}{2} x^\top \Sigma^{-1} x$ is concave.

Proof of Lemma 28.15. By Theorem 28.17, the distribution of X is log-concave. Then

$$\begin{aligned} \mathbb{P}[X \in K] &\stackrel{(a)}{=} \mathbb{P}\left[X \in \frac{1}{2}(K+y) + \frac{1}{2}(K-y)\right] \\ &\stackrel{(b)}{\geq} \sqrt{\mathbb{P}[X \in K-y]\mathbb{P}[X \in K+y]} \\ &\stackrel{(c)}{=} \mathbb{P}[X+y \in K], \end{aligned}$$

where (a) follows from $\frac{1}{2}(K+y) + \frac{1}{2}(K-y) = \frac{1}{2}K + \frac{1}{2}K = K$ since K is convex; (b) follows from the definition of log-concavity in Definition 28.16 with $\lambda = \frac{1}{2}$, $A = K-y = \{x-y : x \in K\}$ and $B = K+y$; (c) follows from $\mathbb{P}[X \in K+y] = \mathbb{P}[X \in -K-y] = \mathbb{P}[X+y \in K]$ since X has a symmetric distribution and K is symmetric ($K = -K$). \square

³ Applying (28.21) to $A' = \text{vol}(A)^{-1/d}A$, $B' = \text{vol}(B)^{-1/d}B$ (both of which have unit volume), and $\lambda = \text{vol}(A)^{1/d}/(\text{vol}(A)^{1/d} + \text{vol}(B)^{1/d})$ yields (28.22).

29

Classical large-sample asymptotics

In this chapter we give an overview of the classical large-sample theory in the setting of iid observations in Section 28.4 focusing again on the minimax risk (28.15). These results pertain to smooth parametric models in fixed dimensions, with the sole asymptotics being the sample size going to infinity. The main result is that, under suitable conditions, the minimax squared error of estimating θ based on $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$ satisfies

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\|\hat{\theta} - \theta\|_2^2] = \frac{1 + o(1)}{n} \sup_{\theta \in \Theta} \text{Tr} J_F^{-1}(\theta). \quad (29.1)$$

where $J_F(\theta)$ is the Fisher information matrix introduced in (2.31) in Chapter 2. This is asymptotic characterization of the minimax risk with sharp constant. In later chapters, we will proceed to high dimensions where such precise results are difficult and rare.

Throughout this chapter, we focus on the quadratic risk and assume that Θ is an open set of the Euclidean space \mathbb{R}^d .

29.1 Statistical lower bound from data processing

In this section we derive several statistical lower bounds from data processing argument. Specifically, we will take a comparison-of-experiment approach by comparing the actual model with a perturbed model. The performance of a given estimator can be then related to the f -divergence via the data processing inequality and the variational representation (Chapter 7).

We start by discussing the Hammersley-Chapman-Robbins lower bound which implies the well-known Cramér-Rao lower bound. Because these results are restricted to unbiased estimators, we will also discuss their Bayesian version; in particular, the Bayesian Cramér-Rao lower bound is responsible for proving the lower bound in (29.1). We focus on explaining how these results can be anticipated from information-theoretic reasoning and postpone the exact statement and assumption of the Bayesian Cramér-Rao bound to Section 29.2.

29.1.1 Hammersley-Chapman-Robbins (HCR) lower bound

The following result due to [149, 60] is a direct consequence of the variational representation of χ^2 -divergence in Section 7.13, which relates it to the mean and variance of test functions.

29.1 Statistical lower bound from data processing 497

Theorem 29.1 (HCR lower bound). *The quadratic loss of any estimator $\hat{\theta}$ at $\theta \in \Theta \subset \mathbb{R}^d$ satisfies*

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \text{Var}_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'} \| P_\theta)}. \quad (29.2)$$

Proof. Let $\hat{\theta}$ be a (possibly randomized) estimator based on X . Fix $\theta' \neq \theta \in \Theta$. Denote by P and Q the probability distribution when the true parameter is θ or θ' , respectively. That is, $P_X = P_\theta$ and $Q_X = P_{\theta'}$. Then

$$\chi^2(P_X \| Q_X) \geq \chi^2(P_{\hat{\theta}} \| Q_{\hat{\theta}}) \geq \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\text{Var}_\theta(\hat{\theta})} \quad (29.3)$$

where the first inequality applies the data processing inequality (Theorem 7.7) and the second inequality the variational representation (7.85) of χ^2 -divergence. \square

Next we apply Theorem 29.1 to unbiased estimators $\hat{\theta}$ that satisfies $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. Then

$$\text{Var}_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

Lower bounding the supremum by the limit of $\theta' \rightarrow \theta$ and recall the asymptotic expansion of χ^2 -divergence from Theorem 7.31, we get, under the regularity conditions in Theorem 7.31, the celebrated Cramér-Rao (CR) lower bound [77, 250]:

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{J_F(\theta)}. \quad (29.4)$$

A few more remarks are as follows:

- Note that the HCR lower bound Theorem 29.1 is based on the χ^2 -divergence. For a version based on Hellinger distance which also implies the CR lower bound, see Exercise VI.5.
- Both the HCR and the CR lower bounds extend to the multivariate case as follows. Let $\hat{\theta}$ be an unbiased estimator of $\theta \in \Theta \subset \mathbb{R}^d$. Assume that its covariance matrix $\text{Cov}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top]$ is positive definite. Fix $a \in \mathbb{R}^d$. Applying Theorem 29.1 to $\langle a, \hat{\theta} \rangle$, we get

$$\chi^2(P_\theta \| P_{\theta'}) \geq \frac{\langle a, \theta - \theta' \rangle^2}{a^\top \text{Cov}_\theta(\hat{\theta}) a}.$$

Optimizing over a yields¹

$$\chi^2(P_\theta \| P_{\theta'}) \geq (\theta - \theta')^\top \text{Cov}_\theta(\hat{\theta})^{-1} (\theta - \theta').$$

Sending $\theta' \rightarrow \theta$ and applying the asymptotic expansion $\chi^2(P_\theta \| P_{\theta'}) = (\theta - \theta')^\top J_F(\theta)(\theta - \theta')(1 + o(1))$ (see Remark 7.32), we get the multivariate version of CR lower bound:

$$\text{Cov}_\theta(\hat{\theta}) \succeq J_F^{-1}(\theta). \quad (29.5)$$

¹ For $\Sigma \succ 0$, $\sup_{x \neq 0} \frac{\langle x, y \rangle^2}{x^\top \Sigma x} = y^\top \Sigma^{-1} y$, attained at $x = \Sigma^{-1}y$.

- For a sample of n iid observations, by the additivity property (2.35), the Fisher information matrix is equal to $nJ_F(\theta)$. Taking the trace on both sides, we conclude the squared error of any unbiased estimators satisfies

$$\mathbb{E}_\theta[\|\hat{\theta} - \theta\|_2^2] \geq \frac{1}{n} \text{Tr}(J_F^{-1}(\theta)).$$

This is already very close to (29.1), except for the fundamental restriction that of unbiased estimators.

29.1.2 Bayesian Cramér-Rao lower bound

The drawback of the HCR and CR lower bounds is that they are confined to unbiased estimators. For the minimax settings in (29.1), there is no sound reason to restrict to unbiased estimators; in fact, it is often wise to trade bias with variance in order to achieve a smaller overall risk.

Next we discuss a lower bound, known as the Bayesian Cramér-Rao (BCR) lower bound [136] or the van Trees inequality [311], for a Bayesian setting that applies to *all* estimators; to apply to the minimax setting, in view of Theorem 28.4, one just needs to choose an appropriate prior. The exact statement and the application to minimax risk are postponed till the next section. Here we continue the previous line of thinking and derive it from the data processing argument.

Fix a prior π on Θ and a (possibly randomized) estimator $\hat{\theta}$. Then we have the Markov chain $\theta \rightarrow X \rightarrow \hat{\theta}$. Consider two joint distributions for (θ, X) :

- Under Q , θ is drawn from π and $X \sim P_\theta$ conditioned on θ ;
- Under P , θ is drawn from $T_\delta\pi$, where T_δ denote the pushforward of shifting by δ , i.e., $T_\delta\pi(A) = \pi(A - \delta)$, and $X \sim P_{\theta-\delta}$ conditioned on θ .

Similar to (29.3), applying data processing and variational representation of χ^2 -divergence yields

$$\chi^2(P_{\theta X} \| Q_{\theta X}) \geq \chi^2(P_{\theta\hat{\theta}} \| Q_{\theta\hat{\theta}}) \geq \chi^2(P_{\theta-\hat{\theta}} \| Q_{\theta-\hat{\theta}}) \geq \frac{(\mathbb{E}_P[\theta - \hat{\theta}] - \mathbb{E}_Q[\theta - \hat{\theta}])^2}{\text{Var}_Q(\hat{\theta} - \theta)}.$$

Note that by design, $P_X = Q_X$ and thus $\mathbb{E}_P[\hat{\theta}] = \mathbb{E}_Q[\hat{\theta}]$; on the other hand, $\mathbb{E}_P[\theta] = \mathbb{E}_Q[\theta] + \delta$. Furthermore, $\mathbb{E}_\pi[(\hat{\theta} - \theta)^2] \geq \text{Var}_Q(\hat{\theta} - \theta)$. Since this applies to any estimators, we conclude that the Bayes risk R_π^* (and hence the minimax risk) satisfies

$$R_\pi^* \triangleq \inf_{\hat{\theta}} \mathbb{E}_\pi[(\hat{\theta} - \theta)^2] \geq \sup_{\delta \neq 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})}, \quad (29.6)$$

which is referred to as the *Bayesian HCR* lower bound in comparison with (29.2).

Similar to the deduction of CR lower bound from the HCR, we can further lower bound this supremum by evaluating the small- δ limit. First note the following chain rule for the χ^2 -divergence:

$$\chi^2(P_{X\theta} \| Q_{X\theta}) = \chi^2(P_\theta \| Q_\theta) + \mathbb{E}_Q \left[\chi^2(P_{X|\theta} \| Q_{X|\theta}) \cdot \left(\frac{dP_\theta}{dQ_\theta} \right)^2 \right].$$

29.2 Bayesian Cramér-Rao lower bounds 499

Under suitable regularity conditions in Theorem 7.31, again applying the local expansion of χ^2 -divergence yields

- $\chi^2(P_\theta || Q_\theta) = \chi^2(T_\delta \pi || \pi) = (J(\pi) + o(1))\delta^2$, where $J(\pi) \triangleq \int \frac{\pi'^2}{\pi}$ is the Fisher information of the prior;
- $\chi^2(P_{X|\theta} || Q_{X|\theta}) = [J_F(\theta) + o(1)]\delta^2$.

Thus from (29.6) we get

$$R_\pi^* \geq \frac{1}{J(\pi) + \mathbb{E}_{\theta \sim \pi}[J_F(\theta)]}. \quad (29.7)$$

We conclude this section by revisiting the Gaussian Location Model (GLM) in Example 28.1.

Example 29.1. Let $X^n = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ and consider the prior $\theta \sim \pi = \mathcal{N}(0, s)$. To apply the Bayesian HCR bound (29.6), note that

$$\begin{aligned} \chi^2(P_{\theta X^n} || Q_{\theta X^n}) &\stackrel{(a)}{=} \chi^2(P_{\theta \bar{X}} || Q_{\theta \bar{X}}) \\ &= \chi^2(P_\theta || Q_\theta) + \mathbb{E}_Q \left[\left(\frac{dP_\theta}{dQ_\theta} \right)^2 \chi^2(P_{\bar{X}|\theta} || Q_{\bar{X}|\theta}) \right] \\ &\stackrel{(b)}{=} e^{\delta^2/s} - 1 + e^{\delta^2/s} (e^{n\delta^2} - 1) \\ &= e^{\delta^2(n+\frac{1}{s})} - 1. \end{aligned}$$

where (a) follows from the sufficiency of $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$; (b) is by $Q_\theta = \mathcal{N}(0, s)$, $Q_{\bar{X}|\theta} = \mathcal{N}(\theta, \frac{1}{n})$, $P_\theta = \mathcal{N}(\delta, s)$, $P_{\bar{X}|\theta} = \mathcal{N}(\theta - \delta, \frac{1}{n})$, and the fact (7.40) for Gaussians. Therefore,

$$R_\pi^* \geq \sup_{\delta \neq 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \lim_{\delta \rightarrow 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \frac{s}{sn+1}.$$

In view of the Bayes risk found in Example 28.1, we see that in this case the Bayesian HCR and Bayesian Cramér-Rao lower bounds are *exact*.

29.2 Bayesian Cramér-Rao lower bounds

In this section we give the rigorous statement of the Bayesian Cramér-Rao lower bound and discuss its extensions and consequences. For the proof, we take a more direct approach as opposed to the data-processing argument in Section 29.1 based on asymptotic expansion of the χ^2 -divergence.

Theorem 29.2 (BCR lower bound). *Let π be a differentiable prior density on the interval $[\theta_0, \theta_1]$ such that $\pi(\theta_0) = \pi(\theta_1) = 0$ and*

$$J(\pi) \triangleq \int_{\theta_0}^{\theta_1} \frac{\pi'(\theta)^2}{\pi(\theta)} d\theta < \infty. \quad (29.8)$$

Let $P_\theta(dx) = p_\theta(x)\mu(dx)$, where the density $p_\theta(x)$ is differentiable in θ for μ -almost every x . Assume that for π -almost every θ ,

$$\int \mu(dx) \partial_\theta p_\theta(x) = 0. \quad (29.9)$$

Then the Bayes quadratic risk $R_\pi^* \triangleq \inf_{\hat{\theta}} \mathbb{E}[(\theta - \hat{\theta})^2]$ satisfies

$$R_\pi^* \geq \frac{1}{\mathbb{E}_{\theta \sim \pi}[J_F(\theta)] + J(\pi)}. \quad (29.10)$$

Proof. In view of Remark 28.3, it loses no generality to assume that the estimator $\hat{\theta} = \hat{\theta}(X)$ is deterministic. For each x , integration by parts yields

$$\int_{\theta_0}^{\theta_1} d\theta (\hat{\theta}(x) - \theta) \partial_\theta (p_\theta(x) \pi(\theta)) = \int_{\theta_0}^{\theta_1} p_\theta(x) \pi(\theta) d\theta.$$

Integrating both sides over $\mu(dx)$ yields

$$\mathbb{E}[(\hat{\theta} - \theta)V(\theta, X)] = 1.$$

where $V(\theta, x) \triangleq \partial_\theta(\log(p_\theta(x)\pi(\theta))) = \partial_\theta \log p_\theta(x) + \partial_\theta \log \pi(\theta)$ and the expectation is over the joint distribution of (θ, X) . Applying Cauchy-Schwarz, we have $\mathbb{E}[(\hat{\theta} - \theta)^2]\mathbb{E}[V(\theta, X)^2] \geq 1$. The proof is completed by noting that $\mathbb{E}[V(\theta, X)^2] = \mathbb{E}[(\partial_\theta \log p_\theta(X))^2] + \mathbb{E}[(\partial_\theta \log \pi(\theta))^2] = \mathbb{E}[J_F(\theta)] + J(\pi)$, thanks to the assumption (29.9). \square

The multivariate version of Theorem 29.2 is the following.

Theorem 29.3 (Multivariate BCR). *Consider a product prior density $\pi(\theta) = \prod_{i=1}^d \pi_i(\theta_i)$ over the box $\prod_{i=1}^d [\theta_{0,i}, \theta_{1,i}]$, where each π_i is differentiable on $[\theta_{0,i}, \theta_{1,i}]$ and vanishes on the boundary. Assume that for π -almost every θ ,*

$$\int \mu(dx) \nabla_\theta p_\theta(x) = 0. \quad (29.11)$$

Then

$$R_\pi^* \triangleq \inf_{\hat{\theta}} \mathbb{E}_\pi[\|\hat{\theta} - \theta\|_2^2] \geq \text{Tr}((\mathbb{E}_{\theta \sim \pi}[J_F(\theta)] + J(\pi))^{-1}), \quad (29.12)$$

where the Fisher information matrices are given by $J_F(\theta) = \mathbb{E}_\theta[\nabla_\theta \log p_\theta(X) \nabla_\theta \log p_\theta(X)^\top]$ and $J(\pi) = \text{diag}(J(\pi_1), \dots, J(\pi_d))$.

Proof. Fix an estimator $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d)$ and a non-zero $u \in \mathbb{R}^d$. For each $i, k = 1, \dots, d$, integration by parts yields

$$\int_{\theta_{0,i}}^{\theta_{1,i}} (\hat{\theta}_k(x) - \theta_k) \partial_{\theta_i} (p_\theta(x) \pi(\theta)) d\theta_i = 1_{\{k=i\}} \int_{\theta_{0,i}}^{\theta_{1,i}} p_\theta(x) \pi(\theta) d\theta_i.$$

Integrating both sides over $\prod_{j \neq i} d\theta_j$ and $\mu(dx)$, multiplying by u_i , and summing over i , we obtain

$$\mathbb{E}[(\hat{\theta}_k(X) - \theta_k) \langle u, \nabla \log(p_\theta(X)\pi(\theta)) \rangle] = \langle u, e_k \rangle$$

29.2 Bayesian Cramér-Rao lower bounds 501

where e_k denotes the k th standard basis. Applying Cauchy-Schwarz and optimizing over u yield

$$\mathbb{E}[(\hat{\theta}_k(X) - \theta_k)^2] \geq \sup_{u \neq 0} \frac{\langle u, e_k \rangle^2}{u^\top \Sigma u} = \Sigma_{kk}^{-1},$$

where $\Sigma \equiv \mathbb{E}[\nabla \log(p_\theta(X)\pi(\theta)) \nabla \log(p_\theta(X)\pi(\theta))^\top] = \mathbb{E}_{\theta \sim \pi}[J_F(\theta)] + J(\pi)$, thanks to (29.11). Summing over k completes the proof of (29.12). \square

Several remarks are in order:

- The above versions of the BCR bound assume a prior density that vanishes at the boundary. If we choose a uniform prior, the same derivation leads to a similar lower bound known as the Chernoff-Rubin-Stein inequality (see Ex. VI.4), which also suffices for proving the optimal minimax lower bound in (29.1).
- For the purpose of the lower bound, it is advantageous to choose a prior density with the minimum Fisher information. The optimal density with a compact support is known to be a squared cosine density [156, 306]:

$$\min_{g \text{ on } [-1,1]} J(g) = \pi^2,$$

attained by

$$g(u) = \cos^2 \frac{\pi u}{2}. \quad (29.13)$$

- Suppose the goal is to estimate a smooth *functional* $T(\theta)$ of the unknown parameter θ , where $T : \mathbb{R}^d \rightarrow \mathbb{R}^s$ is differentiable with $\nabla T(\theta) = (\frac{\partial T_i(\theta)}{\partial \theta_j})$ its $s \times d$ Jacobian matrix. Then under the same condition of Theorem 29.3, we have the following Bayesian Cramér-Rao lower bound for functional estimation:

$$\inf_{\hat{T}} \mathbb{E}_\pi [\|\hat{T}(X) - T(\theta)\|_2^2] \geq \text{Tr}(\mathbb{E}[\nabla T(\theta)](\mathbb{E}[J_F(\theta)] + J(\pi))^{-1} \mathbb{E}[\nabla T(\theta)]^\top), \quad (29.14)$$

where the expectation on the right-hand side is over $\theta \sim \pi$.

As a consequence of the BCR bound, we prove the lower bound part for the asymptotic minimax risk in (29.1).

Theorem 29.4. Assume that $\theta \mapsto J_F(\theta)$ is continuous. Denote the minimax squared error $R_n^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta [\|\hat{\theta} - \theta\|_2^2]$, where \mathbb{E}_θ is taken over $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_\theta$. Then as $n \rightarrow \infty$,

$$R_n^* \geq \frac{1 + o(1)}{n} \sup_{\theta \in \Theta} \text{Tr} J_F^{-1}(\theta). \quad (29.15)$$

Proof. Fix $\theta \in \Theta$. Then for all sufficiently small δ , $B_\infty(\theta, \delta) = \theta + [-\delta, \delta]^d \subset \Theta$. Let $\pi_i(\theta_i) = \frac{1}{\delta} g(\frac{\theta_i - \theta_i}{\delta})$, where g is the prior density in (29.13). Then the product distribution $\pi = \prod_{i=1}^d \pi_i$ satisfies the assumption of Theorem 29.3. By the scaling rule of Fisher information (see (2.34)), $J(\pi_i) = \frac{1}{\delta^2} J(g) = \frac{\pi^2}{\delta^2}$. Thus $J(\pi) = \frac{\pi^2}{\delta^2} I_d$.

It is known that (see [43, Theorem 2, Appendix V]) the continuity of $\theta \mapsto J_F(\theta)$ implies (29.11). So we are ready to apply the BCR bound in Theorem 29.3. Lower bounding the minimax by the Bayes risk and also applying the additivity property (2.35) of Fisher information, we obtain

$$R_n^* \geq \frac{1}{n} \cdot \text{Tr} \left(\left(\mathbb{E}_{\theta \sim \pi}[J_F(\theta)] + \frac{\pi^2}{n\delta^2} I_d \right)^{-1} \right).$$

Finally, choosing $\delta = n^{-1/4}$ and applying the continuity of $J_F(\theta)$ in θ , the desired (29.15) follows. \square

Similarly, for estimating a smooth functional $T(\theta)$, applying (29.14) with the same argument yields

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta}[\|\hat{T} - T(\theta)\|_2^2] \geq \frac{1 + o(1)}{n} \sup_{\theta \in \Theta} \text{Tr}(\nabla T(\theta) J_F^{-1}(\theta) \nabla T(\theta)^\top). \quad (29.16)$$

29.3 Maximum Likelihood Estimator and asymptotic efficiency

Theorem 29.4 shows that in a small neighborhood of each parameter θ , the best estimation error is at best $\frac{1}{n}(\text{Tr}J_F^{-1}(\theta) + o(1))$ when the sample size n grows; this is known as the *information bound* as determined by the Fisher information matrix. Estimators achieving this bound are called *asymptotic efficient*. A cornerstone of the classical large-sample theory is the asymptotic efficiency of the *maximum likelihood estimator* (MLE). Rigorously stating this result requires a lengthy list of technical conditions, and an even lengthier one is needed to make the error uniform so as to attain the minimax lower bound in Theorem 29.4. In this section we give a sketch of the asymptotic analysis of MLE, focusing on the main ideas and how Fisher information emerges from the likelihood optimization.

Suppose we observe a sample $X^n = (X_1, X_2, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} P_{\theta_0}$, where θ_0 stands for the true parameter. The MLE is defined as:

$$\hat{\theta}_{\text{MLE}} \in \arg \max_{\theta \in \Theta} L_\theta(X^n), \quad (29.17)$$

where

$$L_\theta(X^n) = \sum_{i=1}^n \log p_\theta(X_i)$$

is the total log-likelihood and $p_\theta(x) = \frac{dP_\theta}{d\mu}(x)$ is the density of P_θ with respect to some common dominating measure μ . For discrete distribution P_θ , the MLE can also be written as the KL projection² of the empirical distribution \hat{P}_n to the model class: $\hat{\theta}_{\text{MLE}} \in \arg \min_{\theta \in \Theta} D(\hat{P}_n \| P_\theta)$.

² Note that this is the reverse of the information projection studied in Section 15.3.

29.3 Maximum Likelihood Estimator and asymptotic efficiency 503

The main intuition why MLE works is as follows. Assume that the model is identifiable, namely, $\theta \mapsto P_\theta$ is injective. Then for any $\theta \neq \theta_0$, we have by positivity of the KL divergence (Theorem 2.3)

$$\mathbb{E}_{\theta_0} [L_\theta - L_{\theta_0}] = \mathbb{E}_{\theta_0} \left[\sum_{i=1}^n \log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right] = -nD(P_{\theta_0} || P_\theta) < 0.$$

In other words, $L_\theta - L_{\theta_0}$ is an iid sum with a negative mean and thus negative with high probability for large n . From here the consistency of MLE follows upon assuming appropriate regularity conditions, among which is Wald's integrability condition $\mathbb{E}_{\theta_0} [\sup_{\|\theta-\theta_0\| \leq \epsilon} \log \frac{p_\theta}{p_{\theta_0}}(X)] < \infty$ [319, 321].

Assuming more conditions one can obtain the asymptotic normality and efficiency of the MLE. This follows from the local quadratic approximation of the log-likelihood function. Define $V(\theta, x) \triangleq \nabla_\theta p_\theta(x)$ (score) and $H(\theta, x) \triangleq \nabla_\theta^2 p_\theta(x)$. By Taylor expansion,

$$\begin{aligned} L_\theta = & L_{\theta_0} + (\theta - \theta_0)^\top \left(\sum_{i=1}^n V(\theta_0, X_i) \right) + \frac{1}{2} (\theta - \theta_0)^\top \left(\sum_{i=1}^n H(\theta_0, X_i) \right) (\theta - \theta_0) \\ & + o(n(\theta - \theta_0)^2). \end{aligned} \quad (29.18)$$

Recall from Section 2.6.2* that, under suitable regularity conditions, we have

$$\mathbb{E}_{\theta_0}[V(\theta_0, X)] = 0, \quad \mathbb{E}_{\theta_0}[V(\theta_0, X)V(\theta_0, X)^\top] = -\mathbb{E}_{\theta_0}[H(\theta_0, X)] = J_F(\theta_0).$$

Thus, by the Central Limit Theorem and the Weak Law of Large Numbers, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n V(\theta_0, X_i) \xrightarrow{d} \mathcal{N}(0, J_F(\theta_0)), \quad \frac{1}{n} \sum_{i=1}^n H(\theta_0, X_i) \xrightarrow{\mathbb{P}} -J_F(\theta_0).$$

Substituting these quantities into (29.18), we obtain the following stochastic approximation of the log-likelihood:

$$L_\theta \approx L_{\theta_0} + \langle \sqrt{n} J_F(\theta_0)^{-1/2} Z, \theta - \theta_0 \rangle - \frac{n}{2} (\theta - \theta_0)^\top J_F(\theta_0) (\theta - \theta_0),$$

where $Z \sim \mathcal{N}(0, I_d)$. Maximizing the right-hand side yields:

$$\hat{\theta}_{\text{MLE}} \approx \theta_0 + \frac{1}{\sqrt{n}} J_F(\theta_0)^{-1/2} Z.$$

From this asymptotic normality, we can obtain $\mathbb{E}_{\theta_0} [\|\hat{\theta}_{\text{MLE}} - \theta_0\|_2^2] \leq \frac{1}{n} (\text{Tr} J_F(\theta_0)^{-1} + o(1))$, and for smooth functionals by Taylor expanding T at θ_0 (delta method), $\mathbb{E}_{\theta_0} [\|T(\hat{\theta}_{\text{MLE}}) - T(\theta_0)\|_2^2] \leq \frac{1}{n} (\text{Tr} (\nabla T(\theta_0) J_F(\theta_0)^{-1} \nabla T(\theta_0)^\top) + o(1))$, matching the information bounds (29.15) and (29.16).

Of course, the above heuristic derivation requires additional assumptions to justify (for example, Cramér's condition, cf. [124, Theorem 18] and [265, Theorem 7.63]). Even stronger assumptions are needed to ensure the error is uniform in θ in order to achieve the minimax lower bound in Theorem 29.4; see, e.g., Theorem 34.4 (and also Chapters 36-37) of [43] for the exact conditions and statements. A more general and abstract theory of MLE and the attainment of information bound were developed by Hájek and Le Cam; see [148, 189].

Despite its wide applicability and strong optimality properties, the methodology of MLE is not without limitations. We conclude this section with some remarks along this line.

- MLE may not exist even for simple parametric models. For example, consider X_1, \dots, X_n drawn iid from the location-scale mixture of two Gaussians $\frac{1}{2}\mathcal{N}(\mu_1, \sigma_1^2) + \frac{1}{2}\mathcal{N}(\mu_2, \sigma_2^2)$, where $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ are unknown parameters. Then the likelihood can be made arbitrarily large by setting for example $\mu_1 = X_1$ and $\sigma_1 \rightarrow 0$.
- MLE may be inconsistent; see [265, Example 7.61] and [123] for examples, both in one-dimensional parametric family.
- In high dimensions, it is possible that MLE fails to achieve the minimax rate (Exercise VI.14).

29.4 Application: Estimating discrete distributions and entropy

As an application in this section we consider the concrete problems of estimating a discrete distribution or its property (such as Shannon entropy) based on iid observations. Of course, the asymptotic theory developed in this chapter applies only to the classical setting of fixed alphabet and large sample size. Along the way, we will also discuss extensions to large alphabet and what may go wrong with the classical theory.

Throughout this section, let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P \in \mathcal{P}_k$, where $\mathcal{P}_k \equiv \mathcal{P}([k])$ denotes the collection of probability distributions over $[k] = \{1, \dots, k\}$. We first consider the estimation of P under the squared loss.

Theorem 29.5. *For fixed k , the minimax squared error of estimating P satisfies*

$$R_{\text{sq}}^*(k, n) \triangleq \inf_{\widehat{P}} \sup_{P \in \mathcal{P}_k} \mathbb{E}[\|\widehat{P} - P\|_2^2] = \frac{1}{n} \left(\frac{k-1}{k} + o(1) \right), \quad n \rightarrow \infty. \quad (29.19)$$

Proof. Let $P = (P_1, \dots, P_k)$ be parametrized, as in Example 2.6, by $\theta = (P_1, P_2, \dots, P_{k-1})$ and $P_k = 1 - P_1 - \dots - P_{k-1}$. Then $P = T(\theta)$, where $T : \mathbb{R}^{k-1} \rightarrow \mathbb{R}^k$ is an affine functional so that $\nabla T(\theta) = \begin{bmatrix} I_{k-1} \\ -1^\top \end{bmatrix}$, with 1 being the all-ones (column) vector.

The Fisher information matrix and its inverse have been calculated in (2.36) and (2.37): We have $J_F^{-1}(\theta) = \text{diag}(\theta) - \theta\theta^\top$ and

$$\nabla T(\theta) J_F^{-1}(\theta) \nabla T(\theta)^\top = \begin{bmatrix} \text{diag}(\theta) - \theta\theta^\top & -P_k\theta \\ -P_k\theta^\top & P_k(1 - P_k) \end{bmatrix}.$$

So $\text{Tr}(\nabla T(\theta) J_F^{-1}(\theta) \nabla T(\theta)^\top) = \sum_{i=1}^k P_i(1 - P_i) = 1 - \sum_{i=1}^k P_i^2$, which achieves its maximum $1 - \frac{1}{k}$ at the uniform distribution. Applying the functional form of the BCR bound in (29.16), we conclude $R_{\text{sq}}^*(k, n) \geq \frac{1}{n}(1 - \frac{1}{k} + o(1))$.

For the upper bound, consider the MLE, which in this case coincides with the empirical distribution $\hat{P} = (\hat{P}_i)$ (Exercise VI.8). Note that $n\hat{P}_i = \sum_{j=1}^n 1_{\{X_j=i\}} \sim \text{Bin}(n, P_i)$. Then for any P , $\mathbb{E}[\|\hat{P} - P\|_2^2] = \frac{1}{n} \sum_{i=1}^k P_i(1 - P_i) \leq \frac{1}{n}(1 - \frac{1}{k})$. \square

Some remarks on Theorem 29.5 are in order:

29.4 Application: Estimating discrete distributions and entropy 505

- In fact, for any k, n , we have the precise result: $R_{\text{sq}}^*(k, n) = \frac{1-1/k}{(1+\sqrt{n})^2}$ – see Ex. VI.7h. This can be shown by considering a Dirichlet prior (13.15) and applying the corresponding Bayes estimator, which is an additively-smoothed empirical distribution (Section 13.5).
- Note that $R_{\text{sq}}^*(k, n)$ does not grow with the alphabet size k ; this is because squared loss is too weak for estimating probability vectors. More meaningful loss functions include the f -divergences in Chapter 7, such as the total variation, KL divergence, χ^2 -divergence. These minimax rates are worked out in Exercise VI.8 and Exercise VI.9, for both small and large alphabets, and they indeed depend on the alphabet size k . For example, the minimax KL risk satisfies $\Theta(\frac{k}{n})$ for $k \leq n$ and grows as $\Theta(\log \frac{k}{n})$ for $k \gg n$. This agrees with the rule of thumb that consistent estimation requires the sample size to scale faster than the dimension.

As a final application, let us consider the classical problem of *entropy estimation* in information theory and statistics [214, 97, 152], where the goal is to estimate the Shannon entropy, a non-linear functional of P . The following result follows from the functional BCR lower bound (29.16) and analyzing the MLE (in this case the empirical entropy) [25].

Theorem 29.6. *For fixed k , the minimax quadratic risk of entropy estimation satisfies*

$$R_{\text{ent}}^*(k, n) \triangleq \inf_{\hat{H}} \sup_{P \in \mathcal{P}_k} \mathbb{E}[(\hat{H}(X_1, \dots, X_n) - H(P))^2] = \frac{1}{n} \left(\max_{P \in \mathcal{P}_k} V(P) + o(1) \right), \quad n \rightarrow \infty$$

where $H(P) = \sum_{i=1}^k P_i \log \frac{1}{P_i} = \mathbb{E}[\log \frac{1}{P(X)}]$ and $V(P) = \text{Var}[\log \frac{1}{P(X)}]$ are the Shannon entropy and varentropy (cf. (10.4)) of P .

Let us analyze the result of Theorem 29.6 and see how it extends to large alphabets. It can be shown that³ $\max_{P \in \mathcal{P}_k} V(P) \asymp \log^2 k$, which suggests that $R_{\text{ent}}^* \equiv R_{\text{ent}}^*(k, n)$ may satisfy $R_{\text{ent}}^* \asymp \frac{\log^2 k}{n}$ even when the alphabet size k grows with n ; however, this result only holds for sufficiently small alphabet. In fact, back in Lemma 13.4 we have shown that for the empirical entropy which achieves the bound in Theorem 29.6, its bias is on the order of $\frac{k}{n}$, which is no longer negligible on large alphabets. Using techniques of polynomial approximation [323, 164], one can reduce this bias to $\frac{k}{n \log k}$ and further show that consistent entropy estimation is only possible if and only if $n \gg \frac{k}{\log k}$ [308], in which case the minimax rate satisfies

$$R_{\text{ent}}^* \asymp \left(\frac{k}{n \log k} \right)^2 + \frac{\log^2 k}{n}$$

In summary, one needs to exercise caution extending classical large-sample results to high dimensions, especially when bias becomes the dominating factor.

³ Indeed, $\max_{P \in \mathcal{P}_k} V(P) \leq \log^2 k$ for all $k \geq 3$ [231, Eq. (464)]. For the lower bound, consider $P = (\frac{1}{2}, \frac{1}{2(k-1)}, \dots, \frac{1}{2(k-1)})$.

30 Mutual information method

In this chapter we describe a strategy for proving statistical lower bound we call the *Mutual Information Method* (MIM), which entails comparing the amount of information data provides with the minimum amount of information needed to achieve a certain estimation accuracy. Similar to Section 29.2, the main information-theoretical ingredient is the data-processing inequality, this time for mutual information as opposed to f -divergences.

Here is the main idea of the MIM: Fix some prior π on Θ and we aim to lower bound the Bayes risk R_π^* of estimating $\theta \sim \pi$ on the basis of X with respect to some loss function ℓ . Let $\hat{\theta}$ be an estimator such that $\mathbb{E}[\ell(\theta, \hat{\theta})] \leq D$. Then we have the Markov chain $\theta \rightarrow X \rightarrow \hat{\theta}$. Applying the data processing inequality (Theorem 3.7), we have

$$\inf_{P_{\hat{\theta}|\theta}: \mathbb{E}\ell(\theta, \hat{\theta}) \leq D} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X). \quad (30.1)$$

Note that

- The leftmost quantity can be interpreted as the minimum amount of information required to achieve a given estimation accuracy. This is precisely the rate-distortion function $\phi(D) \equiv \phi_\theta(D)$ (recall Section 24.3).
- The rightmost quantity can be interpreted as the amount of information provided by the data about the latent parameter. Sometimes it suffices to further upper-bound it by the capacity of the channel $P_{X|\theta}$ by maximizing over all priors (Chapter 5):

$$I(\theta; X) \leq \sup_{\pi \in \Delta(\Theta)} I(\theta; X) \triangleq C. \quad (30.2)$$

Therefore, we arrive at the following lower bound on the Bayes and hence the minimax risks

$$R_\pi^* \geq \phi^{-1}(I(\theta; X)) \geq \phi^{-1}(C). \quad (30.3)$$

The reasoning of the mutual information method is reminiscent of the converse proof for joint-source channel coding in Section 26.3. As such, the argument here retains the flavor of “source-channel separation”, in that the lower bound in (30.1) depends only on the prior (source) and the loss function, while the capacity upper bound (30.2) depends only on the statistical model (channel).

In the next few sections, we discuss a sequence of examples to illustrate the MIM and its execution:

30.1 GLM revisited and Shannon lower bound 507

- Denoising a vector in Gaussian noise, where we will compute the exact minimax risk;
- Denoising a sparse vector, where we determine the sharp minimax rate;
- Community detection, where the goal is to recover a dense subgraph planted in a bigger Erdős-Rényi graph.

In the next chapter we will discuss three popular approaches for, namely, *Le Cam's method*, *Assouad's lemma*, and *Fano's method*. As illustrated in Fig. 30.1, all three follow from the mutual

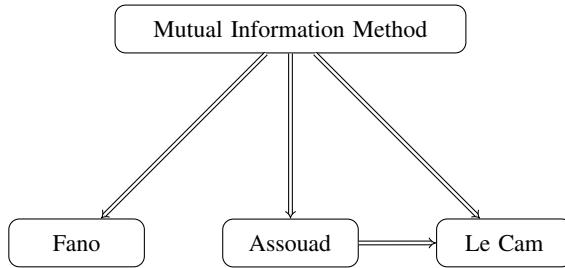


Figure 30.1 The three lower bound techniques as consequences of the Mutual Information Method.

information method, corresponding to different choice of prior π for θ , namely, the uniform distribution over a two-point set $\{\theta_0, \theta_1\}$, the hypercube $\{0, 1\}^d$, and a packing (recall Section 27.1). While these methods are highly useful in determining the minimax rate for many problems, they are often loose with constant factors compared to the MIM. In the last section of this chapter, we discuss the problem of how and when is non-trivial estimation achievable by applying the MIM; for this purpose, none of the three methods in the next chapter works.

30.1 GLM revisited and Shannon lower bound

Consider the d -dimensional GLM, where we observe $X = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} N(\theta, I_d)$ and $\theta \in \Theta$ is the parameter. Denote by $R^*(\Theta)$ the minimax risk with respect to the quadratic loss $\ell(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$.

First, let us consider the unconstrained model where $\Theta = \mathbb{R}^d$. Estimating using the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\theta, \frac{1}{n} I_d)$, we achieve the upper bound $R^*(\mathbb{R}^d) \leq \frac{d}{n}$. This turns out to be the exact minimax risk, as shown in Example 28.2 by computing the Bayes risk for Gaussian priors. Next we apply the mutual information method to obtain the same matching lower bound without evaluating the Bayes risk. Again, let us consider $\theta \sim N(0, sI_d)$ for some $s > 0$. We know from the Gaussian rate-distortion function (Theorem 26.4) that

$$\phi(D) = \inf_{P_{\hat{\theta}|\theta}: \mathbb{E}[\|\hat{\theta} - \theta\|_2^2] \leq D} I(\theta; \hat{\theta}) = \begin{cases} \frac{d}{2} \log \frac{sd}{D} & D < sd \\ 0 & \text{otherwise.} \end{cases}$$

Using the sufficiency of \bar{X} and the formula of Gaussian channel capacity (cf. Theorem 5.9 or Theorem 20.10), the mutual information between the parameter and the data can be computed as

$$I(\theta; X) = I(\theta; \bar{X}) = \frac{d}{2} \log(1 + sn).$$

It then follows from (30.3) that $R_\pi^* \geq \frac{sd}{1+sn}$, which in fact matches the exact Bayes risk in (28.7). Sending $s \rightarrow \infty$ yields the identity

$$R^*(\mathbb{R}^d) = \frac{d}{n}. \quad (30.4)$$

In the above unconstrained GLM, we are able to compute everything in close form when applying the mutual information method. Such exact expressions are rarely available in more complicated models in which case various bounds on the mutual information will prove useful. Next, let us consider the GLM with bounded means, where the parameter space $\Theta = B(\rho) = \{\theta : \|\theta\|_2 \leq \rho\}$ is the ℓ_2 -ball of radius ρ centered at zero. In this case there is no known close-form formula for the minimax quadratic risk even in one dimension. Nevertheless, the next result determines the sharp minimax rate, which characterizes the minimax risk up to universal constant factors.

Theorem 30.1 (Bounded GLM).

$$R^*(B(\rho)) \asymp \frac{d}{n} \wedge \rho^2. \quad (30.5)$$

Remark 30.1. Comparing (30.5) with (30.4), we see that if $\rho \gtrsim \sqrt{d/n}$, it is rate-optimal to ignore the bounded-norm constraint; if $\rho \lesssim \sqrt{d/n}$, we can discard all observations and estimate by zero, because data do not provide a better resolution than the prior information.

Proof. The upper bound $R^*(B(\rho)) \leq \frac{d}{n} \wedge \rho^2$ follows from considering the estimator $\hat{\theta} = \bar{X}$ and $\hat{\theta} = 0$. To prove the lower bound, we apply the mutual information method with a uniform prior $\theta \sim \text{Unif}(B(r))$, where $r \in [0, \rho]$ is to be optimized. The mutual information can be upper bound using the AWGN capacity as follows:

$$I(\theta; X) = I(\theta; \bar{X}) \leq \sup_{P_\theta: \mathbb{E}[\|\theta\|_2^2] \leq r} I(\theta; \theta + \frac{1}{\sqrt{n}}Z) = \frac{d}{2} \log \left(1 + \frac{nr^2}{d} \right) \leq \frac{nr^2}{2}, \quad (30.6)$$

where $Z \sim N(0, I_d)$. Alternatively, we can use Corollary 5.4 to bound the capacity (as information radius) by the KL diameter, which yields the same bound within constant factors:

$$I(\theta; X) \leq \sup_{P_\theta: \|\theta\| \leq r} I(\theta; \theta + \frac{1}{\sqrt{n}}Z) \leq \max_{\theta, \theta' \in B(r)} D(N(\theta, I_d/n) \| N(\theta, I_d/n)) = 2nr^2. \quad (30.7)$$

For the lower bound, due to the lack of close-form formula for the rate-distortion function for uniform distribution over Euclidean balls, we apply the *Shannon lower bound* (SLB) from Section 26.1. Since θ has an isotropic distribution, applying Theorem 26.5 yields

$$\inf_{P_{\hat{\theta}|\theta}: \mathbb{E}\|\theta - \hat{\theta}\|^2 \leq D} I(\theta; \hat{\theta}) \geq h(\theta) + \frac{d}{2} \log \frac{2\pi e d}{D} \geq \frac{d}{2} \log \frac{cr^2}{D},$$

30.1 GLM revisited and Shannon lower bound 509

for some universal constant c , where the last inequality is because for $\theta \sim \text{Unif}(B(r))$, $h(\theta) = \log \text{vol}(B(r)) = d \log r + \log \text{vol}(B(1))$ and the volume of a unit Euclidean ball in d dimensions satisfies (recall (27.14)) $\text{vol}(B(1))^{1/d} \asymp \frac{1}{\sqrt{d}}$.

Finally, applying (30.3) yields $\frac{1}{2} \log \frac{cr^2}{R^*} \leq \frac{nr^2}{2}$, i.e., $R^* \geq cr^2 e^{-nr^2/d}$. Optimizing over r and using the fact that $\sup_{0 < x < 1} xe^{-ax} = \frac{1}{ea}$ if $a \geq 1$ and e^{-a} if $a < 1$, we have

$$R^* \geq \sup_{r \in [0, \rho]} cr^2 e^{-nr^2/d} \asymp \frac{d}{n} \wedge \rho^2 \square$$

Finally, to further demonstrate the usefulness of the SLB, we consider non-quadratic loss $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^r$, the r th power of an arbitrary norm on \mathbb{R}^d , for which the SLB was given in (26.5) (see Exercise V.6). Applying the mutual information method yields the following minimax lower bound.

Theorem 30.2 (GLM with norm loss). *Let $X = (X_1, \dots, X_n) \stackrel{i.i.d.}{\sim} N(\theta, I_d)$ and let $r > 0$ be a constant. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{\theta} [\|\hat{\theta} - \theta\|^r] \geq \frac{d}{re} \left(\frac{2\pi e}{n} \right)^{r/2} \left[V_{\|\cdot\|} \Gamma \left(1 + \frac{d}{r} \right) \right]^{-r/d} \asymp n^{-r/2} V_{\|\cdot\|}^{-r/d}. \quad (30.8)$$

Furthermore,

$$\left(\frac{d}{\mathbb{E}[\|Z\|_*]} \right)^r \lesssim n^{r/2} \cdot \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{\theta} [\|\hat{\theta} - \theta\|^r] \lesssim \mathbb{E}[\|Z\|^r], \quad (30.9)$$

where $Z \sim N(0, I_d)$ and $\|x\|_* = \sup\{\langle x, y \rangle : \|y\| \leq 1\}$ is the dual norm of $\|\cdot\|$.

Proof. Choose a Gaussian prior $\theta \sim \mathcal{N}(0, sI_d)$. Suppose $\mathbb{E}[\|\hat{\theta} - \theta\|^r] \leq D$. By the data processing inequality,

$$\frac{d}{2} \log(1 + ns) \geq I(\theta; X) \geq I(\theta; \hat{\theta}) \geq \frac{d}{2} \log(2\pi es) - \log \left\{ V_{\|\cdot\|} \left(\frac{Dre}{d} \right)^{\frac{d}{r}} \Gamma \left(1 + \frac{d}{r} \right) \right\},$$

where the last inequality follows from (26.5). Rearranging terms and sending $s \rightarrow \infty$ yields the first inequality in (30.8), and the second follows from Stirling's approximation $\Gamma(x)^{1/x} \asymp x$ for $x \rightarrow \infty$. For (30.9), the upper bound follows from choosing $\hat{\theta} = \bar{X}$ and the lower bound follows from applying (30.8) with the following bound of Urysohn (cf., e.g., [229, p. 7]) on the volume of a symmetric convex body. \square

Lemma 30.3. *For any symmetric convex body $K \subset \mathbb{R}^d$, $\text{vol}(K)^{1/d} \lesssim w(K)/d$, where $w(K)$ is the Gaussian width of K defined in (27.19).*

Example 30.1. Recall from Theorem 28.13 that the upper bound in (30.9) is an equality. In view of this, let us evaluate the tightness of the lower bound from SLB. As an example, consider $r = 2$ and the ℓ_q -norm $\|x\|_q = \left(\sum_{i=1}^d |x_i|^q \right)^{1/q}$ with $1 \leq q \leq \infty$. Recall the formula (27.13) for the

volume of a unit ℓ_q -ball:

$$V_{\|\cdot\|_q} = \frac{\left[2\Gamma\left(1 + \frac{1}{q}\right)\right]^d}{\Gamma\left(1 + \frac{d}{q}\right)}.$$

In the special case of $q = 2$, we see that the lower bound in (30.8) is in fact exact and coincides with (30.4). For general $q \in [1, \infty)$, (30.8) shows that the minimax rate is $\frac{d^{2/q}}{n}$. However, for $q = \infty$, the minimax lower bound we get is $1/n$, independent of the dimension d . In fact, the upper bound in (30.9) is tight and since $\mathbb{E}\|Z\|_\infty \asymp \sqrt{\log d}$ (cf. Lemma 27.12), the minimax rate for the squared ℓ_∞ -risk is $\frac{\log d}{n}$. We will revisit this example in Section 31.3 and show how to obtain the sharp logarithmic dependency on the dimension.

Remark 30.2 (SLB versus the volume method). Recall the connection between rate-distortion function and the metric entropy in Section 27.7. As we have seen in Section 27.2, a common lower bound for metric entropy is via the volume bound. In fact, the SLB can be interpreted as a volume-based lower bound to the rate-distortion function. To see this, consider $r = 1$ and let θ be uniformly distributed over some compact set Θ , so that $h(\theta) = \log \text{vol}(\Theta)$ (Theorem 2.7.(a)). Applying Stirling's approximation, the lower bound in (26.5) becomes $\log \frac{\text{vol}(\Theta)}{\text{vol}(B_{\|\cdot\|_q}(c\epsilon))}$ for some constant c , which has the same form as the volume ratio in Theorem 27.4 for metric entropy. We will see later in Section 31.3 that in statistical applications, applying SLB yields basically the same lower bound as applying Fano's method to a packing obtained from the volume bound, although SLB does not rely explicitly on a packing.

30.2 GLM with sparse means

In this section we consider the problem of denoising for a sparse vector. Specifically, consider again the Gaussian location model $N(\theta, I_d)$ where the mean vector θ is known to be k -sparse, taking values in the “ ℓ_0 -ball”

$$B_0(k) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}, \quad k \in [p],$$

where $\|\theta\|_0 = |\{i \in [d] : \theta_i \neq 0\}|$ is the number of nonzero entries of θ , indicating the sparsity of θ . Our goal is to characterize the minimax quadratic risk

$$R_n^*(B_0(k)) = \inf_{\hat{\theta}} \sup_{\theta \in B_0(k)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2.$$

Next we prove an optimal lower bound applying MIM. (For a different proof using Fano's method in Section 31.3, see Exercise VI.11.)

Theorem 30.4.

$$R_n^*(B_0(k)) \gtrsim \frac{k}{n} \log \frac{ed}{k}. \quad (30.10)$$

A few remarks are in order:

30.2 GLM with sparse means 511

Remark 30.3. • The lower bound (30.10) turns out to be tight, achieved by the maximum likelihood estimator

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\|\theta\|_0 \leq k} \|\bar{X} - \theta\|_2, \quad (30.11)$$

which is equivalent to keeping the k entries from \bar{X} with the largest magnitude and setting the rest to zero, or the following hard-thresholding estimator $\hat{\theta}^\tau$ with an appropriately chosen τ (see Exercise VI.12):

$$\hat{\theta}_i^\tau = X_i 1_{\{|X_i| \geq \tau\}}. \quad (30.12)$$

- Sharp asymptotics: For sublinear sparsity $k = o(d)$, we have $R_n^*(B_0(k)) = (2 + o(1))\frac{k}{n} \log \frac{d}{k}$ (Exercise VI.12); for linear sparsity $k = (\eta + o(1))d$ with $\eta \in (0, 1)$, $R_n^*(B_0(k)) = (\beta(\eta) + o(1))d$ for some constant $\beta(\eta)$. For the latter and more refined results, we refer the reader to the monograph [167, Chapter 8].

Proof. First, note that $B_0(k)$ is a union of linear subspace of \mathbb{R}^d and thus homogeneous. Therefore by scaling, we have

$$R_n^*(B_0(k)) = \frac{1}{n} R_1^*(B_0(k)) \triangleq \frac{1}{n} R^*(k, d). \quad (30.13)$$

Thus it suffices to consider $n = 1$. Denote the observation by $X = \theta + Z$.

Next, note that the following oracle lower bound:

$$R^*(k, d) \geq k,$$

which is the optimal risk given the extra information of the support of θ , in view of (30.4). Thus to show (30.10), below it suffices to consider $k \leq d/4$.

We now apply the mutual information method. Recall from (27.10) that S_k^d denotes the Hamming sphere, namely,

$$S_k^d = \{b \in \{0, 1\}^d : w_H(b) = k\},$$

where $w_H(b)$ denotes the Hamming weights of b . Let b be uniformly distributed over S_k^d and let $\theta = \tau b$, where $\tau = \sqrt{\log \frac{d}{k}}$. Given any estimator $\hat{\theta} = \hat{\theta}(X)$, define an estimator $\hat{b} \in \{0, 1\}^d$ for b by

$$\hat{b}_i = \begin{cases} 0 & \hat{\theta}_i \leq \tau/2 \\ 1 & \hat{\theta}_i > \tau/2 \end{cases}, \quad i \in [d].$$

Thus the Hamming loss of \hat{b} can be related to the squared loss of $\hat{\theta}$ as

$$\|\theta - \hat{\theta}\|_2^2 \geq \frac{\tau^2}{4} d_H(b, \hat{b}). \quad (30.14)$$

Let $\mathbb{E} d_H(b, \hat{b}) = \delta k$. Assume that $\delta \leq \frac{1}{4}$, for otherwise, we are done.

512

Note the the following Markov chain $b \rightarrow \theta \rightarrow X \rightarrow \hat{\theta} \rightarrow \hat{b}$ and thus, by the data processing inequality of mutual information,

$$I(b; \hat{b}) \leq I(\theta; X) \leq \frac{d}{2} \log \left(1 + \frac{k\tau^2}{d} \right) \leq \frac{k\tau^2}{2} = \frac{k}{2} \log \frac{d}{k}.$$

where the second inequality follows from the fact that $\|\theta\|_2^2 = k\tau^2$ and the Gaussian channel capacity.

Conversely,

$$\begin{aligned} I(\hat{b}; b) &\geq \min_{\mathbb{E} d_H(b, \hat{b}) \leq \delta d} I(\hat{b}; b) \\ &= H(b) - \max_{\mathbb{E} d_H(b, \hat{b}) \leq \delta d} H(b|\hat{b}) \\ &\geq \log \binom{d}{k} - \max_{\mathbb{E} w_H(b \oplus \hat{b}) \leq \delta k} H(b \oplus \hat{b}) = \log \binom{d}{k} - dh\left(\frac{\delta k}{d}\right), \end{aligned} \quad (30.15)$$

where the last step follows from Exercise I.9.

Combining the lower and upper bound on the mutual information and using $\binom{d}{k} \geq \left(\frac{d}{k}\right)^k$, we get $dh\left(\frac{\delta k}{d}\right) \geq \frac{k}{2}k \log \frac{d}{k}$. Since $h(p) \leq -p \log p + p$ for $p \in [0, 1]$ and $k/d \leq \frac{1}{4}$ by assumption, we conclude that $\delta \geq ck/d$ for some absolute constant c , completing the proof of (30.10) in view of (30.14). \square

30.3 Community detection

As another application of the mutual information method, let us consider the following statistical problem of detecting a single hidden community in random graphs, also known as the *planted dense subgraph model* [212]. Let C^* be drawn uniformly at random from all subsets of $[n]$ of cardinality $k \geq 2$. Let G denote a random graph on the vertex set $[n]$, such that for each $i \neq j$, they are connected independently with probability p if both i and j belong to C^* , and with probability q otherwise. Assuming that $p > q$, the set C^* represents a densely connected community, which forms an Erdős-Rényi graph $G(k, p)$ planted in the bigger $G(n, q)$ graph. Upon observing G , the goal is to reconstruct C^* as accurately as possible. In particular, given an estimator $\hat{C} = \hat{C}(G)$, we say it achieves *almost exact recovery* if $\mathbb{E}|C \triangle \hat{C}| = o(k)$. The following result gives a necessary condition in terms of the parameters (p, q, n, k) :

Theorem 30.5. *Assume that k/n is bounded away from one. If almost exact recovery is possible, then*

$$d(p\|q) \geq \frac{2 + o(1)}{k-1} \log \frac{n}{k}. \quad (30.16)$$

Remark 30.4. In addition to Theorem 30.8, another necessary condition is that

$$d(p\|q) = \omega\left(\frac{1}{k}\right), \quad (30.17)$$

30.4 Estimation better than chance 513

which can be shown by a reduction to testing the membership of two nodes given the rest. It turns out that conditions (30.16) and (30.17) are optimal, in the sense that almost exact recovery can be achieved (via maximum likelihood) provided that (30.17) holds and $d(p\|q) \geq \frac{2+\epsilon}{k-1} \log \frac{n}{k}$ for any constant $\epsilon > 0$. For details, we refer the readers to [147].

Proof. Suppose \hat{C} achieves almost exact recovery of C^* . Let $\xi^*, \hat{\xi} \in \{0, 1\}^k$ denote their indicator vectors, respectively, for example, $\xi_i^* = 1_{\{i \in C^*\}}$ for each $i \in [n]$. Then $\mathbb{E}[d_H(\xi, \hat{\xi})] = \epsilon_n k$ for some $\epsilon_n \rightarrow 0$. Applying the mutual information method as before, we have

$$I(G; \xi^*) \geq I(\hat{\xi}; \xi^*) \stackrel{(a)}{\geq} \log \binom{n}{k} - nh\left(\frac{\epsilon_n k}{n}\right) \stackrel{(b)}{\geq} k \log \frac{n}{k}(1 + o(1)),$$

where (a) follows in exact the same manner as (30.15) did from Exercise I.9; (b) follows from the assumption that $k/n \leq 1 - c$ for some constant c .

On the other hand, we upper bound the mutual information between the hidden community and the graph as follows:

$$I(G; \xi^*) \stackrel{(a)}{=} \min_Q D(P_{G|\xi^*} \| Q | P_{\xi^*}) \stackrel{(b)}{\leq} D(P_{G|\xi^*} \| \text{Ber}(q)^{\otimes \binom{n}{2}} | P_{\xi^*}) \stackrel{(c)}{=} \binom{k}{2} d(p\|q),$$

where (a) is by the variational representation of mutual information in Corollary 4.1; (b) follows from choosing Q to be the distribution of the Erdős-Rényi graph $G(n, q)$; (c) is by the tensorization property of KL divergence for product distributions (see Theorem 2.15). Combining the last two displays completes the proof. \square

30.4 Estimation better than chance

Instead of characterizing the rate of convergence of the minimax risk to zero as the amount of data grows, suppose we are in a regime where this is impossible due to either limited sample size, poor signal to noise ratio, or the high dimensionality; instead, we are concerned with the modest goal of achieving an estimation error strictly better than the trivial error (without data). In the context of clustering, this is known as weak recovery or correlated recovery, where the goal is not to achieve a vanishing misclassification rate but one strictly better than random guessing the labels. It turns out that MIM is particularly suited for this regime. (In fact, we will see in the next chapter that all three popular further relaxations of MIM fall short due to the loss of constant factors.)

As an example, let us continue the setting of Theorem 30.1, where the goal is to estimate a vector in a high-dimensional unit-ball based on noisy observations. Note that the radius of the parameter space is one, so the trivial squared error equals one. The following theorem shows that in high dimensions, non-trivial estimation is achievable if and only if the sample n grows proportionally with the dimension d ; otherwise, when $d \gg n \gg 1$, the optimal estimation error is $1 - \frac{n}{d}(1 + o(1))$.

Theorem 30.6 (Bounded GLM continued). Suppose $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, I_d)$, where θ belongs to B , the unit ℓ_2 -ball in \mathbb{R}^d . Then for some universal constant C_0 ,

$$e^{-\frac{n+C_0}{d-1}} \leq \inf_{\hat{\theta}} \sup_{\theta \in B} \mathbb{E}_{\theta}[\|\hat{\theta} - \theta\|^2] \leq \frac{d}{d+n}.$$

Proof. Without loss of generality, assume that the observation is $X = \theta + \frac{Z}{\sqrt{n}}$, where $Z \sim \mathcal{N}(0, I_d)$. For the upper bound, applying the shrinkage estimator¹ $\hat{\theta} = \frac{1}{1+d/n}X$ yields $\mathbb{E}[\|\hat{\theta} - \theta\|^2] \leq \frac{d}{n+d}$.

For the lower bound, we apply MIM as in Theorem 30.1 with the prior $\theta \sim \text{Unif}(S^{d-1})$. We still apply the AWGN capacity in (30.6) to get $I(\theta; X) \leq n/2$. (Here the constant 1/2 is important and so the diameter-based (30.7) is too loose.) For the rate-distortion function of spherical uniform distribution, applying Theorem 27.20 yields $I(\theta; \hat{\theta}) \geq \frac{d-1}{2} \log \frac{1}{\mathbb{E}[\|\hat{\theta} - \theta\|^2]} - C$. Thus the lower bound on $\mathbb{E}[\|\hat{\theta} - \theta\|^2]$ follows from the data processing inequality. \square

A similar phenomenon also occurs in the problem of estimating a discrete distribution P on k elements based on n iid observations, which has been studied in Section 29.4 for small alphabet in the large-sample asymptotics and extended in Exercise VI.7–VI.9 to large alphabets. In particular, consider the total variation loss, which is at most one. Ex. VI.9f shows that the TV error of any estimator is $1 - o(1)$ if $n \ll k$; conversely, Ex. VI.9b demonstrates an estimator \hat{P} such that $\mathbb{E}[\chi^2(P \parallel \hat{P})] \leq \frac{k-1}{n+1}$. Applying the joint range (7.29) between TV and χ^2 and Jensen's inequality, we have

$$\mathbb{E}[\text{TV}(P, \hat{P})] \leq \begin{cases} \frac{1}{2} \sqrt{\frac{k-1}{n+1}} & n \geq k-2 \\ \frac{k-1}{k+n} & n \leq k-2 \end{cases}$$

which is bounded away from one whenever $n = \Omega(k)$. In summary, non-trivial estimation in total variation is possible if and only if n scales at least proportionally with k .

¹ This corresponds to the Bayes estimator (Example 28.1) when we choose $\theta \sim \mathcal{N}(0, \frac{1}{d}I_d)$, which is approximately concentrated on the unit sphere.

31

Lower bounds via reduction to hypothesis testing

In this chapter we study three commonly used techniques for proving minimax lower bounds, namely, *Le Cam's method*, *Assouad's lemma*, and *Fano's method*. Compared to the results in Chapter 29 geared towards large-sample asymptotics in smooth parametric models, the approach here is more generic, less tied to mean-squared error, and applicable in nonasymptotic settings such as nonparametric or high-dimensional problems.

The common rationale of all three methods is reducing statistical estimation to hypothesis testing. Specifically, to lower bound the minimax risk $R^*(\Theta)$ for the parameter space Θ , the first step is to notice that $R^*(\Theta) \geq R^*(\Theta')$ for any subcollection $\Theta' \subset \Theta$, and Le Cam, Assouad, and Fano's methods amount to choosing Θ' to be a two-point set, a hypercube, or a packing, respectively. In particular, Le Cam's method reduces the estimation problem to binary hypothesis testing. This method is perhaps the easiest to evaluate; however, the disadvantage is that it is frequently loose in estimating high-dimensional parameters. To capture the correct dependency on the dimension, both Assouad's and Fano's method rely on reduction to testing multiple hypotheses.

As illustrated in Fig. 30.1, all three methods in fact follow from the common principle of the mutual information method (MIM) in Chapter 30, corresponding to different choice of priors. The limitation of these methods, compared to the MIM, is that, due to the looseness in constant factors, they are ineffective for certain problems such as estimation better than chance discussed in Section 30.4.

31.1 Le Cam's two-point method

Theorem 31.1. Suppose the loss function $\ell : \Theta \times \Theta \rightarrow \mathbb{R}_+$ satisfies $\ell(\theta, \theta) = 0$ for all $\theta \in \Theta$ and the following α -triangle inequality for some $\alpha > 0$: For all $\theta_0, \theta_1, \theta \in \Theta$,

$$\ell(\theta_0, \theta_1) \leq \alpha(\ell(\theta_0, \theta) + \ell(\theta_1, \theta)). \quad (31.1)$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \ell(\theta, \hat{\theta}) \geq \sup_{\theta_0, \theta_1 \in \Theta} \frac{\ell(\theta_0, \theta_1)}{2\alpha} (1 - \text{TV}(P_{\theta_0}, P_{\theta_1})) \quad (31.2)$$

Proof. Fix $\theta_0, \theta_1 \in \Theta$. Given any estimator $\hat{\theta}$, let us convert it into the following (randomized) test:

$$\tilde{\theta} = \begin{cases} \theta_0 & \text{with probability } \frac{\ell(\theta_1, \hat{\theta})}{\ell(\theta_0, \hat{\theta}) + \ell(\theta_1, \hat{\theta})}, \\ \theta_1 & \text{with probability } \frac{\ell(\theta_0, \hat{\theta})}{\ell(\theta_0, \hat{\theta}) + \ell(\theta_1, \hat{\theta})}. \end{cases}$$

By the α -triangle inequality, we have

$$\mathbb{E}_{\theta_0}[\ell(\tilde{\theta}, \theta_0)] = \ell(\theta_0, \theta_1) \mathbb{E}_{\theta_0} \left[\frac{\ell(\theta_0, \hat{\theta})}{\ell(\theta_0, \hat{\theta}) + \ell(\theta_1, \hat{\theta})} \right] \geq \frac{1}{\alpha} \mathbb{E}_{\theta_0}[\ell(\hat{\theta}, \theta_0)],$$

and similarly for θ_1 . Consider the prior $\pi = \frac{1}{2}(\delta_{\theta_0} + \delta_{\theta_1})$ and let $\theta \sim \pi$. Taking expectation on both sides yields the following lower bound on the Bayes risk:

$$\mathbb{E}_\pi[\ell(\hat{\theta}, \theta)] \geq \frac{\ell(\theta_0, \theta_1)}{\alpha} \mathbb{P}[\tilde{\theta} \neq \theta] \geq \frac{\ell(\theta_0, \theta_1)}{2\alpha} (1 - \text{TV}(P_{\theta_0}, P_{\theta_1}))$$

where the last step follows from the minimum average probability of error in binary hypothesis testing (Theorem 7.12). \square

Remark 31.1. As an example where the bound (31.2) is tight (up to constants), consider a binary hypothesis testing problem with $\Theta = \{\theta_0, \theta_1\}$ and the Hamming loss $\ell(\theta, \hat{\theta}) = 1\{\theta \neq \hat{\theta}\}$, where $\theta, \hat{\theta} \in \{\theta_0, \theta_1\}$ and $\alpha = 1$. Then the left side is the minimax probability of error, and the right side is the optimal average probability of error (cf. (7.17)). These two quantities can coincide (for example for Gaussian location model).

Another special case of interest is the quadratic loss $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, where $\theta, \hat{\theta} \in \mathbb{R}^d$, which satisfies the α -triangle inequality with $\alpha = 2$. In this case, the leading constant $\frac{1}{4}$ in (31.2) makes sense, because in the extreme case of $\text{TV} = 0$ where P_{θ_0} and P_{θ_1} cannot be distinguished, the best estimate is simply $\frac{\theta_0 + \theta_1}{2}$. In addition, the inequality (31.2) can be deduced based on properties of f -divergences and their joint range (Chapter 7). To this end, abbreviate P_θ as P_i for $i = 0, 1$ and consider the prior $\pi = \frac{1}{2}(\delta_{\theta_0} + \delta_{\theta_1})$. Then the Bayes estimator (posterior mean) is $\frac{\theta_0 dP_0 + \theta_1 dP_1}{dP_0 + dP_1}$ and the Bayes risk is given by

$$\begin{aligned} R_\pi^* &= \frac{\|\theta_0 - \theta_1\|^2}{2} \int \frac{dP_0 dP_1}{dP_0 + dP_1} \\ &= \frac{\|\theta_0 - \theta_1\|^2}{4} (1 - \text{LC}(P_0, P_1)) \geq \frac{\|\theta_0 - \theta_1\|^2}{4} (1 - \text{TV}(P_0, P_1)), \end{aligned}$$

where $\text{LC}(P_0, P_1) = \int \frac{(dP_0 - dP_1)^2}{dP_0 + dP_1}$ is the Le Cam divergence defined in (7.6) and satisfies $\text{LC} \leq \text{TV}$.

Example 31.1. As a concrete example, consider the one-dimensional GLM with sample size n . By considering the sufficient statistic $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, the model is simply $\{N(\theta, \frac{1}{n}) : \theta \in \mathbb{R}\}$. Applying Theorem 31.1 yields

$$\begin{aligned} R^* &\geq \sup_{\theta_0, \theta_1 \in \mathbb{R}} \frac{1}{4} |\theta_0 - \theta_1|^2 \left\{ 1 - \text{TV} \left(N \left(\theta_0, \frac{1}{n} \right), N \left(\theta_1, \frac{1}{n} \right) \right) \right\} \\ &\stackrel{(a)}{=} \frac{1}{4n} \sup_{s>0} s^2 (1 - \text{TV}(N(0, 1), N(s, 1))) \stackrel{(b)}{=} \frac{c}{n} \end{aligned} \tag{31.3}$$

31.1 Le Cam's two-point method 517

where (a) follows from the shift and scale invariance of the total variation; in (b) $c \approx 0.083$ is some absolute constant, obtained by applying the formula $\text{TV}(N(0, 1), N(s, 1)) = 2\Phi(\frac{s}{2}) - 1$ from (7.37). On the other hand, we know from Example 28.2 that the minimax risk equals $\frac{1}{n}$, so the two-point method is rate-optimal in this case.

In the above example, for two points separated by $\Theta(\frac{1}{\sqrt{n}})$, the corresponding hypothesis cannot be tested with vanishing probability of error so that the resulting estimation risk (say in squared error) cannot be smaller than $\frac{1}{n}$. This convergence rate is commonly known as the “parametric rate”, which we have studied in Chapter 29 for smooth parametric families focusing on the Fisher information as the sharp constant. More generally, the $\frac{1}{n}$ rate is not improvable for models with locally quadratic behavior

$$H^2(P_{\theta_0}, P_{\theta_0+t}) \asymp t^2, \quad t \rightarrow 0. \quad (31.4)$$

(Recall that Theorem 7.33 gives a sufficient condition for this behavior.) Indeed, pick θ_0 in the interior of the parameter space and set $\theta_1 = \theta_0 + \frac{1}{\sqrt{n}}$, so that $H^2(P_{\theta_0}, P_{\theta_1}) = \Theta(\frac{1}{n})$ thanks to (31.4). By Theorem 7.14, we have $\text{TV}(P_{\theta_0}^{\otimes n}, P_{\theta_1}^{\otimes n}) \leq 1 - c$ for some constant c and hence Theorem 31.1 yields the lower bound $\Omega(1/n)$ for the squared error. Furthermore, later we will show that the same locally quadratic behavior in fact guarantees the achievability of the $1/n$ rate; see Corollary 32.1.

Example 31.2. As a different example, consider the family $\text{Unif}(0, \theta)$. Note that as opposed to the quadratic behavior (31.4), we have

$$H^2(\text{Unif}(0, 1), \text{Unif}(0, 1+t)) = 2(1 - 1/\sqrt{1+t}) \asymp t.$$

Thus an application of Theorem 31.1 yields an $\Omega(1/n^2)$ lower bound. This rate is not achieved by the empirical mean estimator (which only achieves $1/n$ rate), but by the maximum likelihood estimator $\hat{\theta} = \max\{X_1, \dots, X_n\}$. Other types of behavior in t , and hence the rates of convergence, can occur even in compactly supported location families – see Example 7.1.

The limitation of Le Cam's two-point method is that it does not capture the correct dependency on the dimensionality. To see this, let us revisit Example 31.1 for d dimensions.

Example 31.3. Consider the d -dimensional GLM in Corollary 28.1. Again, it is equivalent to consider the reduced model $\{N(\theta, \frac{1}{n}) : \theta \in \mathbb{R}^d\}$. We know from Example 28.2 (see also Theorem 28.10) that for quadratic risk $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, the exact minimax risk is $R^* = \frac{d}{n}$ for any d and n . Let us compare this with the best two-point lower bound. Applying Theorem 31.1 with $\alpha = 2$,

$$\begin{aligned} R^* &\geq \sup_{\theta_0, \theta_1 \in \mathbb{R}^d} \frac{1}{4} \|\theta_0 - \theta_1\|_2^2 \left\{ 1 - \text{TV} \left(N \left(\theta_0, \frac{1}{n} I_d \right), N \left(\theta_1, \frac{1}{n} I_d \right) \right) \right\} \\ &= \sup_{\theta \in \mathbb{R}^d} \frac{1}{4n} \|\theta\|_2^2 \{1 - \text{TV}(N(0, I_d), N(\theta, I_d))\} \\ &= \frac{1}{4n} \sup_{s>0} s^2 (1 - \text{TV}(N(0, 1), N(s, 1))), \end{aligned}$$

where the second step applies the shift and scale invariance of the total variation; in the last step, by rotational invariance of isotropic Gaussians, we can rotate the vector θ align with a coordinate vector (say, $e_1 = (1, 0 \dots, 0)$) which reduces the problem to one dimension, namely,

$$\begin{aligned}\text{TV}(N(0, I_d), N(\theta, I_d)) &= \text{TV}(N(0, I_d), N(\|\theta\|e_1, I_d)) \\ &= \text{TV}(N(0, 1), N(\|\theta\|, 1)).\end{aligned}$$

Comparing the above display with (31.3), we see that the best Le Cam two-point lower bound in d dimensions coincide with that in one dimension.

Let us mention in passing that although Le Cam's two-point method is typically suboptimal for estimating a high-dimensional parameter θ , for functional estimation in high dimensions (e.g. estimating a scalar functional $T(\theta)$), Le Cam's method is much more effective and sometimes even optimal. The subtlety is that as opposed to testing a pair of simple hypotheses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, we need to test $H_0 : T(\theta) = t_0$ versus $H_1 : T(\theta) = t_1$, both of which are composite hypotheses and require a sagacious choice of priors. See Exercise VI.13 for an example.

31.2 Assouad's Lemma

From Example 31.3 we see that Le Cam's two-point method effectively only perturbs one out of d coordinates, leaving the remaining $d - 1$ coordinates unexplored; this is the source of its suboptimality. In order to obtain a lower bound that scales with the dimension, it is necessary to randomize all d coordinates. Our next topic Assouad's Lemma is an extension in this direction.

Theorem 31.2 (Assouad's Lemma). *Assume that the loss function ℓ satisfies the α -triangle inequality (31.1). Suppose Θ contains a subset $\Theta' = \{\theta_b : b \in \{0, 1\}^d\}$ indexed by the hypercube, such that $\ell(\theta_b, \theta_{b'}) \geq \beta \cdot d_H(b, b')$ for all b, b' and some $\beta > 0$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \ell(\theta, \hat{\theta}) \geq \frac{\beta d}{4\alpha} \left(1 - \max_{d_H(b, b')=1} \text{TV}(P_{\theta_b}, P_{\theta_{b'}}) \right) \quad (31.5)$$

Proof. We lower bound the Bayes risk with respect to the uniform prior over Θ' . Given any estimator $\hat{\theta} = \hat{\theta}(X)$, define $\hat{b} \in \operatorname{argmin}_{\hat{\theta}} \ell(\hat{\theta}, \theta_b)$. Then for any $b \in \{0, 1\}^d$,

$$\beta d_H(\hat{b}, b) \leq \ell(\theta_{\hat{b}}, \theta_b) \leq \alpha(\ell(\theta_{\hat{b}}, \hat{\theta}_b) + \ell(\hat{\theta}, \theta_b)) \leq 2\alpha \ell(\hat{\theta}, \theta_b).$$

Let $b \sim \text{Unif}(\{0, 1\}^d)$ and we have $b \rightarrow \theta_b \rightarrow X$. Then

$$\begin{aligned}\mathbb{E}[\ell(\hat{\theta}, \theta_b)] &\geq \frac{\beta}{2\alpha} \mathbb{E}[d_H(\hat{b}, b)] \\ &= \frac{\beta}{2\alpha} \sum_{i=1}^d \mathbb{P}[\hat{b}_i \neq b_i] \\ &\geq \frac{\beta}{4\alpha} \sum_{i=1}^d (1 - \text{TV}(P_{X|b_i=0}, P_{X|b_i=1})),\end{aligned}$$

31.3 Assouad's lemma from the Mutual Information Method 519

where the last step is again by Theorem 7.12, just like in the proof of Theorem 31.1. Each total variation can be upper bounded as follows:

$$\text{TV}(P_{X|b_i=0}, P_{X|b_i=1}) \stackrel{(a)}{=} \text{TV} \left(\frac{1}{2^{d-1}} \sum_{b:b_i=1} P_{\theta_b}, \frac{1}{2^{d-1}} \sum_{b:b_i=0} P_{\theta_b} \right) \stackrel{(b)}{\leq} \max_{d_H(b, b')=1} \text{TV}(P_{\theta_b}, P_{\theta_{b'}})$$

where (a) follows from the Bayes rule, and (b) follows from the convexity of total variation (Theorem 7.8). This completes the proof. \square

Example 31.4. Let us continue the discussion of the d -dimensional GLM in Example 31.3. Consider the quadratic loss first. To apply Theorem 31.3, consider the hypercube $\theta_b = \epsilon b$, where $b \in \{0, 1\}^d$. Then $\|\theta_b - \theta'_b\|_2^2 = \epsilon^2 d_H(b, b')$. Applying Theorem 31.3 yields

$$\begin{aligned} R^* &\geq \frac{\epsilon^2 d}{4} \left\{ 1 - \max_{b, b' \in \{0, 1\}^d, d_H(b, b')=1} \text{TV} \left(N \left(\epsilon b, \frac{1}{n} I_d \right), N \left(\epsilon b', \frac{1}{n} I_d \right) \right) \right\} \\ &= \frac{\epsilon^2 d}{4} \left\{ 1 - \text{TV} \left(N \left(0, \frac{1}{n} \right), N \left(\epsilon, \frac{1}{n} \right) \right) \right\}, \end{aligned}$$

where the last step applies (7.10) for f -divergence between product distributions that only differ in one coordinate. Setting $\epsilon = \frac{1}{\sqrt{n}}$ and by the scale-invariance of TV, we get the desired $R^* \gtrsim \frac{d}{n}$.

Next, let's consider the loss function $\|\theta_b - \theta'_b\|_\infty$. In the same setup, we only $\|\theta_b - \theta'_b\|_\infty \geq \frac{\epsilon}{d} d_H(b, b')$. Then Assouad's lemma yields $R^* \gtrsim \frac{1}{\sqrt{n}}$, which does not depend on d . In fact, $R^* \asymp \sqrt{\frac{\log d}{n}}$ as shown in Corollary 28.1. In the next section, we will discuss Fano's method which can resolve this deficiency.

31.3 Assouad's lemma from the Mutual Information Method

One can integrate the Assouad's idea into the mutual information method. Consider the Bayesian setting of Theorem 31.3, where $b^d = (b_1, \dots, b_d) \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\frac{1}{2})$. From the rate-distortion function of the Bernoulli source (Section 26.1.1), we know that for any \hat{b}^d and $\tau > 0$ there is some $\tau' > 0$ such that

$$I(b^d; X) \leq d(1 - \tau) \log 2 \implies \mathbb{E}[d_H(\hat{b}^d, b^d)] \geq d\tau'. \quad (31.6)$$

Here τ' is related to τ by $\tau \log 2 = h(\tau')$. Thus, using the same ‘‘hypercube embedding $b \rightarrow \theta_b$ ’’, the bound similar to (31.5) will follow once we can bound $I(b^d; X)$ away from $d \log 2$.

Can we use the pairwise total variation bound in (31.5) to do that? Yes! Notice that thanks to the independence of b_i 's we have¹

$$I(b_i; X|b^{i-1}) = I(b_i; X, b^{i-1}) \leq I(b_i; X, b_{\setminus i}) = I(b_i; X|b_{\setminus i}).$$

¹ Equivalently, this also follows from the convexity of the mutual information in the channel (cf. Theorem 5.5).

Applying the chain rule leads to the upper bound

$$I(b^d; X) = \sum_{i=1}^d I(b_i; X|b^{i-1}) \leq \sum_{i=1}^d I(b_i; X|b_{\setminus i}) \leq d \log 2 \max_{d_H(b, b')=1} \text{TV}(P_{X|b^d=b}, P_{X|b^d=b'}) , \quad (31.7)$$

where in the last step we used the fact that whenever $B \sim \text{Ber}(1/2)$,

$$I(B; X) \leq \text{TV}(P_{X|B=0}, P_{X|B=1}) \log 2 , \quad (31.8)$$

which follows from (7.36) by noting that the mutual information is expressed as the Jensen-Shannon divergence as $2I(B; X) = \text{JS}(P_{X|B=0}, P_{X|B=1})$. Combining (31.6) and (31.7), the mutual information method implies the following version of the Assouad's lemma: Under the assumption of Theorem 31.3,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \ell(\theta, \hat{\theta}) \geq \frac{\beta d}{4\alpha} \cdot f\left(\max_{d_H(\theta, \theta')=1} \text{TV}(P_{\theta}, P_{\theta'})\right) , \quad f(t) \triangleq h^{-1}\left(\frac{(1-t)\log 2}{2}\right) \quad (31.9)$$

where $h^{-1} : [0, \log 2] \rightarrow [0, 1/2]$ is the inverse of the binary entropy function. Note that (31.9) is slightly weaker than (31.5). Nevertheless, as seen in Example 31.4, Assouad's lemma is typically applied when the pairwise total variation is bounded away from one by a constant, in which case (31.9) and (31.5) differ by only a constant factor.

In all, we may summarize Assouad's lemma as a convenient method for bounding $I(b^d; X)$ away from the full entropy (d bits) on the basis of distances between $P_{X|b^d}$ corresponding to adjacent b^d 's.

31.4 Fano's method

In this section we discuss another method for proving minimax lower bound by reduction to multiple hypothesis testing. To this end, assume that the loss function is a metric. The idea is to consider an ϵ -packing (Chapter 27) of the parameter space, namely, a finite collection of parameters whose minimum separation is ϵ . Suppose we can show that given data one cannot reliably distinguish these hypotheses. Then the best estimation error is at least proportional to ϵ . The impossibility of testing is often shown by applying Fano's inequality (Corollary 6.1), which bounds the probability of error of testing in terms of the mutual information in Section 6.3. As such, we refer to this program *Fano's method*. The following is a precise statement.

Theorem 31.3. *Let d be a metric on Θ . Fix an estimator $\hat{\theta}$. For any $T \subset \Theta$ and $\epsilon > 0$,*

$$\mathbb{P}\left[d(\theta, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq 1 - \frac{\text{rad}_{\text{KL}}(T) + \log 2}{\log M(T, d, \epsilon)} , \quad (31.10)$$

where $\text{rad}_{\text{KL}}(T) \triangleq \inf_Q \sup_{\theta \in T} D(P_{\theta} \| Q)$ is the KL radius of the set of distributions $\{P_{\theta} : \theta \in T\}$ (recall Corollary 5.4). Consequently,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} [d(\theta, \hat{\theta})^r] \geq \sup_{T \subset \Theta, \epsilon > 0} \left(\frac{\epsilon}{2}\right)^r \left(1 - \frac{\text{rad}_{\text{KL}}(T) + \log 2}{\log M(T, d, \epsilon)}\right) , \quad (31.11)$$

31.4 Fano's method 521

Proof. It suffices to show (31.10). Fix $T \subset \Theta$. Consider an ϵ -packing $T' = \{\theta_1, \dots, \theta_M\} \subset T$ such that $\min_{i \neq j} d(\theta_i, \theta_j) \geq \epsilon$. Let θ be uniformly distributed on this packing and $X \sim P_\theta$ conditioned on θ . Given any estimator $\hat{\theta}$, construct a test by rounding $\hat{\theta}$ to $\tilde{\theta} = \operatorname{argmin}_{\theta \in T'} d(\hat{\theta}, \theta)$. By triangle inequality, $d(\theta, \tilde{\theta}) \leq 2d(\theta, \hat{\theta})$. Thus $\mathbb{P}[\theta \neq \tilde{\theta}] \leq \mathbb{P}[d(\theta, \tilde{\theta}) \geq \epsilon/2]$. On the other hand, applying Fano's inequality (Corollary 6.1) yields

$$\mathbb{P}[\theta \neq \tilde{\theta}] \geq 1 - \frac{I(\theta; X) + \log 2}{\log M}.$$

The proof of (31.10) is completed by noting that $I(\theta; X) \leq \text{rad}_{\text{KL}}(T)$ since the latter equals the maximal mutual information over the distribution of θ (Corollary 5.4). \square

As an application of Fano's method, we revisit the d -dimensional GLM in Corollary 28.1 under the ℓ_q loss ($1 \leq q \leq \infty$), with the particular focus on the dependency on the dimension. (For a different application in sparse setting see Exercise VI.11.)

Example 31.5. Consider GLM with sample size n , where $P_\theta = N(\theta, I_d)^{\otimes n}$. Taking natural logs here and below, we have

$$D(P_\theta \| P_{\theta'}) = \frac{n}{2} \|\theta - \theta'\|_2^2;$$

in other words, KL-neighborhoods are ℓ_2 -balls. As such, let us apply Theorem 31.4 to $T = B_2(\rho)$ for some $\rho > 0$ to be specified. Then $\text{rad}_{\text{KL}}(T) \leq \sup_{\theta \in T} D(P_\theta \| P_0) = \frac{n\rho^2}{2}$. To bound the packing number from below, we applying the volume bound in Theorem 27.4,

$$M(B_2(\rho), \|\cdot\|_q, \epsilon) \geq \frac{\rho^d \text{vol}(B_2)}{\epsilon^d \text{vol}(B_q)} \geq \left(\frac{c_q \rho d^{1/q}}{\epsilon \sqrt{d}} \right)^d$$

for some constant c_q , where the last step follows the volume formula (27.13) for ℓ_q -balls. Choosing $\rho = \sqrt{d/n}$ and $\epsilon = \frac{c_q}{\epsilon^2} \rho d^{1/q-1/2}$, an application of Theorem 31.4 yields the minimax lower bound

$$R_q \equiv \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta [\|\hat{\theta} - \theta\|_q] \geq C_q \frac{d^{1/q}}{\sqrt{n}} \quad (31.12)$$

for some constant C_q depending on q . This is the same lower bound as that in Example 30.1 obtained via the mutual information method plus the Shannon lower bound (which is also volume-based).

For any $q \geq 1$, (31.12) is rate-optimal since we can apply the MLE $\hat{\theta} = \bar{X}$. (Note that at $q = \infty$, the constant C_q is still finite since $\text{vol}(B_\infty) = 2^d$.) However, for the special case of $q = \infty$, (31.12) does not depend on the dimension at all, as opposed to the correct dependency $\sqrt{\log d}$ shown in Corollary 28.1. In fact, this is the same suboptimal result we previously obtained from applying Shannon lower bound in Example 30.1 or Assouad's lemma in Example 31.4. So is it possible to fix this looseness with Fano's method? It turns out that the answer is yes and the suboptimality is due to the volume bound on the metric entropy, which, as we have seen in Section 27.3, can be ineffective if ϵ scales with dimension. Indeed, if we apply the tight bound of $M(B_2, \|\cdot\|_\infty, \epsilon)$

in (27.18),² with $\epsilon = \sqrt{\frac{c \log d}{n}}$ and $\rho = \sqrt{c' \frac{\log d}{n}}$ for some absolute constants c, c' , we do get $R_\infty \gtrsim \sqrt{\frac{\log d}{n}}$ as desired.

We end this section with some comments regarding the application Theorem 31.4:

- It is sometimes convenient to further bound the KL radius by the KL diameter, since $\text{rad}_{\text{KL}}(T) \leq \text{diam}_{\text{KL}}(T) \triangleq \sup_{\theta, \theta' \in T} D(P_{\theta'} \| P_\theta)$ (cf. Corollary 5.4). This suffices for Example 31.5.
- In Theorem 31.4 we actually lower bound the global minimax risk by that restricted on a parameter subspace $T \subset \Theta$ for the purpose of controlling the mutual information, which is often difficult to compute. For the GLM considered in Example 31.5, the KL divergence is proportional to squared ℓ_2 -distance and T is naturally chosen to be a Euclidean ball. For other models such as the covariance model (Exercise VI.15) wherein the KL divergence is more complicated, the KL neighborhood T needs to be chosen carefully. Later in Section 32.4 we will apply the same Fano's method to the infinite-dimensional problem of estimating smooth density.

² In fact, in this case we can also choose the explicit packing $\{\epsilon e_1, \dots, \epsilon e_d\}$.

32 Entropic upper bound for statistical estimation

So far our discussion on information-theoretic methods have been mostly focused on statistical lower bounds (impossibility results), with matching upper bounds obtained on a case-by-case basis. In this chapter, we will discuss three information-theoretic upper bounds for statistical estimation. These three results apply to different loss functions and are obtained using completely different means; however, they take on exactly the same form involving the appropriate metric entropy of the model. Specifically, suppose that we observe X_1, \dots, X_n drawn independently from a distribution P_θ for some unknown parameter $\theta \in \Theta$, and the goal is to produce an estimate \hat{P} for the true distribution P_θ . We have the following entropic minimax upper bounds:

- KL loss (Yang-Barron [329]):

$$\inf_{\hat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[D(P_\theta \| \hat{P})] \lesssim \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{1}{n} \log N_{\text{KL}}(\mathcal{P}, \epsilon) \right\}. \quad (32.1)$$

- Hellinger loss (Le Cam-Birgé [189, 34]):

$$\inf_{\hat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[H^2(P_\theta, \hat{P})] \lesssim \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{1}{n} \log N_H(\mathcal{P}, \epsilon) \right\}. \quad (32.2)$$

- Total variation loss (Yatracos [330]):

$$\inf_{\hat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\text{TV}^2(P_\theta \| \hat{P})] \lesssim \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{1}{n} \log N_{\text{TV}}(\mathcal{P}, \epsilon) \right\}. \quad (32.3)$$

Here $N(\mathcal{P}, \epsilon)$ refers to the metric entropy (cf. Chapter 27) of the model class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ under various distances, which we will formalize along the way.

32.1 Yang-Barron's construction

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions on the space \mathcal{X} . Given $X^n = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} P_\theta$ for some $\theta \in \Theta$, we obtain an estimate $\hat{P} = \hat{P}(\cdot | X^n)$, which is a distribution

depending on X^n . The loss function is the KL divergence $D(P_\theta \parallel \hat{P})$.¹ The average risk is thus

$$\mathbb{E}_\theta D(P_\theta \parallel \hat{P}) = \int D(P_\theta \parallel \hat{P}(\cdot | X^n)) P^{\otimes n}(dx^n).$$

If the family has a common dominating measure μ , the problem is equivalent to estimate the density $p_\theta = \frac{dP_\theta}{d\mu}$, commonly referred to as the problem of *density estimation* in the statistics literature.

Our objective is to prove the upper bound (32.1) for the minimax KL risk

$$R_{\text{KL}}^*(n) \triangleq \inf_{\hat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta D(P_\theta \parallel \hat{P}), \quad (32.4)$$

where the infimum is taken over all estimators $\hat{P} = \hat{P}(\cdot | X^n)$ which is a distribution on \mathcal{X} ; in other words, we allow *improper* estimates in the sense that \hat{P} can step outside the model class \mathcal{P} . Indeed, the construction we will use in this section (such as predictive density estimators (Bayes) or their mixtures) need not be a member of \mathcal{P} . Later we will see in Sections 32.2 and 32.3 that for total variation and Hellinger loss we can always restrict to *proper* estimators;² however these loss functions are weaker than the KL divergence.

The main result of this section is the following.

Theorem 32.1. *Let C_n denotes the capacity of the channel $\theta \mapsto X^n \sim P_\theta^{\otimes n}$, namely*

$$C_n = \sup I(\theta; X^n), \quad (32.5)$$

where the supremum is over all distributions (priors) of θ taking values in Θ . Denote by

$$N_{\text{KL}}(\mathcal{P}, \epsilon) \triangleq \min \{N : \exists Q_1, \dots, Q_N \text{ s.t. } \forall \theta \in \Theta, \exists i \in [N], D(P_\theta \parallel Q_i) \leq \epsilon^2\}. \quad (32.6)$$

the KL covering number for the class \mathcal{P} . Then

$$R_{\text{KL}}^*(n) \leq \frac{C_{n+1}}{n+1} \quad (32.7)$$

$$\leq \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{1}{n+1} \log N_{\text{KL}}(\mathcal{P}, \epsilon) \right\}. \quad (32.8)$$

Conversely,

$$\sum_{t=0}^n R_{\text{KL}}^*(t) \geq C_{n+1}. \quad (32.9)$$

Note that the capacity C_n is precisely the redundancy (13.10) which governs the minimax regret in universal compression; the fact that it bounds the KL risk can be attributed to a generic relation

¹ Note the asymmetry in this loss function. Alternatively the loss $D(\hat{P} \parallel P)$ is typically infinite in nonparametric settings, because it is impossible to estimate the support of the true density exactly.

² This is in fact a generic observation: Whenever the loss function satisfies an approximate triangle inequality, any improper estimate can be converted to a proper one by its project on the model class whose risk is inflated by no more than a constant factor.

32.1 Yang-Barron's construction 525

between individual and cumulative risks which we explain later in Section 32.1.4. As explained in Chapter 13, it is in general difficult to compute the exact value of C_n even for models as simple as Bernoulli ($P_\theta = \text{Ber}(\theta)$). This is where (32.8) comes in: one can use metric entropy and tools from Chapter 27 to bound this capacity, leading to useful (and even optimal) risk bounds. We discuss two types of applications of this result.

Finite-dimensional models Consider a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of smooth parametrized densities, where $\Theta \subset \mathbb{R}^d$ is some compact set. Suppose that the KL-divergence behaves like squared norm, namely, $D(P_\theta \| P_{\theta'}) \asymp \|\theta - \theta'\|^2$ for any $\theta, \theta' \in \Theta$ and some norm $\|\cdot\|$ on \mathbb{R}^d . (For example, for GLM with $P_\theta = N(\theta, I_d)$, we have $D(P_\theta \| P_{\theta'}) = \frac{1}{2}\|\theta - \theta'\|_2^2$.) In this case, the KL covering numbers inherits the usual behavior of metric entropy in finite-dimensional space (cf. Theorem 27.4 and Corollary 27.1) and we have

$$N_{\text{KL}}(\mathcal{P}, \epsilon) \lesssim \left(\frac{1}{\epsilon}\right)^d.$$

Then (32.8) yields

$$C_n \lesssim \inf_{\epsilon > 0} \left\{ n\epsilon^2 + d \log \frac{1}{\epsilon} \right\} \asymp d \log n, \quad (32.10)$$

which is consistent with the typical asymptotics of redundancy $C_n = \frac{d}{2} \log n + o(\log n)$ (recall (13.23) and (13.24)).

Applying the upper bound (32.7) or (32.8) yields

$$R_{\text{KL}}^*(n) \lesssim \frac{d \log n}{n}.$$

As compared to the usual parametric rate of $\frac{d}{n}$ in d dimensions (e.g. GLM), this upper bound is suboptimal only by a logarithmic factor.

Infinite-dimensional models Similar to the results in Section 27.4, for nonparametric models $N_{\text{KL}}(\epsilon)$ typically grows super-polynomially in $\frac{1}{\epsilon}$ and, in turn, the capacity C_n grows super-logarithmically. In fact, whenever we have $C_n \asymp n^\alpha$ for some $\alpha > 0$, Theorem 32.1 yields the *sharp* minimax rate

$$R_{\text{KL}}^*(n) \asymp \frac{C_n}{n} \asymp n^{\alpha-1} \quad (32.11)$$

which easily follows from combining (32.7) and (32.8) – see (32.23) for details.

As a concrete example, consider the class \mathcal{P} of Lipschitz densities on $[0, 1]$ that are bounded away from zero. Using the L_2 -metric entropy previously established in Theorem 27.14, we will show in Section 32.4 that $N_{\text{KL}}(\epsilon) \asymp \epsilon^{-1}$ and thus $C_n \leq \inf_{\epsilon > 0}(n\epsilon^2 + \epsilon^{-1}) \asymp n^{1/3}$ and, in turn, $R_{\text{KL}}^*(n) \lesssim n^{-2/3}$. This rate turns out to be optimal: In Section 32.1.3 we will develop capacity lower bound based on metric entropy that shows $C_n \asymp n^{1/3}$ and hence, in view of (32.11), $R_{\text{KL}}^*(n) \asymp n^{-2/3}$.

Next, we explain the intuition behind and the proof of Theorem 32.1.

32.1.1 Bayes risk as conditional mutual information and capacity bound

To gain some insight, let us start by considering the Bayesian setting with a prior π on Θ . Conditioned on $\theta \sim \pi$, the data $X^n = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} P_\theta$.³ Any estimator, $\hat{P} = \hat{P}(\cdot | X^n)$, is a distribution on \mathcal{X} depending on X^n . As such, \hat{P} can be identified with a *conditional distribution*, say, $Q_{X_{n+1}|X^n}$, and we shall do so henceforth. For convenience, let us introduce an (unseen) observation X_{n+1} that is drawn from the same P_θ and independent of X^n conditioned on θ . In this light, the role of the estimator is to *predict* the distribution of the unseen X_{n+1} .

The following lemma shows that the Bayes KL risk equals the conditional mutual information and the Bayes estimator is precisely $P_{X_{n+1}|X^n}$ (with respect to the joint distribution induced by the prior), known as the *predictive density estimator* in the statistics literature.

Lemma 32.2. *The Bayes risk for prior π is given by*

$$R_{\text{KL}, \text{Bayes}}^*(\pi) \triangleq \inf_{\hat{P}} \int \pi(d\theta) P_\theta^{\otimes n}(dx^n) D(P_\theta \| \hat{P}(\cdot | x^n)) = I(\theta; X_{n+1} | X^n),$$

where $\theta \sim \pi$ and $(X_1, \dots, X_{n+1}) \stackrel{\text{i.i.d.}}{\sim} P_\theta$ conditioned on θ . The Bayes estimator achieving this infimum is given by $\hat{P}_{\text{Bayes}}(\cdot | x^n) = P_{X_{n+1}|X^n=x^n}$. If each P_θ has a density p_θ with respect to some common dominating measure μ , the Bayes estimator has density:

$$\hat{P}_{\text{Bayes}}(x_{n+1} | x^n) = \frac{\int \pi(d\theta) \prod_{i=1}^{n+1} p_\theta(x_i)}{\int \pi(d\theta) \prod_{i=1}^n p_\theta(x_i)}. \quad (32.12)$$

Proof. The Bayes risk can be computed as follows:

$$\begin{aligned} \inf_{Q_{X_{n+1}|X^n}} \mathbb{E}_{\theta, X^n} [D(P_\theta \| Q_{X_{n+1}|X^n})] &= \inf_{Q_{X_{n+1}|X^n}} D(P_{X_{n+1}|\theta} \| \hat{P}_{X_{n+1}|X^n} | P_{\theta, X^n}) \\ &= \mathbb{E}_{X^n} \left[\inf_{Q_{X_{n+1}|X^n}} D(P_{X_{n+1}|\theta} \| \hat{P}_{X_{n+1}|X^n} | P_{\theta|X^n}) \right] \\ &\stackrel{(a)}{=} \mathbb{E}_{X^n} [D(P_{X_{n+1}|\theta} \| P_{X_{n+1}|X^n} | P_{\theta|X^n})] \\ &= D(P_{X_{n+1}|\theta} \| P_{X_{n+1}|X^n} | P_{\theta, X^n}) \\ &\stackrel{(b)}{=} I(\theta; X_{n+1} | X^n). \end{aligned}$$

where (a) follows from the variational representation of mutual information (Theorem 4.1 and Corollary 4.1); (b) invokes the definition of the conditional mutual information (Section 3.4) and the fact that $X^n \rightarrow \theta \rightarrow X_{n+1}$ forms a Markov chain, so that $P_{X_{n+1}|\theta, X^n} = P_{X_{n+1}|\theta}$. In addition, the Bayes optimal estimator is given by $P_{X_{n+1}|X^n}$. \square

Note that the operational meaning of $I(\theta; X_{n+1} | X^n)$ is the information provided by one extra observation about θ having already obtained n observations. In most situations, since X^n will have

³ Throughout this chapter, we continue to use the conventional notation P_θ for a parametric family of distributions and use π to stand for the distribution of θ .

32.1 Yang-Barron's construction 527

already allowed θ to be consistently estimated as $n \rightarrow \infty$, the additional usefulness of X_{n+1} is vanishing. This is made precisely by the following result.

Lemma 32.3 (Diminishing marginal utility in information). $n \mapsto I(\theta; X_{n+1} | X^n)$ is a decreasing sequence. Furthermore,

$$I(\theta; X_{n+1} | X^n) \leq \frac{1}{n} I(\theta; X^{n+1}). \quad (32.13)$$

Proof. In view of the chain rule for mutual information (Theorem 3.7): $I(\theta; X^{n+1}) = \sum_{i=1}^{n+1} I(\theta; X_i | X^{i-1})$, (32.13) follows from the monotonicity. To show the latter, let us consider a “sampling channel” where the input is θ and the output is X sampled from P_θ . Let $I(\pi)$ denote the mutual information when the input distribution is π , which is a concave function in π (Theorem 5.5). Then

$$I(\theta; X_{n+1} | X^n) = \mathbb{E}_{X^n}[I(P_{\theta|X^n})] \leq \mathbb{E}_{X^{n-1}}[I(P_{\theta|X^{n-1}})] = I(\theta; X_n | X^{n-1})$$

where the inequality follows from Jensen’s inequality, since $P_{\theta|X^{n-1}}$ is a mixture of $P_{\theta|X^n}$. \square

Lemma 32.3 allows us to prove the converse bound (32.9): Fix any prior π . Since the minimax risk dominates any Bayes risk (Theorem 28.4), in view of Lemma 32.2, we have

$$\sum_{t=0}^n R_{\text{KL}}^*(t) \geq \sum_{t=0}^n I(\theta; X_{t+1} | X^t) = I(\theta; X^{n+1}).$$

Recall from (32.5) that $C_{n+1} = \sup_{\pi \in \Delta(\Theta)} I(\theta; X^{n+1})$. Optimizing over the prior π yields (32.9).

Now suppose that the minimax theorem holds for (32.4), so that $R_{\text{KL}}^* = \sup_{\pi \in \Delta(\Theta)} R_{\text{KL}, \text{Bayes}}^*(\pi)$. Then Lemma 32.2 allows us to express the minimax risk as the conditional mutual information maximized over the prior π :

$$R_{\text{KL}}^*(n) = \sup_{\pi \in \Delta(\Theta)} I(\theta; X_{n+1} | X^n).$$

Thus Lemma 32.3 implies the desired

$$R_{\text{KL}}^*(n) \leq \frac{1}{n+1} C_{n+1}.$$

Next, we prove this directly without going through the Bayesian route or assuming the minimax theorem. The main idea, due to Yang and Barron [329], is to consider Bayes estimators (of the form (32.12)) but analyze it in the *worst case*. Fix an arbitrary joint distribution $Q_{X^{n+1}}$ on \mathcal{X}^{n+1} , which factorizes as $Q_{X^{n+1}} = \prod_{i=1}^{n+1} Q_{X_i | X^{i-1}}$. (This joint distribution is an auxiliary object used only for constructing an estimator.) For each i , the conditional distribution $Q_{X_i | X^{i-1}}$ defines an estimator taking the sample X^i of size i as the input. Taking their Cesàro mean results in the following estimator operating on the full sample X^n :

$$\hat{P}(\cdot | X^n) \triangleq \frac{1}{n+1} \sum_{i=1}^{n+1} Q_{X_i | X^{i-1}}. \quad (32.14)$$

Let us bound the worst-case KL risk of this estimator. Fix $\theta \in \Theta$ and let X^{n+1} be drawn independently from P_θ so that $P_{X^{n+1}} = P_\theta^{\otimes(n+1)}$. Taking expectations with this law, we have

$$\begin{aligned} \mathbb{E}_\theta[D(P_\theta \parallel \hat{P}(\cdot|X^n))] &= \mathbb{E}\left[D\left(P_\theta \middle\| \frac{1}{n+1} \sum_{i=1}^{n+1} Q_{X_i|X^{i-1}}\right)\right] \\ &\stackrel{(a)}{\leq} \frac{1}{n+1} \sum_{i=1}^{n+1} D(P_\theta \parallel Q_{X_i|X^{i-1}}|P_{X^{i-1}}) \\ &\stackrel{(b)}{=} \frac{1}{n+1} D(P_\theta^{\otimes(n+1)} \parallel Q_{X^{n+1}}), \end{aligned}$$

where (a) and (b) follows from the convexity (Theorem 5.1) and the chain rule for KL divergence (Theorem 2.15(c)). Taking the supremum over $\theta \in \Theta$ bounds the worst-case risk as

$$R_{\text{KL}}^*(n) \leq \frac{1}{n+1} \sup_{\theta \in \Theta} D(P_\theta^{\otimes(n+1)} \parallel Q_{X^{n+1}}).$$

Optimizing over the choice of $Q_{X^{n+1}}$, we obtain

$$R_{\text{KL}}^*(n) \leq \frac{1}{n+1} \inf_{Q_{X^{n+1}}} \sup_{\theta \in \Theta} D(P_\theta^{\otimes(n+1)} \parallel Q_{X^{n+1}}) = \frac{C_{n+1}}{n+1},$$

where the last identity applies Theorem 5.8 of Kemperman, completing the proof of (32.7). Furthermore, Theorem 5.8 asserts that the optimal $Q_{X^{n+1}}$ exists and given uniquely by the capacity-achieving output distribution $P_{X^{n+1}}^*$. Thus the above minimax upper bound can be attained by taking the Cesàro average of $P_{X_1}^*, P_{X_2|X_1}^*, \dots, P_{X_{n+1}|X^n}^*$, namely,

$$\hat{P}^*(\cdot|X^n) = \frac{1}{n+1} \sum_{i=1}^{n+1} P_{X_i|X^{i-1}}^*. \quad (32.15)$$

Note that in general this is an *improper* estimate as it steps outside the class \mathcal{P} .

In the special case where the capacity-achieving input distribution π^* exists, the capacity-achieving output distribution can be expressed as a mixture over product distributions as $P_{X^{n+1}}^* = \int \pi^*(d\theta) P_\theta^{\otimes(n+1)}$. Thus the estimator $\hat{P}^*(\cdot|X^n)$ is in fact the *average of Bayes estimators* (32.12) under prior π^* for sample sizes ranging from 0 to n .

Finally, as will be made clear in the next section, in order to achieve the further upper bound (32.8) in terms of the KL covering numbers, namely $R_{\text{KL}}^*(n) \leq \epsilon^2 + \frac{1}{n+1} \log N_{\text{KL}}(\mathcal{P}, \epsilon)$, it suffices to choose the following $Q_{X^{n+1}}$ as opposed to the exact capacity-achieving output distribution: Pick an ϵ -KL cover Q_1, \dots, Q_N for \mathcal{P} of size $N = N_{\text{KL}}(\mathcal{P}, \epsilon)$ and choose π to be the uniform distribution and define $Q_{X^{n+1}} = \frac{1}{N} \sum_{j=1}^N Q_j^{\otimes(n+1)}$ – this was the original construction in [329]. In this case, applying the Bayes rule (32.12), we see that the estimator is in fact a convex combination $\hat{P}(\cdot|X^n) = \sum_{j=1}^N w_j Q_j$ of the centers Q_1, \dots, Q_N , with data-driven weights given by

$$w_j = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{\prod_{t=1}^{i-1} Q_j(X_t)}{\sum_{j=1}^N \prod_{t=1}^{i-1} Q_j(X_t)}.$$

32.1 Yang-Barron's construction 529

Again, except for the extraordinary case where \mathcal{P} is convex and the centers Q_j belong to \mathcal{P} , the estimate $\hat{P}(\cdot|X^n)$ is improper.

32.1.2 Capacity upper bound via KL covering numbers

As explained earlier, finding the capacity C_n requires solving the difficult optimization problem in (32.5). In this subsection we prove (32.8) which bounds this capacity by metric entropy. Conceptually speaking, both metric entropy and capacity measure the complexity of a model class. The following result, which applies to a more general setting than (32.5), makes precise their relations.

Theorem 32.4. *Let $\mathcal{Q} = \{P_{B|A=a} : a \in \mathcal{A}\}$ be a collection of distributions on some space \mathcal{B} and denote the capacity $C = \sup_{P_A \in \Delta(\mathcal{A})} I(A; B)$. Then*

$$C = \inf_{\epsilon > 0} \{\epsilon^2 + \log N_{\text{KL}}(\mathcal{Q}, \epsilon)\}, \quad (32.16)$$

where N_{KL} is the KL covering number defined in (32.6).

Proof. Fix ϵ and let $N = N_{\text{KL}}(\mathcal{Q}, \epsilon)$. Then there exist Q_1, \dots, Q_N that form an ϵ -KL cover, such that for any $a \in \mathcal{A}$ there exists $i(a) \in [N]$ such that $D(P_{B|A=a} \| Q_{i(a)}) \leq \epsilon^2$. Fix any P_A . Then

$$\begin{aligned} I(A; B) &= I(A, i(A); B) = I(i(A); B) + I(A; B|i(A)) \\ &\leq H(i(A)) + I(A; B|i(A)) \leq \log N + \epsilon^2. \end{aligned}$$

where the last inequality follows from that $i(A)$ takes at most N values and, by applying Theorem 4.1,

$$I(A; B|i(A)) \leq D(P_{B|A} \| Q_{i(A)} | P_{i(A)}) \leq \epsilon^2.$$

For the lower bound, note that if $C = \infty$, then in view of the upper bound above, $N_{\text{KL}}(\mathcal{Q}, \epsilon) = \infty$ for any ϵ and (32.16) holds with equality. If $C < \infty$, Theorem 5.8 shows that C is the KL radius of \mathcal{Q} , namely, there exists P_B^* , such that $C = \sup_{P_A \in \Delta(\mathcal{A})} D(P_{B|A} \| P_B^* | P_A) = \sup_{x \in \mathcal{A}} D(P_{B|A} \| P_B^* | P_A)$. In other words, $N_{\text{KL}}(\mathcal{Q}, \sqrt{C + \delta}) = 1$ for any $\delta > 0$. Sending $\delta \rightarrow 0$ proves the equality of (32.16). \square

Next we specialize Theorem 32.4 to our statistical setting (32.5) where the input A is θ and the output B is $X^n \stackrel{\text{i.i.d.}}{\sim} P_\theta$. Recall that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $\mathcal{P}_n \triangleq \{P_\theta^{\otimes n} : \theta \in \Theta\}$. By tensorization of KL divergence (Theorem 2.15(d)), $D(P_\theta^{\otimes n} \| P_{\theta'}^{\otimes n}) = nD(P_\theta \| P_{\theta'})$. Thus

$$N_{\text{KL}}(\mathcal{P}_n, \epsilon) \leq N_{\text{KL}}\left(\mathcal{P}, \frac{\epsilon}{\sqrt{n}}\right).$$

Combining this with Theorem 32.4, we obtain the following upper bound on the capacity C_n in terms of the KL metric entropy of the (single-letter) family \mathcal{P} :

$$C_n \leq \inf_{\epsilon > 0} \{n\epsilon^2 + \log N_{\text{KL}}(\mathcal{P}, \epsilon)\}. \quad (32.17)$$

This proves (32.8), completing the proof of Theorem 32.1.

32.1.3 Capacity lower bound via Hellinger packing number

Recall that in order to deduce from (32.9) concrete lower bound on the minimax KL risk, such as (32.11), one needs to have matching upper and lower bounds on the capacity C_n . Although Theorem 32.4 characterizes capacity in terms of the KL covering numbers, lower bounding the latter is not easy. On the other hand, it is much easier to lower bound the packing number (since any explicit packing works). One may attempt to use Theorem 27.3 to relate packing and covering, but, alas, KL divergence is not a distance. Thus, it is much easier to use the following method of reducing to Hellinger distance.

Theorem 32.5. *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $M_H(\epsilon) \equiv M(\mathcal{P}, H, \epsilon)$ the Hellinger packing number of the set \mathcal{P} , cf. (27.2). Then C_n defined in (32.5) satisfies*

$$C_n \geq \min \left(\frac{\log e}{2} n \epsilon^2, \log M_H(\epsilon) \right) - \log 2 \quad (32.18)$$

Proof. The idea of the proof is simple. Given a packing $\theta_1, \dots, \theta_M \in \Theta$ with pairwise distances $H^2(Q_i, Q_j) \geq \epsilon^2$ for $i \neq j$, where $Q_i \equiv P_{\theta_i}$, we know that one can test $Q_i^{\otimes n}$ vs $Q_j^{\otimes n}$ with error $e^{-\frac{n\epsilon^2}{2}}$, cf. Theorem 7.14 and Theorem 32.8. Then by the union bound, if $Me^{-\frac{n\epsilon^2}{2}} < \frac{1}{2}$, we can distinguish these M hypotheses with error $< \frac{1}{2}$. Let $\theta \sim \text{Unif}(\theta_1, \dots, \theta_M)$. Then from Fano's inequality we get $I(\theta; X^n) \gtrsim \log M$.

To get sharper constants, though, we will proceed via the inequality shown in Ex. I.47. In the notation of that exercise we take $\lambda = 1/2$ and from Definition 7.34 we get that

$$D_{1/2}(Q_i, Q_j) = -2 \log \left(1 - \frac{1}{2} H^2(Q_i, Q_j) \right) \geq H^2(Q_i, Q_j) \log e \geq \epsilon^2 \log e \quad i \neq j.$$

By the tensorization property (7.73) for Rényi divergence, $D_{1/2}(Q_i^{\otimes n}, Q_j^{\otimes n}) = n D_{1/2}(Q_i, Q_j)$ and we get by Ex. I.47

$$\begin{aligned} I(\theta; X^n) &\geq - \sum_{i=1}^M \frac{1}{M} \log \left(\sum_{j=1}^M \frac{1}{M} \exp \left\{ -\frac{n}{2} D_{1/2}(Q_i, Q_j) \right\} \right) \\ &\stackrel{(a)}{\geq} - \sum_{i=1}^M \frac{1}{M} \log \left(\frac{M-1}{M} e^{-\frac{n\epsilon^2}{2}} + \frac{1}{M} \right) \\ &\geq - \sum_{i=1}^M \frac{1}{M} \log \left(e^{-\frac{n\epsilon^2}{2}} + \frac{1}{M} \right) = - \log \left(e^{-\frac{n\epsilon^2}{2}} + \frac{1}{M} \right), \end{aligned}$$

where in (a) we used the fact that pairwise distances are all $\geq n\epsilon^2$ except when $i = j$. Finally, since $\frac{1}{A} + \frac{1}{B} \leq \frac{2}{\min(A, B)}$ we conclude the result. \square

We note that since $D \gtrsim H^2$ (cf. (7.30)), a different (weaker) lower bound on the KL risk also follows from Section 32.2.4 below.

32.1.4 General bounds between cumulative and individual (one-step) risks

In summary, we can see that the beauty of the Yang-Barron method lies in two ideas:

- Instead of directly studying the risk $R_{KL}^*(n)$, (32.7) relates it to a cumulative risk C_n
- The cumulative risk turns out to be equal to a capacity, which can be conveniently bounded in terms of covering numbers.

In this subsection we want to point out that while the second step is very special to KL (log-loss), the first idea is generic. Namely, we have the following result.

Proposition 32.6. *Fix a loss function $\ell : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \bar{\mathbb{R}}$ and a class Π of distributions on \mathcal{X} . Define cumulative and one-step minimax risks as follows:*

$$C_n = \inf_{\{\hat{P}_t(\cdot)\}} \sup_{P \in \Pi} \mathbb{E} \left[\sum_{t=1}^n \ell(P, \hat{P}_t(X^{t-1})) \right] \quad (32.19)$$

$$R_n^* = \inf_{\hat{P}(\cdot)} \sup_{P \in \Pi} \mathbb{E} \left[\ell(P, \hat{P}(X^n)) \right] \quad (32.20)$$

where both infima are over measurable (possibly randomized) estimators $\hat{P}_t : \mathcal{X}^{t-1} \rightarrow \mathcal{P}(\mathcal{X})$, and the expectations are over $X_i \stackrel{i.i.d.}{\sim} P$ and the randomness of the estimators. Then we have

$$nR_{n-1}^* \leq C_n \leq \sum_{t=0}^{n-1} R_t^*. \quad (32.21)$$

Thus, if the sequence $\{R_n^*\}$ satisfies $R_n^* \asymp \frac{1}{n} \sum_{t=0}^{n-1} R_t^*$ then $C_n \asymp nR_n^*$. Conversely, if $n^{\alpha_-} \lesssim C_n \lesssim n^{\alpha_+}$ for all n and some $\alpha_+ \geq \alpha_- > 0$, then

$$n^{(\alpha_- - 1) \frac{\alpha_+}{\alpha_-}} \lesssim R_n^* \lesssim n^{\alpha_+ - 1}. \quad (32.22)$$

Remark 32.1. The meaning of the above is that $R_n^* \approx \frac{1}{n} C_n$ within either constant or polylogarithmic factors, for most cases of interest.

Proof. To show the first inequality in (32.21), given predictors $\{\hat{P}_t(X^{t-1}) : t \in [n]\}$ for C_n , consider a randomized predictor $\hat{P}(X^{n-1})$ for R_{n-1}^* that equals each of the $\hat{P}_t(X^{t-1})$ with equal probability. The second inequality follows from interchanging \sup_P and \sum_t .

To derive (32.22) notice that the upper bound on R_n^* follows from (32.21). For the lower bound, notice that the sequence R_n^* is monotone and hence we have for any $n < m$

$$C_m \leq \sum_{t=0}^{m-1} R_t^* \leq \sum_{t=0}^{n-1} \frac{C_t}{t} + (m-n)R_n^*. \quad (32.23)$$

Setting $m = an^{\frac{\alpha_+}{\alpha_-}}$ with some appropriate constant a yields the lower bound. \square

32.2 Pairwise comparison à la Le Cam-Birgé

When we proved the lower bound in Theorem 31.4, we applied the reasoning that if an ϵ -packing of the parameter space Θ cannot be tested, then $\theta \in \Theta$ cannot be estimated more than precision ϵ , thereby establishing a minimax *lower bound* in terms of the KL metric entropy. Conversely, we can ask the following question:

Is it possible to construct an estimator based on tests, and produce a minimax *upper bound* in terms of the metric entropy?

For Hellinger loss, the answer is yes, although the metric entropy involved is with respect to the Hellinger distance not KL divergence. The basic construction is due to Le Cam and further developed by Birgé. The main idea is as follows: Fix an ϵ -covering $\{P_1, \dots, P_N\}$ of the set of distributions \mathcal{P} . Given n samples drawn from $P \in \mathcal{P}$, let us test which ball P belongs to; this allows us to estimate P up to Hellinger loss ϵ . This can be realized by a pairwise comparison argument of testing the (composite) hypothesis $P \in B(P_i, \epsilon)$ versus $P \in B(P_j, \epsilon)$. This program can be further refined to involve on the local entropy of the model.

32.2.1 Composite hypothesis testing and Hellinger distance

Recall the problem of composite hypothesis testing introduced in Section 16.4. Let \mathcal{P} and \mathcal{Q} be two (*not* necessarily convex) classes of distributions. Given iid samples X_1, \dots, X_n drawn from some distribution P , we want to test, according some decision rule $\phi = \phi(X_1, \dots, X_n) \in \{0, 1\}$, whether $P \in \mathcal{P}$ (indicated by $\phi = 0$) or $P \in \mathcal{Q}$ (indicated by $\phi = 1$). By the minimax theorem, the optimal error is given by the total variation between the worst-case mixtures:

$$\min_{\phi} \left\{ \sup_{P \in \mathcal{P}} P(\phi = 1) + \sup_{Q \in \mathcal{Q}} Q(\phi = 0) \right\} = 1 - \text{TV}(\text{co}(\mathcal{P}^{\otimes n}), \text{co}(\mathcal{Q}^{\otimes n})), \quad (32.24)$$

wherein the notations are explained as follows:

- $\mathcal{P}^{\otimes n} \triangleq \{P^{\otimes n} : P \in \mathcal{P}\}$ consists of all n -fold products of distributions in \mathcal{P} ;
- $\text{co}(\cdot)$ denotes the convex hull, that is, the set of all mixtures. For example, for a parametric family, $\text{co}(\{P_\theta : \theta \in \Theta\}) = \{P_\pi : \pi \in \Delta(\Theta)\}$, where $P_\pi = \int P_\theta \pi(d\theta)$ is the mixture under the mixing distribution π , and $\Delta(\Theta)$ denotes the collection of all probability distributions (priors) on Θ .

The optimal test that achieves (32.24) is the likelihood ratio given by the worst-case mixtures, that is, the closest⁴ pair of mixture (P_n^*, Q_n^*) such that $\text{TV}(P_n^*, Q_n^*) = \text{TV}(\text{co}(\mathcal{P}^{\otimes n}), \text{co}(\mathcal{Q}^{\otimes n}))$.

The exact result (32.24) is unwieldy as the RHS involves finding the least favorable priors over the n -fold product space. However, there are several known examples where much simpler and

⁴ In case the closest pair does not exist, we can replace it by an infimizing sequence.

32.2 Pairwise comparison à la Le Cam-Birgé 533

explicit results are available. In the case when \mathcal{P} and \mathcal{Q} are TV-balls around P_0 and Q_0 , Huber [157] showed that the minimax optimal test has the form

$$\phi(x^n) = 1 \left\{ \sum_{i=1}^n \min(c', \max(c'', \log \frac{dP_0}{dQ_0}(X_i))) > t \right\}.$$

(See also Ex. III.20.) However, there are few other examples where minimax optimal tests are known explicitly. Fortunately, as was shown by Le Cam, there is a general “single-letter” upper bound in terms of the Hellinger separation between \mathcal{P} and \mathcal{Q} . It is the consequence of the more general tensorization property of Rényi divergence in Proposition 7.35 (of which Hellinger is a special case).

Theorem 32.7.

$$\min_{\phi} \left\{ \sup_{P \in \mathcal{P}} P(\phi = 1) + \sup_{Q \in \mathcal{Q}} Q(\phi = 0) \right\} \leq e^{-\frac{n}{2} \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} H^2(P, Q)}, \quad (32.25)$$

Remark 32.2. For the case when \mathcal{P} and \mathcal{Q} are Hellinger balls of radius r around P_0 and Q_0 , respectively, Birgé [35] constructed an explicit test. Namely, under the assumption $H(P_0, Q_0) > 2.01r$, there is a test $\phi(x^n) = 1 \left\{ \sum_{i=1}^n \log \frac{\alpha + \beta \psi(X_i)}{\beta + \alpha \psi(X_i)} > t \right\}$ attaining error $e^{-n\Omega(r^2)}$, where $\psi(x) = \sqrt{\frac{dP_0}{dQ_0}(x)}$ and $\alpha, \beta > 0$ depend only on $H(P_0, Q_0)$.

Proof. We start by restating the special case of Proposition 7.35:

$$1 - \frac{1}{2} H^2 \left(\text{co} \left(\bigotimes_{i=1}^n \mathcal{P}_i \right), \text{co} \left(\bigotimes_{i=1}^n \mathcal{Q}_i \right) \right) \leq \prod_{i=1}^n \left(1 - \frac{1}{2} H^2(\text{co}(\mathcal{P}_i), \text{co}(\mathcal{Q}_i)) \right). \quad (32.26)$$

The From (32.24) we get

$$\begin{aligned} 1 - \text{TV}(\text{co}(\mathcal{P}^{\otimes n}), \text{co}(\mathcal{Q}^{\otimes n})) &\stackrel{(a)}{\leq} 1 - \frac{1}{2} H^2(\text{co}(\mathcal{P}^{\otimes n}), \text{co}(\mathcal{Q}^{\otimes n})) \\ &\stackrel{(b)}{\leq} \left(1 - \frac{1}{2} H^2(\text{co}(\mathcal{P}), \text{co}(\mathcal{Q})) \right)^n \leq \exp \left(-\frac{n}{2} H^2(\text{co}(\mathcal{P}), \text{co}(\mathcal{Q})) \right) \end{aligned}$$

where (a) follows from (7.20); (b) follows from (32.26). \square

In the sequel we will apply Theorem 32.8 to two disjoint Hellinger balls (both are convex).

32.2.2 Hellinger guarantee on Le Cam-Birgé’s pairwise comparison estimator

The idea of constructing estimator based on pairwise tests is due to Le Cam ([189], see also [309, Section 10]) and Birgé [34]. We are given n i.i.d. observations X_1, \dots, X_n generated from P , where $P \in \mathcal{P}$ is the distribution to be estimated. Here let us emphasize that \mathcal{P} need *not* be a convex set. Let the loss function between the true distribution P and the estimated distribution \hat{P} be their squared Hellinger distance, i.e.

$$\ell(P, \hat{P}) = H^2(P, \hat{P}).$$

Then, we have the following result:

Theorem 32.8 (Le Cam-Birgé). *Denote by $N_H(\mathcal{P}, \epsilon)$ the ϵ -covering number of the set \mathcal{P} under the Hellinger distance (cf. (27.1)). Let ϵ_n be such that*

$$n\epsilon_n^2 \geq \log N_H(\mathcal{P}, \epsilon_n) \vee 1.$$

Then there exists an estimator $\hat{P} = \hat{P}(X_1, \dots, X_n)$ taking values in \mathcal{P} such that for any $t \geq 1$,

$$\sup_{P \in \mathcal{P}} P[H(P, \hat{P}) > 4t\epsilon_n] \lesssim e^{-t^2} \quad (32.27)$$

and, consequently,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \lesssim \epsilon_n^2 \quad (32.28)$$

Proof of Theorem 32.10. It suffices to prove the high-probability bound (32.27). Abbreviate $\epsilon = \epsilon_n$ and $N = N_H(\mathcal{P}, \epsilon_n)$. Let P_1, \dots, P_N be a maximal ϵ -packing of \mathcal{P} under the Hellinger distance, which also serves as an ϵ -covering (cf. Theorem 27.3). Thus, $\forall i \neq j$,

$$H(P_i, P_j) \geq \epsilon,$$

and for $\forall P \in \mathcal{P}, \exists i \in [N]$, s.t.

$$H(P, P_i) \leq \epsilon,$$

Denote $B(P, \epsilon) = \{Q : H(P, Q) \leq \epsilon\}$ denote the ϵ -Hellinger ball centered at P . Crucially, *Hellinger ball is convex*⁵ thanks to the convexity of squared Hellinger distance as an f -divergence (cf. Theorem 7.8). Indeed, for any $P', P'' \in B(P, \epsilon)$ and $\alpha \in [0, 1]$,

$$H^2(\bar{\alpha}P' + \alpha P'', P) \leq \bar{\alpha}H^2(P', P) + \alpha H^2(P'', P) \leq \epsilon^2.$$

Next, consider the following *pairwise comparison problem*, where we test two Hellinger balls (composite hypothesis) against each other:

$$\begin{cases} H_i : P \in B(P_i, \epsilon) \\ H_j : P \in B(P_j, \epsilon) \end{cases}$$

for all $i \neq j$, s.t. $H(P_i, P_j) \geq \delta = 4\epsilon$.

Since both $B(P_i, \epsilon)$ and $B(P_j, \epsilon)$ are convex, applying Theorem 32.8 yields a test $\psi_{ij} = \psi_{ij}(X_1, \dots, X_n)$, with $\psi_{ij} = 0$ corresponding to declaring $P \in B(P_i, \epsilon)$, and $\psi_{ij} = 1$ corresponding to declaring $P \in B(P_j, \epsilon)$, such that $\psi_{ij} = 1 - \psi_{ji}$ and the following large deviation bound holds: for all i, j , s.t. $H(P_i, P_j) \geq \delta$,

$$\sup_{P \in B(P_i, \epsilon)} P(\psi_{ij} = 1) \leq e^{-\frac{n}{8}H(P_i, P_j)^2}, \quad (32.29)$$

⁵ Note that this is not entirely obvious because $P \mapsto H(P, Q)$ is not convex (for example, consider $p \mapsto H(\text{Ber}(p), \text{Ber}(0.1))$).

32.2 Pairwise comparison à la Le Cam-Birgé 535

where we used the triangle inequality of Hellinger distance: for any $P \in B(P_i, \epsilon)$ and any $Q \in B(P_j, \epsilon)$,

$$H(P, Q) \geq H(P_i, P_j) - 2\epsilon \geq H(P_i, P_j)/2 \geq 2\epsilon.$$

For $i \in [N]$, define the random variable

$$T_i \triangleq \begin{cases} \max_{j \in [N]} H^2(P_i, P_j) & \text{s.t. } \psi_{ij} = 1, \quad H(P_i, P_j) > \delta; \\ 0, & \text{no such } j \text{ exists.} \end{cases}$$

Basically, T_i records the maximum distance from P_i to those P_j outside the δ -neighborhood of P_i that is confusable with P_i given the present sample. Our density estimator is defined as

$$\hat{P} = P_{i^*}, \quad \text{where } i^* \in \operatorname{argmin}_{i \in [N]} T_i. \quad (32.30)$$

Now for the proof of correctness, assume that $P \in B(P_1, \epsilon)$. The intuition is that, we should expect, typically, that $T_1 = 0$, and furthermore, $T_j \geq \delta^2$ for all j such that $H(P_1, P_j) \geq \delta$. Note that by the definition of T_i and the symmetry of the Hellinger distance, for any pair i, j such that $H(P_i, P_j) \geq \delta$, we have

$$\max\{T_i, T_j\} \geq H(P_i, P_j).$$

Consequently,

$$\begin{aligned} H(\hat{P}, P_1) \mathbf{1}_{\{H(\hat{P}, P_1) \geq \delta\}} &= H(P_{i_*}, P_1) \mathbf{1}_{\{H(P_{i_*}, P_1) \geq \delta\}} \\ &\leq \max\{T_{i_*}, T_1\} \mathbf{1}_{\{\max\{T_{i_*}, T_1\} \geq \delta\}} = T_1 \mathbf{1}_{\{T_1 \geq \delta\}}, \end{aligned}$$

where the last equality follows from the definition of i_* as a global minimizer in (32.30). Thus, for any $t \geq 1$,

$$\begin{aligned} P[H(\hat{P}, P_1) \geq t\delta] &\leq P[T_1 \geq t\delta] \\ &\leq N(\epsilon) e^{-2n\epsilon^2 t^2} \end{aligned} \quad (32.31)$$

$$\lesssim e^{-t^2}, \quad (32.32)$$

where (32.31) follows from (32.29) and (32.32) uses the assumption that $n\epsilon^2 \geq 1$ and $N \leq e^{n\epsilon^2}$.

□

32.2.3 Refinement using local entropy

Just like Theorem 32.1, while they are often tight for nonparametric problems with superlogarithmically metric entropy, for finite-dimensional models a direct application of Theorem 32.10 results in a slack by a log factor. For example, for a d -dimensional parametric family, e.g., the Gaussian location model or its finite mixtures, the metric entropy usually behaves as $\log N_H(\epsilon) \asymp d \log \frac{1}{\epsilon}$. Thus when $n \gtrsim d$, Theorem 32.10 entails choosing $\epsilon_n^2 \asymp \frac{d}{n} \log \frac{n}{d}$, which falls short of the parametric rate $\mathbb{E}[H^2(\hat{P}, P)] \lesssim \frac{d}{n}$ which are typically achievable.

As usual, such a log factor can be removed using the local entropy argument. To this end, define the local Hellinger entropy:

$$N_{\text{loc}}(\mathcal{P}, \epsilon) \triangleq \sup_{P \in \mathcal{P}} \sup_{\eta \geq \epsilon} N_H(B(P, \eta) \cap \mathcal{P}, \eta/2). \quad (32.33)$$

Theorem 32.9 (Le Cam-Birgé: local entropy version). *Let ϵ_n be such that*

$$n\epsilon_n^2 \geq \log N_{\text{loc}}(\mathcal{P}, \epsilon_n) \vee 1. \quad (32.34)$$

Then there exists an estimator $\hat{P} = \hat{P}(X_1, \dots, X_n)$ taking values in \mathcal{P} such that for any $t \geq 2$,

$$\sup_{P \in \mathcal{P}} P[H(P, \hat{P}) > 4t\epsilon_n] \leq e^{-t^2} \quad (32.35)$$

and hence

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \lesssim \epsilon_n^2 \quad (32.36)$$

Remark 32.3 (Doubling dimension). Suppose that for some $d > 0$, $\log N_{\text{loc}}(\mathcal{P}, \epsilon) \leq d \log \frac{1}{\epsilon}$ holds for all sufficiently large small ϵ ; this is the case for finite-dimensional models where the Hellinger distance is comparable with the vector norm by the usual volume argument (Theorem 27.4). Then we say the *doubling dimension* (also known as the *Le Cam dimension* [309]) of \mathcal{P} is at most d ; this terminology comes from the fact that the local entropy concerns covering Hellinger balls using balls of half the radius. Then Theorem 32.11 shows that it is possible to achieve the “parametric rate” $O(\frac{d}{n})$. In this sense, the doubling dimension serves as the effective dimension of the model \mathcal{P} .

Lemma 32.10. *For any $P \in \mathcal{P}$ and $\eta \geq \epsilon$ and $k \geq \mathbb{Z}_+$,*

$$N_H(B(P, 2^k \eta) \cap \mathcal{P}, \eta/2) \leq N_{\text{loc}}(\mathcal{P}, \epsilon)^k \quad (32.37)$$

Proof. We proceed by induction on k . The base case of $k = 0$ follows from the definition (32.33). For $k \geq 1$, assume that (32.37) holds for $k - 1$ for all $P \in \mathcal{P}$. To prove it for k , we construct a cover of $B(P, 2^k \eta) \cap \mathcal{P}$ as follows: first cover it with $2^{k-1} \eta$ -balls, then cover each ball with $\eta/2$ -balls. By the induction hypothesis, the total number of balls is at most

$$N_H(B(P, 2^k \eta) \cap \mathcal{P}, 2^{k-1} \eta) \cdot \sup_{P' \in \mathcal{P}} N_H(B(P', 2^{k-1} \eta) \cap \mathcal{P}, \eta/2) \leq N_{\text{loc}}(\epsilon) \cdot N_{\text{loc}}(\epsilon)^{k-1}$$

completing the proof. \square

We now prove Theorem 32.11:

Proof. We analyze the same estimator (32.30) following the proof of Theorem 32.10, except that the estimate (32.31) is improved as follows: Define the Hellinger shell $A_k \triangleq \{P : 2^k \delta \leq$

32.2 Pairwise comparison à la Le Cam-Birgé 537

$H(P_1, P) < 2^{k+1}\delta$ and $G_k \triangleq \{P_1, \dots, P_N\} \cap A_k$. Recall that $\delta = 4\epsilon$. Given $t \geq 2$, let $\ell = \lfloor \log_2 t \rfloor$ so that $2^\ell \leq t < 2^{\ell+1}$. Then

$$\begin{aligned} P[T_1 \geq t\delta] &\leq \sum_{k \geq \ell} P[2^k\delta \leq T_1 < 2^{k+1}\delta] \\ &\stackrel{(a)}{\leq} \sum_{k \geq \ell} |G_k| e^{-\frac{n}{8}(2^k\delta)^2} \\ &\stackrel{(b)}{\leq} \sum_{k \geq \ell} N_{\text{loc}}(\epsilon)^{k+3} e^{-2n\epsilon^2 4^k} \\ &\stackrel{(c)}{\lesssim} e^{-4^\ell} \leq e^{-t^2} \end{aligned}$$

where (a) follows from from (32.29); (c) follows from the assumption that $\log N_{\text{loc}} \leq n\epsilon^2$ and $k \geq \ell \geq \log_2 t \geq 1$; (b) follows from the following reasoning: since $\{P_1, \dots, P_N\}$ is an ϵ -packing, we have

$$|G_k| \leq M(A_k, \epsilon) \leq N(A_k, \epsilon/2) \leq N(B(P_1, 2^{k+1}\delta) \cap \mathcal{P}, \epsilon/2) \leq N_{\text{loc}}(\epsilon)^{k+3}$$

where the first and the last inequalities follow from Theorem 27.3 and Lemma 32.13 respectively. \square

As an application of Theorem 32.11, we show that parametric rate (namely, dimension divided by the sample size) is achievable for models with locally quadratic behavior, such as those smooth parametric models (cf. Section 7.11 and in particular Theorem 7.33).

Corollary 32.11. *Consider a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ and \mathcal{P} is totally bounded in Hellinger distance. Suppose that there exists a norm $\|\cdot\|$ and constants t_0, c, C such that for all $\theta_0, \theta_1 \in \Theta$ with $\|\theta_0 - \theta_1\| \leq t_0$,*

$$c\|\theta_0 - \theta_1\| \leq H(P_{\theta_0}, P_{\theta_1}) \leq C\|\theta_0 - \theta_1\|. \quad (32.38)$$

Then there exists an estimator $\hat{\theta}$ based on $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_\theta$, such that

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta[H^2(P_\theta, P_{\hat{\theta}})] \lesssim \frac{d}{n}.$$

Proof. It suffices to bound the local entropy $N_{\text{loc}}(\mathcal{P}, \epsilon)$ in (32.33). Fix $\theta_0 \in \Theta$. Indeed, for any $\eta > t_0$, we have $N_H(B(P_{\theta_0}, \eta) \cap \mathcal{P}, \eta/2) \leq N_H(\mathcal{P}, t_0) \lesssim 1$. For $\epsilon \leq \eta \leq t_0$,

$$\begin{aligned} N_H(B(P_{\theta_0}, \eta) \cap \mathcal{P}, \eta/2) &\stackrel{(a)}{\leq} N_{\|\cdot\|}(B_{\|\cdot\|}(\theta_0, \eta/c), \eta/(2C)) \\ &\stackrel{(b)}{\leq} \frac{\text{vol}(B_{\|\cdot\|}(\theta_0, \eta/c + \eta/(2C)))}{\text{vol}(B_{\|\cdot\|}(\theta_0, \eta/(2C)))} = \left(1 + \frac{2C}{c}\right)^d \end{aligned}$$

where (a) and (b) follow from (32.38) and Theorem 27.4 respectively. This shows that $\log N_{\text{loc}}(\mathcal{P}, \epsilon) \lesssim d$, completing the proof by applying Theorem 32.11. \square

32.2.4 Lower bound using local Hellinger packing

It turns out that under certain regularity assumptions we can prove an almost matching lower bound (typically within a logarithmic term) on the Hellinger-risk. First we define the local packing number as follows:

$$M_{\text{loc}}(\epsilon) \equiv M_{\text{loc}}(\mathcal{P}, H, \epsilon) = \max \{M : \exists R, P_1, \dots, P_M \in \mathcal{P} : H(P_i, R) \leq \epsilon, H(P_i, P_j) \geq \frac{\epsilon}{2} \quad \forall i \neq j\}.$$

Note that unlike the definition of N_{loc} in (32.33) we are not taking the supremum over the scale $\eta \geq \epsilon$. For this reason, we cannot generally apply Theorem 27.3 to conclude that $N_{\text{loc}}(\epsilon) \geq M_{\text{loc}}(\epsilon)$. In all instances known to us we have $\log N_{\text{loc}} \asymp \log M_{\text{loc}}$, in which case the following general result provides a minimax lower bound that matches the upper bound in Theorem 32.11 up to logarithmic factors.

Theorem 32.12. *Suppose that the family \mathcal{P} has a finite D_λ radius for some $\lambda > 1$, i.e.*

$$R_\lambda(\mathcal{P}) \triangleq \inf_U \sup_{P \in \mathcal{P}} D_\lambda(P \| U) < \infty, \quad (32.39)$$

where D_λ is the Rényi divergence of order λ (see Definition 7.34). There exists constants $c = c(\lambda)$ and $\epsilon < \epsilon_0(\lambda)$ such that whenever n and $\epsilon < \epsilon_0$ are such that

$$c(\lambda)n\epsilon^2 \left(\log \frac{1}{\epsilon^2} + R_\lambda(\mathcal{P}) \right) + 2\log 2 < \log M_{\text{loc}}(\epsilon), \quad (32.40)$$

any estimator $\hat{P} = \hat{P}(\cdot; X^n)$ must satisfy

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \geq \frac{\epsilon^2}{32},$$

where \mathbb{E}_P is taken with respect to $X^n \stackrel{i.i.d.}{\sim} P$.

Remark 32.4. When $\log M_{\text{loc}}(\epsilon) \asymp \epsilon^{-p}$, a minimax lower bound for the squared Hellinger risk on the order of $(n \log n)^{-\frac{2}{p+2}}$ follows. Consider the special case of \mathcal{P} being the class of β -smooth densities on the unit cube $[0, 1]^d$ as defined in Theorem 27.15. The χ^2 -radius of this class is finite since each density therein is bounded from above and the uniform distribution works as a center for (32.39). In this case we have $p = \frac{d}{\beta}$ and hence the lower bound $\Omega((n \log n)^{-\frac{2\beta}{d+2\beta}})$. Here, however, we can argue differently by considering the subcollection $\mathcal{P}' = \frac{1}{2}\text{Unif}([0, 1]^d) + \frac{1}{2}\mathcal{P}$, which has (up to a constant factor) the same minimax risk, but has the advantage that $D(P \| P') \asymp H^2(P, P')$ for all $P, P' \in \mathcal{P}'$ (see Section 32.4). Repeating the argument in the proof below, then, yields the optimal lower bound $\Omega(n^{-\frac{2\beta}{d+2\beta}})$ removing the unnecessary logarithmic factors.

Proof. Let $M = M_{\text{loc}}(\mathcal{P}, \epsilon)$. From the definition there exists an $\epsilon/2$ -packing P_1, \dots, P_M in some Hellinger ball $B(R, \epsilon)$.

Let $\theta \sim \text{Unif}([M])$ and $X^n \stackrel{i.i.d.}{\sim} P_\theta$ conditioned on θ . Then from Fano's inequality in the form of Theorem 31.4 we get

32.2 Pairwise comparison à la Le Cam-Birgé 539

$$\sup_{P \in \mathcal{P}} \mathbb{E}[H^2(P, \hat{P})] \geq \left(\frac{\epsilon}{4}\right)^2 \left(1 - \frac{I(\theta; X^n) + \log 2}{\log M}\right)$$

It remains to show that

$$\frac{I(\theta; X^n) + \log 2}{\log M} \leq \frac{1}{2}. \quad (32.41)$$

To that end for an arbitrary distribution U define

$$Q = \epsilon^2 U + (1 - \epsilon^2)R.$$

We first notice that from Ex. I.48 we have that for all $i \in [M]$

$$D(P_i \| Q) \leq 8(H^2(P_i, R) + 2\epsilon^2) \left(\frac{\lambda}{\lambda - 1} \log \frac{1}{\epsilon^2} + D_\lambda(P_i \| U) \right)$$

provided that $\epsilon < 2^{-\frac{5\lambda}{2(\lambda-1)}} \triangleq \epsilon_0$. Since $H^2(P_i, R) \leq \epsilon^2$, by optimizing U (as the D_λ -center of \mathcal{P}) we obtain

$$\inf_U \max_{i \in [M]} D(P_i \| Q) \leq 24\epsilon^2 \left(\frac{\lambda}{\lambda - 1} \log \frac{1}{\epsilon^2} + R_\lambda \right) \leq \frac{c(\lambda)}{2} \epsilon^2 \left(\log \frac{1}{\epsilon^2} + R_\lambda \right).$$

By Theorem 4.1 we have

$$I(\theta; X^n) \leq \max_{i \in [M]} D(P_i^{\otimes n} \| Q^{\otimes n}) \leq \frac{nc(\lambda)}{2} \epsilon^2 \left(\log \frac{1}{\epsilon^2} + R_\lambda \right).$$

This final bound and condition (32.40) then imply (32.41) and the statement of the theorem. \square

Finally, we mention that for *sufficiently regular* models wherein the KL divergence and the squared Hellinger distances are comparable, the upper bound in Theorem 32.11 based on local entropy gives the *exact* minimax rate. Models of this type include GLM and more generally Gaussian local mixtures with bounded centers in arbitrary dimensions.

Corollary 32.13. *Assume that*

$$H^2(P, P') \asymp D(P \| P'), \quad \forall P, P' \in \mathcal{P}.$$

Then

$$\inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \asymp \epsilon_n^2$$

where ϵ_n was defined in (32.34).

Proof. By assumption, KL neighborhoods coincide with Hellinger balls up to constant factors. Thus the lower bound follows from apply Fano's method in Theorem 31.4 to a Hellinger ball of radius $O(\epsilon_n)$. \square

32.3 Yatracos' class and minimum distance estimator

In this section we prove (32.3), the third entropy upper bound on statistical risk. Paralleling the result (32.1) of Yang-Barron (for KL divergence) and (32.2) of Le Cam-Birgé (for Hellinger distance), the following result bounds the minimax total variation risk using the metric entropy of the parameter space in total variation:

Theorem 32.14 (Yatracos [330]). *There exists a universal constant C such that the following holds. Let $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P \in \mathcal{P}$, where \mathcal{P} is a collection of distributions on a common measurable space $(\mathcal{X}, \mathcal{E})$. For any $\epsilon > 0$, there exists a proper estimator $\hat{P} = \hat{P}(X_1, \dots, X_n) \in \mathcal{P}$, such that*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[\text{TV}(\hat{P}, P)^2] \leq C \left(\epsilon^2 + \frac{1}{n} \log N(\mathcal{P}, \text{TV}, \epsilon) \right) \quad (32.42)$$

For loss function that is a distance, a natural idea for obtaining proper estimator is the *minimum distance estimator*. In the current context, we compute the minimum-distance projection of the empirical distribution on the model class \mathcal{P} .⁶

$$P_{\min\text{-dist}} = \operatorname{argmin}_{P \in \mathcal{P}} \text{TV}(\hat{P}_n, P)$$

where $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ is the empirical distribution. However, since the empirical distribution is discrete, this strategy does not make sense if elements of \mathcal{P} have densities. The reason for this degeneracy is because the total variation distance is too strong. The key idea is to replace TV, which compares two distributions over all measurable sets, by a proxy, which only inspects a “low-complexity” family of sets.

To this end, let $\mathcal{A} \subset \mathcal{E}$ be a finite collection of measurable sets to be specified later. Define a pseudo-distance

$$\text{dist}(P, Q) \triangleq \sup_{A \in \mathcal{A}} |P(A) - Q(A)|. \quad (32.43)$$

(Note that if $\mathcal{A} = \mathcal{E}$, then this is just TV.) One can verify that dist satisfies the triangle inequality. As a result, the estimator

$$\tilde{P} \triangleq \operatorname{argmin}_{P \in \mathcal{P}} \text{dist}(P, \hat{P}_n), \quad (32.44)$$

as a minimizer, satisfies

$$\text{dist}(\tilde{P}, P) \leq \text{dist}(\tilde{P}, \hat{P}_n) + \text{dist}(P, \hat{P}_n) \leq 2\text{dist}(P, \hat{P}_n). \quad (32.45)$$

In addition, applying the binomial tail bound and the union bound, we have

$$\mathbb{E}[\text{dist}(P, \hat{P}_n)^2] \leq \frac{C_0 \log |\mathcal{A}|}{n}. \quad (32.46)$$

for some absolute constant C_0 .

⁶ Here and below, if the minimizer does not exist, we can replace it by an infimizing sequence.

32.3 Yatracos' class and minimum distance estimator 541

The main idea of Yatracos [330] boils down to the following choice of \mathcal{A} : Consider an ϵ -covering $\{Q_1, \dots, Q_N\}$ of \mathcal{P} in TV. Define the set

$$A_{ij} \triangleq \left\{ x : \frac{dQ_i}{d(Q_i + Q_j)}(x) \geq \frac{dQ_j}{d(Q_i + Q_j)}(x) \right\}$$

and the collection (known as the *Yatracos class*)

$$\mathcal{A} \triangleq \{A_{ij} : i \neq j \in [N]\}. \quad (32.47)$$

Then the corresponding dist approximates the TV on \mathcal{P} , in the sense that

$$\text{dist}(P, Q) \leq \text{TV}(P, Q) \leq \text{dist}(P, Q) + 4\epsilon, \quad \forall P, Q \in \mathcal{P}. \quad (32.48)$$

To see this, we only need to justify the upper bound. For any $P, Q \in \mathcal{P}$, there exists $i, j \in [N]$, such that $\text{TV}(P, P_i) \leq \epsilon$ and $\text{TV}(Q, Q_j) \leq \epsilon$. By the key observation that $\text{dist}(Q_i, Q_j) = \text{TV}(Q_i, Q_j)$, we have

$$\begin{aligned} \text{TV}(P, Q) &\leq \text{TV}(P, Q_i) + \text{TV}(Q_i, Q_j) + \text{TV}(Q_j, Q) \\ &\leq 2\epsilon + \text{dist}(Q_i, Q_j) \\ &\leq 2\epsilon + \text{dist}(Q_i, P) + \text{dist}(P, Q) + \text{dist}(Q, Q_j) \\ &\leq 4\epsilon + \text{dist}(P, Q). \end{aligned}$$

Finally, we analyze the estimator (32.44) with \mathcal{A} given in (32.47). Applying (32.48) and (32.45) yields

$$\begin{aligned} \text{TV}(\tilde{P}, P) &\leq \text{dist}(P, \tilde{P}) + 4\epsilon \\ &\leq 2\text{dist}(P, \hat{P}_n) + 4\epsilon. \end{aligned}$$

Squaring both sides, taking expectation and applying (32.46), we have

$$\mathbb{E}[\text{TV}(\tilde{P}, P)^2] \leq 32\epsilon^2 + 8\mathbb{E}[\text{dist}(P, \hat{P}_n)^2] \leq 32\epsilon^2 + \frac{8C_0 \log |N|}{n}.$$

Choosing the optimal TV-covering completes the proof of (32.42).

Remark 32.5 (Robust version). Note that Yatracos' scheme idea works even if the data generating distribution $P \notin \mathcal{P}$ but close to \mathcal{P} . Indeed, denote $Q^* = \operatorname{argmin}_{Q \in \{\mathcal{Q}_i\}} \text{TV}(P, Q)$ and notice that

$$\text{dist}(Q^*, \hat{P}_n) \leq \text{dist}(Q^*, P) + \text{dist}(P, \hat{P}_n) \leq \text{TV}(P, Q^*) + \text{dist}(P, \hat{P}_n),$$

since $\text{dist}(Q, Q') \leq \text{TV}(Q, Q')$ for any pair of distributions. Then we have:

$$\begin{aligned} \text{TV}(\tilde{P}, P) &\leq \text{TV}(\tilde{P}, Q^*) + \text{TV}(Q^*, P) = \text{dist}(\tilde{P}, Q^*) + \text{TV}(Q^*, P) \\ &\leq \text{dist}(\tilde{P}, \hat{P}_n) + \text{dist}(\hat{P}_n, P) + \text{dist}(P, Q^*) + \text{TV}(Q^*, P) \\ &\leq \text{dist}(Q^*, \hat{P}_n) + \text{dist}(\hat{P}_n, P) + 2\text{TV}(P, Q^*) \\ &\leq 2\text{dist}(P, \hat{P}_n) + 3\text{TV}(P, Q^*). \end{aligned}$$

Since $3\text{TV}(P, Q^*) \leq 3\epsilon + 3\min_{P' \in \mathcal{P}} \text{TV}(P, P')$ we can see that the estimator also works for “misspecified case”. Surprisingly, the multiplier 3 is not improvable if the estimator is required to be proper (inside \mathcal{P}), cf. [46].

32.4 Application: Estimating smooth densities

As a concrete application, in this section we study a nonparametric density estimation problem under smoothness constraint. Let \mathcal{F} denote the collection of 1-Lipschitz densities (with respect to Lebesgue) on the unit interval $[0, 1]$. (In this case the parameter is simply the density f , so we shall refrain from writing a parametrized form.) Given $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, an estimator of the unknown density f is a function $\hat{f}(\cdot) = \hat{f}(\cdot; X_1, \dots, X_n)$. Let us consider the conventional quadratic risk $\|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$.

Theorem 32.15. *Given $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f \in \mathcal{F}$, the minimax quadratic risk over \mathcal{F} satisfies*

$$R_{L_2}^*(n; \mathcal{F}) \triangleq \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E} \|f - \hat{f}\|_2^2 \asymp n^{-\frac{2}{3}}. \quad (32.49)$$

Capitalizing on the metric entropy of smooth densities studied in Section 27.4, we will prove this result by applying the entropic upper bound in Theorem 32.1 and the minimax lower bound based on Fano's inequality in Theorem 31.4. However, Theorem 32.18 pertains to the L_2 rather than KL risk. This can be fixed by a simple reduction.

Lemma 32.16. *Let \mathcal{F}' denote the collection of $f \in \mathcal{F}$ which is bounded from below by $1/2$. Then*

$$R_{L_2}^*(n; \mathcal{F}') \leq R_{L_2}^*(n; \mathcal{F}) \leq 4R_{L_2}^*(n; \mathcal{F}').$$

Proof. The left inequality follows because $\mathcal{F}' \subset \mathcal{F}$. For the right inequality, we apply a simulation argument. Fix some $f \in \mathcal{F}$ and we observe $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f$. Let us sample U_1, \dots, U_n independently and uniformly from $[0, 1]$. Define

$$Z_i = \begin{cases} U_i & \text{w.p. } \frac{1}{2}, \\ X_i & \text{w.p. } \frac{1}{2}. \end{cases}$$

Then $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} g = \frac{1}{2}(1 + f) \in \mathcal{F}'$. Let \hat{g} be an estimator that achieves the minimax risk $R_{L_2}^*(n; \mathcal{F}')$ on \mathcal{F}' . Consider the estimator $\hat{f} = 2\hat{g} - 1$. Then $\|f - \hat{f}\|_2^2 = 4\|g - \hat{g}\|_2^2$. Taking the supremum over $f \in \mathcal{F}$ proves $R_{L_2}^*(n; \mathcal{F}) \leq 4R_{L_2}^*(n; \mathcal{F}')$. \square

Lemma 32.19 allows us to focus on the subcollection \mathcal{F}' , where each density is lower bounded by $1/2$. In addition, each 1-Lipschitz density is also upper bounded by an absolute constant. Therefore, the KL divergence and squared L_2 distance are in fact equivalent on \mathcal{F}' , i.e.,

$$D(f||g) \asymp \|f - g\|_2^2, \quad f, g \in \mathcal{F}', \quad (32.50)$$

as shown by the following lemma:

Lemma 32.17. *Suppose both $f = \frac{dP}{d\mu}$ and $g = \frac{dQ}{d\mu}$ are upper and lower bounded by absolute constants c and C respectively. Then*

$$\frac{1}{C} \int d\mu(f - g)^2 \leq 2H^2(f||g) \leq D(P||Q) \leq \chi^2(P||Q) \leq \frac{1}{c} \int d\mu(f - g)^2.$$

32.4 Application: Estimating smooth densities 543

Proof. For the upper bound, applying (7.31), $D(P\|Q) \leq \chi^2(P\|Q) = \int d\mu \frac{(f-g)^2}{g} \leq \frac{1}{c} \int d\mu \frac{(f-g)^2}{g}$. For the lower bound, applying (7.30), $D(P\|Q) \geq 2H^2(f\|g) = 2 \int d\mu \frac{(f-g)^2}{(\sqrt{f}+\sqrt{g})^2} \geq \frac{1}{C} \int d\mu (f-g)^2$. \square

We now prove Theorem 32.18:

Proof. In view of Lemma 32.19, it suffices to consider $R_{L_2}^*(n; \mathcal{F}')$. For the upper bound, we have

$$\begin{aligned} R_{L_2}^*(n; \mathcal{F}') &\stackrel{(a)}{\asymp} R_{KL}^*(n; \mathcal{F}') \\ &\stackrel{(b)}{\lesssim} \inf_{\epsilon>0} \left\{ \epsilon^2 + \frac{1}{n} \log N_{KL}(\mathcal{F}', \epsilon) \right\} \\ &\stackrel{(c)}{\asymp} \inf_{\epsilon>0} \left\{ \epsilon^2 + \frac{1}{n} \log N(\mathcal{F}', \|\cdot\|_2, \epsilon) \right\} \\ &\stackrel{(d)}{\asymp} \inf_{\epsilon>0} \left\{ \epsilon^2 + \frac{1}{n\epsilon} \right\} \asymp n^{-2/3}. \end{aligned}$$

where both (a) and (c) apply (32.50), so that both the risk and the metric entropy are equivalent for KL and L_2 distance; (b) follows from Theorem 32.1; (d) applies the metric entropy (under L_2) of the Lipschitz class from Theorem 27.14 and the fact that the metric entropy of the subclass \mathcal{F}' is at most that of the full class \mathcal{F} .

For the lower bound, we apply Fano's inequality. Applying Theorem 27.14 and the relation between covering and packing numbers in Theorem 27.3, we have $\log N(\mathcal{F}, \|\cdot\|_2, \epsilon) \asymp \log M(\mathcal{F}, \|\cdot\|_2, \epsilon) \asymp \frac{1}{\epsilon}$. Fix ϵ to be specified and let f_1, \dots, f_M be an ϵ -packing in \mathcal{F} , where $M \geq \exp(C/\epsilon)$. Then g_1, \dots, g_M is an $\frac{\epsilon}{2}$ -packing in \mathcal{F}' , with $g_i = (f_i+1)/2$. Applying Fano's inequality in Theorem 31.4, we have

$$R_{L_2}^*(n; \mathcal{F}) \gtrsim \epsilon^2 \left(1 - \frac{C_n}{\log M} \right).$$

Using (32.17), we have $C_n \leq \inf_{\epsilon>0} (n\epsilon^2 + \epsilon^{-1}) \asymp n^{1/3}$. Thus choosing $\epsilon = cn^{-1/3}$ for sufficiently small c ensures $C_n \leq \frac{1}{2} \log M$ and hence $R_{L_2}^*(n; \mathcal{F}) \gtrsim \epsilon^2 \asymp n^{-2/3}$. \square

Remark 32.6. Note that the above proof of Theorem 32.18 relies on the entropic risk bound (32.1), which, though rate-optimal, is not attained by a computationally efficient estimator. (The same criticism also applies to (32.2) and (32.3) for Hellinger and total variation.) To remedy this, for the squared loss, a classical idea is to apply the kernel density estimator (KDE) – cf. Section 7.9. Specifically, one compute the convolution of the empirical distribution $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ with a kernel function $K(\cdot)$ whose shape and bandwidth are chosen according to the smooth constraint. For Lipschitz density, the optimal rate in Theorem 32.18 can be attained by a box kernel $K(\cdot) = \frac{1}{2h} \mathbf{1}_{\{|\cdot| \leq h\}}$ with bandwidth $h = n^{-1/3}$ (cf. e.g. [304, Sec. 1.2]).

33 Strong data processing inequality

In this chapter we explore statistical implications of the following effect. For any Markov chain

$$U \rightarrow X \rightarrow Y \rightarrow V \quad (33.1)$$

we know from the data-processing inequality (Theorem 3.7) that

$$I(U; Y) \leq I(U; X), \quad I(X; V) \geq I(Y; V).$$

However, something stronger can often be said. Namely, if the Markov chain (33.1) factor through a *known* noisy channel $P_{Y|X} : \mathcal{X} \rightarrow \mathcal{Y}$, then oftentimes we can prove strong data processing inequalities (SDPI):

$$I(U; Y) \leq \eta I(U; X), \quad \eta^{(p)} I(X; V) \geq I(Y; V),$$

where coefficients $\eta = \eta(P_{Y|X})$, $\eta^{(p)}(P_{Y|X}) < 1$ only depend on the channel and not the (generally unknown or very complex) $P_{U,X}$ or $P_{Y,V}$. The coefficients η and $\eta^{(p)}$ approach 0 for channels that are very noisy (for example, η is always up to a constant factor equal to the Hellinger-squared diameter of the channel).

The purpose of this chapter is twofold. First, we want to introduce general properties of the SDPI coefficients. Second, we want to show how SDPIs help prove sharp lower (impossibility) bounds on statistical estimation questions. The flavor of the statistical problems in this chapter is different from the rest of the book in that here the information about unknown parameter θ is *thinly distributed* across a high dimensional vector (as in spiked Wigner and tree-coloring examples), or across different terminals (as in correlation and mean estimation examples).

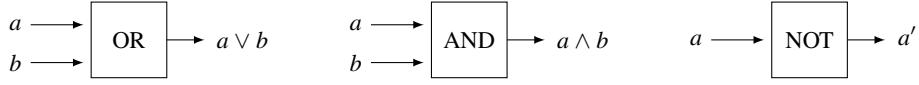
We point out that SDPIs are an area of current research and multiple topics are not covered by our brief exposition here. For more, we recommend surveys [241] and [248], of which the latter explores the functional-theoretic side of SDPIs and their close relation to logarithmic Sobolev inequalities – a topic we omitted entirely.

33.1 Computing a boolean function with noisy gates

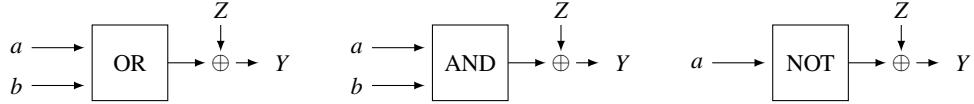
A boolean function with n inputs is defined as $f : \{0, 1\}^n \rightarrow \{0, 1\}$. Note that a boolean function can be described as a network of primitive logic gates of the three following kinds:

In 1938, Shannon has shown how any boolean function f can be represented with primitive logic gates [267].

33.1 Computing a boolean function with noisy gates 545



Now suppose there are additive noise components on the output of each primitive gate. In this case, we have a network of the following noisy gates.



Here, $Z \sim \text{Bern}(\delta)$ and assumed to be independent of the inputs. In other words, with probability δ , the output of a gate will be flipped no matter what input is given to that gate. Hence, we sometimes refer to these gates as δ -noisy gates.

In 1950s John von Neumann was laying the groundwork for the digital computers, and he was bothered by the following question. Can we compute any boolean function f with δ -noisy gates? Note that any circuit that consists of noisy gates necessarily has noisy (non-deterministic) output. Therefore, when we say that a noisy gate circuit C computes f we require the existence of some $\epsilon_0 = \epsilon_0(\delta)$ (that cannot depend on f) such that

$$\mathbb{P}[C(x_1, \dots, x_n) \neq f(x_1, \dots, x_n)] \leq \frac{1}{2} - \epsilon_0 \quad (33.2)$$

where $C(x_1, \dots, x_n)$ is the output of the noisy circuit inputs x_1, \dots, x_n . If we build the circuit according to the classical (Shannon) methods, we would obviously have catastrophic error accumulation so that deep circuits necessarily have $\epsilon_0 \rightarrow 0$. At the same time, von Neumann was bothered by the fact that evidently our brains operate with very noisy gates and yet are able to carry very long computations without mistakes. His thoughts culminated in the following ground-breaking result.

Theorem 33.1 (von Neumann, 1957). *There exists $\delta^* > 0$ such that for all $\delta < \delta^*$ it is possible to compute every boolean function f via δ -noisy 3-majority gates.*

von Neumann's original estimate $\delta^* \approx 0.087$ was subsequently improved by Pippenger. The main (still open) question of this area is to find the largest δ^* for which the above theorem holds.

Condition in (33.2) implies the output should be correlated with the inputs. This requires the mutual information between the inputs (if they are random) and the output to be greater than zero. We now give a theorem of Evans and Schulman that gives an upper bound to the mutual information between any of the inputs and the output. We will prove the theorem in Section 33.3 as a consequence of the more general *directed information percolation* theory.

Theorem 33.2 ([116]). *Suppose an n -input noisy boolean circuit composed of gates with at most K inputs and with noise components having at most δ probability of error. Then, the mutual information between any input X_i and output Y is upper bounded as*

$$I(X_i; Y) \leq (K(1 - 2\delta)^2)^{d_i} \log 2$$

546

where d_i is the minimum length between X_i and Y (i.e., the minimum number of gates required to be passed through until reaching Y).

Theorem 33.2 implies that noisy computation is only possible for $\delta < \frac{1}{2} - \frac{1}{2\sqrt{K}}$. This is the best known threshold. An illustration is given below:

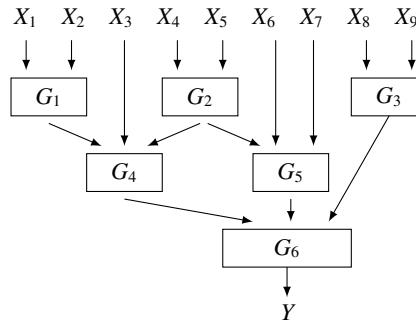


Figure 33.1 An example of a 9-input Boolean Circuit

The above 9-input circuit has gates with at most 3 inputs. The 3-input gates are G_4 , G_5 and G_6 . The minimum distance between X_3 and Y is $d_3 = 2$, and the minimum distance between X_5 and Y is $d_5 = 3$. If G_i 's are δ -noisy gates, we can invoke Theorem 33.2 between any input and the output.

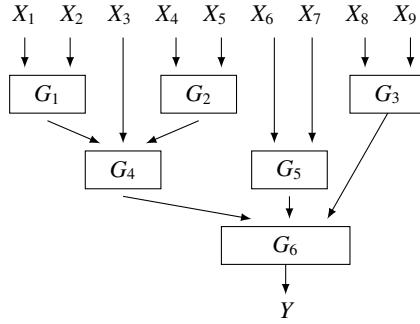
Unsurprisingly, Theorem 33.2 also tells us that there are some circuits that are not computable with δ -noisy gates. For instance, take $f(X_1, \dots, X_n) = \text{XOR}(X_1, \dots, X_n)$. Then for at least one input X_i , we have $d_i \geq \frac{\log n}{\log K}$. This shows that $I(X_i; Y) \rightarrow 0$ as $n \rightarrow \infty$, hence X_i and Y will be almost independent for large n . Note that $\text{XOR}(X_1, \dots, X_n) = \text{XOR}(\text{XOR}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n), X_i)$. Therefore, it is impossible to compute an n -input XOR with δ -noisy gates for large n .

Computation with formulas: Note that the graph structure given in Figure 33.1 contains some undirected loops. A *formula* is a type of boolean circuits that does not contain any undirected loops unlike the case in Figure 33.1. In other words, for a formula the underlying graph structure forms a tree. Removing one of the outputs of G_2 of Figure 33.1, we obtain a formula as given below.

In Theorem 1 of [115], it is shown that we can compute reliably any boolean function f that is represented with a formula with at most K -input gates with K odd and every gate are at most δ -noisy and $\delta < \delta_f^*$, and no such computation is possible for $\delta > \delta_f^*$, where

$$\delta_f^* = \frac{1}{2} - \frac{2^{K-1}}{K \binom{K-1}{\frac{K-1}{2}}}$$

33.2 Strong Data Processing Inequality 547



where the approximation holds for large K . This threshold is better than the upper-bound on the threshold given by Theorem 33.2 for general boolean circuits. However, for large K we have

$$\delta_f^* \approx \frac{1}{2} - \frac{\sqrt{\pi/2}}{2\sqrt{K}}, K \gg 1$$

showing that the estimate of Evans-Schulman $\delta^* \leq \frac{1}{2} - \frac{1}{2\sqrt{K}}$ is order-tight for large K . This demonstrates the tightness of Theorem 33.2.

33.2 Strong Data Processing Inequality

Definition 33.3. (Contraction coefficient for $P_{Y|X}$) For a fixed conditional distribution (or kernel) $P_{Y|X}$, define

$$\eta_f(P_{Y|X}) = \sup \frac{D_f(P_Y||Q_Y)}{D_f(P_X||Q_X)}, \quad (33.3)$$

where $P_Y = P_{Y|X} \circ P_X$, $Q_Y = P_{Y|X} \circ Q_X$ and supremum is over all pairs (P_X, Q_X) satisfying $0 < D_f(P_X||Q_X) < \infty$.

Recall that the DPI (Theorem 7.7) states that $D_f(P_X||Q_X) \geq D_f(P_Y||Q_Y)$. The concept of the Strong DPI introduced above quantifies the multiplicative decrease between the two f -divergences.

Example 33.1. Suppose $P_{Y|X}$ is a kernel for a time-homogeneous Markov chain with stationary distribution π (i.e., $P_{Y|X} = P_{X_{t+1}|X_t}$). Then for any initial distribution q , SDPI gives the following bound:

$$D_f(qP^n||\pi) \leq \eta_f^n D_f(q||\pi)$$

These type of exponential decreases are frequently encountered in the Markov chains literature, especially for KL and χ^2 divergences. For example, for reversible Markov chains, we have [90, Prop. 3]

$$\chi^2(P_{X_n}||\pi) \leq \gamma_*^{2n} \chi^2(P_{X_1}||\pi) \quad (33.4)$$

where γ_* is the absolute spectral gap of P . See Exercise VI.18.

We note that in general $\eta_f(P_{Y|X})$ is hard to compute. However, total variation is an exception.

Theorem 33.4 ([92]). $\eta_{\text{TV}} = \sup_{x \neq x'} \text{TV}(P_{Y|X=x}, P_{Y|X=x'})$.

Proof. We consider two directions separately.

- $\eta_{\text{TV}} \geq \sup_{x_0 \neq x'_0} \text{TV}(P_{Y|X=x_0}, P_{Y|X=x'_0})$:

This case is obvious. Take $P_X = \delta_{x_0}$ and $Q_X = \delta_{x'_0}$.¹ Then from the definition of η_{TV} , we have $\eta_{\text{TV}} \geq \text{TV}(P_{Y|X=x_0}, P_{Y|X=x'_0})$ for any x_0 and x'_0 , $x_0 \neq x'_0$.

- $\eta_{\text{TV}} \leq \sup_{x_0 \neq x'_0} \text{TV}(P_{Y|X=x_0}, P_{Y|X=x'_0})$:

Define $\tilde{\eta} \triangleq \sup_{x_0 \neq x'_0} \text{TV}(P_{Y|X=x_0}, P_{Y|X=x'_0})$. We consider the discrete alphabet case for simplicity. Fix any P_X , Q_X and $P_Y = P_X \circ P_{Y|X}$, $Q_Y = Q_X \circ P_{Y|X}$. Observe that for any $E \subseteq \mathcal{Y}$

$$P_{Y|X=x_0}(E) - P_{Y|X=x'_0}(E) \leq \tilde{\eta} \mathbf{1}\{x_0 \neq x'_0\}. \quad (33.5)$$

Now suppose there are random variables X_0 and X'_0 having some marginals P_X and Q_X respectively. Consider any coupling π_{X_0, X'_0} with marginals P_X and Q_X respectively. Then averaging (33.5) and taking the supremum over E , we obtain

$$\sup_{E \subseteq \mathcal{Y}} P_Y(E) - Q_Y(E) \leq \tilde{\eta} \pi[X_0 \neq X'_0]$$

Now the left-hand side equals $\text{TV}(P_Y, Q_Y)$ by Theorem 7.12(a). Taking the infimum over couplings π the right-hand side evaluates to $\text{TV}(P_X, Q_X)$ by Theorem 7.12(b).

□

Example 33.2 (η_{TV} of a Binary Symmetric Channel). The η_{TV} of the BSC_δ is given by

$$\begin{aligned} \eta_{\text{TV}}(\text{BSC}_\delta) &= \text{TV}(\text{Bern}(\delta), \text{Bern}(1-\delta)) \\ &= \frac{1}{2}(|\delta - (1-\delta)| + |1-\delta - \delta|) = |1-2\delta|. \end{aligned}$$

We sometimes want to relate η_f with the f -mutual informations instead of f -divergences. This relation is given in the following theorem.

Theorem 33.5.

$$\eta_f(P_{Y|X}) = \sup_{P_{UX}: U \rightarrow X \rightarrow Y} \frac{I_f(U; Y)}{I_f(U; X)}.$$

¹ δ_{x_0} is the probability distribution with $\mathbb{P}(X = x_0) = 1$

33.2 Strong Data Processing Inequality 549

Recall that for any Markov chain $U \rightarrow X \rightarrow Y$, DPI states that $I_f(U; Y) \leq I_f(U; X)$ and Theorem 33.5 gives the stronger bound

$$I_f(U; Y) \leq \eta_f(P_{Y|X})I_f(U; X). \quad (33.6)$$

Proof. First, notice that for any u_0 , we have $D_f(P_{Y|U=u_0} \| P_Y) \leq \eta_f D_f(P_{X|U=u_0} \| P_X)$. Averaging the above expression over any P_U , we obtain

$$I_f(U; Y) \leq \eta_f I_f(U; X)$$

Second, fix \tilde{P}_X, \tilde{Q}_X and let $U \sim \text{Bern}(\lambda)$ for some $\lambda \in [0, 1]$. Define the conditional distribution $P_{X|U}$ as $P_{X|U=1} = \tilde{P}_X, P_{X|U=0} = \tilde{Q}_X$. Take $\lambda \rightarrow 0$, then (see [241] for technical subtleties)

$$\begin{aligned} I_f(U; X) &= \lambda D_f(\tilde{P}_X \| \tilde{Q}_X) + o(\lambda) \\ I_f(U; Y) &= \lambda D_f(\tilde{P}_Y \| \tilde{Q}_Y) + o(\lambda) \end{aligned}$$

The ratio $\frac{I_f(U; Y)}{I_f(U; X)}$ will then converge to $\frac{D_f(\tilde{P}_Y \| \tilde{Q}_Y)}{D_f(\tilde{P}_X \| \tilde{Q}_X)}$. Thus, optimizing over \tilde{P}_X and \tilde{Q}_X we can get ratio of I_f 's arbitrarily close to η_f . \square

We next state some of the fundamental properties of contraction coefficients.

Theorem 33.6. *In the statements below η_f (and others) corresponds to $\eta_f(P_{Y|X})$ for some fixed $P_{Y|X}$.*

- (a) For any f , $\eta_f \leq \eta_{TV}$.
- (b) $\eta_{KL} = \eta_{H^2} = \eta_{\chi^2}$. More generally, for any operator-convex and twice continuously differentiable f we have $\eta_f = \eta_{\chi^2}$.
- (c) η_{χ^2} equals the maximal correlation: $\eta_{\chi^2} = \sup_{P_{X,f,g}} \rho(f(X), g(Y))$, where $\rho(X, Y) \triangleq \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ is the correlation coefficient between X and Y .
- (d) For binary-input channels, denote $P_0 = P_{Y|X=0}$ and $P_1 = P_{Y|X=1}$. Then we have

$$\eta_{KL} = LC_{max}(P_0, P_1) \triangleq \sup_{0 < \beta < 1} LC_\beta(P_0 \| P_1)$$

where (recall $\bar{\beta} \triangleq 1 - \beta$)

$$LC_\beta(P \| Q) = D_f(P \| Q), \quad f(x) = \bar{\beta}\beta \frac{(1-x)^2}{\bar{\beta}x + \beta}$$

is the Le Cam divergence of order β (recall (7.6) for $\beta = 1/2$).

- (e) Consequently,

$$\frac{1}{2} H^2(P_0, P_1) \leq \eta_{KL} \leq H^2(P_0, P_1) - \frac{H^4(P_0, P_1)}{4}. \quad (33.7)$$

- (f) If the binary-input channel is also input-symmetric (or BMS, see Section 19.4*) then $\eta_{KL} = I_{\chi^2}(X; Y)$ for $X \sim \text{Bern}(1/2)$.

(g) For any channel the supremum in (33.3) can be restricted to P_X, Q_X with a common binary support. In other words, $\eta_f(P_{Y|X})$ coincides with that of the least contractive binary subchannel. Consequently, from (e) we conclude

$$\frac{1}{2} \text{diam}_{H^2} \leq \eta_{\text{KL}}(P_{Y|X}) = \text{diam}_{LC_{\max}} \leq \text{diam}_{H^2} - \frac{\text{diam}_{H^2}}{4},$$

(in particular $\eta_{\text{KL}} \asymp \text{diam}_{H^2}$), where $\text{diam}_{H^2}(P_{Y|X}) = \sup_{x, x' \in \mathcal{X}} H^2(P_{Y|X=x}, P_{Y|X=x'})$, $\text{diam}_{LC_{\max}} = \sup_{x, x'} LC_{\max}(P_{Y|X=x}, P_{Y|X=x'})$ are Hellinger and Le Cam diameters of the channel.

Proof. Most proofs in full generality can be found in [241]. For (a) one first shows that $\eta_f \leq \eta_{\text{TV}}$ for the so-called \mathcal{E}_γ divergences corresponding to $f(x) = |x - \gamma|_+ - |1 - \gamma|_+$, which is not hard to believe since \mathcal{E}_γ is piecewise linear. Then the general result follows from the fact that any convex function f can be approximated (as $N \rightarrow \infty$) in the form

$$\sum_{j=1}^N a_j |x - c_j|_+ + a_0 x + c_0.$$

For (b) see [65, Theorem 1] and [69, Proposition II.6.13 and Corollary II.6.16]. The idea of this proof is as follows: :

- $\eta_{\text{KL}} \geq \eta_{\chi^2}$ by locality. Recall that every f -divergence behaves locally as χ^2 – Theorem 7.27.
- Using the identity $D(P\|Q) = \int_0^\infty \chi^2(P\|Q_t) dt$ where $Q_t = \frac{tP + Q}{1+t}$, we have

$$D(P_Y\|Q_Y) = \int_0^\infty \chi^2(P_Y\|Q_{Y_t}) dt \leq \eta_{\chi^2} \int_0^\infty \chi^2(P_X\|Q_{X_t}) dt = \eta_{\chi^2} D(P_X\|Q_X).$$

For (c), we fix Q_X (and thus $Q_{X,Y} = Q_X P_{Y|X}$). If $g = \frac{dP_X}{dQ_X}$ then $Tg(y) = \frac{dP_Y}{dQ_Y} = \mathbb{E}_{Q_{X|Y}}[g(X)|Y=y]$ is a linear operator. $\eta_{\chi^2}(P_{Y|X})$ is then nothing else than the maximal singular value (spectral norm squared) of $T : L_2(Q_X) \rightarrow L_2(Q_Y)$ when restricted to $\{g : \mathbb{E}_{Q_X}[g] = 0\}$. The adjoint of T is $T^*h(x) = \mathbb{E}_{P_{Y|X}}[h(Y)|X=x]$. Since the spectral norms of T and T^* coincide, and the spectral norm of T^* is precisely the maximal correlation we get the result.

The (d) follows from the definition of $\eta_{\chi^2} = \sup_{\alpha, \beta} \frac{\chi^2(\alpha P_1 + \bar{\alpha} P_0 \| \beta P_1 + \bar{\beta} P_0)}{\chi^2(\text{Ber}(\alpha) \| \text{Ber}(\beta))}$ and some algebra.

Next, (e) follows from bounding (via Cauchy-Schwarz etc) LC_{\max} in terms of H^2 ; see [241, Appendix B].

The (f) follows from the fact that every BMS channel can be represented as $X \mapsto Y = (Y_\Delta, \Delta)$ where $\Delta \in [0, 1/2]$ is independent of X and $Y_\delta = \text{BSC}_\delta(X)$. In other words, every BMS channel is a mixture of BSCs; see [255, Section 4.1]. Thus, we have for any $U \rightarrow X \rightarrow Y = (Y_\Delta, \Delta)$ and $\Delta \perp (U, X)$ the following chain

$$I(U; Y) = I(U; Y|\Delta) \leq \mathbb{E}_{\delta \sim P_\Delta} [(1 - 2\delta)^2 I(U; X|\Delta = \delta)] = \mathbb{E}[(1 - 2\Delta)^2] I(U; X),$$

where we used the fact that $I(U; X|\Delta = \delta) = I(U; X)$ and Example 33.3 below.

For (g) see Ex. VI.19. □

33.3 Directed Information Percolation 551

Example 33.3 (Computing $\eta_{\text{KL}}(\text{BSC}_\delta)$). Consider the BSC_δ channel. In Example 33.2 we computed η_{TV} . Here we have $\text{diam}_{H^2} = 2 - 4\sqrt{\delta(1-\delta)}$ and thus the bound (33.7) we get $\eta_{\text{KL}} \leq (1-2\delta)^2$. On the other hand taking $U = \text{Ber}(1/2)$ and $P_{X|U} = \text{Ber}(\alpha)$ we get

$$\eta_{\text{KL}} \geq \frac{I(U; Y)}{I(U; X)} = \frac{\log 2 - h(\alpha + (1-2\alpha)\delta)}{\log 2 - h(\alpha)} \rightarrow (1-2\delta)^2 \quad \alpha \rightarrow \frac{1}{2}.$$

Thus we have shown:

$$\eta_{\text{KL}}(\text{BSC}_\delta) = \eta_{H^2}(\text{BSC}_\delta) = \eta_{\chi^2} = (1-2\delta)^2.$$

This example has the following consequence for the KL-divergence geometry.

Proposition 33.7. *Consider any distributions P_0 and P_1 on \mathcal{X} and let us consider the interval in $\mathcal{P}(\mathcal{X})$: $P_\lambda = \lambda P_1 + (1-\lambda)P_0$ for $\lambda \in [0, 1]$. Then divergence (with respect to the midpoint) behaves subquadratically:*

$$D(P_\lambda \| P_{1/2}) + D(P_{1-\lambda} \| P_{1/2}) \leq (1-2\lambda)^2 \{D(P_0 \| P_{1/2}) + D(P_1 \| P_{1/2})\}.$$

The same statement holds with D replaced by χ^2 (and any other D_f satisfying Theorem 33.6(b)).

Proof. Let $X \sim \text{Ber}(1/2)$ and $Y = \text{BSC}_\lambda(X)$. Let $U \leftarrow X \rightarrow Y$ be defined with $U \sim P_0$ if $X = 0$ and $U \sim P_1$ if $X = 1$. Then

$$I_f(U; Y) = \frac{1}{2}D_f(P_\lambda \| P_{1/2}) + \frac{1}{2}D_f(P_{1-\lambda} \| P_{1/2}).$$

Thus, applying SDPI (33.6) completes the proof. \square

Remark 33.1. Let us introduce $d_{JS}(P, Q) = \sqrt{\text{JS}(P, Q)}$ and $d_{LC} = \sqrt{\text{LC}(P, Q)}$ – the Jensen-Shannon (7.7) and Le Cam (7.6) metrics. Then the proposition can be rewritten as

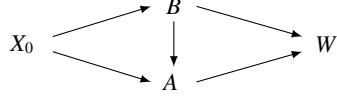
$$\begin{aligned} d_{JS}(P_\lambda, P_{1-\lambda}) &\leq |1-2\lambda|d_{JS}(P_0, P_1) \\ d_{LC}(P_\lambda, P_{1-\lambda}) &\leq |1-2\lambda|d_{LC}(P_0, P_1). \end{aligned}$$

Notice that for any metric $d(P, Q)$ on $\mathcal{P}(\mathcal{X})$ that is induced from the norm on the vector space $\mathcal{M}(\mathcal{X})$ of all signed measures (such as TV), we must necessarily have $d(P_\lambda, P_{1-\lambda}) = |1-2\lambda|d(P_0, P_1)$. Thus, the $\eta_{\text{KL}}(\text{BSC}_\lambda) = (1-2\lambda)^2$ which yields the inequality is rather natural.

33.3 Directed Information Percolation

In this section, we are concerned about the amount of information decay experienced in a directed acyclic graph (DAG) $G = (V, E)$. In the following context the vertex set V refers to a set of vertices v , each associated with a random variable X_v and the edge set E refers to a set of directed edges whose configuration allows us to factorize the joint distribution over X_V by Throughout the section, we consider Shannon mutual information, i.e., $f = x \log x$. Let us give a detailed example below.

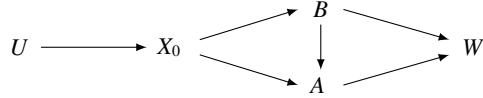
552



Example 33.4. Suppose we have a graph $G = (V, E)$ as below. This means that we have a joint distribution factorizing as

$$P_{X_0, A, B, W} = P_{X_0} P_{B|X_0} P_{A, B|X_0} P_{W|A, B}.$$

Then every node has a channel from its parents to itself, for example W corresponds to a noisy channel $P_{W|A, B}$, and we can define $\eta \triangleq \eta_{\text{KL}}(P_{W|A, B})$. Now, prepend another random variable $U \sim \text{Bern}(\lambda)$ at the beginning, the new graph $G' = (V', E')$ is shown below: We want to verify the



relation

$$I(U; B, W) \leq \bar{\eta}I(U; B) + \eta I(U; A, B). \quad (33.8)$$

Recall that from chain rule we have $I(U; B, W) = I(U; B) + I(U; W|B) \geq I(U; B)$. Hence, if (33.8) is correct, then $\eta \rightarrow 0$ implies $I(U; B, W) \approx I(U; B)$ and symmetrically $I(U; A, W) \approx I(U; A)$. Therefore for small δ , observing W, A or W, B does not give advantage over observing solely A or B , respectively.

Observe that G' forms a Markov chain $U \rightarrow X_0 \rightarrow (A, B) \rightarrow W$, which allows us to factorize the joint distribution over E' as

$$P_{U, X_0, A, B, W} = P_U P_{X_0|U} P_{A, B|X_0} P_{W|A, B}.$$

Now consider the joint distribution conditioned on $B = b$, i.e., $P_{U, X_0, A, W|B}$. We claim that the conditional Markov chain $U \rightarrow X_0 \rightarrow A \rightarrow W|B = b$ holds. Indeed, given B and A , X_0 is independent of W , that is $P_{X_0|A, B} P_{W|A, B} = P_{X_0, W|AB}$, from which follows the mentioned conditional Markov chain. Using the conditional Markov chain, SDPI gives us for any b ,

$$I(U; W|B = b) \leq \eta I(U; A|B = b).$$

Averaging over b and adding $I(U; B)$ to both sides we obtain

$$\begin{aligned} I(U; W, B) &\leq \eta I(U; A|B) + I(U; B) \\ &= \eta I(U; A, B) + \bar{\eta} I(U; B). \end{aligned}$$

From the characterization of η in Theorem 33.5 we conclude

$$\eta_{\text{KL}}(P_{W, B|X_0}) \leq \eta \cdot \eta_{\text{KL}}(P_{A, B|X_0}) + (1 - \eta) \cdot \eta_{\text{KL}}(P_{B|X_0}). \quad (33.9)$$

Now, we provide another example which has in some sense an analogous setup to Example 33.4.

33.3 Directed Information Percolation 553

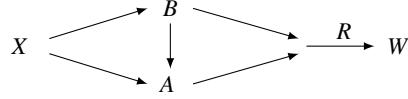


Figure 33.2 Illustration for Example 33.5.

Example 33.5 (Percolation). Take the graph $G = (V, E)$ in example 33.4 with a small modification. See Fig. 33.2. Now, suppose X, A, B, W are some cities and the edge set E represents the roads between these cities. Let R be a random variable denoting the state of the road connecting to W with $\mathbb{P}(R \text{ is open}) = \eta$ and $\mathbb{P}(R \text{ is closed}) = \bar{\eta}$. For any $Y \in V$, let the event $\{X \rightarrow Y\}$ indicate that one can drive from X to Y . Then

$$\mathbb{P}[X \rightarrow B \text{ or } W] = \eta \mathbb{P}[X \rightarrow A \text{ or } B] + \bar{\eta} \mathbb{P}[X \rightarrow B]. \quad (33.10)$$

Observe the resemblance between (33.9) and (33.10).

We will now give a theorem that relates η_{KL} to percolation probability on a DAG under the following setting: Consider a DAG $G = (V, E)$.

- All edges are open
- Every vertex is open with probability $p(v) = \eta_{KL}(P_{X_v|X_{Pa(v)}})$ where $Pa(v)$ denotes the set of parents of v .

Under this model, for two subsets $T, S \subset V$ we define $\text{perc}[T \rightarrow S] = \mathbb{P}[\exists \text{ open path } T \rightarrow S]$.

Note that $P_{X_v|X_{Pa(v)}}$ describe the stochastic recipe for producing X_v based on its parent variables. We assume that in addition to a DAG we also have been given all these constituent channels (or at least bounds on their η_{KL} coefficients).

Theorem 33.8 ([241]). *Let $G = (V, E)$ be a DAG and let 0 be a node with in-degree equal to zero (i.e. a source node). Note that for any $0 \not\ni S \subset V$ we can inductively stitch together constituent channels $P_{X_v|X_{Pa(v)}}$ and obtain $P_{X_S|X_0}$. Then we have*

$$\eta_{KL}(P_{X_S|X_0}) \leq \text{perc}(0 \rightarrow S). \quad (33.11)$$

Proof. For convenience let us denote $\eta(T) = \eta_{KL}(P_{X_T|X_0})$ and $\eta_v = \eta_{KL}(P_{X_v|X_{Pa(v)}})$. The proof follows from an induction on the size of G . The statement is clear for the $|V(G)| = 1$ since $S = \emptyset$ or $S = \{X_0\}$. Now suppose the statement is already shown for all graphs smaller than G . Let v be the node with out-degree 0 in G . If $v \notin S$ then we can exclude it from G and the statement follows from induction hypothesis. Otherwise, define $S_A = Pa(v) \setminus S$ and $S_B = S \setminus \{v\}$, $A = X_{S_A}, B = X_{S_B}, W = X_v$. (If $0 \in A$ then we can create a fake $0'$ with $X_{0'} = X_0$ and retain $0' \in A$ while moving 0 out of A . So without loss of generality, $0 \notin A$.) Prepending arbitrary U to the graph as $U \rightarrow X_0$, the joint DAG of random variables (X_0, A, B, W) is then given by precisely the graph in (33.8). Thus, we obtain from (33.9) the estimate

$$\eta(S) \leq \eta_v \eta(S_A \cup S_B) + (1 - \eta_v) \eta_{KL}(S_B). \quad (33.12)$$

From induction hypothesis $\eta(S_A \cup S_B) \leq \text{perc}(0 \rightarrow S_A)$ and $\eta(S_B) \leq \text{perc}(0 \rightarrow S_B)$ (they live on a graph $G \setminus \{v\}$). Thus, from computation (33.10) we see that the right-hand side of (33.12) is precisely $\text{perc}(0 \rightarrow S)$ and thus $\eta(S) \leq \text{perc}(S)$ as claimed. \square

We are now in position to complete the postponed proof.

Proof of Theorem 33.2. First observe the noisy boolean circuit is a form of DAG. Since the gates are δ -noisy contraction coefficients of constituent channels η_v in the DAG can be bounded by $(1 - 2\delta)^2$. Thus, in the percolation question all vertices are open with probability $(1 - 2\delta)^2$

From SDPI, for each i , we have $I(X_i; Y) \leq \eta_{\text{KL}}(P_{Y|X_i})H(X_i)$. From Theorem 33.9, we know $\eta_{\text{KL}}(P_{Y|X_i}) \leq \text{perc}(X_i \rightarrow Y)$. We now want to upper bound $\text{perc}(X_i \rightarrow Y)$. Recall that the minimum distance between X_i and Y is d_i . For any path π of length $\ell(\pi)$ from X_i to Y , therefore, the probability that it will be open is $\leq (1 - 2\delta)^{2\ell(\pi)}$. We can thus bound

$$\text{perc}(X_i \rightarrow Y) \leq \sum_{\pi: X_i \rightarrow Y} (1 - 2\delta)^{2\ell(\pi)}. \quad (33.13)$$

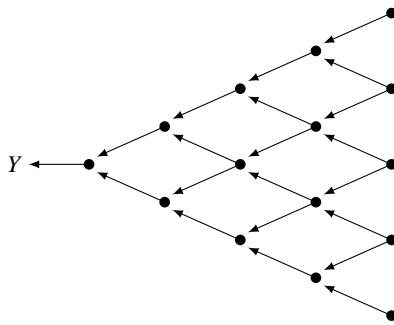
Let us now build paths backward starting from Y , which allows us to represent paths $X \rightarrow Y_i$ as vertices of a K -ary tree with root Y_i . By labeling all vertices on a K -ary tree corresponding to paths $X \rightarrow Y_i$ we observe two facts: the labeled set V is prefix-free (two labeled vertices are never in ancestral relation) and the depth of each labeled set is at least d_i . It is easy to see that $\sum_{u \in V} c^{\text{depth}(u)} \leq (Kc)^{d_i}$ provided $Kc \leq 1$ and attained by taking V to be set of all vertices in the tree at depth d_i . We conclude that whenever $K(1 - 2\delta)^2 \leq 1$ the right-hand side of (33.13) is bounded by $(K(1 - 2\delta)^2)^{d_i}$, which concludes the proof by upper bounding $H(X_i) \leq \log 2$ as

$$I(X_i; Y) \leq \eta_{\text{KL}}(P_{Y|X_i})H(X_i) \leq K^{d_i}(1 - 2\delta)^{2d_i} \log 2$$

\square

We conclude the section with an example illustrating that Theorem 33.9 may give stronger bounds when compared to Theorem 33.2.

Example 33.6. Suppose we have the topological restriction on the placement of gates (namely that the inputs to each gate should be from nearest neighbors to the left), resulting in the following circuit of 2-input δ -noisy gates. Note that each gate may be a simple passthrough (i.e. serve as



router) or a constant output. Theorem 33.2 states that if $(1 - 2\delta)^2 < \frac{1}{2}$, then noisy computation

within arbitrary topology is not possible. Theorem 33.9 improves this to $(1 - 2\delta)^2 < p_c$, where p_c is the oriented site-percolation threshold for the particular graph we have. Namely, if each vertex is open with probability $p < p_c$ then with probability 1 the connected component emanating from any given node (and extending to the right) is finite. For the example above the site percolation threshold is estimated as $p_c \approx 0.705$ (so called Stavskaya automata).

33.4 Input-dependent SDPI

Previously we have defined contraction coefficient $\eta_f(P_{Y|X})$, as the maximum contraction of an f -divergences over all channel input distributions. We now define an analogous concept for a specific input distribution P_X .

Definition 33.9 (Input Dependent Contraction Coefficient). For any input distribution P_X , Markov kernel $P_{Y|X}$ and convex function f , we define

$$\eta_f(P_X, P_{Y|X}) \triangleq \sup \frac{D_f(Q_Y \| P_Y)}{D_f(Q_X \| P_X)}$$

where $P_Y = P_{Y|X} \circ P_X$, $Q_Y = P_{Y|X} \circ Q_X$ and supremum is over Q_X satisfying $0 < D_f(P_X \| Q_X) < \infty$.

We refer to $\eta_f(P_X, P_{Y|X})$ as the input dependent contraction coefficient, to contrast it with the input independent contraction coefficient $\eta_f(P_{Y|X})$.

Remarks:

- As for $\eta_{KL}(P_{Y|X})$, we also have a corresponding mutual information characterization of $\eta_{KL}(P_X, P_{Y|X})$ as

$$\eta_{KL}(P_X, P_{Y|X}) = \sup_{P_{U|X}: U \rightarrow X \rightarrow Y} \frac{I(U; Y)}{I(U; X)}.$$

- From the definition, the following inequality holds

$$\eta_f(P_X, P_{Y|X}) \leq \eta_f(P_{Y|X}).$$

- Although we have the equality $\eta_{KL}(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X})$ when $P_{Y|X}$ is a BMS channel, we do not have the same equality for $\eta_{KL}(P_X, P_{Y|X})$.

Example 33.7. ($\eta_{KL}(P_X, P_{Y|X})$ for Erasure Channel) We define EC_τ as the following channel,

$$Y = \begin{cases} X & \text{w.p. } 1 - \tau \\ ? & \text{w.p. } \tau. \end{cases}$$

Let us define an auxiliary random variable $B = 1\{Y = ?\}$. Thus we have the following equality,

$$I(U; Y) = I(U; Y, B) = \underbrace{I(U; B)}_{0, B \perp U} + I(U; Y|B) = (1 - \tau)I(U; X).$$

where the last equality is due to the fact that $I(U; Y|B = 1) = 0$ and $I(U; Y|B = 0) = I(U; X)$. By the mutual information characterization of $\eta_{\text{KL}}(P_X, P_{Y|X})$, we have $\eta_{\text{KL}}(P_X, \text{EC}_\tau) = 1 - \tau$.

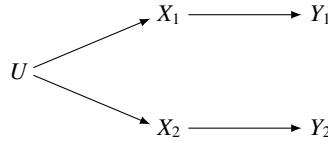
Proposition 33.10 (Tensorization of η_{KL}). *For a given number n , two measures P_X and $P_{Y|X}$ we have*

$$\eta_{\text{KL}}(P_X^{\otimes n}, P_{Y|X}^{\otimes n}) = \eta_{\text{KL}}(P_X, P_{Y|X})$$

In particular, if $(X_i, Y_i) \stackrel{i.i.d.}{\sim} P_{X,Y}$, then $\forall P_{U|X^n}$

$$I(U; Y^n) \leq \eta_{\text{KL}}(P_X, P_{Y|X})I(U; X^n)$$

Proof. Without loss of generality (by induction) it is sufficient to prove the proposition for $n = 2$. It is always useful to keep in mind the following diagram Let $\eta = \eta_{\text{KL}}(P_X, P_{Y|X})$



$$\begin{aligned} I(U; Y_1, Y_2) &= I(U; Y_1) + I(U; Y_2 | Y_1) \\ &\leq \eta [I(U; X_1) + I(U; X_2 | Y_1)] \tag{33.14} \\ &= \eta [I(U; X_1) + I(U; X_2 | X_1) + I(U; X_1 | Y_1) - I(U; X_1 | Y_1, X_2)] \tag{33.15} \\ &\leq \eta [I(U; X_1) + I(U; X_2 | Y_1)] \tag{33.16} \\ &= \eta I(U; X_1, X_2) \end{aligned}$$

where (33.14) is due to the fact that conditioned on Y_1 , $U - X_2 - Y_2$ is still a Markov chain, (33.15) is because $U - X_1 - Y_1$ is a Markov chain and (33.16) follows from the fact that $X_2 - U - X_1$ is a Markov chain even when condition Y_1 . \square

33.5 Application: Broadcasting and coloring on trees

Consider an infinite b -ary tree $G = (\mathcal{V}, \mathcal{E})$. We assign a random variable X_v for each $v \in \mathcal{V}$. These random variables X_v 's are defined on the same alphabet \mathcal{X} . In this model, the joint distribution is induced by the distribution on the root vertex π , i.e., $X_\rho \sim \pi$, and the edge kernel $P_{X'|X}$, i.e. $\forall (p, c) \in \mathcal{E}, P_{X_c|X_p} = P_{X'|X}$.

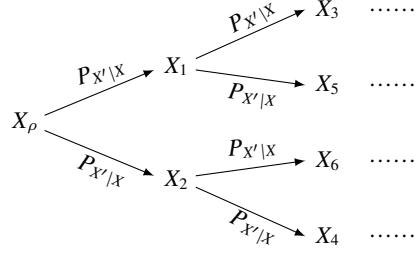
To simplify our discussion, we will assume that π is a reversible measure on kernel $P_{X'|X}$, i.e.,

$$P_{X'|X}(a|b)\pi(b) = P_{X'|X}(b|a)\pi(a). \tag{33.17}$$

By standard result on Markov chain, this also implies that π is a stationary distribution of the reversed Markov kernel $P_{X|X'}$.

We make the following observations:

33.5 Application: Broadcasting and coloring on trees 557



- We can think of this model as a broadcasting scenario, where the root broadcasts its message X_ρ to the leaves through noisy channels $P_{X'|X}$. The condition (33.17) here is only made to avoid defining the reverse channel. In general, one only requires that π is a stationary distribution of $P_{X'|X}$, in which case the (33.19) should be replaced with $\eta_{\text{KL}}(\pi, P_{X|X'})b < 1$.
- This model arises frequently in community detection, sparse codes and statistical physics.
- Under the assumption (33.17), the joint distribution of this tree can also be written as a Gibbs distribution

$$P_{X_{\text{all}}} = \frac{1}{Z} \exp \left(\sum_{(p,c) \in \mathcal{E}} f(X_p, X_c) + \sum_{v \in \mathcal{V}} g(X_v) \right), \quad (33.18)$$

where Z is the normalization constant, $f(x_p, x_c) = f(x_c, x_p)$ is symmetric. When $\mathcal{X} = \{0, 1\}$, this model is known as the Ising model (on a tree). Note, however, that not every measure factorizing as (33.18) (with symmetric f) can be written as a broadcasting process for some P and π .

We can define a corresponding inference problem, where we want to reconstruct the root variable X_ρ given the observations $X_{L_d} = \{X_v : v \in L_d\}$, with $L_d = \{v : v \in \mathcal{V}, \text{depth}(v) = d\}$. A natural question is to upper bound the performance of any inference algorithm on this problem. The following theorem shows that there exists a phase transition depending on the branching factor b and the contraction coefficient of the kernel $P_{X'|X}$.

Theorem 33.11. Consider the broadcasting problem on infinite b -ary tree ($b > 1$), with root distribution π and edge kernel $P_{X'|X}$. If π is a reversible measure of $P_{X'|X}$ such that

$$\eta_{\text{KL}}(\pi, P_{X'|X})b < 1, \quad (33.19)$$

then $I(X_\rho; X_{L_d}) \rightarrow 0$ as $d \rightarrow 0$.

Proof. For every $v \in L_1$, we define the set $L_{d,v} = \{u : u \in L_d, v \in \text{ancestor}(u)\}$. We can upper bound the mutual information between the root vertex and leaves at depth d

$$I(X_\rho; X_{L_d}) \leq \sum_{v \in L_1} I(X_\rho; X_{L_{d,v}}).$$

For each term in the summation, we consider the Markov chain

$$X_{L_{d,v}} \rightarrow X_v \rightarrow X_\rho.$$

Due to our assumption on π and $P_{X'|X}$, we have $P_{X_\rho|X_v} = P_{X'|X}$ and $P_{X_v} = \pi$. By the definition of the contraction coefficient, we have

$$I(X_{L_{d,v}}; X_\rho) \leq \eta_{\text{KL}}(\pi, P_{X'|X}) I(X_{L_{d,v}}; X_v).$$

Observe that because $P_{X_v} = \pi$ and all edges have the same kernel, then $I(X_{L_{d,v}}; X_v) = I(X_{L_{d-1}}; X_\rho)$. This gives us the inequality

$$I(X_\rho; X_{L_d}) \leq \eta_{\text{KL}}(\pi, P_{X'|X}) b I(X_\rho; X_{L_{d-1}}),$$

which implies

$$I(X_\rho; X_{L_d}) \leq (\eta_{\text{KL}}(\pi, P_{X'|X}) b)^d H(X_\rho).$$

Therefore if $\eta_{\text{KL}}(\pi, P_{X'|X}) b < 1$ then $I(X_\rho; X_{L_d}) \rightarrow 0$ exponentially fast as $d \rightarrow \infty$. \square

Note that a weaker version of this theorem (non-reconstruction when $\eta_{\text{KL}}(P_{X'|X}) b \leq 1$) is implied by the directed information percolation theorem. The k -coloring example (see below) demonstrates that this strengthening is essential; see [145] for details.

Example 33.8. (Broadcasting on BSC tree.) Consider a broadcasting problem on b -ary tree with vertex alphabet $\mathcal{X} = \{0, 1\}$, edge kernel $P_{X'|X} = \text{BSC}_\delta$, and $\pi = \text{Unif}$. Note that uniform distribution is a reversible measure for BSC_δ . In Example 33.3, we calculated $\eta_{\text{KL}}(\text{BSC}_\delta) = (1 - 2\delta)^2$. Therefore, using theorem 33.12, we can deduce that if

$$b(1 - 2\delta)^2 < 1$$

then no inference algorithm can recover the root nodes as depth of the tree goes to infinity. This result is originally proved in [39].

Example 33.9 (k -coloring on tree). Given a b -ary tree, we assign a k -coloring $X_{v_{\text{all}}}$ by sampling uniformly from the ensemble of all valid k -coloring. For this model, we can define a corresponding inference problem, namely given all the colors of the leaves at a certain depth, i.e., X_{L_d} , determine the color of the root node, i.e., X_ρ .

This problem can be modeled as a broadcasting problem on tree where the root distribution π is given by the uniform distribution on k colors, and the edge kernel $P_{X'|X}$ is defined as

$$P_{X'|X}(a|b) = \begin{cases} \frac{1}{k-1} & a \neq b \\ 0, & a = b. \end{cases}$$

It can be shown, see Ex. VI.23, that $\eta_{\text{KL}}(\text{Unif}, P_{X'|X}) = \frac{1}{k \log k(1+o(1))}$. By Theorem 33.12, this implies that if $b < k \log k(1+o(1))$ then reliable reconstruction of the root node is not possible. This result is originally proved in [278] and [32].

The other direction $b > k \log k(1+o(1))$ can be shown by observing that if $b > k \log k(1+o(1))$ then the probability of the children of a node taking all available colors (except its own) is close to 1. Thus, an inference algorithm can always determine the color of a node by finding a color that is not assigned to any of its children. Similarly, when $b > (1 + \epsilon)k \log k$ even observing $(1 - \epsilon)$ -fraction of the node's children is sufficient to reconstruct its color exactly. Proceeding recursively

33.6 Application: distributed correlation estimation 559

from bottom up, such a reconstruction algorithm will succeed with high probability. In this regime with positive probability (over the leaf variables) the posterior distribution of the root color is a delta-function (deterministic). This effect is known as “freezing” of the root given the boundary.

33.6 Application: distributed correlation estimation

Tensorization property can be used for correlation estimation. Suppose Alice have samples $\{X_i\}_{i \geq 1} \stackrel{\text{i.i.d.}}{\sim} B(1/2)$ and Bob have samples $\{Y_i\}_{i \geq 1} \stackrel{\text{i.i.d.}}{\sim} B(1/2)$ such that the (X_i, Y_i) are i.i.d. with $\mathbb{E}[X_i Y_i] = \rho \in [-1, 1]$. The goal is for Bob to send W to Alice with $H(W) = B$ bits and for Alice to estimate $\hat{\rho} = \hat{\rho}(X^\infty, W)$ with objective

$$R^*(B) = \inf_{W, \hat{\rho}} \sup_{\rho} \mathbb{E}[(\rho - \hat{\rho})^2].$$

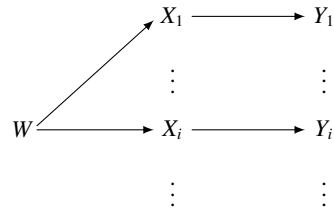
Notice that in this problem we are not sample-limited (each party has infinitely many samples), but communication-limited (only B bits can be exchanged).

Here is a trivial attempt to solve it. Notice that if Bob sends $W = (Y_1, \dots, Y_B)$ then the optimal estimator is $\hat{\rho}(X^\infty, W) = \frac{1}{n} \sum_{i=1}^B X_i Y_i$ which has minimax error $\frac{1}{B}$, hence $R^*(B) \leq \frac{1}{B}$. Surprisingly, this can be improved.

Theorem 33.12 ([146]). *The optimal rate when $B \rightarrow \infty$ is given by*

$$R^*(X^\infty, W) = \frac{1 + o(1)}{2 \ln 2} \cdot \frac{1}{B}$$

Proof. Fix $P_{W|Y^\infty}$, we get the following decomposition Note that once the messages W are fixed



we have a parameter estimation problem $\{Q_\rho, \rho \in [-1, 1]\}$ where Q_ρ is a distribution of (X^∞, W) when A^∞, B^∞ are ρ -correlated. Since we minimize MMSE, we know from the van Trees inequality (Theorem 29.2)² that $R^*(B) \geq \frac{1+o(1)}{\min_\rho J_F(\rho)} \geq \frac{1+o(1)}{J_F(0)}$ where $J_F(\rho)$ is the Fisher Information of the family $\{Q_\rho\}$.

Recall, that we also know from the local approximation that

$$D(Q_\rho \| Q_0) = \frac{\rho^2 \log e}{2} J_F(0) + o(\rho^2)$$

² This requires some technical justification about smoothness of the Fisher information $J_F(\rho)$.

Furthermore, notice that under $\rho = 0$ we have X^∞ and W independent and thus

$$\begin{aligned} D(Q_\rho \| Q_0) &= D(P_{X^\infty, W}^\rho \| P_{X^\infty, W}^0) \\ &= D(P_{X^\infty, W}^\rho \| P_{X^\infty}^\rho \times P_W^\rho) \\ &= I(W; X^\infty) \\ &\leq \rho^2 I(W; Y^\infty) \\ &\leq \rho^2 B \log 2 \end{aligned}$$

hence $J_F(0) \leq (2 \ln 2)B + o(1)$ which in turns implies the theorem. For full details and the extension to interactive communication between Alice and Bob see [146].

We comment on the upper bound next. First, notice that by taking blocks of $m \rightarrow \infty$ consecutive bits and setting $\tilde{X}_i = \frac{1}{\sqrt{m}} \sum_{j=(i-1)m}^{im-1} X_j$ and similarly for \tilde{Y}_i , Alice and Bob can replace ρ -correlated bits with ρ -correlated standard Gaussians $(\tilde{X}_i, \tilde{Y}_i) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$. Next, fix some very large N and let

$$W = \underset{1 \leq j \leq N}{\operatorname{argmax}} Y_j.$$

From standard concentration results we know that $\mathbb{E}[Y_W] = \sqrt{2 \ln N}(1 + o(1))$ and $\operatorname{Var}[Y_W] = O(\frac{1}{\ln N})$. Therefore, knowing W Alice can estimate

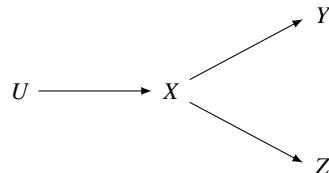
$$\hat{\rho} = \frac{X_W}{\mathbb{E}[Y_W]}.$$

This is an unbiased estimator and $\operatorname{Var}_\rho[\hat{\rho}] = \frac{1-\rho^2+o(1)}{2 \ln N}$. Finally, setting $N = 2^B$ completes the argument. \square

33.7 Channel comparison: degradation, less noisy, more capable

It turns out that the η_{KL} coefficient is intimately related to the concept of less noisy partial order on channels. We define several such partial orders together.

Definition 33.13 (partial orders on channels). Let $P_{Y|X}$ and $P_{Z|X}$ be two channels. We say that $P_{Y|X}$ is a degradation of $P_{Z|X}$, denoted $P_{Y|X} \leq_{\text{deg}} P_{Z|X}$, if there exists $P_{Y|Z}$ such that $P_{Y|X} = P_{Y|Z} \circ P_{Z|X}$. We say that $P_{Z|X}$ is less noisy than $P_{Y|X}$, $P_{Y|X} \leq_{\text{ln}} P_{Z|X}$, iff for every $P_{U,X}$ on the following Markov chain we have $I(U; Y) \leq I(U; Z)$. We say that $P_{Z|X}$ is more capable than $P_{Y|X}$, denoted $P_{Y|X} \leq_{\text{mc}} P_{Z|X}$ if



33.7 Channel comparison: degradation, less noisy, more capable 561

for any P_X we have $I(X; Y) \leq I(X; Z)$.

We make some remarks, see [241] for proofs:

- $P_{Y|X} \leq_{\text{deg}} P_{Z|X} \implies P_{Y|X} \leq_{\text{ln}} P_{Z|X} \implies P_{Y|X} \leq_{\text{mc}} P_{Z|X}$. Counter examples for reverse implications can be found in [80, Problem 15.11].
- For less noisy we also have the equivalent definition in terms of the divergence, namely $P_{Y|X} \leq_{\text{ln}} P_{Z|X}$ if and only if for all P_X, Q_X we have $D(Q_Y||P_Y) \leq D(Q_Z||P_Z)$. We refer to [203, Sections I.B, II.A] and [241, Section 6] for alternative useful characterizations of the less-noisy order.
- For BMS channels (see Section 19.4*) it turns out that among all channels with a given $I_{\chi^2}(X; Y) = \eta$ (with $X \sim \text{Ber}(1/2)$) the BSC and BEC are the minimal and maximal elements in the poset of \leq_{ln} ; see Ex. VI.20 for details.

Proposition 33.14. $\eta_{\text{KL}}(P_{Y|X}) \leq 1 - \tau$ if and only if $P_{Y|X} \leq_{LN} \text{EC}_\tau$, where EC_τ was defined in Example 33.7.

Proof. For EC_τ we always have

$$I(U; Z) = (1 - \tau)I(U; X).$$

By the mutual information characterization of η_{KL} we have,

$$I(U; Y) \leq (1 - \tau)I(U; X).$$

Combining these two inequalities gives us

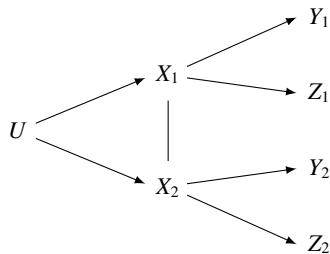
$$I(U; Y) \leq I(U; Z).$$

□

This proposition gives us an intuitive interpretation of contraction coefficient as the worst erasure channel that still dominates the channel.

Proposition 33.15. (Tensorization of Less Noisy Ordering) If for all $i \in [n]$, $P_{Y_i|X_i} \leq_{LN} P_{Z_i|X_i}$, then $P_{Y_1|X_1} \otimes P_{Y_2|X_2} \leq_{LN} P_{Z_1|X_1} \otimes P_{Z_2|X_2}$. Note that $P \otimes Q$ refers to the product channel of P and Q .

Proof. Consider the following Markov chain.



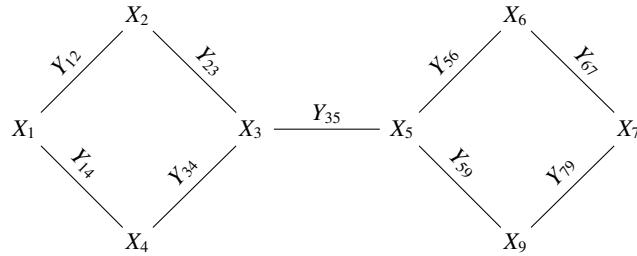
It can be seen from the Markov chain that $I(U; Y_1, Y_2) \leq I(U; Y_1, Z_2)$ implies $I(U; Y_1, Y_2) \leq I(U; Z_1, Z_2)$. Consider the following inequalities,

$$\begin{aligned} I(U; Y_1, Y_2) &= I(U; Y_1) + I(U; Y_2|Y_1) \\ &\leq I(U; Y_1) + I(U; Z_2|Y_1) \\ &= I(U; Y_1, Z_2). \end{aligned}$$

Hence $I(U; Y_1, Y_2) \leq I(U; Y_1, Z_2)$ for any $P_{X_1, X_2, U}$. \square

33.8 Undirected information percolation

In this section we will study the problem of inference on undirected graph. Consider an undirected graph $G = (\mathcal{V}, \mathcal{E})$. We assign a random variable X_v on the alphabet \mathcal{X} to each vertex v . For each $e = (u, v) \in \mathcal{E}$, we assign Y_e sampled according to the kernel $P_{Y_e|X_e}$ with $X_e = (X_u, X_v)$. The goal of this inference model is to determine the value of X_v 's given the value of Y_e 's.



Example 33.10 (Community Detection). In this model, we consider a complete graph with n vertices, i.e. K_n , and the random variables X_v representing the membership of each vertex to one of the m communities. We assume that X_v is sampled uniformly from $[m]$ and independent of the other vertices. The observation $Y_{u,v}$ is defined as

$$Y_{uv} \sim \begin{cases} \text{Ber}(a/n) & X_u = X_v \\ \text{Ber}(b/n) & X_u \neq X_v. \end{cases}$$

Example 33.11 (\mathbb{Z}_2 Synchronization). For any graph G , we sample X_v uniformly from $\{-1, +1\}$ and $Y_e = \text{BSC}_\delta(X_u X_v)$.

Example 33.12 (Spiked Wigner Model). We consider the inference problem of determining the value of vector $(X_i)_{i \in [n]}$ given the observation $(Y_{ij})_{i,j \in [n], i \leq j}$. The X_i 's and Y_{ij} 's are related by a linear model

$$Y_{ij} = \sqrt{\frac{\lambda}{n}} X_i X_j + W_{ij},$$

33.8 Undirected information percolation 563

where X_i is sampled uniformly from $\{-1, +1\}$ and $W_{ij} \sim N(0, 1)$. This model can also be written in matrix form as

$$\mathbf{Y} = \sqrt{\frac{\lambda}{n}} \mathbf{X} \mathbf{X}^T + \mathbf{W}$$

where \mathbf{W} is the Wigner matrix, hence the name of the model. It is used as a probabilistic model for principle component analysis (PCA).

This problem can also be treated as a problem of inference on undirected graph. In this case, the underlying graph is a complete graph, and we assign X_i to the i th vertex. Under this model, the edge observations is given by $Y_{ij} = \text{BIAWGN}_{\lambda/n}(X_i X_j)$.

Although seemingly different, these problems share similar characteristics, namely:

- X_i 's are uniformly distributed,
- If we define an auxiliary random variable $B = 1\{X_u \neq X_v\}$ for any edges $e = (u, v)$, then the following Markov chain holds

$$(X_u, X_v) \rightarrow B \rightarrow Y_e.$$

In other words, the observation on each edge only depends on whether the random variables on its endpoints are similar.

We will refer to the problem which have this characteristics as the Special Case (S.C.). Due to S.C., the reconstructed X_v 's is symmetric up to any permutation on \mathcal{X} . In the case of alphabet $\mathcal{X} = \{-1, +1\}$, this implies that for any realization σ then $P_{X_{\text{all}}|Y_{\text{all}}}(\sigma|b) = P_{X_{\text{all}}|Y_{\text{all}}}(-\sigma|b)$. Consequently, our reconstruction metric also needs to accommodate this symmetry. For $\mathcal{X} = \{-1, +1\}$, this leads to the use of $\frac{1}{n} |\sum_{i=1}^n X_i \hat{X}_i|$ as our reconstruction metric.

Our main theorem for undirected inference problem can be seen as the analogue of the information percolation theorem for DAG. However, instead of controlling the contraction coefficient, the percolation probability is used to directly control the conditional mutual information between any subsets of vertices in the graph.

Before stating our main theorem, we will need to define the corresponding percolation model for inference on undirected graph. For any undirected graph $G = (\mathcal{V}, \mathcal{E})$ we define a percolation model on this graph as follows :

- Every edge $e \in \mathcal{E}$ is open with the probability $\eta_{\text{KL}}(P_{Y_e|X_e})$, independent of the other edges,
- For any $v \in \mathcal{V}$ and $S \subset \mathcal{V}$, we define the $v \leftrightarrow S$ as the event that there exists an open path from v to any vertex in S ,
- For any $S_1, S_2 \subset \mathcal{V}$, we define the function $\text{perc}_u(S_1, S_2)$ as

$$\text{perc}_u(S_1, S_2) \triangleq \sum_{v \in S_1} P(v \leftrightarrow S_2).$$

Notice that this function is different from the percolation function for information percolation in DAG. Most importantly, this function is not equivalent to the exact percolation probability.

Instead, it is an upper bound on the percolation probability by union bounding with respect to S_1 . Hence, it is natural that this function is not symmetric, i.e. $\text{perc}_u(S_1, S_2) \neq \text{perc}_u(S_2, S_1)$.

Theorem 33.16 (Undirected information percolation [243]). *Consider an inference problem on undirected graph $G = (\mathcal{V}, \mathcal{E})$. For any $S_1, S_2 \subset \mathcal{V}$, then*

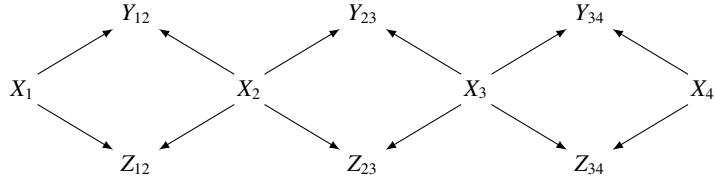
$$I(X_{S_1}; X_{S_2} | Y) \leq \text{perc}_u(S_1, S_2) \log |\mathcal{X}|.$$

Instead of proving theorem 33.17 in its full generality, we will prove the theorem under S.C. condition. The main step of the proof utilizes the fact we can upper bound the mutual information of any channel by its less noisy upper bound.

Theorem 33.17. *Consider the problem of inference on undirected graph $G = (\mathcal{V}, \mathcal{E})$ with X_1, \dots, X_n are not necessarily independent. If $P_{Y_e|X_e} \leq_{LN} P_{Z_e|X_e}$, then for any $S_1, S_2 \subset \mathcal{V}$ and $E \subset \mathcal{E}$*

$$I(X_{S_1}; Y_E | X_{S_2}) \leq I(X_{S_1}; Z_E | X_{S_2})$$

Proof. Consider the following Markov chain.



From our assumption and the tensorization property of less noisy ordering (Proposition 33.16), we have $P_{Y_E|X_{S_1}, X_{S_2}} \leq_{LN} P_{Z_E|X_{S_1}, X_{S_2}}$. This implies that for σ as a valid realization of X_{S_2} we will have

$$I(X_{S_1}; Y_E | X_{S_2} = \sigma) = I(X_{S_1}, X_{S_2}; Y_E | X_{S_2} = \sigma) \leq I(X_{S_1}, X_{S_2}; Z_E | X_{S_2} = \sigma) = I(X_{S_1}; Z_E | X_{S_2} = \sigma).$$

As this inequality holds for all realization of X_{S_2} , then the following inequality also holds

$$I(X_{S_1}; Y_E | X_{S_2}) \leq I(X_{S_1}; Z_E | X_{S_2}).$$

□

Proof of Theorem 33.17. We only give a proof under the S.C. condition above and only for the case $S_1 = \{i\}$. For the full proof (that proceeds by induction and does not leverage the less noisy idea), see [243]. We have the following equalities

$$I(X_i; X_{S_2} | Y_E) = I(X_i; X_{S_2}, Y_E) = I(X_i; Y_E | X_{S_2}) \tag{33.20}$$

where the first inequality is due to the fact $B_E \perp\!\!\!\perp X_i$ under S.C, and the second inequality is due to $X_i \perp\!\!\!\perp X_{S_2}$ under S.C.

33.9 Application: Spiked Wigner model 565

Due to our previous result, if $\eta_{\text{KL}}(P_{Y_e|X_e}) = 1 - \tau$ then $P_{Y_e|X_e} \leq_{\text{LN}} P_{Z_e|X_e}$ where $P_{Z_e|X_e} = \mathbb{E} C_\tau$. By tensorization property, this ordering also holds for the channel $P_{Y_E|X_E}$, thus we have

$$I(X_i; Y_E|X_{S_2}) \leq I(X_j; Z_E|X_{S_2}).$$

Let us define another auxiliary random variable $D = 1\{i \leftrightarrow S_2\}$, namely it is the indicator that there is an open path from i to S_2 . Notice that D is fully determined by Z_E . By the same argument as in (33.20), we have

$$\begin{aligned} I(X_i; Z_E|X_{S_2}) &= I(X_i; X_{S_2}|Z_E) \\ &= I(X_i; X_{S_2}|Z_E, D) \\ &= (1 - \mathbb{P}[i \leftrightarrow S_2]) \underbrace{I(X_i; X_{S_2}|Z_E, D = 0)}_0 + \mathbb{P}[i \leftrightarrow S_2] \underbrace{I(X_i; X_{S_2}|Z_E, D = 1)}_{\leq \log |\mathcal{X}|} \\ &\leq \mathbb{P}[i \leftrightarrow S_2] \log |\mathcal{X}| \\ &= \text{perc}_u(i, S_2) \end{aligned}$$

□

33.9 Application: Spiked Wigner model

The following theorem shows how the undirected information percolation concept allows us to derive a converse result for spiked Wigner model, which we described in Example 33.12.

Theorem 33.18. *Consider the spiked Wigner model. If $\lambda \leq 1$, then for any sequence of estimators $\hat{X}^n(Y)$,*

$$\frac{1}{n} \mathbb{E} \left[\left| \sum_{i=1}^n X_i \hat{X}_i \right| \right] \rightarrow 0 \quad (33.21)$$

as $n \rightarrow \infty$.

Proof of Theorem 33.19. Note that by $\mathbb{E}[|T|] \leq \sqrt{\mathbb{E}[T^2]}$ the left-hand side of (33.21) can be upper-bounded by

$$\frac{1}{n} \sqrt{\mathbb{E} \left[\sum_{i,j} X_i X_j \hat{X}_i \hat{X}_j \right]}.$$

Next, it is clear that we can simplify the task of maximizing (over \hat{X}^n) by allowing to separately estimate each product by $\hat{T}_{i,j}$, i.e.

$$\max_{\hat{X}^n} \mathbb{E} \left[\sum_{i,j} X_i X_j \hat{X}_i \hat{X}_j \right] \leq \sum_{i,j} \max_{\hat{T}_{i,j}} \mathbb{E} [X_i X_j \hat{T}_{i,j}] .$$

The latter maximization is easy to solve:

$$\hat{T}_{i,j}(Y) = \operatorname{argmax}_{\sigma} \mathbb{P}[X_i X_j = \sigma | Y].$$

Since each $X_i \sim \text{Ber}(1/2)$ it is easy to see that

$$I(X_i; X_j | Y) \rightarrow 0 \iff \max_{\hat{T}_{i,j}} \mathbb{E}[X_i X_j \hat{T}_{i,j}] \rightarrow 0.$$

(For example, we may notice $I(X_i; X_j | Y) = I(X_i, X_j; Y) \geq I(X_i X_j; Y)$ and apply Fano's inequality). Thus, from symmetry of the problem it is sufficient to prove $I(X_1; X_2 | Y) \rightarrow 0$ as $n \rightarrow \infty$.

By using the undirected information percolation theorem, we have

$$I(X_2; X_1 | Y) \leq \text{perc}_u(\{1\}, \{2\})$$

in which the percolation model is defined on a complete graph with edge probability $\frac{\lambda+o(1)}{n}$ as $\eta_{KL}(\text{BIAWGN}_{\lambda/n}) = \frac{\lambda}{n}(1+o(1))$. We only treat the case of $\lambda < 1$ below. For such λ we can over-bound $\frac{\lambda+o(1)}{n}$ by $\frac{\lambda'}{n}$ with $\lambda' < 1$. This percolation random graph is equivalent to the Erdős-Rényi random graph with n vertices and λ'/n edge probability, i.e., $ER(n, \lambda'/n)$. Using this observation, the inequality can be rewritten as

$$I(X_2; X_1 | Y) \leq P(\text{Vertex 1 and 2 is connected on } ER(n, \lambda'/n)).$$

The largest components on $ER(n, \lambda'/n)$ contains $O(\log n)$ if $\lambda' < 1$. This implies that the probability that two specific vertices are connected is $o(1)$, hence $I(X_2; X_1 | Y) \rightarrow 0$ as $n \rightarrow \infty$. To treat the case of $\lambda = 1$ we need slightly more refined information about behavior of giant component of $ER(n, \frac{1+o(1)}{n})$ graph, see [243]. \square

Remark 33.2 (Dense-Sparse equivalence). This reduction changes the underlying structure of the graph. Instead of dealing with a complete graph, the percolation problem is defined on an Erdős-Rényi random graph. Moreover, if η_{KL} is small enough, then the underlying percolation graph tends to have a locally tree-like structure. This is demonstration of the ubiquitous effect: dense inference (such as spiked Wigner or sparse regression) with very weak signals ($\eta_{KL} \approx 1$) is similar to sparse inference (broadcasting on trees) with moderate signals ($\eta_{KL} \in (\epsilon, 1 - \epsilon)$).

33.10 Strong data post-processing inequality (Post-SDPI)

We introduce the following version of the SDPI constant.

Definition 33.19 (Post-SDPI constant). Given a conditional measure $P_{Y|X}$, define the input-dependent and input-free contraction coefficients as

$$\begin{aligned} \eta_{KL}^{(p)}(P_X, P_{Y|X}) &= \sup_{P_{U|Y}} \left\{ \frac{I(U; X)}{I(U; Y)} : X \rightarrow Y \rightarrow U \right\} \\ \eta_{KL}^{(p)}(P_{Y|X}) &= \sup_{P_X, P_{U|Y}} \left\{ \frac{I(U; X)}{I(U; Y)} : X \rightarrow Y \rightarrow U \right\} \end{aligned}$$

33.10 Strong data post-processing inequality (Post-SDPI) 567

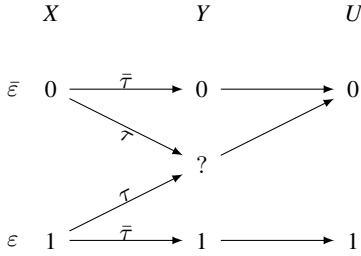


Figure 33.3 Post-SDPI coefficient of BEC equals to 1.

To get characterization in terms of KL-divergence we simply notice that

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) = \eta_{\text{KL}}(P_Y, P_{X|Y}) \quad (33.22)$$

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) = \sup_{P_X} \eta_{\text{KL}}(P_Y, P_{X|Y}), \quad (33.23)$$

where $P_Y = P_{Y|X} \circ P_X$ and $P_{X|Y}$ is the conditional measure corresponding to $P_X P_{Y|X}$. From (33.22) and Prop. 33.11 we also get tensorization property for input-dependent post-SDPI:

$$\eta_{\text{KL}}^{(p)}(P_X^n, (P_{Y|X})^n) = \eta_{\text{KL}}^{(p)}(P_X^n, (P_{Y|X})^n) \quad (33.24)$$

It is easy to see that by the data processing inequality, $\eta_{\text{KL}}^{(p)}(P_{Y|X}) \leq 1$. Unlike the η_{KL} coefficient the $\eta_{\text{KL}}^{(p)}$ can equal to 1 even for a noisy channel $P_{Y|X}$.

Example 33.13 ($\eta_{\text{KL}}^{(p)} = 1$). Let $P_{Y|X} = \text{BEC}_\tau$ and $X \rightarrow Y \rightarrow U$ be defined as on Fig. 33.3 Then we can compute $I(Y; U) = H(U) = h(\varepsilon\bar{\tau})$ and $I(X; U) = H(U) - H(U|X) = h(\varepsilon\bar{\tau}) - \varepsilon h(\tau)$ hence

$$\begin{aligned} \eta_{\text{KL}}^{(p)}(P_{Y|X}) &\geq \frac{I(X; U)}{I(Y; U)} \\ &= 1 - h(\tau) \frac{\varepsilon}{h(\varepsilon\bar{\tau})} \end{aligned}$$

This last term tends to 1 when ε tends to 0 hence

$$\eta_{\text{KL}}^{(p)}(\text{BEC}_\tau) = 1$$

even though Y is not a one to one function of X .

Note that this example also shows that even for an input-constrained version of $\eta_{\text{KL}}^{(p)}$ the natural conjecture $\eta_{\text{KL}}^{(p)}(\text{Unif}, \text{BMS}) = \eta_{\text{KL}}(\text{BMS})$ is *incorrect*. Indeed, by taking $\varepsilon = \frac{1}{2}$, we have that $\eta_{\text{KL}}^{(p)}(\text{Unif}, \text{BEC}_\tau) > 1 - \tau$ for $\tau \rightarrow 1$.

Nevertheless, the post-SDPI constant is often non-trivial, most importantly for the BSC:

Theorem 33.20.

$$\eta_{\text{KL}}^{(p)}(\text{BSC}_\delta) = (1 - 2\delta)^2$$

To prove this theorem, the following lemma is useful.

Lemma 33.21. *If for any X and Y in $\{0, 1\}$ we have*

$$p_{X,Y}(x, y) = f(x) \left(\frac{\delta}{1 - \delta} \right)^{1(x \neq y)} g(Y)$$

for some functions f and g , then $\eta_{\text{KL}}(P_{Y|X}) \leq (1 - 2\delta)^2$.

Proof. It is known that for binary input channels $P_{Y|X}$ [241]

$$\eta_{\text{KL}}(P_{Y|X}) \leq H^2(P_{Y|X=0} \| P_{Y|X=1}) - \frac{H^4(P_{Y|X=0} \| P_{Y|X=1})}{4}$$

If we let $\phi = \frac{g(0)}{g(1)}$, then we have $p_{Y|X=0} = B\left(\frac{\lambda}{\phi + \lambda}\right)$ and $p_{Y|X=1} = B\left(\frac{1}{1 + \phi\lambda}\right)$ and a simple check shows that

$$\begin{aligned} \max_{\phi} H^2(P_{Y|X=0} \| P_{Y|X=1}) - \frac{H^4(P_{Y|X=0} \| P_{Y|X=1})}{4} &\stackrel{\phi=1}{=} H_{\phi=1}^2(P_{Y|X=0} \| P_{Y|X=1}) - \frac{H_{\phi=1}^4(P_{Y|X=0} \| P_{Y|X=1})}{4} \\ &= (1 - 2\delta)^2 \end{aligned}$$

Now observe that $P_{X,Y}$ in Theorem 33.22 satisfies the property of the lemma with X and Y exchanged, hence $\eta_{\text{KL}}(P_Y, P_{X|Y}) \leq (1 - 2\delta)^2$ which implies that $\eta_{\text{KL}}^{(p)}(P_{Y|X}) = \sup_{P_X} \eta_{\text{KL}}(P_Y, P_{X|Y}) \leq (1 - 2\delta)^2$ with equality if P_X is uniform. \square

Theorem 33.22 (Post-SDPI for BI-AWGN). *Let $0 \leq \epsilon \leq 1$ and consider the channel $P_{Y|X}$ with $X \in \{\pm 1\}$ given by*

$$Y = \epsilon X + Z, \quad Z \sim \mathcal{N}(0, 1).$$

Then for any $\pi \in (0, 1)$ taking $P_X = \text{Ber}(\pi)$ we have for some absolute constant K the estimate

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) \leq K \frac{\epsilon^2}{\pi(1 - \pi)}.$$

Proof. In this proof we assume all information measures are used to base- e . First, notice that

$$v(y) \triangleq P[X = 1 | Y = y] = \frac{1}{1 + \frac{1-\pi}{\pi} e^{-2y\epsilon}}.$$

Then, the optimization defining $\eta_{\text{KL}}^{(p)}$ can be written as

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) \leq \sup_{Q_Y} \frac{d(\mathbb{E}_{Q_Y}[v(Y)] \| \pi)}{D(Q_Y \| P_Y)}. \quad (33.25)$$

From (7.31) we have

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) \leq \frac{1}{\pi(1 - \pi)} \sup_{Q_Y} \frac{(\mathbb{E}_{Q_Y}[v(Y)] - \pi)^2}{D(Q_Y \| P_Y)}. \quad (33.26)$$

To proceed, we need to introduce a new concept. The T_1 -transportation inequality for the measure P_Y states the following: For every Q_Y we have for some $c = c(P_Y)$

$$W_1(Q_Y, P_Y) \leq \sqrt{2cD(Q_Y \| P_Y)}, \quad (33.27)$$

33.11 Application: Distributed Mean Estimation 569

where $W_1(Q_Y, P_Y)$ is the 1-Wasserstein distance defined as

$$\begin{aligned} W_1(Q_Y, P_Y) &= \sup\{\mathbb{E}_{Q_Y}[f] - \mathbb{E}_{P_Y}[f] : f \text{ 1-Lipschitz}\} \\ &= \inf\{\mathbb{E}[|A - B|] : A \sim Q_Y, B \sim P_Y\}. \end{aligned} \quad (33.28)$$

The constant $c(P_Y)$ in (33.27) has been characterized in [40, 91] in terms of properties of P_Y . One such estimate is the following:

$$c(P_Y) \leq \frac{2}{\delta} \sup_{k \geq 1} \left(\frac{G(\delta)}{\binom{2k}{k}} \right)^{1/k},$$

where $G(\delta) = \mathbb{E}[e^{\delta(Y-Y')^2}]$ where $Y, Y' \stackrel{\text{i.i.d.}}{\sim} P_Y$. Using the estimate $\binom{2k}{k} \geq \frac{4^k}{\sqrt{\pi(k+1/2)}}$ and the fact that $\frac{1}{k} \ln(k+1/2) \leq \frac{1}{2}$ we get a further bound

$$c(P_Y) \leq \frac{2}{\delta} G(\delta) \frac{\pi \sqrt{e}}{4} \leq \frac{6G(\delta)}{\delta}.$$

Next notice that $Y - Y' \stackrel{d}{=} B_\epsilon + \sqrt{2}Z$ where $B_\epsilon \perp\!\!\!\perp Z \sim \mathcal{N}(0, 1)$ and B_ϵ is symmetric and $|B_\epsilon| \leq 2\epsilon$. Thus, we conclude that for any $\delta < 1/4$ we have $\bar{c} \triangleq \frac{6}{\delta} \sup_{\epsilon \leq 1} G(\delta) < \infty$. In the end, we have inequality (33.27) with constant $c = \bar{c}$ that holds uniformly for all $0 \leq \epsilon \leq 1$.

Now, notice that $\left| \frac{d}{dy} v(y) \right| \leq \frac{\epsilon}{2}$ and therefore v is $\frac{\epsilon}{2}$ -Lipschitz. From (33.27) and (33.28) we obtain then

$$|\mathbb{E}_{Q_Y}[v(Y)] - \mathbb{E}_{P_Y}[v(Y)]| \leq \frac{\epsilon}{2} \sqrt{2\bar{c}D(Q_Y||P_Y)}.$$

Squaring this inequality and plugging back into (33.26) completes the proof. \square

Remark 33.3. Notice that we can also compute the exact value of $\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X})$ by noticing the following. From (33.25) it is evident that among all measures Q_Y with a given value of $\mathbb{E}_{Q_Y}[v(Y)]$ we are interested in the one minimizing $D(Q_Y||P_Y)$. From Theorem 15.14 we know that such Q_Y is given by $dQ_Y = e^{bv(y)-\psi_V(b)} dP_Y$, where $\psi_V(b) \triangleq \ln \mathbb{E}_{P_Y}[e^{bv(Y)}]$. Thus, by defining the convex dual $\psi_V^*(\lambda)$ we can get the exact value in terms of the following single-variable optimization:

$$\eta_{\text{KL}}^{(p)}(P_X, P_{Y|X}) = \sup_{\lambda \in [0, 1]} \frac{d(\lambda||\pi)}{\psi_V^*(\lambda)}.$$

Numerically, for $\pi = 1/2$ it turns out that the optimal value is $\lambda \rightarrow \frac{1}{2}$, justifying our overbounding of d by χ^2 , and surprisingly giving

$$\eta_{\text{KL}}^{(p)}(\text{Ber}(1/2), P_{Y|X}) = 4 \mathbb{E}_{P_Y}[\tanh^2(\epsilon Y)] = \eta_{\text{KL}}(P_{Y|X}),$$

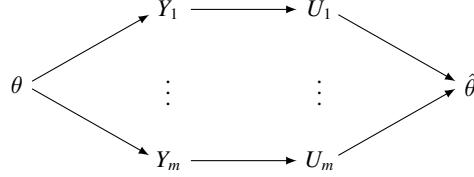
where in the last equality we used Theorem 33.6(f)).

33.11 Application: Distributed Mean Estimation

We want to estimate $\theta \in [-1, 1]^d$ and we have m machines observing $Y_i = \theta + \sigma Z_i$ where $Z_i \sim \mathcal{N}(0, I_d)$ independently. They can send a total of B bits to a remote estimator. The goal of the

570

estimator is to minimize $\sup_{\theta} \mathbb{E}[\|\theta - \hat{\theta}\|^2]$ over $\hat{\theta}$. If we denote by $U_i \in \mathcal{U}_i$ the messages then $\sum_i \log_2 |\mathcal{U}_i| \leq B$ and the diagram is



Finally, let

$$R^*(m, d, \sigma^2, B) = \inf_{U_1, \dots, U_m, \hat{\theta}} \sup_{\theta} \mathbb{E}[\|\theta - \hat{\theta}\|^2]$$

Observations:

- Without constraint on the magnitude of $\theta \in [-1, 1]^d$, we could give $\theta \sim \mathcal{N}(0, bI_d)$ and from rate-distortion quickly conclude that estimating θ within risk R requires communicating at least $\frac{d}{2} \log \frac{bd}{R}$ bits, which diverges as $b \rightarrow \infty$. Thus, restricting the magnitude of θ is necessary in order to be able to estimate it with finitely many bits communicated.
- It is easy to establish that $R^*(m, d, \sigma^2, \infty) = \mathbb{E} \left[\left\| \frac{\sigma}{m} \sum_i Z_i \right\|^2 \right] = \frac{d\sigma^2}{m}$ by taking $U_i = Y_i$ and $\hat{\theta} = \frac{1}{m} \sum_i U_i$.
- In order to approach the risk of order $\frac{d\sigma^2}{m}$ we could do the following. Let $U_i = \text{sign}(Y_i)$ (coordinate-wise sign). This yields $B = md$ and it is easy to show that the achievable risk is $O(\frac{d\sigma^2}{m})$. Indeed, notice that by averaging the signs we can estimate (within $O_p(\frac{1}{\sqrt{m}})$) all quantities $\Phi(\theta_i)$ with Φ denoting the CDF of $\mathcal{N}(0, 1)$. Since Φ has derivative bounded away from 0 on $[-1, 1]$, we get the estimate we claimed by applying Φ^{-1} .
- Our main result is that this strategy is (order) optimal in terms of communicated bits. This simplifies the proofs of [101, 49]. (However, in those works authors also consider a more general setting with interaction between the machines and the “fusion center”).
- In addition, all of these results (again in the non-interactive case, but with optimal constants) are contained in the long line of work in the information theoretic literature on the so-called *Gaussian CEO problem*. We recommend consulting [114]; in particular, Theorem 3 there implies the $B \gtrsim dm$ lower bound. The Gaussian CEO work uses a lot more sophisticated machinery (the entropy power inequality and related results). The advantage of our SDPI proof is its simplicity.

Our goal is to show that $R^* \lesssim \frac{d}{m}$ implies $B \gtrsim md$.

Notice, first of all, that this is completely obvious for $d = 1$. Indeed, if $B \leq \tau m$ then less than τm machines are communicating anything at all, and hence $R^* \geq \frac{K}{\tau m}$ for some universal constant K (which is not 1 because θ is restricted to $[-1, 1]$). However, for $d \gg 1$ it is not clear whether each machine is required to communicate $\Omega(d)$ bits. Perhaps sending $\ll d$ single-bit measurements taken in different orthogonal bases could work? Hopefully, this (incorrect) guess demonstrates why the following result is interesting and non-trivial.

33.11 Application: Distributed Mean Estimation 571

Theorem 33.23. *There exists a constant $c_1 > 0$ such that if $R^*(m, d, \sigma^2, B) \leq \frac{d\epsilon^2}{9}$ then $B \geq \frac{c_1 d}{\epsilon^2}$.*

Proof. Let $X \sim \text{Unif}(\{\pm 1\}^d)$ and set $\theta = \epsilon X$. Given an estimate $\hat{\theta}$ we can convert it into an estimator of X via $\hat{X} = \text{sign}(\hat{\theta})$ (coordinatewise). Then, clearly

$$\mathbb{E}[d_H(X, \hat{X})] \frac{\epsilon^2}{4} \leq \mathbb{E}[\|\hat{\theta} - \theta\|^2] \leq \frac{d\epsilon^2}{9}.$$

Thus, we have an estimator of X within Hamming distance $\frac{4}{9}d$. From Rate-Distortion (Theorem 26.1) we conclude that $I(X; \hat{X}) \geq cd$ for some constant $c > 0$. On the other hand, from the standard DPI we have

$$cd \leq I(X; \hat{X}) \leq I(X; U_1, \dots, U_m) \leq \sum_{j=1}^m I(X; U_j), \quad (33.29)$$

where we also applied Theorem 6.1. Next we estimate $I(X; U_j)$ via $I(Y_j; U_j)$ by applying the Post-SUPI. To do this we need to notice that the channel $X \rightarrow Y_j$ for each j is just a memoryless extension of the binary-input AWGN channel with SNR ϵ . Since each coordinate of X is uniform, we can apply Theorem 33.24 (with $\pi = 1/2$) together with tensorization (33.24) to conclude that

$$I(X; U_j) \leq 4K\epsilon^2 I(Y_j; U_j) \leq 4K\epsilon^2 \log |\mathcal{U}_j|$$

Together with (33.29) we thus obtain

$$cd \leq I(X; \hat{X}) \leq 4K\epsilon^2 B \log 2 \quad (33.30)$$

□

Exercises for Part VI

VI.1 Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\exp(\theta))$, where θ follows the Cauchy distribution π with parameter s , whose pdf is given by $p(\theta) = \frac{1}{\pi s(1 + \frac{\theta^2}{s^2})}$ for $\theta \in \mathbb{R}$. Show that the Bayes risk

$$R_\pi^* \triangleq \inf_{\hat{\theta}} \mathbb{E}_{\theta \sim \pi} \mathbb{E}(\hat{\theta}(X^n) - \theta)^2$$

satisfies $R_\pi^* \geq \frac{2s^2}{2ns^2+1}$.

VI.2 (System identification) Let $\theta \in \mathbb{R}$ be an unknown parameter of a dynamical system:

$$X_t = \theta X_{t-1} + Z_t, \quad Z_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad X_0 = 0.$$

Learning parameters of dynamical systems is known as “system identification”. Denote the law of (X_1, \dots, X_n) corresponding to θ by P_θ .

1. Compute $D(P_\theta \| P_{\theta_0})$. (Hint: chain rule saves a lot of effort.)
2. Show that Fisher information

$$J_F(\theta) = \sum_{1 \leq t \leq n-1} \theta^{2t-2}(n-t).$$

3. Argue that the hardest regime for system identification is when $\theta \approx 0$, and that instability ($|\theta| > 1$) is in fact helpful.

VI.3 (Linear regression) Consider the model

$$Y = X\beta + Z$$

where the design matrix $X \in \mathbb{R}^{n \times d}$ is known and $Z \sim N(0, I_n)$. Define the minimax mean-square error of estimating the regression coefficient $\beta \in \mathbb{R}^d$ based on X and Y as follows:

$$R_{\text{est}}^* = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{R}^d} \mathbb{E} \|\hat{\beta} - \beta\|_2^2. \tag{VI.1}$$

- (a) Show that if $\text{rank}(X) < d$, then $R_{\text{est}}^* = \infty$;
- (b) Show that if $\text{rank}(X) = d$, then

$$R_{\text{est}}^* = \text{tr}((X^\top X)^{-1})$$

and identify which estimator achieves the minimax risk.

- (c) As opposed to the estimation error in (VI.1), consider the *prediction error*:

$$R_{\text{pred}}^* = \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{R}^d} \mathbb{E} \|X\hat{\beta} - X\beta\|_2^2. \tag{VI.2}$$

Redo (a) and (b) by finding the value of R_{pred}^* and identify the minimax estimator. Explain intuitively why R_{pred}^* is always finite even when d exceeds n .

Exercises for Part VI 573

VI.4 (Chernoff-Rubin-Stein lower bound.) Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$ and $\theta \in [-a, a]$.

(a) State the appropriate regularity conditions and prove the following minimax lower bound:

$$\inf_{\hat{\theta}} \sup_{\theta \in [-a, a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \geq \min_{0 < \epsilon < 1} \max \left\{ \epsilon^2 a^2, \frac{(1 - \epsilon)^2}{n \bar{J}_F} \right\},$$

where $\bar{J}_F = \frac{1}{2a} \int_{-a}^a J_F(\theta) d\theta$ is the average Fisher information. (Hint: Consider the uniform prior on $[-a, a]$ and proceed as in the proof of Theorem 29.2 by applying integration by parts.)

(b) Simplify the above bound and show that

$$\inf_{\hat{\theta}} \sup_{\theta \in [-a, a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \geq \frac{1}{(a^{-1} + \sqrt{n \bar{J}_F})^2}. \quad (\text{VI.3})$$

(c) Assuming the continuity of $\theta \mapsto J_F(\theta)$, show that the above result also leads to the optimal local minimax lower bound in Theorem 29.4 obtained from Bayesian Cramér-Rao:

$$\inf_{\hat{\theta}} \sup_{\theta \in [\theta_0 \pm n^{-1/4}]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \geq \frac{1 + o(1)}{n J_F(\theta_0)}.$$

Note: (VI.3) is an improvement of the inequality given in [64, Lemma 1] without proof and credited to Rubin and Stein.

VI.5 In this exercise we give a Hellinger-based lower bound analogous to the χ^2 -based HCR lower bound in Theorem 29.1. Let $\hat{\theta}$ be an unbiased estimator for $\theta \in \Theta \subset \mathbb{R}$.

(a) For any $\theta, \theta' \in \Theta$, show that [273]

$$\frac{1}{2}(\text{Var}_\theta(\hat{\theta}) + \text{Var}_{\theta'}(\hat{\theta})) \geq \frac{(\theta - \theta')^2}{4} \left(\frac{1}{H^2(P_\theta, P_{\theta'})} - 1 \right). \quad (\text{VI.4})$$

(Hint: For any c , $\theta - \theta' = \int (\hat{\theta} - c)(\sqrt{p_\theta} + \sqrt{p_{\theta'}})(\sqrt{p_\theta} - \sqrt{p_{\theta'}})$. Apply Cauchy-Schwarz and optimize over c .)

(b) Show that

$$H^2(P_\theta, P_{\theta'}) \leq \frac{1}{4}(\theta - \theta')^2 \bar{J}_F \quad (\text{VI.5})$$

where $\bar{J}_F = \frac{1}{\theta' - \theta} \int_{\theta}^{\theta'} J_F(u) du$ is the average Fisher information.

(c) State the needed regularity conditions and deduce the Cramér-Rao lower bound from (VI.4) and (VI.5) with $\theta' \rightarrow \theta$.

(d) Extend the previous parts to the multivariate case.

VI.6 (Bayesian distribution estimation.) Let $\{P_\theta : \theta \in \Theta\}$ be a family of distributions on \mathcal{X} with a common dominating measure μ and density $p_\theta(x) = \frac{dP_\theta}{du}(x)$. Given a sample $X^n = (X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} P_\theta$ for some $\theta \in \Theta$, the goal is to estimate the data-generating distribution P_θ by some estimator $\hat{P}(\cdot) = \hat{P}(\cdot; X^n)$ with respect to some loss function $\ell(P, \hat{P})$. Suppose we are in a Bayesian setting where θ is drawn from a prior π . Let's find the form of the Bayes estimator and the Bayes risk

574 Exercises for Part VI

- (a) For convenience, let X_{n+1} denote a test data point (unseen) drawn from P_θ and independent of the observed data X^n . Convince yourself that every estimator \hat{P} can be formally identified as a conditional distribution $Q_{X_{n+1}|X^n}$.
- (b) Consider the KL loss $\ell(P, \hat{P}) = D(P||\hat{P})$. Using Corollary 4.1, show that the Bayes estimator minimizing the average KL risk is the posterior (conditional mean). Namely,

$$\min \underbrace{D(P_{X_{n+1}|\theta} \| Q_{X_{n+1}|X^n} | P_{\theta, X^n})}_{\text{average KL risk}}$$

is achieved at $Q_{X_{n+1}|X^n} = P_{X_{n+1}|X^n}$. In other words, the estimated density value at x_{n+1} is

$$\frac{dQ_{X_{n+1}|X^n}(x_{n+1}|X^n)}{d\mu} = \frac{\mathbb{E}_{\theta \sim \pi}[\prod_{i=1}^{n+1} p_\theta(x_i)]}{\mathbb{E}_{\theta \sim \pi}[\prod_{i=1}^n p_\theta(x_i)]}. \quad (\text{VI.6})$$

(Hint: The risk conditioned on X^n is $D(P_{X_{n+1}|\theta} \| Q_{X_{n+1}|X^n} | P_{\theta|X^n})$.)

- (c) Conclude that the Bayes KL risk equals $I(\theta; X_{n+1}|X^n)$.
- (d) Now, consider the χ^2 loss $\ell(P, \hat{P}) = \chi^2(P||\hat{P})$. Using (I.13) in Exercise I.35, show that

$$\min \underbrace{\chi^2(P_{X_{n+1}|\theta} \| Q_{X_{n+1}|X^n} | P_{\theta, X^n})}_{\text{average } \chi^2 \text{ risk}}$$

is achieved by

$$\frac{dQ_{X_{n+1}|X^n}(x_{n+1}|X^n)}{d\mu} \propto \mathbb{E}_\theta[p_\theta(x_{n+1})^2 | X^n = x^n] \propto \mathbb{E}_{\theta \sim \pi} \left[\prod_{i=1}^n p_\theta(x_i) p_\theta(x_{n+1})^2 \right]. \quad (\text{VI.7})$$

Furthermore, the Bayes χ^2 risk equals

$$\mathbb{E}_{X^n} \left[\left(\int \mu(dx_{n+1}) \sqrt{\mathbb{E}_\theta[p_\theta(x_{n+1})^2 | X^n]} \right)^2 \right] - 1. \quad (\text{VI.8})$$

- (e) Consider the discrete alphabet $[k]$ and $X^n \stackrel{\text{i.i.d.}}{\sim} P$, where $P = (P_1, \dots, P_k)$ is drawn from the Dirichlet prior $\text{Dirichlet}(\alpha, \dots, \alpha)$. Applying (VI.6) and (VI.7) (with μ the counting measure), show that the Bayes estimator for the KL loss is the add- α estimator (Section 13.5):

$$\hat{P}_j = \frac{n_j + \alpha}{n + k\alpha}, \quad (\text{VI.9})$$

where $n_j = \sum_{i=1}^n 1_{\{X_i=j\}}$ is the empirical count, and for the χ^2 loss is

$$\hat{P}_j = \frac{\sqrt{(n_j + \alpha)(n_j + \alpha + 1)}}{\sum_{j=1}^k \sqrt{(n_j + \alpha)(n_j + \alpha + 1)}}. \quad (\text{VI.10})$$

- VI.7 (Coin flips)** Given $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\theta)$ with $\theta \in \Theta = [0, 1]$, we aim to estimate θ with respect to the quadratic loss function $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Denote the minimax risk by R_n^* .

Exercises for Part VI 575

- (a) Use the empirical frequency $\hat{\theta}_{\text{emp}} = \bar{X}$ to estimate θ . Compute and plot the risk $R_\theta(\hat{\theta})$ and show that

$$R_n^* \leq \frac{1}{4n}.$$

- (b) Compute the Fisher information of $P_\theta = \text{Ber}(\theta)^{\otimes n}$ and $Q_\theta = \text{Bin}(n, \theta)$. Explain why they are equal.
(c) Invoke the Bayesian Cramér-Rao lower bound Theorem 29.2 to show that

$$R_n^* = \frac{1 + o(1)}{4n}.$$

- (d) Notice that the risk of $\hat{\theta}_{\text{emp}}$ is maximized at 1/2 (fair coin), which suggests that it might be possible to hedge against this situation by the following randomized estimator

$$\hat{\theta}_{\text{rand}} = \begin{cases} \hat{\theta}_{\text{emp}}, & \text{with probability } \delta \\ \frac{1}{2}, & \text{with probability } 1 - \delta \end{cases} \quad (\text{VI.11})$$

Find the worst-case risk of $\hat{\theta}_{\text{rand}}$ as a function of δ . Optimizing over δ , show the improved upper bound:

$$R_n^* \leq \frac{1}{4(n+1)}.$$

- (e) As discussed in Remark 28.3, randomized estimator can always be improved if the loss is convex; so we should average out the randomness in (VI.11) by considering the estimator

$$\hat{\theta}^* = \mathbb{E}[\hat{\theta}_{\text{rand}}|X] = \bar{X}\delta + \frac{1}{2}(1 - \delta). \quad (\text{VI.12})$$

Optimizing over δ to minimize the worst-case risk, find the resulting estimator $\hat{\theta}^*$ and its risk, show that it is constant (independent of θ), and conclude

$$R_n^* \leq \frac{1}{4(1 + \sqrt{n})^2}.$$

- (f) Next we show $\hat{\theta}^*$ found in part (e) is exactly minimax and hence

$$R_n^* = \frac{1}{4(1 + \sqrt{n})^2}.$$

Consider the following prior Beta(a, b) with density:

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0, 1],$$

where $\Gamma(a) \triangleq \int_0^\infty x^{a-1} e^{-x} dx$. Show that if $a = b = \frac{\sqrt{n}}{2}$, $\hat{\theta}^*$ coincides with the Bayes estimator for this prior, which is therefore least favorable. (Hint: work with the sufficient statistic $S = X_1 + \dots + X_n$.)

- (g) Show that the least favorable prior is not unique; in fact, there is a continuum of them. (Hint: consider the Bayes estimator $\mathbb{E}[\theta|X]$ and show that it only depends on the first $n+1$ moments of π .)

576 Exercises for Part VI

- (h) (Larger alphabet) Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$ on $[k]$. Show that for any k, n , the minimax squared risk of estimating P in Theorem 29.5 is exactly

$$R_{\text{sq}}^*(k, n) = \inf_{\hat{P}} \sup_{P \in \mathcal{P}_k} \mathbb{E}[\|\hat{P} - P\|_2^2] = \frac{1}{(\sqrt{n} + 1)^2} \frac{k - 1}{k}, \quad (\text{VI.13})$$

achieved by the add- $\frac{\sqrt{n}}{k}$ estimator. (Hint: For the lower bound, show that the Bayes estimator for the squared loss and the KL loss coincide, then apply (VI.9) in Exercise VI.6.)

- (i) (Nonparametric extension) Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$, where P is an arbitrary probability distribution on $[0, 1]$. The goal is to estimate the *mean* of P under the quadratic loss. Show that the minimax risk equals $\frac{1}{4(1+\sqrt{n})^2}$.

VI.8 (Distribution estimation in TV) Continuing (VI.13), we show that the minimax rate for estimating P with respect to the total variation loss is

$$R_{\text{TV}}^*(k, n) \triangleq \inf_{\hat{P}} \sup_{P \in \mathcal{P}_k} \mathbb{E}_P[\text{TV}(\hat{P}, P)] \asymp \sqrt{\frac{k}{n}} \wedge 1, \quad \forall k \geq 2, n \geq 1, \quad (\text{VI.14})$$

- (a) Show that the MLE coincides with the empirical distribution.
- (b) Show that the MLE achieves the RHS of (VI.14) within constant factors.
- (c) Establish the minimax lower bound. (Hint: apply Assouad's lemma, or Fano's inequality (with volume method or explicit construction of packing), or the mutual information method directly.)

VI.9 (Distribution estimation in KL and χ^2) Continuing Exercise VI.8, let us consider estimating the distribution P in KL and χ^2 divergence, which are unbounded loss. We show that

$$R_{\text{KL}}^*(k, n) \triangleq \inf_{\hat{P}} \sup_{P \in \mathcal{P}_k} \mathbb{E}_P[D(\hat{P} \| P)] \asymp \log\left(1 + \frac{k}{n}\right) \asymp \begin{cases} \frac{k}{n} & k \leq n \\ \log \frac{k}{n} & k \geq n \end{cases} \quad (\text{VI.15})$$

and

$$R_{\chi^2}^*(k, n) \triangleq \inf_{\hat{P}} \sup_{P \in \mathcal{P}_k} \mathbb{E}_P[\chi^2(\hat{P} \| P)] \asymp \frac{k}{n}. \quad (\text{VI.16})$$

To this end, we will apply results on Bayes risk in Exercise VI.6 as well as multiple inequalities between f -divergences from Chapter 7.

- (a) Show that the empirical distribution, which has been shown optimal for the TV loss in Exercise VI.8, achieves infinite KL and χ^2 loss in the worst case.
- (b) To show the upper bound in (VI.16), consider the add- α estimator \hat{P} in (VI.9) with $\alpha = 1$. Show that

$$\mathbb{E}[\chi^2(P \| \hat{P})] \leq \frac{k - 1}{n + 1}.$$

Using $D \leq \log(1 + \chi^2) - \text{cf. (7.31)}$, conclude the upper bound part of (VI.15). (Hint: $\mathbb{E}_{N \sim \text{Bin}(n, p)}[\frac{1}{N+1}] = \frac{1}{(n+1)p}(1 - \bar{p}^{n+1})$.)

- (c) Show that for the small alphabet regime of $k \lesssim n$, the lower bound follows from that of (VI.15) and Pinsker's inequality (7.25).

Exercises for Part VI 577

- (d) Next assume $k \geq 4n$. Consider a $\text{Dirichlet}(\alpha, \dots, \alpha)$ prior in (13.15). Applying the formula (VI.7) and (VI.8) for the Bayes χ^2 risk and choosing $\alpha = n/k$, show the lower bound in (VI.16).
- (e) Finally, we prove the lower bound in (VI.15). Consider the prior under which P is uniform over a set S chosen uniformly at random from all s -subsets of $[k]$ and s is some constant to be specified. Applying (VI.6), show that for this prior the Bayes estimator for KL loss takes a natural form:

$$\hat{P}_j = \begin{cases} \frac{1}{s} & i \in \hat{S} \\ \frac{1-\hat{s}/s}{k-\hat{s}} & i \notin \hat{S} \end{cases}$$

where $\hat{S} = \{i : n_i \geq 1\}$ is the support of the empirical distribution and $\hat{s} = |\hat{S}|$.

- (f) Choosing $s = \sqrt{nk}$, conclude $\text{TV}(P, \hat{P}) \geq 1 - 2\sqrt{\frac{n}{k}}$. (Hint: Show that $\text{TV}(P, \hat{P}) \geq (1 - \frac{\hat{s}}{s})(1 - \frac{s}{k})$ and $\hat{s} \leq n$.)
- (g) Using (7.28), show that $D(P\| \hat{P}) \geq \Omega(\log \frac{k}{n})$. (Note that (7.28) is convex in TV so Jensen's inequality applies.)

Note: The following refinement of (VI.15) was known:

- For fixed k , a deep result of [48, 47] is that $R_{\text{KL}}^*(k, n) = \frac{k-1+o(1)}{2n}$, achieved by an add- c estimator where c is a function of the empirical count chosen using polynomial approximation arguments.
- When $k \gg n$, $R_{\text{KL}}^*(k, n) = \log \frac{k}{n}(1 + o(1))$, shown in [224] by a careful analysis of the Dirichlet prior.

VI.10 (Nonparametric location model) In this exercise we consider some nonparametric extensions of the Gaussian location model and the Bernoulli model. Observing $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P$ for some $P \in \mathcal{P}$, where \mathcal{P} is a collection of distributions on the real line, our goal is to estimate the *mean* of the distribution P : $\mu(P) \triangleq \int xP(dx)$, which is a linear functional of P . Denote the minimax quadratic risk by

$$R_n^* = \inf_{\hat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(\hat{\mu}(X_1, \dots, X_n) - \mu(P))^2].$$

- (a) Let \mathcal{P} be the class of distributions (which need not have a density) on the real line with variance at most σ^2 . Show that $R_n^* = \frac{\sigma^2}{n}$.
- (b) Let $\mathcal{P} = \mathcal{P}([0, 1])$, the collection of all probability distributions on $[0, 1]$. Show that $R_n^* = \frac{1}{4(1+\sqrt{n})^2}$. (Hint: For the upper bound, using the fact that, for any $[0, 1]$ -valued random variable Z , $\text{Var}(Z) \leq \mathbb{E}[Z](1 - \mathbb{E}[Z])$, mimic the analysis of the estimator (VI.12) in Ex. VI.7e.)

VI.11 Prove Theorem 30.6 using Fano's method. (Hint: apply Theorem 31.4 with $T = \epsilon \cdot S_k^d$, where S_k^d denotes the Hamming sphere of radius k in d dimensions. Choose ϵ appropriately and apply the Gilbert-Varshamov bound for the packing number of S_k^d in Theorem 27.8.)

VI.12 (Sharp minimax rate in sparse denoising) Continuing Theorem 30.6, in this exercise we determine the sharp minimax risk for denoising a high-dimensional sparse vector. In the notation of (30.13), we show that, for the d -dimensional GLM model $X \sim \mathcal{N}(\theta, I_d)$, the following minimax

578 Exercises for Part VI

risk satisfies, as $d \rightarrow \infty$ and $k/d \rightarrow 0$,

$$R^*(k, d) \triangleq \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \leq k} \mathbb{E}_{\theta}[\|\hat{\theta} - \theta\|_2^2] = (2 + o(1))k \log \frac{d}{k}. \quad (\text{VI.17})$$

(a) We first consider 1-sparse vectors and prove

$$R^*(1, d) \triangleq \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \leq 1} \mathbb{E}_{\theta}[\|\hat{\theta} - \theta\|_2^2] = (2 + o(1)) \log d, \quad d \rightarrow \infty. \quad (\text{VI.18})$$

For the lower bound, consider the prior π under which θ is uniformly distributed over $\{\tau e_1, \dots, \tau e_d\}$, where e_i 's denote the standard basis. Let $\tau = \sqrt{(2 - \epsilon) \log d}$. Show that for any $\epsilon > 0$, the Bayes risk is given by

$$\inf_{\hat{\theta}} \mathbb{E}_{\theta \sim \pi}[\|\hat{\theta} - \theta\|_2^2] = \tau^2(1 + o(1)), \quad d \rightarrow \infty.$$

(Hint: either apply the mutual information method, or directly compute the Bayes risk by evaluating the conditional mean and conditional variance.)

- (b) Demonstrate an estimator $\hat{\theta}$ that achieves the RHS of (VI.18) asymptotically. (Hint: consider the hard-thresholding estimator (30.13) or the MLE (30.11).)
- (c) To prove the lower bound part of (VI.17), prove the following generic result

$$R^*(k, d) \geq k R^*\left(1, \frac{d}{k}\right)$$

and then apply (VI.18). (Hint: consider a prior of d/k blocks each of which is 1-sparse.)

- (d) Similar to the 1-sparse case, demonstrate an estimator $\hat{\theta}$ that achieves the RHS of (VI.17) asymptotically.

Note: For both the upper and lower bound, the normal tail bound in Exercise V.8 is helpful.

VI.13 Consider the following functional estimation problem in GLM. Observing $X \sim N(\theta, I_d)$, we intend to estimate the maximal coordinate of θ : $T(\theta) = \theta_{\max} \triangleq \max\{\theta_1, \dots, \theta_d\}$. Prove the minimax rate:

$$\inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{\theta}(\hat{T} - \theta_{\max})^2 \asymp \log d. \quad (\text{VI.19})$$

- (a) Prove the upper bound by considering $\hat{T} = X_{\max}$, the plug-in estimator with the MLE.
- (b) For the lower bound, consider two hypotheses:

$$H_0 : \theta = 0, \quad H_1 : \theta \sim \text{Unif}\{\tau e_1, \tau e_2, \dots, \tau e_d\}.$$

where e_i 's are the standard bases and $\tau > 0$. Then under H_0 , $X \sim P_0 = N(0, I_d)$; under H_1 , $X \sim P_1 = \frac{1}{d} \sum_{i=1}^d N(\tau e_i, I_d)$. Show that

$$\chi^2(P_1 \| P_0) = \frac{e^{\tau^2} - 1}{d}.$$

- (c) Applying the joint range (7.29) (or (7.35)) to bound $\text{TV}(P_0, P_1)$, conclude the lower bound part of (VI.19) via Le Cam's method (Theorem 31.1).

Exercises for Part VI 579

(d) By improving both the upper and lower bound prove the sharp version:

$$\inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{\theta}(\hat{T} - \theta_{\max})^2 = \left(\frac{1}{2} + o(1) \right) \log d, \quad d \rightarrow \infty. \quad (\text{VI.20})$$

VI.14 (Suboptimality of MLE in high dimensions) Consider the d -dimensional GLM: $X \sim \mathcal{N}(\theta, I_d)$, where θ belongs to the parameter space

$$\Theta = \left\{ \theta \in \mathbb{R}^d : |\theta_1| \leq d^{1/4}, \|\theta_{\setminus 1}\|_2 \leq 2(1 - d^{-1/4}|\theta_1|) \right\}$$

with $\theta_{\setminus 1} \equiv (\theta_2, \dots, \theta_d)$. For the square loss, prove the following for sufficiently large d .

(a) The minimax risk is bounded:

$$\inf_{\theta} \sup_{\theta \in \Theta} \mathbb{E}_{\theta}[\|\hat{\theta} - \theta\|_2^2] \lesssim 1.$$

(b) The worst-case risk of maximal likelihood estimator

$$\theta_{\text{MLE}} \triangleq \underset{\theta \in \Theta}{\operatorname{argmin}} \|X - \tilde{\theta}\|_2$$

is unbounded:

$$\sup_{\theta \in \Theta} \mathbb{E}_{\theta}[\|\hat{\theta}_{\text{MLE}} - \theta\|_2^2] \gtrsim \sqrt{d}.$$

VI.15 (Covariance model) Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma)$, where Σ is a $d \times d$ covariance matrix. Let us show that the minimax quadratic risk of estimating Σ using X_1, \dots, X_n satisfies

$$\inf_{\hat{\Sigma}} \sup_{\|\Sigma\|_{\text{F}} \leq r} \mathbb{E}[\|\hat{\Sigma} - \Sigma\|_{\text{F}}^2] \asymp \left(\frac{d}{n} \wedge 1 \right) r^2, \quad \forall r > 0, d, n \in \mathbb{N}. \quad (\text{VI.21})$$

where $\|\hat{\Sigma} - \Sigma\|_{\text{F}}^2 = \sum_{ij} (\hat{\Sigma}_{ij} - \Sigma_{ij})^2$.

- (a) Show that unlike location model, without restricting to a compact parameter space for Σ , the minimax risk in (VI.21) is infinite.
(b) Consider the sample covariance matrix $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$. Show that

$$\mathbb{E}[\|\hat{\Sigma} - \Sigma\|_{\text{F}}^2] = \frac{1}{n} (\|\Sigma\|_{\text{F}}^2 + \text{Tr}(\Sigma)^2)$$

and use this to deduce the minimax upper bound in (VI.21).

- (c) To prove the minimax lower bound, we can proceed in several steps. Show that for any positive semidefinite (PSD) $\Sigma_0, \Sigma_1 \succeq 0$, the KL divergence satisfies

$$D(N(0, I_d + \Sigma_0) \| N(0, I_d + \Sigma_1)) \leq \frac{1}{2} \|\Sigma_0^{1/2} - \Sigma_1^{1/2}\|_{\text{F}}^2, \quad (\text{VI.22})$$

where I_d is the $d \times d$ identity matrix.

- (d) Let $B(\delta)$ denote the Frobenius ball of radius δ centered at the zero matrix. Let $\text{PSD} = \{X : X \succeq 0\}$ denote the collection of $d \times d$ PSD matrices. Show that

$$\frac{\text{vol}(B(\delta) \cap \text{PSD})}{\text{vol}(B(\delta))} = \mathbb{P}[Z \succeq 0], \quad (\text{VI.23})$$

580 Exercises for Part VI

where Z is a GOE matrix, that is, Z is symmetric with independent diagonals $Z_{ii} \stackrel{\text{i.i.d.}}{\sim} N(0, 2)$ and off-diagonals $Z_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$.

- (e) Show that $\mathbb{P}[Z \succeq 0] \geq c^{d^2}$ for some absolute constant c .³
- (f) Prove the following lower bound on the packing number on the set of PSD matrices:

$$M(B(\delta) \cap \text{PSD}, \|\cdot\|_F, \epsilon) \geq \left(\frac{c'\delta}{\epsilon}\right)^{d^2/2} \quad (\text{VI.24})$$

for some absolute constant c' . (Hint: Use the volume bound and the result of Part (d) and (e).)

- (g) Complete the proof of lower bound of (VI.21). (Hint: WLOG, we can consider $r \asymp \sqrt{d}$ and aim for the lower bound $\Omega(\frac{d^2}{n} \wedge d)$. Take the packing from (VI.24) and shift by the identity matrix I . Then apply Fano's method and use (VI.22).)

VI.16 For a family of probability distributions \mathcal{P} and a functional $T : \mathcal{P} \rightarrow \mathbb{R}$ define its χ^2 -modulus of continuity as

$$\delta_{\chi^2}(t) = \sup_{P_1, P_2 \in \mathcal{P}} \{T(P_1) - T(P_2) : \chi^2(P_1 \| P_2) \leq t\}.$$

When the functional T is affine, and continuous, and \mathcal{P} is compact⁴ it can be shown [242] that

$$\frac{1}{7} \delta_{\chi^2}(1/n)^2 \leq \inf_{\hat{T}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{X_i \stackrel{\text{i.i.d.}}{\sim} P} (T(P) - \hat{T}_n(X_1, \dots, X_n))^2 \leq \delta_{\chi^2}(1/n)^2. \quad (\text{VI.25})$$

Consider the following problem (interval censored model): In i -th mouse a tumour develops at time $A_i \in [0, 1]$ with $A_i \stackrel{\text{i.i.d.}}{\sim} \pi$ where π is a pdf on $[0, 1]$ bounded by $\frac{1}{2} \leq \pi \leq 2$. For each i the existence of tumour is checked at another random time $B_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ with $B_i \perp A_i$. Given observations $X_i = (1\{A_i \leq B_i\}, B_i)$ one is trying to estimate $T(\pi) = \pi[A \leq 1/2]$. Show that

$$\inf_{\hat{T}_n} \sup_{\pi} \mathbb{E}[(T(\pi) - \hat{T}_n(X_1, \dots, X_n))^2] \asymp n^{-2/3}.$$

VI.17 (Comparison between contraction coefficients.) Let X be a random variable with distribution P_X , and let $P_{Y|X}$ be a Markov kernel. For an f -divergence, define

$$\eta_f(P_{Y|X}, P_X) \triangleq \sup_{Q_X: 0 < D_f(Q_X \| P_X) < \infty} \frac{D_f(P_{Y|X} \circ Q_X \| P_{Y|X} \circ P_X)}{D_f(Q_X \| P_X)}.$$

Prove that

$$\eta_{\chi^2}(P_{Y|X}, P_X) \leq \eta_{\text{KL}}(P_{Y|X}, P_X).$$

Hint: Use local behavior of f -divergences (Proposition 2.19).

³ Getting the exact exponent is a difficult result (cf. [17]). Here we only need some crude estimate.

⁴ Both under the same, but otherwise arbitrary topology on \mathcal{P} .

Exercises for Part VI 581

VI.18 (χ^2 -contraction for Markov chains.) In this exercise we prove (33.4). Let $P = (P(x, y))$ denote the transition matrix of a time-reversible Markov chain with finite state space \mathcal{X} and stationary distribution π , so that $\pi(x)P(x, y) = \pi(y)P(y, x)$ for all $x, y \in \mathcal{X}$. It is known that the $k = |\mathcal{X}|$ eigenvalues of P satisfy $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq -1$. Define by $\gamma_* \triangleq \max\{\lambda_2, |\lambda_k|\}$ the absolute spectral gap.

(a) Show that

$$\chi^2(P_{X_1} \| \pi) \leq \chi^2(P_{X_0} \| \pi) \gamma_*^{2n}$$

from which (33.4) follows.

(b) Conclude that for any initial state x ,

$$\chi^2(P_{X_n | X_0=x} \| \pi) \leq \frac{1 - \pi(x)}{\pi(x)} \gamma_*^{2n}.$$

(c) Compute γ_* for the BSC_δ channel and compare with the η_{χ^2} contraction coefficients.

For a continuous-time version, see [90].

VI.19 (Input-independent contraction coefficient is achieved by binary inputs [?]) Let $K : \mathcal{X} \rightarrow \mathcal{Y}$ be a Markov kernel with countable \mathcal{X} . Prove that for all f -divergence, we have

$$\eta_f(K) = \sup_{\substack{P, Q: |\text{supp}(P) \cup \text{supp}(Q)| \leq 2 \\ 0 < D_f(P \| Q) < \infty}} \frac{D_f(K \circ P \| K \circ Q)}{D_f(P \| Q)}.$$

Hint: Define function

$$L_\lambda(P, Q) = D_f(K \circ P \| K \circ Q) - \lambda D_f(P \| Q)$$

and prove that $L_\lambda(\frac{P}{Q}\hat{Q}, \hat{Q})$ is convex on the set

$$\left\{ \hat{Q} \in \mathcal{P}(\mathcal{X}) : \text{supp}(\hat{Q}) \subseteq \text{supp}(Q), \frac{P}{Q}\hat{Q} \in \mathcal{P}(\mathcal{X}) \right\}.$$

VI.20 (BMS channel comparison [260]) Below $X \sim \text{Ber}(1/2)$ and $P_{Y|X}$ is an input-symmetric channel (BMS). It turns out that BSC and BEC are extremal for various partial orders. Prove the following statements.

(a) If $I_{TV}(X; Y) = \frac{1}{2}(1 - 2\delta)$, then

$$\text{BSC}_\delta \leq_{deg} P_{Y|X} \leq_{deg} \text{BEC}_{2\delta}.$$

(b) If $I(X; Y) = C$, then

$$\text{BSC}_{h^{-1}(\log 2 - C)} \leq_{mc} P_{Y|X} \leq_{mc} \text{BEC}_{1-C/\log 2}.$$

(c) If $I_{\chi^2}(X; Y) = \eta$, then

$$\text{BSC}_{1/2 - \sqrt{\eta}/2} \leq_{ln} P_{Y|X} \leq_{ln} \text{BEC}_{1-\eta}.$$

582 Exercises for Part VI

VI.21 (Broadcasting on Trees with BSC [?]) We have seen that Broadcasting on Trees with BSC_δ has non-reconstruction when $b(1 - 2\delta)^2 < 1$. In this exercise we prove the achievability bound (known as the Kesten-Stigum bound [172]) using channel comparison.

We work with an infinite b -ary tree with BSC_δ edge channels. Let ρ be the root and L_k be the set of nodes at distance k to ρ . Let M_k denote the channel $X_\rho \rightarrow X_{L_k}$.

In the following, assume that $b(1 - 2\delta)^2 > 1$.

- (a) Prove that there exists $\tau < \frac{1}{2}$ such that

$$\text{BSC}_\tau \leq_{ln} \text{BSC}_\tau^{\otimes b} \circ M_1.$$

Hint: Use Ex. VI.20.

- (b) Prove $\text{BSC}_\tau \leq_{ln} M_k$ by induction on k . Conclude that reconstruction holds.

Hint: Use tensorization of less noisy ordering.

VI.22 (Broadcasting on a 2D Grid) Consider the following broadcasting model on a 2D grid:

- Nodes are labeled with (i, j) for $i, j \in \mathbb{Z}$;
- $X_{i,j} = 0$ when $i < 0$ or $j < 0$;
- $X_{0,0} \sim \text{Ber}(\frac{1}{2})$;
- $X_{i,j} = f_{i,j}(X_{i-1,j} \oplus Z_{i,j,1}, X_{i,j-1} \oplus Z_{i,j,2})$ for $i, j \geq 0$ and $(i, j) \neq (0, 0)$, where $Z_{i,j,k} \stackrel{\text{i.i.d.}}{\sim} \text{Ber}(\delta)$, and $f_{i,j}$ is any function $\{0, 1\} \times \{0, 1\} \rightarrow \{0, 1\}$.

Let $L_n = \{(n - i, i) : 0 \leq i \leq n\}$ be the set of nodes at level n .

Let p_c be directed bond percolation threshold from $(0, 0)$ to L_n for $n \rightarrow \infty$. Prove that when $(1 - 2\delta)^2 < p_c$, we have

$$\lim_{n \rightarrow \infty} I(X_{0,0}; X_{L_n}) = 0.$$

Note: It is known that $p_c \approx 0.645$ (e.g. [163]). Using site percolation we can prove non-reconstruction whenever

$$1 - 2\delta + 4\delta^3 - 2\delta^4 - 2\delta(1 - \delta)\sqrt{\delta(1 + \delta)(1 - \delta)(2 - \delta)} < p'_c,$$

where $p'_c \approx 0.705$ is the directed site percolation threshold. One can check that the bound from bond percolation is stronger.

VI.23 (Input-dependent contraction coefficient for coloring channel [145]) Fix an integer $q \geq 3$ and let $\mathcal{X} = [q]$. Consider the following coloring channel $K : \mathcal{X} \rightarrow \mathcal{X}$:

$$K(y|x) = \begin{cases} 0 & y = x, \\ \frac{1}{q-1} & y \neq x. \end{cases}$$

Let π be uniform distribution on \mathcal{X} .

- (a) Compute $\eta_{KL}(\pi, K)$.
(b) Conclude that there exists a function $f(q) = (1 - o(1))q \log q$ as $q \rightarrow \infty$ such that for all $d < f(q)$, BOT with the coloring channel on a d -ary tree has non-reconstruction.

Note: This bound is tight up to the first order: there exists a function $g(q) = (1 + o(1))q \log q$ such that for all $d > g(q)$, BOT with coloring channel on a d -ary tree has reconstruction.

Exercises for Part VI 583

VI.24 ([145]) Fix an integer $q \geq 2$ and let $\mathcal{X} = [q]$. Let $\lambda \in [-\frac{1}{q-1}, 1]$ be a real number. Define the Potts channel $P_\lambda : \mathcal{X} \rightarrow \mathcal{X}$ as

$$P_\lambda(y|x) = \begin{cases} \lambda + \frac{1-\lambda}{q} & y = x, \\ \frac{1-\lambda}{q} & y \neq x. \end{cases}$$

Prove that

$$\eta_{KL}(P_\lambda) = \frac{q\lambda^2}{(q-2)\lambda+2}.$$

VI.25 (Spectral Independence) Say a probability distribution $\mu = \mu_{X^n}$ supported on $[q]^n$ is c -pairwise independent if for every $T \subset [n]$, $\sigma_T \in [q]^T$ the conditional measure $\mu^{(\sigma_T)} \triangleq \mu_{X_T | X_T = \sigma_T}$ and every $\nu_{X_T^c}$ satisfies

$$\sum_{i \neq j \in T^c} D(\nu_{X_{i,j}} || \mu_{X_{i,j}}^{(\sigma_T)}) \geq \left(2 - \frac{c}{n - |T| - 1}\right) \sum_{i \in T^c} D(\nu_{X_i} || \mu_{X_i}^{(\sigma_T)}).$$

Prove that for such a measure μ we have

$$\eta_{KL}(\mu, \text{EC}_\tau) \leq 1 - \tau^{c+1},$$

where EC_τ is the erasure channel, cf. Example 33.7. (*Hint:* Define $f(\tau) = D(\text{EC}_\tau \circ \nu || \text{EC}_\tau \circ \mu)$ and prove $f'(\tau) \geq \frac{c}{\tau} f'(\tau)$.)

Remark: Applying the above with $\tau = \frac{1}{n}$ shows that a Markov chain known as (small-block) Glauber dynamics for μ is mixing in $O(n^{c+1} \log n)$ time. It is known that c -pairwise independence is implied (under some additional conditions on μ and $q = 2$) by the uniform boundedness of the operator norms of the covariance matrices of all $\mu^{(\sigma_T)}$ (see [63] for details).

References

- [1] M. C. Abbott and B. B. Machta, “A scaling law from discrete to continuous solutions of channel capacity problems in the low-noise limit,” *Journal of Statistical Physics*, vol. 176, no. 1, pp. 214–227, 2019.
- [2] I. Abou-Faycal, M. Trott, and S. Shamai, “The capacity of discrete-time memoryless rayleigh-fading channels,” *IEEE Transaction Information Theory*, vol. 47, no. 4, pp. 1290 – 1301, 2001.
- [3] R. Ahlswede, “Extremal properties of rate distortion functions,” *IEEE transactions on information theory*, vol. 36, no. 1, pp. 166–171, 1990.
- [4] R. Ahlswede, B. Balkenhol, and L. Khachatrian, *Some properties of fix free codes*. Citeseer, 1997.
- [5] S. M. Alamouti, “A simple transmit diversity technique for wireless communications,” *IEEE Journal on selected areas in communications*, vol. 16, no. 8, pp. 1451–1458, 1998.
- [6] P. H. Algoet and T. M. Cover, “A sandwich proof of the Shannon-Mcmillan-Breiman theorem,” *The annals of probability*, pp. 899–909, 1988.
- [7] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: a Hitchhiker’s Guide*, 3rd ed. Berlin: Springer-Verlag, 2006.
- [8] N. Alon, “On the number of subgraphs of prescribed type of graphs with a given number of edges,” *Israel J. Math.*, vol. 38, no. 1-2, pp. 116–130, 1981. [Online]. Available: <http://dx.doi.org/10.1007/BF02761855>
- [9] N. Alon and A. Orlitsky, “A lower bound on the expected length of one-to-one codes,” *IEEE Transactions on Information Theory*, vol. 40, no. 5, pp. 1670–1672, 1994.
- [10] N. Alon and J. H. Spencer, *The Probabilistic Method*, 3rd ed. John Wiley & Sons, 2008.
- [11] S.-i. Amari and H. Nagaoka, *Methods of information geometry*. American Mathematical Soc., 2007, vol. 191.
- [12] G. Aminian, Y. Bu, L. Toni, M. R. Rodrigues, and G. Wornell, “Characterizing the generalization error of Gibbs algorithm with symmetrized KL information,” *arXiv preprint arXiv:2107.13656*, 2021.
- [13] T. W. Anderson, “The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities,” *Proceedings of the American Mathematical Society*, vol. 6, no. 2, pp. 170–176, 1955.
- [14] A. Antos and I. Kontoyiannis, “Convergence properties of functional estimates for discrete distributions,” *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.
- [15] E. Arikan, “Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,” *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [16] S. Arimoto, “On the converse to the coding theorem for discrete memoryless channels (corresp.),” *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 357–359, 1973.
- [17] G. B. Arous and A. Guionnet, “Large deviations for wigner’s law and voiculescu’s non-commutative entropy,” *Probability theory and related fields*, vol. 108, no. 4, pp. 517–542, 1997.

References 585

- [18] S. Artstein, V. Milman, and S. J. Szarek, “Duality of metric entropy,” *Annals of mathematics*, pp. 1313–1328, 2004.
- [19] R. B. Ash, *Information Theory*. New York, NY: Dover Publications Inc., 1965.
- [20] A. V. Banerjee, “A simple model of herd behavior,” *The quarterly journal of economics*, vol. 107, no. 3, pp. 797–817, 1992.
- [21] A. Barg and G. D. Forney, “Random codes: Minimum distances and error exponents,” *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2568–2573, 2002.
- [22] A. Barg and A. McGregor, “Distance distribution of binary codes and the error probability of decoding,” *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4237–4246, 2005.
- [23] S. Barman and O. Fawzi, “Algorithmic aspects of optimal channel coding,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1038–1045, 2017.
- [24] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, 1993.
- [25] G. Basharin, “On a statistical estimate for the entropy of a sequence of independent random variables,” *Theory of Probability & Its Applications*, vol. 4, no. 3, pp. 333–336, 1959.
- [26] A. Beirami and F. Fekri, “Fundamental limits of universal lossless one-to-one compression of parametric sources,” in *Information Theory Workshop (ITW), 2014 IEEE*. IEEE, 2014, pp. 212–216.
- [27] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, “Entanglement-assisted classical capacity of noisy quantum channels,” *Physical Review Letters*, vol. 83, no. 15, p. 3081, 1999.
- [28] W. R. Bennett, “Spectra of quantized signals,” *The Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.
- [29] J. M. Bernardo, “Reference posterior distributions for Bayesian inference,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 113–128, 1979.
- [30] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near shannon limit error-correcting coding and decoding: Turbo-codes. 1,” in *Proceedings of ICC’93-IEEE International Conference on Communications*, vol. 2. IEEE, 1993, pp. 1064–1070.
- [31] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Belmont, MA, USA: Athena Scientific, 2003.
- [32] N. Bhatnagar, J. Vera, E. Vigoda, and D. Weitz, “Reconstruction for colorings on trees,” *SIAM Journal on Discrete Mathematics*, vol. 25, no. 2, pp. 809–826, 2011. [Online]. Available: <https://doi.org/10.1137/090755783>
- [33] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.
- [34] L. Birgé, “Approximation dans les espaces métriques et théorie de l'estimation,” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 65, no. 2, pp. 181–237, 1983.
- [35] ———, “Robust tests for model selection,” *From probability to statistics and back: high-dimensional models and processes—A Festschrift in honor of Jon A. Wellner, IMS Collections, Volume 9*, pp. 47–64, 2013.
- [36] M. Š. Birman and M. Solomjak, “Piecewise-polynomial approximations of functions of the classes,” *Mathematics of the USSR-Sbornik*, vol. 2, no. 3, p. 295, 1967.
- [37] D. Blackwell, L. Breiman, and A. Thomasian, “The capacity of a class of channels,” *The Annals of Mathematical Statistics*, pp. 1229–1241, 1959.
- [38] R. E. Blahut, “Hypothesis testing and information theory,” *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, 1974.
- [39] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov, “On the purity of the limiting gibbs state for

586 Strong data processing inequality

- the ising model on the bethe lattice,” *Journal of Statistical Physics*, vol. 79, no. 1, pp. 473–482, Apr 1995. [Online]. Available: <https://doi.org/10.1007/BF02179399>
- [40] S. G. Bobkov and F. Götze, “Exponential integrability and transportation cost related to logarithmic sobolev inequalities,” *Journal of Functional Analysis*, vol. 163, no. 1, pp. 1–28, 1999.
- [41] S. Bobkov and G. P. Chistyakov, “Entropy power inequality for the Rényi entropy.” *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 708–714, 2015.
- [42] T. Bohman, “A limit theorem for the shannon capacities of odd cycles i,” *Proceedings of the American Mathematical Society*, vol. 131, no. 11, pp. 3559–3569, 2003.
- [43] H. F. Bohnenblust, “Convex regions and projections in Minkowski spaces,” *Ann. Math.*, vol. 39, no. 2, pp. 301–308, 1938.
- [44] A. Borovkov, *Mathematical Statistics*. CRC Press, 1999.
- [45] S. Boucheron, G. Lugosi, and O. Bousquet, “Concentration inequalities,” in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds. Springer, 2004, pp. 208–240.
- [46] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.
- [47] O. Bousquet, D. Kane, and S. Moran, “The optimal approximation factor in density estimation,” in *Conference on Learning Theory*. PMLR, 2019, pp. 318–341.
- [48] D. Braess and T. Sauer, “Bernstein polynomials and learning theory,” *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.
- [49] D. Braess, J. Forster, T. Sauer, and H. U. Simon, “How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution,” in *Algorithmic Learning Theory*. Springer, 2002, pp. 380–394.
- [50] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, “Communication lower bounds for statistical estimation problems via a distributed data processing inequality,” in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 2016, pp. 1011–1020.
- [51] L. M. Bregman, “Some properties of non-negative matrices and their permanents,” *Soviet Math. Dokl.*, vol. 14, no. 4, pp. 945–949, 1973.
- [52] L. Breiman, “The individual ergodic theorem of information theory,” *Ann. Math. Stat.*, vol. 28, no. 3, pp. 809–811, 1957.
- [53] L. Brillouin, *Science and information theory*, 2nd Ed. Academic Press, 1962.
- [54] L. D. Brown, “Fundamentals of statistical exponential families with applications in statistical decision theory,” in *Lecture Notes-Monograph Series*, S. S. Gupta, Ed. Hayward, CA: Institute of Mathematical Statistics, 1986, vol. 9.
- [55] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [56] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrák, “Maximizing a monotone submodular function subject to a matroid constraint,” *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1740–1766, 2011.
- [57] M. X. Cao and M. Tomamichel, “On the quadratic decaying property of the information rate function,” *arXiv preprint arXiv:2208.12945*, 2022.
- [58] O. Catoni, “PAC-Bayesian supervised classification: the thermodynamics of statistical learning,” *Lecture Notes-Monograph Series. IMS*, vol. 1277, 2007.
- [59] E. Çinlar, *Probability and Stochastics*. New York: Springer, 2011.
- [60] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.
- [61] D. G. Chapman and H. Robbins, “Minimum variance estimation without regularity

References 587

- assumptions,” *The Annals of Mathematical Statistics*, vol. 22, no. 4, pp. 581–586, 1951.
- [62] S. Chatterjee, “An error bound in the Sudakov-Fernique inequality,” *arXiv preprint arXiv:0510424*, 2005.
- [63] S. Chatterjee and P. Diaconis, “The sample size required in importance sampling,” *The Annals of Applied Probability*, vol. 28, no. 2, pp. 1099–1135, 2018.
- [64] Z. Chen, K. Liu, and E. Vigoda, “Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion,” in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 1537–1550.
- [65] H. Chernoff, “Large-sample theory: Parametric case,” *The Annals of Mathematical Statistics*, vol. 27, no. 1, pp. 1–22, 1956.
- [66] M. Choi, M. Ruskai, and E. Seneta, “Equivalence of certain entropy contraction coefficients,” *Linear algebra and its applications*, vol. 208, pp. 29–36, 1994.
- [67] N. Chomsky, “Three models for the description of language,” *IRE Trans. Inform. Th.*, vol. 2, no. 3, pp. 113–124, 1956.
- [68] B. S. Clarke and A. R. Barron, “Information-theoretic asymptotics of Bayes methods,” *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, 1990.
- [69] ——, “Jeffreys’ prior is asymptotically least favorable under entropy risk,” *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.
- [70] J. E. Cohen, J. H. B. Kempermann, and G. Zbăganu, *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*. Springer, 1998.
- [71] A. Collins and Y. Polyanskiy, “Coherent multiple-antenna block-fading channels at finite blocklength,” *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 380–405, 2018.
- [72] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. Springer Science & Business Media, 1999, vol. 290.
- [73] D. J. Costello and G. D. Forney, “Channel coding: The road to channel capacity,” *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1150–1177, 2007.
- [74] T. M. Cover, “Universal data compression and portfolio selection,” in *Proceedings of 37th Conference on Foundations of Computer Science*. IEEE, 1996, pp. 534–538.
- [75] T. M. Cover and B. Gopinath, *Open problems in communication and computation*. Springer Science & Business Media, 2012.
- [76] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd Ed. New York, NY, USA: Wiley-Interscience, 2006.
- [77] H. Cramér, “Über eine eigenschaft der normalen verteilungsfunktion,” *Mathematische Zeitschrift*, vol. 41, no. 1, pp. 405–414, 1936.
- [78] ——, *Mathematical methods of statistics*. Princeton university press, 1946.
- [79] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observation,” *Studia Sci. Math. Hungar.*, vol. 2, pp. 229–318, 1967.
- [80] I. Csiszár and J. Körner, “Graph decomposition: a new key to coding theorems,” *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, 1981.
- [81] ——, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [82] I. Csiszár and G. Tusnády, “Information geometry and alternating minimization problems,” *Statistics & Decision, Supplement Issue No*, vol. 1, 1984.
- [83] I. Csiszár, “ i -divergence geometry of probability distributions and minimization problems,” *The Annals of Probability*, pp. 146–158, 1975.
- [84] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, 2nd ed. Cambridge University Press, 2011.
- [85] P. Cuff, “Distributed channel synthesis,” *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7071–7096, 2013.

588 Strong data processing inequality

- [86] M. Cuturi, “Sinkhorn distances: Light-speed computation of optimal transport,” *Advances in neural information processing systems*, vol. 26, pp. 2292–2300, 2013.
- [87] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. New York: Springer Verlag, 2009.
- [88] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [89] P. Diaconis and L. Saloff-Coste, “Logarithmic Sobolev inequalities for finite Markov chains,” *Ann. Probab.*, vol. 6, no. 3, pp. 695–750, 1996.
- [90] P. Diaconis and D. Freedman, “Finite exchangeable sequences,” *The Annals of Probability*, vol. 8, no. 4, pp. 745–764, 1980.
- [91] P. Diaconis and D. Stroock, “Geometric bounds for eigenvalues of Markov chains,” *The Annals of Applied Probability*, vol. 1, no. 1, pp. 36–61, 1991.
- [92] H. Djellout, A. Guillin, and L. Wu, “Transportation cost-information inequalities and applications to random dynamical systems and diffusions,” *The Annals of Probability*, vol. 32, no. 3B, pp. 2702–2732, 2004.
- [93] R. Dobrushin, “Central limit theorem for nonstationary Markov chains, I,” *Theory Probab. Appl.*, vol. 1, no. 1, pp. 65–80, 1956.
- [94] R. Dobrushin and B. Tsybakov, “Information transmission with additional noise,” *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.
- [95] R. L. Dobrushin, “A general formulation of the fundamental theorem of Shannon in the theory of information,” *Uspekhi Mat. Nauk*, vol. 14, no. 6, pp. 3–104, 1959, english translation in *Eleven Papers in Analysis: Nine Papers on Differential Equations, Two on Information Theory*, American Mathematical Society Translations: Series 2, Volume 33, 1963.
- [96] ——, “Mathematical problems in the Shannon theory of optimal coding of information,” in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 1, Berkeley, CA, USA, 1961, pp. 211–252.
- [97] ——, “Asymptotic bounds on error probability for transmission over DMC with symmetric transition probabilities,” *Theor. Probability Appl.*, vol. 7, pp. 283–311, 1962.
- [98] R. Dobrushin, “A simplified method of experimentally evaluating the entropy of a stationary sequence,” *Theory of Probability & Its Applications*, vol. 3, no. 4, pp. 428–430, 1958.
- [99] D. L. Donoho, “Wald lecture I: Counting bits with Kolmogorov and Shannon,” *Note for the Wald Lectures, IMS Annual Meeting*, July 1997.
- [100] M. D. Donsker and S. S. Varadhan, “Asymptotic evaluation of certain markov process expectations for large time. iv,” *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.
- [101] J. L. Doob, *Stochastic Processes*. New York Wiley, 1953.
- [102] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, “Optimality guarantees for distributed statistical estimation,” *arXiv preprint arXiv:1405.0782*, 2014.
- [103] J. Duda, “Asymmetric numeral systems: entropy coding combining speed of huffman coding with compression rate of arithmetic coding,” *arXiv preprint arXiv:1311.2540*, 2013.
- [104] R. M. Dudley, *Uniform central limit theorems*. Cambridge university press, 1999, no. 63.
- [105] G. Dueck, “The strong converse to the coding theorem for the multiple-access channel,” *J. Comb. Inform. Syst. Sci.*, vol. 6, no. 3, pp. 187–196, 1981.
- [106] G. Dueck and J. Korner, “Reliability function of a discrete memoryless channel at rates above capacity (corresp.),”

References 589

- [107] N. Dunford and J. T. Schwartz, *Linear operators, part 1: general theory*. John Wiley & Sons, 1988, vol. 10.
- [108] R. Durrett, *Probability: Theory and Examples*, 4th ed. Cambridge University Press, 2010.
- [109] A. Dytso, S. Yagli, H. V. Poor, and S. S. Shitz, “The capacity achieving distribution for the amplitude constrained additive gaussian channel: An upper bound on the number of mass points,” *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2006–2022, 2019.
- [110] H. G. Eggleston, *Convexity*, ser. Tracts in Math and Math. Phys. Cambridge University Press, 1958, vol. 47.
- [111] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.
- [112] P. Elias, “The efficient construction of an unbiased random sequence,” *Annals of Mathematical Statistics*, vol. 43, no. 3, pp. 865–870, 1972.
- [113] ——, “Coding for noisy channels,” *IRE Convention Record*, vol. 3, pp. 37–46, 1955.
- [114] D. M. Endres and J. E. Schindelin, “A new metric for probability distributions,” *IEEE Transactions on Information theory*, vol. 49, no. 7, pp. 1858–1860, 2003.
- [115] K. Eswaran and M. Gastpar, “Remote source coding under Gaussian noise: Dueling roles of power and entropy power,” *IEEE Transactions on Information Theory*, 2019.
- [116] W. Evans and N. Pippenger, “On the maximum tolerable noise for reliable computation by formulas,” *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1299–1305, May 1998.
- [117] W. S. Evans and L. J. Schulman, “Signal propagation and noisy circuits,” *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2367–2373, Nov 1999.
- [118] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.
- [119] M. Feder and Y. Polyanskiy, “Sequential prediction under log-loss and misspecification,” *arXiv preprint arXiv:2102.00050*, 2021.
- [120] A. A. Fedotov, P. Harremoës, and F. Topsøe, “Refinements of Pinsker’s inequality,” *Information Theory, IEEE Transactions on*, vol. 49, no. 6, pp. 1491–1498, Jun. 2003.
- [121] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: Wiley, 1970, vol. I.
- [122] ——, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York: Wiley, 1971, vol. II.
- [123] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. New York, NY: Academic Press, 1967.
- [124] ——, “An inconsistent maximum likelihood estimate,” *Journal of the American Statistical Association*, vol. 77, no. 380, pp. 831–834, 1982.
- [125] ——, *A course in large sample theory*. CRC Press, 1996.
- [126] R. A. Fisher, “The logic of inductive inference,” *Journal of the royal statistical society*, vol. 98, no. 1, pp. 39–82, 1935.
- [127] B. M. Fitingof, “The compression of discrete information,” *Problemy Peredachi Informatsii*, vol. 3, no. 3, pp. 28–36, 1967.
- [128] P. Fleisher, “Sufficient conditions for achieving minimum distortion in a quantizer,” *IEEE Int. Conv. Rec.*, pp. 104–111, 1964.
- [129] G. D. Forney, “Concatenated codes,” *MIT RLE Technical Rep.*, vol. 440, 1965.
- [130] E. Friedgut and J. Kahn, “On the number of copies of one hypergraph in another,” *Israel J. Math.*, vol. 105, pp. 251–256, 1998. [Online]. Available: <http://dx.doi.org/10.1007/BF02780332>
- [131] R. G. Gallager, “A simple derivation of the coding theorem and some applications,”

590 Strong data processing inequality

- IEEE Trans. Inf. Theory*, vol. 11, no. 1, pp. 3–18, 1965.
- [132] ——, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [133] R. Gallager, “The random coding bound is tight for the average code (corresp.),” *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 244–246, 1973.
- [134] R. Gardner, “The Brunn-Minkowski inequality,” *Bulletin of the American Mathematical Society*, vol. 39, no. 3, pp. 355–405, 2002.
- [135] I. M. Gel’fand, A. N. Kolmogorov, and A. M. Yaglom, “On the general definition of the amount of information,” *Dokl. Akad. Nauk. SSSR*, vol. 11, pp. 745–748, 1956.
- [136] G. L. Gilardoni, “On a Gel’fand-Yaglom-Peres theorem for f -divergences,” *arXiv preprint arXiv:0911.1934*, 2009.
- [137] ——, “On pinsker’s and vajda’s type inequalities for csiszár’s-divergences,” *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5377–5386, 2010.
- [138] R. D. Gill and B. Y. Levit, “Applications of the van Trees inequality: a Bayesian Cramér-Rao bound,” *Bernoulli*, vol. 1, no. 1–2, pp. 59–79, 1995.
- [139] C. Giraud, *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2014.
- [140] O. Goldreich, *Introduction to property testing*. Cambridge University Press, 2017.
- [141] V. Goodman, “Characteristics of normal samples,” *The Annals of Probability*, vol. 16, no. 3, pp. 1281–1290, 1988.
- [142] V. D. Goppa, “Codes and information,” *Russian Mathematical Surveys*, vol. 39, no. 1, p. 87, 1984.
- [143] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.
- [144] R. M. Gray, *Entropy and Information Theory*. New York, NY: Springer-Verlag, 1990.
- [145] U. Grenander and G. Szegö, *Toeplitz forms and their applications*, 2nd ed. New York: Chelsea Publishing Company, 1984.
- [146] L. Gross, “Logarithmic sobolev inequalities,” *American Journal of Mathematics*, vol. 97, no. 4, pp. 1061–1083, 1975.
- [147] Y. Gu and Y. Polyanskiy, “Non-linear log-sobolev inequalities for the potts semigroup and applications to reconstruction problems,” *arXiv preprint arXiv:2005.05444*, 2020.
- [148] U. Hadar, J. Liu, Y. Polyanskiy, and O. Shayevitz, “Communication complexity of estimating correlations,” in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2019, pp. 792–803.
- [149] B. Hajek, Y. Wu, and J. Xu, “Information limits for recovering a hidden community,” *IEEE Trans. on Information Theory*, vol. 63, no. 8, pp. 4729 – 4745, 2017.
- [150] J. Hájek, “Local asymptotic minimax and admissibility in estimation,” in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1972, pp. 175–194.
- [151] J. M. Hammersley, “On estimating restricted parameters,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 12, no. 2, pp. 192–240, 1950.
- [152] T. S. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, 1993.
- [153] P. Harremoës and I. Vajda, “On pairs of f -divergences and their joint range,” *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3230–3235, Jun. 2011.
- [154] B. Harris, “The statistical estimation of entropy in the non-parametric case,” in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Springer Netherlands, 1975, vol. 16, pp. 323–355.
- [155] D. Haussler and M. Opper, “Mutual information, metric entropy and cumulative relative entropy risk,” *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.

References 591

- [156] M. Hayashi, “General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1562–1575, 2006.
- [157] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *The Annals of Mathematical Statistics*, pp. 369–401, 1965.
- [158] P. J. Huber, “Fisher information and spline interpolation,” *Annals of Statistics*, pp. 1029–1033, 1974.
- [159] ———, “A robust version of the probability ratio test,” *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.
- [160] I. A. Ibragimov and R. Z. Khas'minskij, *Statistical Estimation: Asymptotic Theory*. Springer, 1981.
- [161] S. Ihara, “On the capacity of channels with additive non-Gaussian noise,” *Information and Control*, vol. 37, no. 1, pp. 34–39, 1978.
- [162] ———, *Information theory for continuous systems*. World Scientific, 1993, vol. 2.
- [163] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. New York, NY: Springer, 2003.
- [164] Y. I. Ingster, “Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics,” *Theory of Probability & Its Applications*, vol. 31, no. 2, pp. 333–337, 1987.
- [165] I. Jensen and A. J. Guttmann, “Series expansions of the percolation probability for directed square and honeycomb lattices,” *Journal of Physics A: Mathematical and General*, vol. 28, no. 17, p. 4813, 1995.
- [166] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Minimax estimation of functionals of discrete distributions,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.
- [167] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, “Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences,” in *Advances in neural information processing systems*, 2016, pp. 4116–4124.
- [168] W. B. Johnson, G. Schechtman, and J. Zinn, “Best constants in moment inequalities for linear combinations of independent and exchangeable random variables,” *The Annals of Probability*, pp. 234–253, 1985.
- [169] I. Johnstone, *Gaussian estimation: Sequence and wavelet models*, 2011, available at <http://www-stat.stanford.edu/~imj/>.
- [170] L. K. Jones, “A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training,” *The Annals of Statistics*, pp. 608–613, 1992.
- [171] A. B. Juditsky and A. S. Nemirovski, “Nonparametric estimation by convex programming,” *The Annals of Statistics*, vol. 37, no. 5A, pp. 2278–2300, 2009.
- [172] T. Kawabata and A. Dembo, “The rate-distortion dimension of sets and measures,” *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1564 – 1572, Sep. 1994.
- [173] M. Keane and G. O’Brien, “A Bernoulli factory,” *ACM Transactions on Modeling and Computer Simulation*, vol. 4, no. 2, pp. 213–219, 1994.
- [174] H. Kesten and B. P. Stigum, “Additional limit theorems for indecomposable multidimensional galton-watson processes,” *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1463–1481, 1966.
- [175] T. Koch, “The Shannon lower bound is asymptotically tight,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6155–6161, 2016.
- [176] Y. Kochman, O. Ordentlich, and Y. Polyanskiy, “A lower bound on the expected distortion of joint source-channel coding,” *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4722–4741, 2020.
- [177] A. Kolchinsky and B. D. Tracey, “Estimating mixture entropy with pairwise distances,” *Entropy*, vol. 19, no. 7, p. 361, 2017.

592 Strong data processing inequality

- [178] A. N. Kolmogorov and V. M. Tikhomirov, “ ε -entropy and ε -capacity of sets in function spaces,” *Uspekhi Matematicheskikh Nauk*, vol. 14, no. 2, pp. 3–86, 1959, reprinted in Shirayev, A. N., ed. *Selected Works of AN Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*, Vol. 27, Springer Netherlands, 1993, pp 86–170.
- [179] I. Kontoyiannis and S. Verdú, “Optimal lossless data compression: Non-asymptotics and asymptotics,” *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, 2014.
- [180] J. Körner and A. Orlitsky, “Zero-error information theory,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2207–2229, 1998.
- [181] V. Koshelev, “Quantization with minimal entropy,” *Probl. Pered. Inform.*, vol. 14, pp. 151–156, 1963.
- [182] O. Kosut and L. Sankar, “Asymptotics and non-asymptotics for universal fixed-to-variable source coding,” *arXiv preprint arXiv:1412.4444*, 2014.
- [183] A. Krause and D. Golovin, “Submodular function maximization,” *Tractability*, vol. 3, pp. 71–104, 2014.
- [184] J. Kuelbs, “A strong convergence theorem for banach space valued random variables,” *The Annals of Probability*, vol. 4, no. 5, pp. 744–771, 1976.
- [185] J. Kuelbs and W. V. Li, “Metric entropy and the small ball problem for Gaussian measures,” *Journal of Functional Analysis*, vol. 116, no. 1, pp. 133–157, 1993.
- [186] S. Kullback, *Information theory and statistics*. Mineola, NY: Dover publications, 1968.
- [187] C. Külske and M. Formentin, “A symmetric entropy bound on the non-reconstruction regime of Markov chains on Galton-Watson trees,” *Electronic Communications in Probability*, vol. 14, pp. 587–596, 2009.
- [188] A. Lapidot, *A foundation in digital communication*. Cambridge University Press, 2017.
- [189] A. Lapidot and S. M. Moser, “Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2426–2467, 2003.
- [190] L. Le Cam, “Convergence of estimates under dimensionality restrictions,” *Annals of Statistics*, vol. 1, no. 1, pp. 38 – 53, 1973.
- [191] ——, *Asymptotic methods in statistical decision theory*. New York, NY: Springer-Verlag, 1986.
- [192] C. C. Leang and D. H. Johnson, “On the asymptotics of m -hypothesis Bayesian detection,” *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 280–282, 1997.
- [193] K. Lee, Y. Wu, and Y. Bresler, “Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization,” *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1666–1698, Mar. 2018.
- [194] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY: Springer, 1998.
- [195] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.
- [196] W. V. Li and W. Linde, “Approximation, metric entropy and small ball estimates for Gaussian measures,” *The Annals of Probability*, vol. 27, no. 3, pp. 1556–1578, 1999.
- [197] W. V. Li and Q.-M. Shao, “Gaussian processes: inequalities, small ball probabilities and applications,” *Handbook of Statistics*, vol. 19, pp. 533–597, 2001.
- [198] T. Linder and R. Zamir, “On the asymptotic tightness of the Shannon lower bound,” *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2026–2031, 1994.
- [199] S. Litsyn, “New upper bounds on error exponents,” *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 385–398, 1999.
- [200] S. Lloyd, “Least squares quantization in pcm,” *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

References 593

- [201] G. G. Lorentz, M. v. Golitschek, and Y. Makovoz, *Constructive approximation: advanced problems*. Springer, 1996, vol. 304.
- [202] L. Lovász, “On the shannon capacity of a graph,” *IEEE Transactions on Information theory*, vol. 25, no. 1, pp. 1–7, 1979.
- [203] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [204] M. Madiman and P. Tetali, “Information inequalities for joint distributions, with interpretations and applications,” *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2699–2713, 2010.
- [205] M. Mahoney, “Large text compression benchmark,” <http://www.mattmahoney.net/dc/text.html>, Aug. 2021.
- [206] A. Makur and Y. Polyanskiy, “Comparison of channels: Criteria for domination by a symmetric channel,” *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5704–5725, 2018.
- [207] B. Mandelbrot, “An informational theory of the statistical structure of language,” *Communication theory*, vol. 84, pp. 486–502, 1953.
- [208] J. Massey, “On the fractional weight of distinct binary n -tuples (corresp.),” *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 131–131, 1974.
- [209] ———, “Causality, feedback and directed information,” in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, 1990, pp. 303–305.
- [210] W. Matthews, “A linear program for the finite block length converse of polyanskiy–poor–verdú via nonsignaling codes,” *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7036–7044, 2012.
- [211] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, “Maximizing the information learned from finite data selects a simple model,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1760–1765, 2018.
- [212] R. McEliece, E. Rodemich, H. Rumsey, and L. Welch, “New upper bounds on the rate of a code via the Delsarte-Macwilliams inequalities,” *IEEE transactions on Information Theory*, vol. 23, no. 2, pp. 157–166, 1977.
- [213] R. J. McEliece and E. C. Posner, “Hide and seek, data storage, and entropy,” *The Annals of Mathematical Statistics*, vol. 42, no. 5, pp. 1706–1716, 1971.
- [214] B. McMillan, “The basic theorems of information theory,” *Ann. Math. Stat.*, pp. 196–219, 1953.
- [215] F. McSherry, “Spectral partitioning of random graphs,” in *42nd IEEE Symposium on Foundations of Computer Science*, Oct. 2001, pp. 529 – 537.
- [216] N. Merhav and M. Feder, “Universal prediction,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.
- [217] G. A. Miller, “Note on the bias of information estimates,” *Information theory in psychology: Problems and methods*, vol. 2, pp. 95–100, 1955.
- [218] M. Mitzenmacher, “A brief history of generative models for power law and lognormal distributions,” *Internet mathematics*, vol. 1, no. 2, pp. 226–251, 2004.
- [219] E. Mossel and Y. Peres, “New coins from old: computing with unknown bias,” *Combinatorica*, vol. 25, no. 6, pp. 707–724, 2005.
- [220] X. Mu, L. Pomatto, P. Strack, and O. Tamuz, “From Blackwell dominance in large samples to rényi divergences and back again,” *Econometrica*, vol. 89, no. 1, pp. 475–506, 2021.
- [221] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions—I,” *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.
- [222] J. Neveu, *Mathematical foundations of the calculus of probability*. Holden-day, 1965.
- [223] M. E. Newman, “Power laws, pareto distributions and zipf’s law,” *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.
- [224] M. Okamoto, “Some inequalities relating to the partial sum of binomial probabilities,”

594 Strong data processing inequality

- Annals of the institute of Statistical Mathematics*, vol. 10, no. 1, pp. 29–35, 1959.
- [225] B. Oliver, J. Pierce, and C. Shannon, “The philosophy of PCM,” *Proceedings of the IRE*, vol. 36, no. 11, pp. 1324–1331, 1948.
- [226] Y. Oohama, “On two strong converse theorems for discrete memoryless channels,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 98, no. 12, pp. 2471–2475, 2015.
- [227] L. Paninski, “Variational minimax estimation of discrete distributions under kl loss,” *Advances in Neural Information Processing Systems*, vol. 17, 2004.
- [228] P. Panter and W. Dite, “Quantization distortion in pulse-count modulation with nonuniform spacing of levels,” *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.
- [229] M. Pardo and I. Vajda, “About distances of discrete distributions satisfying the data processing theorem of information theory,” *IEEE transactions on information theory*, vol. 43, no. 4, pp. 1288–1293, 1997.
- [230] Y. Peres, “Iterating von Neumann’s procedure for extracting random bits,” *Annals of Statistics*, vol. 20, no. 1, pp. 590–597, 1992.
- [231] M. S. Pinsker, “Optimal filtering of square-integrable signals in Gaussian noise,” *Problemy Peredachi Informatsii*, vol. 16, no. 2, pp. 52–68, 1980.
- [232] G. Pisier, *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1999.
- [233] Y. Polyanskiy, “Channel coding: non-asymptotic fundamental limits,” Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010.
- [234] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
- [235] ——, “Dispersion of the Gilbert-Elliott channel,” *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, Apr. 2011.
- [236] ——, “Feedback in the non-asymptotic regime,” *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 4903 – 4925, Apr. 2011.
- [237] ——, “Minimum energy to send k bits with and without feedback,” *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4880–4902, Aug. 2011.
- [238] Y. Polyanskiy and S. Verdú, “Arimoto channel coding converse and Rényi divergence,” in *Proceedings of the Forty-eighth Annual Allerton Conference on Communication, Control, and Computing*, 2010, pp. 1327–1333.
- [239] Y. Polyanskiy and S. Verdú, “Arimoto channel coding converse and Rényi divergence,” in *Proc. 2010 48th Allerton Conference*, Allerton Retreat Center, Monticello, IL, USA, Sep. 2010.
- [240] Y. Polyanskiy and S. Verdú, “Binary hypothesis testing with feedback,” in *Information Theory and Applications Workshop (ITA)*, 2011.
- [241] Y. Polyanskiy and S. Verdú, “Empirical distribution of good channel codes with non-vanishing error probability,” *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 5–21, Jan. 2014.
- [242] Y. Polyanskiy and Y. Wu, “Peak-to-average power ratio of good codes for Gaussian channel,” *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7655–7660, Dec. 2014.
- [243] Y. Polyanskiy, “Saddle point in the minimax converse for channel coding,” *IEEE Transactions on Information Theory*, vol. 59, no. 5, pp. 2576–2595, 2012.
- [244] ——, “On dispersion of compound dmcs,” in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2013, pp. 26–32.
- [245] Y. Polyanskiy and Y. Wu, “Strong data-processing inequalities for channels and Bayesian networks,” in *Convexity and Concentration. The IMA Volumes in Mathematics and its Applications*, vol 161, E. Carlen, M. Madiman, and E. M. Werner, Eds. New York, NY: Springer, 2017, pp. 211–249.

- [246] ——, “Dualizing Le Cam’s method for functional estimation, with applications to estimating the unseens,” *arXiv preprint arXiv:1902.05616*, 2019.
- [247] ——, “Application of the information-percolation method to reconstruction problems on graphs,” *Mathematical Statistics and Learning*, vol. 2, no. 1, pp. 1–24, 2020.
- [248] ——, “Self-regularizing property of non-parametric maximum likelihood estimator in mixture models,” *arXiv preprint arXiv:2008.08244*, 2020.
- [249] E. C. Posner and E. R. Rodemich, “Epsilon entropy and data compression,” *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 2079–2125, Dec. 1971.
- [250] A. Prékopa, “Logarithmic concave measures with application to stochastic programming,” *Acta Scientiarum Mathematicarum*, vol. 32, pp. 301–316, 1971.
- [251] J. Radhakrishnan, “An entropy proof of Bregman’s theorem,” *J. Combin. Theory Ser. A*, vol. 77, no. 1, pp. 161–164, 1997.
- [252] M. Raginsky, “Strong data processing inequalities and ϕ -Sobolev inequalities for discrete channels,” *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.
- [253] M. Raginsky and I. Sason, “Concentration of measure inequalities in information theory, communications, and coding,” *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.
- [254] C. R. Rao, “Information and the accuracy attainable in the estimation of statistical parameters,” *Bull. Calc. Math. Soc.*, vol. 37, pp. 81–91, 1945.
- [255] A. H. Reeves, “The past present and future of PCM,” *IEEE Spectrum*, vol. 2, no. 5, pp. 58–62, 1965.
- [256] A. Rényi, “On the dimension and entropy of probability distributions,” *Acta Mathematica Hungarica*, vol. 10, no. 1 – 2, Mar. 1959.
- [257] R. B. Reznikova Zh., “Analysis of the language of ants by information-theoretical methods,” *Problemi Peredachi Informatsii*, vol. 22, no. 3, pp. 103–108, 1986, english translation: <http://reznikova.net/R-R-entropy-09.pdf>.
- [258] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, “Design of capacity-approaching irregular low-density parity-check codes,” *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, 2001.
- [259] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge University Press, 2008.
- [260] P. Rigollet and J.-C. Hütter, “High dimensional statistics,” *Lecture Notes for 18.657, MIT*, 2017, <https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf>.
- [261] Y. Rinott, “On convexity of measures,” *Annals of Probability*, vol. 4, no. 6, pp. 1020–1026, 1976.
- [262] J. J. Rissanen, “Fisher information and stochastic complexity,” *IEEE transactions on information theory*, vol. 42, no. 1, pp. 40–47, 1996.
- [263] C. Rogers, *Packing and Covering*, ser. Cambridge tracts in mathematics and mathematical physics. Cambridge University Press, 1964.
- [264] H. Roozbehani and Y. Polyanskiy, “Low density majority codes and the problem of graceful degradation,” *arXiv preprint arXiv:1911.12263*, 2019.
- [265] H. P. Rosenthal, “On the subspaces of $\ell^p (p > 2)$ spanned by sequences of independent random variables,” *Israel Journal of Mathematics*, vol. 8, no. 3, pp. 273–303, 1970.
- [266] D. Russo and J. Zou, “Controlling bias in adaptive data analysis using information theory,” in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1232–1240.
- [267] I. Sason and S. Verdu, “ f -divergence inequalities,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

596 Strong data processing inequality

- [268] G. Schechtman, “Extremal configurations for moments of sums of independent positive random variables,” in *Banach Spaces and their Applications in Analysis*. De Gruyter, 2011, pp. 183–192.
- [269] M. J. Schervish, *Theory of statistics*. Springer-Verlag New York, 1995.
- [270] A. Schrijver, *Theory of linear and integer programming*. John Wiley & Sons, 1998.
- [271] C. E. Shannon, “A symbolic analysis of relay and switching circuits,” *Electrical Engineering*, vol. 57, no. 12, pp. 713–723, Dec 1938.
- [272] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, Jul./Oct. 1948.
- [273] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels i,” *Inf. Contr.*, vol. 10, pp. 65–103, 1967.
- [274] C. Shannon, “The zero error capacity of a noisy channel,” *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.
- [275] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE Nat. Conv. Rec.*, vol. 4, no. 142–163, p. 1, 1959.
- [276] O. Shayevitz, “On Rényi measures and hypothesis testing,” in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 894–898.
- [277] O. Shayevitz and M. Feder, “Optimal feedback communication via posterior matching,” *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, 2011.
- [278] G. Simons and M. Woodroffe, “The Cramér-Rao inequality holds almost everywhere,” in *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*. Academic, New York, 1983, pp. 69–93.
- [279] R. Sinkhorn, “A relationship between arbitrary positive matrices and doubly stochastic matrices,” *Ann. Math. Stat.*, vol. 35, no. 2, pp. 876–879, 1964.
- [280] M. Sion, “On general minimax theorems,” *Pacific J. Math*, vol. 8, no. 1, pp. 171–176, 1958.
- [281] M.-K. Siu, “Which latin squares are cayley tables?” *Amer. Math. Monthly*, vol. 98, no. 7, pp. 625–627, Aug. 1991.
- [282] D. Slepian and H. O. Pollak, “Prolate spheroidal wave functions, fourier analysis and uncertainty—i,” *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.
- [283] A. Sly, “Reconstruction of random colourings,” *Communications in Mathematical Physics*, vol. 288, no. 3, pp. 943–961, Jun 2009. [Online]. Available: <https://doi.org/10.1007/s00220-009-0783-7>
- [284] B. Smith, “Instantaneous companding of quantized signals,” *Bell System Technical Journal*, vol. 36, no. 3, pp. 653–709, 1957.
- [285] J. G. Smith, “The information capacity of amplitude and variance-constrained scalar Gaussian channels,” *Information and Control*, vol. 18, pp. 203 – 219, 1971.
- [286] Spectre, “SPECTRE: Short packet communication toolbox,” <https://github.com/yp-mit/spectre>, 2015, GitHub repository.
- [287] R. Speer, J. Chin, A. Lin, S. Jewett, and L. Nathan, “Luminosoinsight/word-freq: v2.2,” Oct. 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1443582>
- [288] A. J. Stam, “Distance between sampling with and without replacement,” *Statistica Neerlandica*, vol. 32, no. 2, pp. 81–91, 1978.
- [289] M. Steiner, “The strong simplex conjecture is false,” *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 721–731, 1994.
- [290] V. Strassen, “Asymptotische Abschätzungen in Shannon’s Informationstheorie,” in *Trans. 3d Prague Conf. Inf. Theory*, Prague, 1962, pp. 689–723.
- [291] —, “The existence of probability measures with given marginals,” *Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 423–439, 1965.
- [292] H. Strasser, *Mathematical theory of statistics: Statistical experiments and asymptotic*

References 597

- decision theory.* Berlin, Germany: Walter de Gruyter, 1985.
- [293] S. Szarek, “Nets of Grassmann manifold and orthogonal groups,” in *Proceedings of Banach Space Workshop*. University of Iowa Press, 1982, pp. 169–185.
- [294] ——, “Metric entropy of homogeneous spaces,” *Banach Center Publications*, vol. 43, no. 1, pp. 395–410, 1998.
- [295] W. Szpankowski and S. Verdú, “Minimum expected length of fixed-to-variable lossless compression without prefix constraints,” *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4017–4025, 2011.
- [296] I. Tal and A. Vardy, “List decoding of polar codes,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2213–2226, 2015.
- [297] M. Talagrand, *Upper and lower bounds for stochastic processes*. Springer, 2014.
- [298] G. Taricco and M. Elia, “Capacity of fading channel with no side information,” *Electronics Letters*, vol. 33, no. 16, pp. 1368–1370, 1997.
- [299] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, “Space-time block codes from orthogonal designs,” *IEEE Transactions on Information theory*, vol. 45, no. 5, pp. 1456–1467, 1999.
- [300] V. Tarokh, N. Seshadri, and A. R. Calderbank, “Space-time codes for high data rate wireless communication: Performance criterion and code construction,” *IEEE transactions on information theory*, vol. 44, no. 2, pp. 744–765, 1998.
- [301] H. Te Sun, *Information-spectrum methods in information theory*. Springer Science & Business Media, 2003.
- [302] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *European trans. telecom.*, vol. 10, no. 6, pp. 585–595, 1999.
- [303] ——, “Wringing lemmas and multiple descriptions,” 2016, unpublished draft.
- [304] V. N. Temlyakov, “On estimates of ϵ -entropy and widths of classes of functions with a bounded mixed derivative or difference,” *Doklady Akademii Nauk*, vol. 301, no. 2, pp. 288–291, 1988.
- [305] F. Topsøe, “Some inequalities for information divergence and related measures of discrimination,” *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1602–1609, 2000.
- [306] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005. [Online]. Available: <http://www.eecs.berkeley.edu/~dtse/book.html>
- [307] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY: Springer Verlag, 2009.
- [308] B. P. Tunstall, “Synthesis of noiseless compression codes,” Ph.D. dissertation, Georgia Institute of Technology, 1967.
- [309] E. Uhrmann-Klingen, “Minimal Fisher information distributions with compact-supports,” *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 360–374, 1995.
- [310] I. Vajda, “Note on discrimination information and variation (corresp.),” *IEEE Transactions on Information Theory*, vol. 16, no. 6, pp. 771–773, 1970.
- [311] G. Valiant and P. Valiant, “Estimating the unseen: an $n/\log(n)$ -sample estimator for entropy and support size, shown optimal via new CLTs,” in *Proceedings of the 43rd annual ACM symposium on Theory of computing*, 2011, pp. 685–694.
- [312] A. van der Vaart, “The statistical work of Lucien Le Cam,” *Annals of Statistics*, pp. 631–682, 2002.
- [313] T. Van Erven and P. Harremoës, “Rényi divergence and kullback-leibler divergence,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.
- [314] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. Wiley, New York, 1968.
- [315] S. Verdú, “On channel capacity per unit cost,” *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1019–1030, Sep. 1990.

598 Strong data processing inequality

- [316] ——, *Multiuser Detection*. Cambridge, UK: Cambridge Univ. Press, 1998.
- [317] A. G. Vitushkin, “On the 13th problem of Hilbert,” *Dokl. Akad. Nauk SSSR*, vol. 95, no. 4, pp. 701–704, 1954.
- [318] ——, “On hilbert’s thirteenth problem and related questions,” *Russian Mathematical Surveys*, vol. 59, no. 1, p. 11, 2004.
- [319] ——, *Theory of the Transmission and Processing of Information*. Pergamon Press, 1961.
- [320] J. von Neumann, “Various techniques used in connection with random digits,” *Monte Carlo Method, National Bureau of Standards, Applied Math Series*, no. 12, pp. 36–38, 1951.
- [321] V. G. Vovk, “Aggregating strategies,” *Proc. of Computational Learning Theory*, 1990, 1990.
- [322] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.
- [323] A. Wald, “Sequential tests of statistical hypotheses,” *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.
- [324] ——, “Note on the consistency of the maximum likelihood estimate,” *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, 1949.
- [325] A. Wald and J. Wolfowitz, “Optimum character of the sequential probability ratio test,” *The Annals of Mathematical Statistics*, pp. 326–339, 1948.
- [326] M. M. Wilde, *Quantum information theory*. Cambridge University Press, 2013.
- [327] J. Wolfowitz, “On wald’s proof of the consistency of the maximum likelihood estimate,” *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 601–602, 1949.
- [328] Y. Wu and J. Xu, “Statistical problems with planted structures: Information-theoretical and computational limits,” in *Information-Theoretic Methods in Data Science*, Y. Eldar and M. Rodrigues, Eds. Cambridge University Press, 2020, arXiv:1806.00118.
- [329] Y. Wu and P. Yang, “Minimax rates of entropy estimation on large alphabets via best polynomial approximation,” *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.
- [330] A. Wyner, “The common information of two dependent random variables,” *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.
- [331] Q. Xie and A. R. Barron, “Minimax redundancy for the class of memoryless sources,” *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, 1997.
- [332] A. Xu and M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” *arXiv preprint arXiv:1705.07809*, 2017.
- [333] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, “Quasi-static multiple-antenna fading channels at finite blocklength,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.
- [334] W. Yang, G. Durisi, and Y. Polyanskiy, “Minimum energy to send k bits over multiple-antenna fading channels,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 6831–6853, 2016.
- [335] Y. Yang and A. R. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.
- [336] Y. G. Yatracos, “Rates of convergence of minimum distance estimators and Kolmogorov’s entropy,” *The Annals of Statistics*, pp. 768–774, 1985.
- [337] S. Yekhanin, “Improved upper bound for the redundancy of fix-free codes,” *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2815–2818, 2004.
- [338] P. L. Zador, “Development and evaluation of procedures for quantizing multivariate distributions,” Ph.D. dissertation, Stanford University, Department of Statistics, 1963.
- [339] ——, “Asymptotic quantization error of continuous signals and the quantization

References 599

- dimension,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 139–149, 1982.
- [340] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.
- [341] Z. Zhang and R. W. Yeung, “A non-Shannon-type conditional inequality of information quantities,” *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1982–1986, 1997.
- [342] ——, “On characterization of entropy function via information inequalities,” *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1440–1452, 1998.
- [343] L. Zheng and D. N. C. Tse, “Communication on the grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel,” *IEEE transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002.
- [344] G. Zipf, *Selective Studies and the Principle of Relative Frequency in Language*. Cambridge MA: Harvard University Press, 1932.