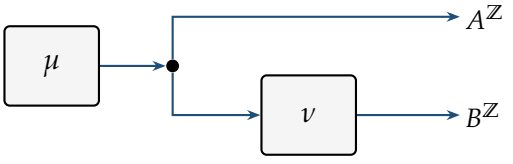
Higher Categorical Semantics

Entropy, Sources, and Channels in Higher Categories

Cocycles, Markov categories, and coding in the mathematical foundations of information theory



Anonymous

2026

About this text. This monograph is a higher-categorical treatment of the territory mapped by A. I. Khinchin's *Mathematical Foundations of Information Theory* (Dover, 1957; the English translation, by R. A. Silverman and M. D. Friedman, of two papers of 1953 and 1956): the entropy of finite schemes and its uniqueness, stationary and ergodic sources, McMillan's theorem, channels with memory, Feinstein's lemma, and Shannon's coding theorems. It is an independent work. It reproduces none of that book's text or proofs verbatim; it is not affiliated with or endorsed by its publisher. Appendix B gives a concordance so that the two can be read side by side.

Status of results. Every numbered statement carries a tag: classical (standard information theory, probability or ergodic theory, attributed), literature (established in the categorical or homological literature, attributed), reformulation (a known result restated in a different language, with the restatement proved or sketched here), observation (an elementary remark we have not found stated in this form), or programme (a conjecture or a direction, not a result). No cited theorem is claimed as new.

Authorship and assistance. The text was drafted and typeset with a large language model (Anthropic's Claude) under the direction of the submitting author, who is responsible for it. Generative AI assisted with drafting, literature recall, mathematical consistency checks, computation, and typesetting. The numerical examples and a number of finite structural claims are recomputed by the accompanying script `verify.py` (Appendix C). Proofs and bibliographic details have not been independently refereed; corrections are welcome.

Open access. Rights remain with the author.

Contents

Preface vii

Introduction ix

I Entropy 1

1 Finite Schemes 2
- 1.1 Schemes and their maps . . . 2
- 1.2 Entropy . . . 3
- 1.3 The entropy of a map . . . 4
- 1.4 Mixtures and the operad of simplices . . . 5
- 1.5 Kernels, noise and equivocation . . . 6

2 The Uniqueness Theorem 8
- 2.1 Khinchin's theorem . . . 8
- 2.2 Faddeev's theorem and the operadic form . . . 9
- 2.3 The functorial characterization . . . 10
- 2.4 Entropy as a derivation . . . 11
- 2.5 What the chain rule excludes . . . 11

3 Entropy as a Cocycle 14
- 3.1 Observables . . . 14
- 3.2 The conditioning action . . . 14
- 3.3 The complex . . . 15
- 3.4 The fundamental equation . . . 16
- 3.5 Deformation . . . 18
- 3.6 A derivation and its defect; entropy modulo a prime . . . 19
- 3.7 The general setting, and higher degrees . . . 20

4 Divergence and the Elementary Inequalities 22
- 4.1 Relative entropy . . . 22
- 4.2 Composition is short . . . 23
- 4.3 Entropy as distance from determinism . . . 24
- 4.4 Shannon's inequality and its two generalizations . . . 24
- 4.5 Feinstein's three inequalities . . . 25

4.6 Relative entropy as a functor . . . 26
4.7 Beyond the elementary inequalities . . . 26

II Sources **28**

5 Sources **29**
5.1 Sequence spaces . . . 29
5.2 Stationary sources . . . 29
5.3 The entropy rate . . . 30
5.4 Stationary codes and cofreeness . . . 31
5.5 Markov chains . . . 32
5.6 Finite-state sources as coalgebras . . . 32
5.7 Entropy as an invariant . . . 33

6 Ergodic Sources **36**
6.1 Invariant sets . . . 36
6.2 The simplex of stationary sources . . . 37
6.3 Ergodicity as determinism . . . 37
6.4 Entropy is affine . . . 38
6.5 An example . . . 39

7 Martingales **41**
7.1 The setting . . . 41
7.2 Probability spaces and kernels . . . 41
7.3 Convergence in mean . . . 42
7.4 Almost sure convergence . . . 43
7.5 Limits against colimits . . . 44
7.6 The mean ergodic theorem . . . 44

8 McMillan's Theorem **47**
8.1 The statement . . . 47
8.2 Information as a sum over time . . . 47
8.3 Proof of the theorem . . . 48
8.4 The almost sure theorem . . . 49
8.5 Typical sets and the counting theorem . . . 49
8.6 The theorem as a statement about retractions . . . 50

III Channels and Coding **53**

9 Channels **54**

9.1 Khinchin's definitions . . . 54
9.2 Connecting a channel to a source . . . 54
9.3 Anticipation and naturality . . . 55
9.4 Ergodicity: a gap and its repair . . . 56
9.5 Rate and capacity . . . 57
9.6 Memoryless channels . . . 58
9.7 Capacity as a function on morphisms . . . 58
9.8 Joinings and the backward channel . . . 59

10 Feinstein's Lemma **61**
10.1 Codes . . . 61
10.2 The one-shot lemma . . . 61
10.3 Memoryless channels . . . 62
10.4 Channels with memory . . . 63
10.5 Codes as retractions . . . 64
10.6 What the lemma is . . . 65

11 Shannon's Theorems **66**
11.1 Coding without noise . . . 66
11.2 Codes and channels . . . 67
11.3 The first theorem . . . 67
11.4 The second theorem . . . 68
11.5 The converse, and the two kinds of statement . . . 69
11.6 Separation . . . 70
11.7 Polarization: the martingale returns . . . 70

12 Rates and Spectra **73**
12.1 Coding without error . . . 73
12.2 Asymptotic spectra . . . 74
12.3 Sources: the Rényi family . . . 74
12.4 Infinite time: entropy as a rank . . . 76
12.5 Channels without noise . . . 76
12.6 Other entropies . . . 77

IV Assessment **79**

13 What the Language Buys **80**
13.1 The idea . . . 80
13.2 What became formal . . . 80
13.3 What did not . . . 81
13.4 How much of it is higher . . . 82

13.5 Khinchin, seventy years on 82
13.6 Open problems 82

A Categorical Toolkit **85**
A.1 Markov categories 85
A.2 Operads 86
A.3 Cohomology of a monoid 86
A.4 Enrichment 87
A.5 Dagger categories and idempotents 87
A.6 Preordered semirings 87

B Concordance with Khinchin **88**

C Computational Checks **90**

D Solutions to the Exercises **93**

References **103**

Preface

Khinchin's two papers of 1953 and 1956 did for Shannon's theory what Kolmogorov's *Grundbegriffe* had done for probability: they said what the objects are. A source is a stationary measure on a space of sequences. A channel is a family of measures on output sequences indexed by input sequences. Entropy is the function on finite schemes singled out by three properties. The coding theorem is a statement about these and about nothing else, and it has a proof. The papers are short, they contain almost no examples, and seventy years on they are still a good place to learn what the theorems say.

They were written in the mathematical language available at the time, which was that of measures on product spaces. This book rewrites them in a language that was not available. The claim is not that category theory proves Shannon's theorems; for the most part it does not, and the places where it does not are marked. The claim is that three features of Khinchin's text, each of which he states in words and then handles by computation, are structure that can now be named.

The first is the additivity axiom $H(AB) = H(A) + H_A(B)$. It is the condition for a function on observables to be a 1-cocycle for the action of observables on functions by conditioning, and the uniqueness theorem computes a cohomology group: $H^1 \cong \mathbb{R}$, generated by entropy (Chapter 3). Symmetry of entropy is then not an axiom but the commutativity of a monoid; the Tsallis entropies are the coboundaries of a deformed action; and Shannon entropy is what survives when the deformation is switched off.

The second is Khinchin's remark that "formally every code can be regarded as a channel", together with his construction of the compound source. Sources, codes and channels are the states and the morphisms of one category, that of shift-equivariant Markov kernels. The compound source is a composite, and his theorem that it is stationary is the statement that composition preserves equivariance (Chapter 9). The same category, taken up to almost sure equality, carries a dagger. Conditional expectations are its idempotents, and the two limits in McMillan's theorem, over longer pasts and over longer times, are limits of idempotents (Chapters 7 and 8).

The third is the form of the two asymptotic theorems on which everything rests. McMillan's theorem says that a long block from an ergodic source is, up to ε, a uniform choice among 2^{nH} sequences. Feinstein's lemma says that a long block of channel uses is, up to ε, an identity map on 2^{nC} messages.

Both are statements about approximate retractions, entropy and capacity are the corresponding ranks, and Shannon's theorem is that retractions compose (Chapters 10 and 11). What the tolerance ε buys can be said exactly: without it a whole spectrum of invariants is needed, and with it the spectrum collapses to a point (Chapter 12).

The reader will find that the amount of genuinely higher structure varies. The first part uses homological algebra and operads in earnest. The second and third use monoidal, enriched and dagger categories, which are low-dimensional. Chapter 13 says where we think the language earned its place and where it only relabelled.

Introduction

What is assumed

We assume basic category theory (functors, natural transformations, limits, adjunctions, monoidal categories) and measure-theoretic probability at the level of a first graduate course. Markov categories, operads, monoid cohomology, enrichment and dagger categories are introduced as needed and collected in Appendix A. No information theory is assumed; every classical notion is defined, and every classical theorem is either proved or cited.

Logarithms are to base 2 unless written ln, so entropies are in bits. We write $h(x) = -x \log x - (1-x)\log(1-x)$ for the binary entropy function and use the convention $0 \log 0 = 0$.

How the book is organized

Figure 1 shows the one picture to keep in mind. A source is a state μ, a channel is a morphism ν, codes are deterministic morphisms, and using a channel means composing with it. Everything in Parts II and III is about this diagram and its finite truncations.

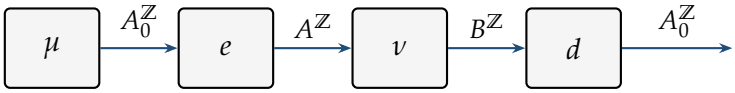


Figure 1: A source μ, an encoder e, a channel ν and a decoder d. Each box is a shift-equivariant Markov kernel; the encoder and decoder are deterministic. Shannon's theorem asks when the composite $d \circ \nu \circ e$ can be made close to the identity on μ-typical inputs.

Part I is about entropy as a function. Chapter 1 sets up finite schemes as a category and shows that conditional entropy is the entropy of a morphism, so that the chain rule is functoriality. Chapter 2 proves Khinchin's uniqueness theorem and places it among its successors: Faddeev's theorem, the functorial characterization of Baez, Fritz and Leinster, and the operadic one. Chapter 3 develops information cohomology in a finite and explicit form. Chapter 4 treats Khinchin's elementary inequalities through relative entropy and the data-processing inequality, which is the statement that composition is short for a divergence.

Part II is about sources. Chapter 5 constructs sequence spaces as infinite tensor products, defines the entropy rate, and reads finite-state sources as coalgebras. Chapter 6 treats ergodicity and the ergodic decomposition. Chapter 7 is about conditional expectations as idempotents and the convergence of martingales. Chapter 8 proves McMillan's theorem in mean from two soft convergence statements and one elementary estimate.

Part III is about channels. Chapter 9 defines channels with memory, repairs a gap in Khinchin's treatment of ergodicity that was found soon after publication, and studies capacity as a function on a monoidal category. Chapter 10 proves Feinstein's lemma in its one-shot form and reads codes as approximate retractions. Chapter 11 proves the coding theorems and their converses, and shows the martingale returning as the engine of an explicit construction. Chapter 12 places entropy and capacity among asymptotic invariants of preordered semirings.

Part IV, Chapter 13, is an assessment and a list of open problems. The appendices contain the categorical toolkit, a concordance with Khinchin's text, a description of the computational checks, and solutions to all exercises.

A dictionary

Khinchin	Here
finite scheme	object of FinProb; state $1 \to X$ in FinStoch
conditional entropy $H_A(B)$	entropy of the projection $(A \times B, r) \to (A, p)$
$H(AB) = H(A) + H_A(B)$	functoriality; the 1-cocycle condition
uniqueness theorem	$H^1 \cong \mathbb{R}$; initiality of FinProb
Shannon's inequality and its generalizations	positivity of relative entropy; submodularity
source $[A, \mu]$	shift-invariant state on a Kolmogorov product
ergodic source	invariant state that is deterministic on the invariant σ-algebra
conditional probability given the past	dagger-split idempotent
the E property	the block A^n is an ε-retract of a set of 2^{nH} points
channel $[A, \nu_x, B]$	equivariant Markov kernel $A^{\mathbb{Z}} \to B^{\mathbb{Z}}$
without anticipation	natural transformation of towers
connection of channel to source	composition after copying
distinguishable group of sequences	ε-retraction of the channel onto a finite set
ergodic capacity	asymptotic ε-rank
a code is a channel	deterministic morphisms are morphisms

How to read the tags

A reader who wants only the mathematics of information theory can read the statements tagged classical and their proofs; they are complete in themselves. A reader who knows that mathematics and wants the categorical account can read the statements tagged literature and reformulation. The tags observation and programme mark where we are on our own: the first for elementary remarks with proofs, the second for directions without them.

Part I

Entropy

Chapter 1

Finite Schemes

Khinchin's first paper begins with a *finite scheme*: a complete system of mutually exclusive events $A_1, \dots, A_n$ together with their probabilities $p_1, \dots, p_n$. He defines its entropy, lists the properties of entropy that will be needed later, and observes that one of them, the rule for the entropy of a product scheme, is the important one. This chapter does the same, with one change of emphasis. Finite schemes form a category, and the important property says that entropy is a functor on it. The entropy of a scheme is the value of that functor on a particular map.

1.1 Schemes and their maps

Definition 1.1. A ***finite probability space***, or finite scheme, is a pair (X, p) of a finite set X and a probability distribution p on it. A morphism $f\colon (X,p) \to (Y,q)$ is a map of sets with $f_* p = q$, that is, $q(y) = \sum_{x \in f^{-1}(y)} p(x)$. These form the category FinProb. The category FinStoch has finite sets as objects and stochastic matrices as morphisms: a morphism $f\colon X \to Y$ is a family of distributions $f(\,\cdot \mid x)$ on Y, one for each $x \in X$, and $(g \circ f)(z \mid x) = \sum_y g(z \mid y) f(y \mid x)$.

A distribution on X is a morphism $1 \to X$ in FinStoch, a ***state***. A map of sets is a stochastic matrix with entries in $\{0,1\}$, so FinSet is a subcategory of FinStoch, and a morphism of FinProb is a map of sets that carries one state to another. We write Δ_n for the simplex of distributions on $\{1, \dots, n\}$ and u_n for the uniform one.

FinStoch is a symmetric monoidal category under the product of sets and the tensor product of matrices, $(f \otimes g)(y, y' \mid x, x') = f(y \mid x)\, g(y' \mid x')$. Each object has a copy map $\mathrm{copy}_X\colon X \to X \times X$ and a delete map $\mathrm{del}_X\colon X \to 1$, both coming from FinSet. These data make it a *Markov category* (Appendix A). A morphism f is ***deterministic*** when it commutes with copying, $\mathrm{copy}_Y \circ f = (f \otimes f) \circ \mathrm{copy}_X$; in FinStoch these are exactly the maps of sets. A state is deterministic exactly when it is a point mass. We shall see in Chapter 4 that entropy measures by how much a state fails this equation.

Khinchin's product scheme AB needs a comment. Given schemes A and B he considers the events $A_i B_k$ and their probabilities, which presupposes that

A and B are defined on a common probability space. In our terms there is a space (Ω, P) and two morphisms out of it,

$$(X,p) \xleftarrow{a} (\Omega,P) \xrightarrow{b} (Y,q),$$

and AB is the image of the pairing, a distribution r on $X \times Y$ with marginals p and q. It is the product distribution $p \otimes q$ only when A and B are independent. Chapter 3 takes the common space Ω as the basic datum.

1.2 Entropy

Definition 1.2. The ***entropy*** of a distribution p on a finite set X is

$$H(p) = -\sum_{x \in X} p(x) \log p(x),$$

with $0 \log 0 = 0$ and logarithms to base 2.

For a joint distribution r on $X \times Y$ with marginals p and q, write $r(\,\cdot \mid x)$ for the conditional distribution on Y given x, defined when $p(x) > 0$, and put

$$H_A(B) = \sum_x p(x)\, H\big(r(\,\cdot \mid x)\big).$$

This is Khinchin's conditional entropy of B given A; the modern notation is $H(B \mid A)$.

Proposition 1.3 (Khinchin's properties). [classical; [Sha48, Khi53]]

(i) $H(p) \geq 0$, *with equality iff p is a point mass.*

(ii) $H(p) \leq \log |X|$, *with equality iff p is uniform.*

(iii) $H(r) = H(p) + H_A(B)$.

(iv) $H_A(B) \leq H(q)$, *with equality iff* $r = p \otimes q$. *In particular* $H(p \otimes q) = H(p) + H(q)$.

Proof. Let $\varphi(t) = -t \log t$, a strictly concave function on $[0,1]$ vanishing at 0 and 1. (i) is clear. (ii) By Jensen's inequality $\frac{1}{n}\sum_x \varphi(p(x)) \leq \varphi(\frac{1}{n})$ with $n = |X|$. (iii) Write $r(x,y) = p(x)\, r(y \mid x)$ and take logarithms. (iv) Since $q(y) = \sum_x p(x)\, r(y \mid x)$, Jensen's inequality gives $\varphi(q(y)) \geq \sum_x p(x)\, \varphi(r(y \mid x))$ for each y; sum over y. Equality requires $r(y \mid x)$ to be independent of x. □

Example 1.4. For $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ we have $H(p) = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + 2 \cdot \frac{1}{8} \cdot 3 = 1.75$ bits. A uniform distribution on four points has 2 bits. One can identify an outcome of p by asking "is it the first?", then "is it the second?", then "is it the third?", which takes 1, 2, 3 or 3 questions and 1.75 on average.

1.3 The entropy of a map

Property (iii) is usually read as a rule for computing the entropy of a joint distribution. Read in the other direction, it says what a difference of entropies is.

Definition 1.5. Let $f\colon (X,p) \to (Y,q)$ be a morphism of $\mathsf{FinProb}$. For y with $q(y) > 0$ let p_y be p conditioned on the fibre $f^{-1}(y)$. The ***entropy of*** f is

$$H(f) = \sum_y q(y)\, H(p_y).$$

Proposition 1.6. [classical; the chain rule] $H(f) = H(p) - H(q)$.

Proof. For x in the fibre over y we have $p(x) = q(y)\, p_y(x)$, so $-\sum_x p(x)\log p(x) = -\sum_y q(y)\log q(y) - \sum_y q(y) \sum_{x\in f^{-1}y} p_y(x)\log p_y(x)$. □

Theorem 1.7 (Entropy is a functor)**.** [literature; [BFL11]] *Regard the additive monoid $[0,\infty)$ as a category with one object. Then $f \mapsto H(f)$ is a functor $\mathsf{FinProb} \to [0,\infty)$:*

$$H(g \circ f) = H(f) + H(g), \qquad H(\mathrm{id}) = 0.$$

Moreover $H(f) = 0$ iff f is injective on the support of p.

Proof. By Proposition 1.6 the sum telescopes. $H(f) = 0$ iff each p_y is a point mass. □

The entropy of an object is recovered as the entropy of its unique map to the one-point space, $H(p) = H\big((X,p) \to 1\big)$, and Khinchin's conditional entropy as the entropy of a projection, $H_A(B) = H\big(\pi_X\colon (X\times Y, r) \to (X,p)\big)$. Property (iii) is then functoriality along

$$(X\times Y, r) \xrightarrow{\ \pi_X\ } (X,p) \longrightarrow 1.$$

So the statement Khinchin singles out as the essential property of entropy says that entropy is a functor, and the definition of conditional entropy as an average over fibres says what that functor does to a map that is not a projection to a point. $H(f)$ is the information lost by applying f; equivalently, it is the average information needed to recover x from $f(x)$.

Example 1.8. Let p be as in Example 1.4. The map f that merges the last two outcomes has image $q = (\frac12, \frac14, \frac14)$ with $H(q) = 1.5$, so $H(f) = 0.25$. Directly: only the fibre over the third point is not a singleton, it has probability $\frac14$ and conditional distribution $(\frac12,\frac12)$, so $H(f) = \frac14 \cdot 1$. The map f' that merges the first two outcomes has $H(f') = \frac34\, h(\frac23) \approx 0.689$. Merging likely outcomes loses more.

1.4 Mixtures and the operad of simplices

For $q \in \Delta_m$ and $p_i \in \Delta_{k_i}$, $i = 1, \dots, m$, define the composite

$$q \circ (p_1, \dots, p_m) = (q_1 p_{11}, \dots, q_1 p_{1k_1}, \ \dots, \ q_m p_{m1}, \dots, q_m p_{mk_m}) \in \Delta_{k_1 + \cdots + k_m}.$$

This is the distribution of a two-stage experiment: choose i according to q, then an outcome according to p_i. Composition is associative and has the unit $u_1 \in \Delta_1$, and permuting the entries of q permutes blocks. The simplices therefore form a symmetric topological operad $\Delta = (\Delta_n)_{n \geq 1}$ (Appendix A).

Proposition 1.9 (Chain rule, operadic form). [classical] $H\big(q \circ (p_1, \dots, p_m)\big) = H(q) + \sum_{i=1}^{m} q_i \, H(p_i)$.

Proof. The composite is a distribution p on a set mapping to $\{1, \dots, m\}$ with image q and conditional distributions p_i on the fibres. Apply Proposition 1.6. □

Conversely every morphism $f \colon (X, p) \to (Y, q)$ exhibits p as a composite $q \circ (p_y)_{y \in Y}$, up to the order of the points. Morphisms of FinProb are operadic composites, and $H(f) = \sum_y q(y) H(p_y)$ is the second term of the chain rule.

There is a second structure. For $\lambda \in [0, 1]$ and two objects, let $\lambda(X, p) \oplus (1 - \lambda)(X', p')$ be the disjoint union $X \sqcup X'$ with the distribution λp on X and $(1 - \lambda)p'$ on X'. The same formula on maps defines $\lambda f \oplus (1 - \lambda) f'$.

Proposition 1.10. [literature; [BFL11]] *H is convex-linear on morphisms:*

$$H\big(\lambda f \oplus (1 - \lambda) f'\big) = \lambda H(f) + (1 - \lambda) H(f').$$

On objects, $H\big(\lambda p \oplus (1 - \lambda) p'\big) = \lambda H(p) + (1 - \lambda) H(p') + h(\lambda)$.

Proof. A fibre of $\lambda f \oplus (1 - \lambda) f'$ is a fibre of f or of f', with the same conditional distribution and with weight multiplied by λ or $1 - \lambda$. The second formula is Proposition 1.9 for $(\lambda, 1 - \lambda) \circ (p, p')$. □

The two statements do not conflict. The entropy of an object is the entropy of its map to 1, and the mixture of the maps $(X, p) \to 1$ and $(X', p') \to 1$ is a map to the two-point space with distribution $(\lambda, 1 - \lambda)$, not to 1. The term $h(\lambda)$ is the entropy of the remaining map to 1. A function on objects that is not affine comes from a function on morphisms that is.

1.5 Kernels, noise and equivocation

Let $f\colon X \to Y$ be a morphism of $\mathsf{FinStoch}$ and p a state on X. The joint distribution $r(x,y) = p(x)f(y \mid x)$ on $X \times Y$ has marginals p and $q = f \circ p$. Its two conditional entropies have names:

$$H(Y \mid X) = \sum_x p(x)\, H\big(f(\,\cdot \mid x)\big) \qquad \text{(noise)},$$

$$H(X \mid Y) = H(\pi_Y) \qquad \text{(equivocation)}.$$

The ***mutual information*** is $I(X;Y) = H(p) + H(q) - H(r) = H(q) - H(Y \mid X) = H(p) - H(X \mid Y)$. By Proposition 1.3(iv) it is nonnegative and vanishes iff $r = p \otimes q$.

When q has full support there is a unique kernel $f^\dagger\colon Y \to X$ with

$$p(x)\, f(y \mid x) = q(y)\, f^\dagger(x \mid y),$$

the ***Bayesian inverse*** of f with respect to p. Noise and equivocation are exchanged by $f \mapsto f^\dagger$. If f is a map of sets its noise is zero and its Bayesian inverse is $f^\dagger(\,\cdot \mid y) = p_y$, so that

$$H(f) = \sum_y q(y)\, H\big(f^\dagger(\,\cdot \mid y)\big).$$

Proposition 1.11. [reformulation; cf. [BF14, FuP21]] *The entropy of a morphism of* $\mathsf{FinProb}$ *is the noise of its Bayesian inverse. For a general kernel f and state p,*

$$H(X \mid Y) = H(p) - H(q) + H(Y \mid X).$$

So the functor of Theorem 1.7 is the restriction to deterministic maps of the equivocation, which is defined for all kernels. The equivocation is not additive under composition of kernels, only subadditive (Exercise 1.5); it becomes additive on maps because their noise is zero. This is the first appearance of a pattern that will recur. A statement that is an equality for deterministic morphisms weakens to an inequality for stochastic ones, and the inequality is the useful part.

Notes

Entropy and its properties are due to Shannon [Sha48]; Khinchin's account is [Khi53, §1]. Categories of probability spaces and stochastic maps go back to Lawvere [Law62], Čencov [Cen82] and Giry [Gir82]. Markov categories were introduced by Cho and Jacobs [CJ19] and Fritz [Fri20]. The observation that entropy is naturally a function of a morphism, and Theorem 1.7, are due to Baez, Fritz and Leinster [BFL11]; the operad of simplices and its role are developed in Leinster's book [Lei21]. The extension of the information-loss functor to stochastic maps is studied by Fullwood and Parzygnat [FuP21].

Exercises

Exercise 1.1. For $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$, let g merge the first two outcomes and then h merge the two remaining light outcomes. Compute $H(g)$, $H(h)$ and $H(h \circ g)$, and check Theorem 1.7.

Exercise 1.2. Show that $H(f) = H(p)$ iff $f_* p$ is a point mass, and that isomorphic objects of $\mathsf{FinProb}$ have the same entropy. Is the converse true?

Exercise 1.3. Let $p \in \Delta_n$ with $p_1 = t$. Show that $H(p) \leq h(t) + (1 - t)\log(n - 1)$, with equality iff the remaining mass is spread uniformly.

Exercise 1.4. Let W be a stochastic matrix from a finite set X to itself. Show that $H(W \circ p) \geq H(p)$ for every state p if and only if W is doubly stochastic. (For the direct implication you may use that a doubly stochastic matrix is a convex combination of permutation matrices.)

Exercise 1.5. Let $X \to Y \to Z$ be composable kernels with a state on X. Show that $H(X \mid Z) \leq H(X \mid Y) + H(Y \mid Z)$, with equality when both kernels are deterministic. Give an example of strict inequality.

Chapter 2

The Uniqueness Theorem

Shannon remarked that his formula for entropy is forced by a few natural requirements, and gave an argument. Khinchin's first paper turned the remark into a theorem with a complete proof, and Faddeev sharpened it three years later. The result has been reproved and reformulated many times since. This chapter proves Khinchin's version, states its main descendants, and asks which hypothesis does what. The answer prepares Chapter 3: the hypothesis that matters is a cocycle condition, and the others are regularity.

2.1 Khinchin's theorem

Let H assign a real number to each $p \in \Delta_n$ for every $n \geq 1$. We assume throughout that H is symmetric, that is, invariant under permutations of $p_1, \dots, p_n$; Khinchin does not state this, but it is implicit in his treating H as a function of the scheme. Consider the conditions:

(K1) for each n, H is largest at the uniform distribution u_n;

(K2) for every joint distribution r on $X \times Y$ with first marginal p, $H(r) = H(p) + \sum_x p(x) H\big(r(\cdot \mid x)\big)$;

(K3) $H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$.

Theorem 2.1 (Khinchin). [classical; [Khi53]] *If H is continuous on each Δ_n and satisfies* (K1)–(K3)*, then there is a constant $\lambda \geq 0$ with*

$$H(p) = -\lambda \sum_i p_i \log p_i.$$

Khinchin says that λ is positive. The zero function satisfies the hypotheses, so $\lambda = 0$ must be allowed.

Lemma 2.2. [classical] *Let $L \colon \{1, 2, \dots\} \to \mathbb{R}$ be nondecreasing with $L(mn) = L(m) + L(n)$. Then $L(n) = \lambda \log n$ with $\lambda = L(2) \geq 0$.*

Proof. $L(1) = 0$, so $L \geq 0$. Fix $n \geq 2$ and $k \geq 1$ and choose m with $2^m \leq n^k < 2^{m+1}$. Monotonicity and additivity give $mL(2) \leq kL(n) \leq (m+1)L(2)$, and also $m \leq k \log n < m + 1$. If $L(2) = 0$ then $L(n) = 0$. Otherwise $|L(n)/L(2) - \log n| \leq 1/k$ for every k. □

Proof of Theorem 2.1. Put $L(n) = H(u_n)$. By (K3) and (K1), $L(n) = H(\frac{1}{n}, \dots, \frac{1}{n}, 0) \le L(n+1)$. Apply (K2) to the product of u_m and u_n: every conditional distribution is u_n, so $L(mn) = L(m) + L(n)$. By Lemma 2.2, $L(n) = \lambda \log n$.

Now let p have rational entries $p_i = g_i/g$ with positive integers g_i summing to g. Let Y be a set of g points divided into blocks of sizes $g_1, \dots, g_n$, and let r be the joint distribution on $\{1, \dots, n\} \times Y$ under which i has distribution p and, given i, the second coordinate is uniform on the i-th block. Then r is uniform on g pairs and zero elsewhere, so $H(r) = L(g)$ by (K3) and symmetry. Each conditional distribution is uniform on g_i points, so (K2) reads $L(g) = H(p) + \sum_i p_i L(g_i)$, that is,

$$H(p) = \lambda \sum_i p_i (\log g - \log g_i) = -\lambda \sum_i p_i \log p_i.$$

Rational points with positive entries are dense in Δ_n and H is continuous. □

The proof has two steps of different character. The second, which passes from uniform distributions to all rational ones, uses only (K2) and (K3) and is pure algebra. The first, which determines L, is where a regularity hypothesis is needed, and (K1) is used there and nowhere else. It supplies the monotonicity that makes Lemma 2.2 a three-line argument.

2.2 Faddeev's theorem and the operadic form

Faddeev showed that (K1) can be dropped and (K2) weakened to its simplest instance.

Theorem 2.3 (Faddeev). [classical; [Fad56]] *Let H be symmetric, with $H(t, 1-t)$ continuous in t, and suppose that whenever $p_n = q_1 + q_2 > 0$,*

$$H(p_1, \dots, p_{n-1}, q_1, q_2) = H(p_1, \dots, p_n) + p_n \, H\Big(\frac{q_1}{p_n}, \frac{q_2}{p_n}\Big).$$

Then $H(p) = -\lambda \sum_i p_i \log p_i$ for a real constant λ.

The displayed condition is called recursivity. By induction it implies the chain rule of Proposition 1.9 in full (Exercise 2.2), and expansibility (K3) follows by taking $q_2 = 0$. The price of dropping (K1) is that $L(n) = H(u_n)$ is no longer known to be monotone. Faddeev replaces Lemma 2.2 by a harder statement about additive arithmetic functions, in the spirit of a theorem of Erdős [Erd46]:

Lemma 2.4. [classical; [Fad56, Ren61]] *If $a(mn) = a(m) + a(n)$ for all $m, n \ge 1$ and $a(n+1) - \frac{n}{n+1} a(n) \to 0$, then $a(n) = c \log n$.*

In the operadic language of Chapter 1 the theorem takes its cleanest form. Symmetry need not be assumed.

Theorem 2.5 (Faddeev–Leinster). [literature; [Lei21], Thm. 2.5.1] *Let $I = (I_n \colon \Delta_n \to \mathbb{R})_{n \geq 1}$ be a sequence of functions. The following are equivalent:*

(i) each I_n is continuous and $I\big(q \circ (p_1, \dots, p_m)\big) = I(q) + \sum_i q_i I(p_i)$ for all composable $q, p_1, \dots, p_m$;

(ii) $I = cH$ for some $c \in \mathbb{R}$.

Proof sketch. Assume (i). From $u_1 \circ (u_1) = u_1$ we get $I(u_1) = 0$. From $(1,0) \circ \big((1,0), u_1\big) = (1,0,0) = (1,0) \circ \big(u_1, (1,0)\big)$ we get $I(1,0) + I(1,0) = I(1,0)$, so $I(1,0) = 0$. Since $u_m \circ (u_n, \dots, u_n) = u_{mn}$, the sequence $a(n) = I(u_n)$ is additive. Since $u_{n+1} = (\frac{n}{n+1}, \frac{1}{n+1}) \circ (u_n, u_1)$,

$$a(n+1) - \tfrac{n}{n+1}\, a(n) = I\big(\tfrac{n}{n+1}, \tfrac{1}{n+1}\big) \longrightarrow I(1,0) = 0$$

by continuity. Lemma 2.4 gives $a(n) = c \log n$. For rational p with $p_i = g_i / g$ we have $p \circ (u_{g_1}, \dots, u_{g_n}) = u_g$, so $I(p) = a(g) - \sum_i p_i a(g_i) = cH(p)$, and continuity finishes the proof. □

The identity $p \circ (u_{g_1}, \dots, u_{g_n}) = u_g$ is Khinchin's auxiliary scheme B. His proof and this one are the same after the determination of L.

2.3 The functorial characterization

Theorem 1.7 and Proposition 1.10 say that entropy is a convex-linear functor. Baez, Fritz and Leinster showed that this characterizes it. Call a functor $F \colon \mathsf{FinProb} \to [0, \infty)$ ***continuous*** if $F(f_k) \to F(f)$ whenever $f_k \colon (X, p_k) \to (Y, q_k)$ have the same underlying map as $f \colon (X, p) \to (Y, q)$ and $p_k \to p$.

Theorem 2.6 (Baez–Fritz–Leinster). [literature; [BFL11]] *Let $F \colon \mathsf{FinProb} \to [0, \infty)$ be a functor that is convex-linear and continuous. Then there is a constant $c \geq 0$ such that for every $f \colon (X, p) \to (Y, q)$,*

$$F(f) = c\,\big(H(p) - H(q)\big).$$

Proof. For an object (X, p) let $I(p) = F\big((X, p) \to 1\big)$. Functoriality along $(X, p) \to (Y, q) \to 1$ gives $F(f) = I(p) - I(q)$. An isomorphism ϕ has $F(\phi) + F(\phi^{-1}) = 0$ with both terms nonnegative, so I is invariant under relabelling. Now let $p = q \circ (p_1, \dots, p_m)$ and let f be the corresponding map to $(\{1, \dots, m\}, q)$. It is the convex combination, with weights q_i, of the maps from the i-th block to a point, so convex-linearity gives $F(f) = \sum_i q_i I(p_i)$. Hence $I(p) = I(q) + \sum_i q_i I(p_i)$. I is continuous, so $I = cH$ by Theorem 2.5, and $c \geq 0$ because $F \geq 0$. □

Three things have changed relative to Khinchin's formulation. Nothing is assumed about maxima. Symmetry is not an axiom; it is the fact that functors send isomorphisms to invertible elements. And expansibility has disappeared, because adjoining an outcome of probability zero is an isomorphism in FinProb once morphisms are taken up to null sets, or else follows from the chain rule as in Theorem 2.3. What remains is the chain rule in two halves, functoriality and convex-linearity, and continuity.

2.4 Entropy as a derivation

There is one more way to write the chain rule. For $p \in \Delta_n$, $q \in \Delta_k$ and $1 \leq i \leq n$ let $p \circ_i q \in \Delta_{n+k-1}$ be the result of replacing the i-th outcome of p by a copy of q scaled by p_i. The full composite is an iterated partial composite, and the chain rule is equivalent to

$$H(p \circ_i q) = H(p) + p_i\, H(q). \tag{2.1}$$

Let the operad Δ act on $\mathbb{R}$ on the left by $p \cdot_i r = p_i r$ and on the right trivially. Then (2.1) is the Leibniz rule $d(p \circ_i q) = dp \cdot q + p \cdot_i dq$.

Theorem 2.7. [literature; [Bra21, Lei21]] *Entropy is a derivation of the operad of simplices with values in $\mathbb{R}$, and every continuous derivation is a real multiple of it.*

The first statement is due to Bradley, who also gives a pointwise form of the converse; the second as stated here is Theorem 2.5. Leinster proves a categorified version: FinProb is the free categorical algebra for Δ containing an internal algebra, and the internal algebras in the one-object category $(\mathbb{R}, +)$ are the multiples of entropy [Lei21, Ch. 12], from which Theorem 2.6 follows.

A derivation is a 1-cocycle in Hochschild cohomology, and one expects a cohomology theory in which entropy is a class. The next chapter constructs it, for a monoid in place of the operad.

2.5 What the chain rule excludes

How much of the theorem is due to the conditional form of (K2)? For $\alpha > 0$, $\alpha \neq 1$, the ***Rényi entropy*** of order α is

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha.$$

It is continuous and symmetric, it satisfies (K1) and (K3), and it is additive on independent products, $H_\alpha(p \otimes q) = H_\alpha(p) + H_\alpha(q)$. As $\alpha \to 1$ it tends to $H(p)$; as $\alpha \to 0$ it tends to the ***Hartley entropy*** $\log |\operatorname{supp} p|$, and as $\alpha \to \infty$ to the min-entropy $-\log \max_i p_i$.

Example 2.8. Take $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = (\frac{1}{2}, \frac{1}{2}) \circ (u_1, u_2)$. Then $H_2(p) = -\log(\frac{1}{4} + \frac{1}{16} + \frac{1}{16}) = \log \frac{8}{3} \approx 1.415$, while the chain rule would predict $H_2(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2} H_2(u_2) = 1.5$.

So among Khinchin's hypotheses, everything except the conditional additivity (K2) for dependent schemes is satisfied by a one-parameter family. The chain rule is the hypothesis with content, and it is strong: it fixes the function up to a scalar. Chapter 12 returns to the Rényi family, which reappears when coding is required to be exact.

Regularity cannot be dropped altogether. If $\ell \colon (0,1] \to \mathbb{R}$ satisfies $\ell(xy) = \ell(x) + \ell(y)$, then $p \mapsto -\sum_i p_i \ell(p_i)$ satisfies the chain rule, by the same computation as for the logarithm. Using the axiom of choice one can take $\ell = A \circ \log$ with A a discontinuous additive function, and the result is not a multiple of entropy. Measurability is enough to exclude this [Lee64].

Notes

Shannon's axiomatic remark is in [Sha48, §6]; Khinchin's theorem is [Khi53, §2]. Faddeev's theorem is [Fad56]; Rényi [Ren61] simplified the proof and introduced the entropies named after him. The monograph of Aczél and Daróczy [AD75] surveys the functional-equation literature. The functorial characterization is [BFL11]; the operadic and categorical-algebra viewpoint is in [Lei21], and the derivation viewpoint in [Bra21]. Parzygnat [Par22] gives a functorial characterization of von Neumann entropy in the same spirit.

Exercises

Exercise 2.1. Carry out the second step of Khinchin's proof for $p = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$: describe the joint distribution r, and verify numerically that $\log g - \sum_i p_i \log g_i = H(p)$.

Exercise 2.2. Show that symmetry and Faddeev's recursivity imply $H(p \circ_i q) = H(p) + p_i H(q)$ for all $q \in \Delta_k$, by induction on k, and deduce the full chain rule.

Exercise 2.3. Show that H_α is additive on products and satisfies (K1) and (K3). Compute $\lim_{\alpha \to 1} H_\alpha(p)$.

Exercise 2.4. Any function I on objects of $\mathsf{FinProb}$ that is invariant under isomorphism defines a functor to $(\mathbb{R}, +)$ by $F(f) = I(p) - I(q)$. Show that for $I = H_2$ this functor is continuous and takes values in $[0, \infty)$, and exhibit a failure of convex-linearity. Which step of the proof of Theorem 2.6 breaks?

Exercise 2.5. Let I satisfy (2.1) for all p, q and i. Show that $I(u_1) = 0$, that $I(1,0) = 0$, and that I satisfies the full chain rule.

Chapter 3

Entropy as a Cocycle

The chain rule $H(AB) = H(A) + H_A(B)$ has the shape of a cocycle condition: the value on a product is the value on the first factor plus the value on the second, transported by the first. Baudot and Bennequin made this precise. Observables form a monoid, the monoid acts on functions of probability laws by conditioning, and entropy is a 1-cocycle for this action. The uniqueness theorem becomes the computation of a cohomology group. This chapter develops the theory in the simplest setting in which it is nontrivial, a finite sample space with all its partitions, where everything can be done by hand.

3.1 Observables

Let Ω be a finite set. An ***observable*** is a partition X of Ω; one thinks of the quotient map $\Omega \to \Omega/X$ as a random variable, and of two random variables as the same observable when each is a function of the other. Say that X ***refines*** Y if every block of X lies in a block of Y, so that Y is a function of X. The ***product*** XY is the coarsest common refinement, whose blocks are the nonempty intersections $B \cap C$ of a block of X with a block of Y; it is the observable "X and Y jointly". The trivial partition 1 is a unit. Under this product the set $\Pi(\Omega)$ of observables is a commutative monoid in which every element is idempotent:

$$XY = YX, \qquad XX = X, \qquad 1X = X.$$

X refines Y iff $XY = X$.

A ***law*** is a probability measure P on Ω. The law of an observable X under P is the distribution X_*P on the set of blocks of X. For a block B with $P(B) > 0$ we write $P|_B$ for the conditional law.

3.2 The conditioning action

Let $\mathcal{F}$ be the vector space of measurable real functions on the simplex $\Delta(\Omega)$ of laws. For an observable Y and $f \in \mathcal{F}$ define

$$(Y.f)(P) = \sum_{B \in Y,\ P(B)>0} P(B)\, f\big(P|_B\big). \tag{3.1}$$

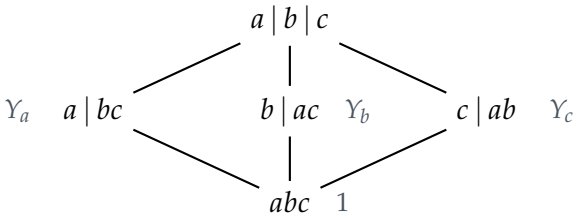


Figure 3.1: The five observables on a three-point set, ordered by refinement. The product of any two of the binary observables Y_a, Y_b, Y_c is the atomic observable at the top. This is the smallest structure on which the chain rule has content, and Theorem 3.5 is proved by reduction to it.

This is the average of f after Y has been observed.

Lemma 3.1. [literature; [BB15]] *Formula* (3.1) *is an action of the monoid* $\Pi(\Omega)$ *on* $\mathcal{F}$*:* $1.f = f$ *and* $Y.(Z.f) = (YZ).f$*.*

Proof. $(Y.(Z.f))(P) = \sum_B P(B) \sum_C P|_B(C)\, f\big((P|_B)|_C\big) = \sum_{B,C} P(B \cap C)\, f\big(P|_{B\cap C}\big)$, the sums being over blocks with positive probability. □

Call $f \in \mathcal{F}$ X***-local*** if $f(P)$ depends only on the law X_*P of X, and write $\mathcal{F}_X$ for the space of such functions. A function that is local for an observable is local for every refinement of it. If f is Z-local then $Y.f$ is YZ-local, because the law of YZ determines the probabilities of the blocks of Y and the law of Z under each conditional law.

3.3 The complex

Definition 3.2. An n***-cochain*** is a family $f = \big(f[Y_1|\cdots|Y_n]\big)$, indexed by n-tuples of observables, with $f[Y_1|\cdots|Y_n] \in \mathcal{F}_{Y_1\cdots Y_n}$. The coboundary of f is the $(n+1)$-cochain

$$(\delta f)[Y_1|\cdots|Y_{n+1}] = Y_1.f[Y_2|\cdots|Y_{n+1}] + \sum_{k=1}^{n} (-1)^k f[Y_1|\cdots|Y_kY_{k+1}|\cdots|Y_{n+1}] + (-1)^{n+1} f[Y_1|\cdots|Y_n].$$

The cohomology of the complex $(C^\bullet, \delta)$ is the ***information cohomology*** $H^\bullet(\Omega)$ of the full structure on Ω.

This is the standard complex computing the cohomology of a monoid with coefficients in a module, cut down by the locality condition. That $\delta\delta = 0$ is the usual computation, using Lemma 3.1, and the remark on locality shows

that δ preserves the condition. A 0-cochain is a function local for the trivial observable, that is, a constant, so $C^0 = \mathbb{R}$.

In low degrees, for a constant c and a 1-cochain f,

$$(\delta c)[Y] = Y.c - c = 0, \qquad (\delta f)[Y|Z] = Y.f[Z] - f[YZ] + f[Y].$$

The first vanishes because $\sum_B P(B) = 1$. Thus $H^0(\Omega) = \mathbb{R}$, there are no nonzero 1-coboundaries, and a 1-cochain is a cocycle exactly when

$$f[YZ] = f[Y] + Y.f[Z] \qquad \text{for all observables } Y, Z. \tag{3.2}$$

Proposition 3.3. [literature; [BB15]] *The family $H[Y](P) = H(Y_*P)$ is a 1-cocycle.*

Proof. It is local by definition, and (3.2) for it is Proposition 1.3(iii) applied to the joint law of Y and Z: the term $(Y.H[Z])(P) = \sum_B P(B)\, H\big(Z_*(P|_B)\big)$ is Khinchin's $H_Y(Z)$. □

So Khinchin's property 2 is literally the cocycle condition. Two features of the formalism deserve notice before we compute. The cocycle condition is imposed for a fixed sample space and all pairs of observables on it, whereas Khinchin's theorem quantifies over all finite schemes; we shall see that three points suffice. And since $YZ = ZY$, a cocycle satisfies

$$f[Y] + Y.f[Z] = f[Z] + Z.f[Y], \tag{3.3}$$

which is a nontrivial constraint relating the values on Y and on Z. Symmetry of entropy in its arguments, an axiom in Chapter 2, will come out of the commutativity of the monoid.

3.4 The fundamental equation

Theorem 3.4 (Tverberg, Kendall, Lee). [classical; [Tve58, Ken64, Lee64]; see [AD75]] *Let $h\colon [0,1] \to \mathbb{R}$ be measurable with $h(1) = 0$, and suppose that*

$$h(x) + (1-x)\, h\Big(\frac{y}{1-x}\Big) = h(y) + (1-y)\, h\Big(\frac{x}{1-y}\Big) \tag{3.4}$$

for all $x, y \in [0,1)$ with $x + y \le 1$. Then $h = c\, h$ for a constant c, where $h(x) = -x \log x - (1-x)\log(1-x)$.

Equation (3.4) is called the fundamental equation of information theory. The condition $h(1) = 0$ cannot be omitted, since $h(x) = x$ is a solution. The binary entropy satisfies the equation because both sides compute $H(x, y, 1-x-y)$, by splitting off first one outcome and then the other.

Theorem 3.5. [literature; [BB15, Vig20]; in this form proved here] *Let $|\Omega| \geq 3$. Every 1-cocycle is a constant multiple of entropy. Hence*

$$H^0(\Omega) \cong \mathbb{R}, \qquad H^1(\Omega) \cong \mathbb{R} \cdot [H].$$

Proof. Let f be a 1-cocycle.

Step 0: normalization. Taking $Y = Z$ in (3.2) and using $YY = Y$ gives $Y.f[Y] = 0$, that is, $\sum_B P(B)\, f[Y](P|_B) = 0$ for all P. Under $P|_B$ the observable Y is constant, so by locality $f[Y](P|_B)$ is a number c_B depending only on B. Then $\sum_B P(B) c_B = 0$ for all P forces $c_B = 0$: a cocycle vanishes at laws under which its argument is deterministic. In particular $f[1] = 0$.

Step 1: three points. Let $\Omega = \{a, b, c\}$ and let Y_a be the partition $\{a\} \mid \{b, c\}$, and similarly Y_b, Y_c. By locality $f[Y_a](P) = h_a(P(a))$ for a measurable function h_a on $[0, 1]$, and by Step 0, $h_a(0) = h_a(1) = 0$. Evaluate (3.3) for Y_a, Y_c at the law $P = (x, 1 - x - y, y)$. Conditioning on $\{a\}$ makes Y_c deterministic; conditioning on $\{b, c\}$ gives c the probability $y/(1 - x)$. So

$$h_a(x) + (1 - x)\, h_c\Big(\frac{y}{1 - x}\Big) = h_c(y) + (1 - y)\, h_a\Big(\frac{x}{1 - y}\Big).$$

Putting $y = 1 - x$ gives $h_a(x) = h_c(1 - x)$. The same argument for the other two pairs gives $h_a(x) = h_b(1 - x)$ and $h_b(x) = h_c(1 - x)$. Hence $h_a = h_b = h_c = h$ with $h(x) = h(1 - x)$, and h satisfies (3.4). By Theorem 3.4, $h = c\, h$, so $f[Y] = cH[Y]$ for the three binary observables. The atomic observable is $Y_a Y_c$, and (3.2) gives $f[Y_a Y_c] = c\,(H[Y_a] + Y_a.H[Y_c]) = cH[Y_a Y_c]$.

Step 2: the general case. Let W be a partition of Ω into three blocks. The observables refined by W, with the laws of W, form a copy of the three-point structure, because all the functions involved are W-local and every distribution on three points is the law of W under some P. So there is a constant c_W with $f[Y] = c_W H[Y]$ for every Y refined by W. If two such partitions W, W' both refine a binary observable Y, then $c_W = c_{W'}$, since $H[Y] \neq 0$. Any two binary observables $\{A, A^c\}$ and $\{B, B^c\}$ are connected in this way: if the two are nested, say $A \subsetneq B$, both are refined by $\{A, B \setminus A, B^c\}$; otherwise $A \cap B$ is a proper nonempty subset of both A and B and $\{A \cap B, (A \cap B)^c\}$ is nested with each. Since $|\Omega| \geq 3$, every binary observable is refined by some W. Therefore $f[Y] = cH[Y]$ for all binary Y with one constant c. Finally an observable with blocks $B_1, \ldots, B_k$ is the product of the binary observables $\{B_j, B_j^c\}$, and (3.2) determines the value of a cocycle on a product from its values on the factors. Since f and cH agree on binary observables, they agree everywhere. □

Remark 3.6. The hypothesis $|\Omega| \geq 3$ is necessary. For $|\Omega| = 2$ there is one nontrivial observable X, the only cocycle conditions are $f[1] = 0$ and $X.f[X] = 0$, and Z^1 is the infinite-dimensional space of measurable functions on $[0, 1]$

vanishing at the endpoints (Exercise 3.3). The chain rule constrains a function only when there are two different observables to chain.

Compare the hypotheses with those of Theorem 2.1. Continuity has been weakened to measurability. The maximum property (K1) and expansibility (K3) are not assumed. Symmetry is not assumed; it was derived in Step 1 from $YZ = ZY$. What is assumed is locality, which says that the quantity attached to an observable depends only on its law, and the cocycle condition.

3.5 Deformation

For $\alpha > 0$ replace (3.1) by

$$(Y \cdot_\alpha f)(P) = \sum_{B \in Y,\ P(B)>0} P(B)^\alpha f\big(P|_B\big).$$

The proof of Lemma 3.1 goes through, since $P(B)^\alpha\, P|_B(C)^\alpha = P(B \cap C)^\alpha$. Let δ_α be the coboundary built from this action and $H^\bullet_\alpha(\Omega)$ its cohomology. A constant is no longer invariant:

$$(\delta_\alpha c)[Y](P) = c\left(\sum_B P(B)^\alpha - 1\right).$$

Theorem 3.7. [literature; [Vig20]; cf. [Dar70]] *For $\alpha \neq 1$ the **Tsallis entropy***

$$S_\alpha[Y](P) = \frac{1}{\alpha - 1}\Big(1 - \sum_{B \in Y} P(B)^\alpha\Big)$$

is the coboundary of the constant $-1/(\alpha-1)$, hence a 1-cocycle for the α-action: $S_\alpha[YZ] = S_\alpha[Y] + Y \cdot_\alpha S_\alpha[Z]$. For $|\Omega| \geq 3$ every measurable α-cocycle is a multiple of it, so that $H^0_\alpha(\Omega) = 0$ and $H^1_\alpha(\Omega) = 0$.

As $\alpha \to 1$ the coboundary $\delta_\alpha 1$ tends to zero and S_α tends to entropy measured in natural units, $-\sum_B P(B) \ln P(B)$. This has a general explanation.

Proposition 3.8 (Entropy as an obstruction). [observation] *Let δ_α be a family of differentials on a fixed graded vector space, differentiable in α, and write δ' for the derivative at $\alpha = 1$. If $\delta_1 c = 0$, then $\delta' c$ is a δ_1-cocycle, and its class in cohomology vanishes iff c extends to first order to a family of cocycles $c + (\alpha - 1)c'$. For the conditioning actions and $c = 1$,*

$$\delta' 1 = -\ln 2 \cdot H,$$

so the class of entropy in $H^1(\Omega)$ is the obstruction to deforming the invariant constant 1 along the family of α-actions.

Proof. Differentiating $\delta_\alpha \delta_\alpha = 0$ at $\alpha = 1$ gives $\delta'\delta_1 + \delta_1\delta' = 0$, so $\delta_1(\delta' c) = -\delta'(\delta_1 c) = 0$. The family $c + (\alpha - 1)c'$ is closed to first order iff $\delta' c + \delta_1 c' = 0$. Finally $\frac{d}{d\alpha} \sum_B P(B)^\alpha \big|_{\alpha=1} = \sum_B P(B) \ln P(B)$. □

The dimensions fit: $(\dim H^0_\alpha, \dim H^1_\alpha)$ is $(0,0)$ for $\alpha \neq 1$ and jumps to $(1,1)$ at $\alpha = 1$, as cohomology does at a special parameter value, with the alternating sum unchanged.

3.6 A derivation and its defect; entropy modulo a prime

The function $\partial(x) = -x \log x$ satisfies the Leibniz rule

$$\partial(xy) = x\,\partial(y) + \partial(x)\,y,$$

but it is not additive, and entropy is its defect of additivity: for a distribution, $H(p) = \sum_i \partial(p_i) - \partial\big(\sum_i p_i\big)$, the last term being $\partial(1) = 0$. The chain rule is then one line,

$$\sum_{i,j} \partial(q_i p_{ij}) = \sum_i q_i \sum_j \partial(p_{ij}) + \sum_i \partial(q_i) \sum_j p_{ij}.$$

This way of seeing entropy does not need the real numbers. Let p be a prime. On the integers the Fermat quotient $\partial_p(a) = (a - a^p)/p$ satisfies $\partial_p(ab) \equiv a\,\partial_p(b) + \partial_p(a)\,b \pmod p$ and fails to be additive.

Definition 3.9. Let $\pi \in \mathbb{F}_p^n$ with $\sum_i \pi_i = 1$, and choose integers a_i representing the π_i. The ***entropy of π modulo p*** is

$$H_p(\pi) = \frac{1}{p}\Big(1 - \sum_i a_i^p\Big) \bmod p \ \in \mathbb{F}_p.$$

It is well defined because a^p modulo p^2 depends only on a modulo p.

Theorem 3.10. [literature; [Kon95, Lei21p]] *H_p satisfies the chain rule $H_p\big(\gamma \circ (\pi^1, \dots, \pi^n)\big) = H_p(\gamma) + \sum_i \gamma_i H_p(\pi^i)$, and every function on such "distributions" with values in $\mathbb{F}_p$ satisfying the chain rule is a multiple of it; no regularity hypothesis is needed. For odd p,*

$$H_p(x, 1-x) = \sum_{k=1}^{p-1} \frac{x^k}{k},$$

a truncated logarithm, which satisfies (3.4) for all $x, y \in \mathbb{F}_p \setminus \{1\}$.

The polynomial was found by Kontsevich, who observed that it satisfies the fundamental equation; Leinster identified the function it comes from and proved the uniqueness. The real and the mod p theories are parallel. Whether there is a single cohomological statement specializing to both is, as far as we know, open (Chapter 13).

3.7 The general setting, and higher degrees

We have worked with all partitions of one finite set. Baudot and Bennequin, and in greater generality Vigneaux, consider an ***information structure***: a poset $\mathcal{S}$ of observables in which the product of two elements exists whenever they have a common refinement in $\mathcal{S}$, with a functor of laws and a presheaf of functions on it. For each X the observables refined by X form a monoid $\mathcal{S}_X$; the monoid algebras $\mathcal{A}_X = \mathbb{R}[\mathcal{S}_X]$ form a presheaf of rings $\mathcal{A}$ on $\mathcal{S}$; and the functions form a presheaf of $\mathcal{A}$-modules.

Theorem 3.11. [literature; [Vig20]] *Information cohomology is* $\mathrm{Ext}^\bullet_{\mathcal{A}}(\mathbb{R}, \mathcal{F})$*, computed in the abelian category of* $\mathcal{A}$*-modules in the topos of presheaves on* $\mathcal{S}$*. A relative bar resolution of the trivial module* $\mathbb{R}$ *yields the complex of Definition 3.2. If every observable in* $\mathcal{S}$ *can be written as a product of two others in a nondegenerate way,* H^1 *is a product of copies of* $\mathbb{R}$*, one for each connected component of* $\mathcal{S}$*, each generated by a multiple of entropy.*

The generality matters when not all observables are jointly measurable, as in quantum theory or in the contextual situations studied by sheaf-theoretic methods. For our purposes the finite full structure is enough.

About degrees above one much less is known. If the conditioning action is replaced by the trivial action, the coboundary of entropy is $H[Y] + H[Z] - H[YZ]$, the mutual information, and the higher alternating sums are the multivariate mutual informations [BB15]; these fit together into Hu's theorem that entropy extends to a signed measure on the Boolean algebra generated by the observables [Hu62, Yeu08]. For the conditioning action itself we know of no general computation of $H^n(\Omega)$ for $n \geq 2$.

Notes

Information cohomology was introduced by Baudot and Bennequin [BB15]; the topos-theoretic foundations, the bar resolution, and the treatment of the Tsallis case are due to Vigneaux [Vig20, Vig19], who also treats Gaussian laws, where differential entropy and the dimension appear as cocycles. The fundamental equation has a long history, for which see [AD75]; its generalization of degree α is due to Daróczy [Dar70], and the entropies S_α to Havrda and Charvát [HC67] and Tsallis [Tsa88]. Cathelineau [Cat88] met the equation in the homology of SL_2, which is the origin of the connection with polylogarithms pursued in [Kon95, EVG02]. Entropy modulo a prime is developed by Leinster [Lei21p].

Exercises

Exercise 3.1. Show that $Y \cdot_\alpha (Z \cdot_\alpha f) = (YZ) \cdot_\alpha f$ for every $\alpha > 0$, and that the constant function 1 is invariant under the α-action iff $\alpha = 1$.

Exercise 3.2. Verify directly that $S_\alpha[YZ] = S_\alpha[Y] + Y \cdot_\alpha S_\alpha[Z]$. For independent Y and Z deduce $S_\alpha[YZ] = S_\alpha[Y] + S_\alpha[Z] - (\alpha - 1) S_\alpha[Y] S_\alpha[Z]$.

Exercise 3.3. Let $|\Omega| = 2$. Describe all 1-cocycles and compute $H^1(\Omega)$.

Exercise 3.4. Show that $h(x) = x$ satisfies (3.4), and that h satisfies it by computing $H(x, y, 1 - x - y)$ in two ways. Why does $h(x) = x$ not contradict Theorem 3.5?

Exercise 3.5. Prove the congruence $\partial_p(ab) \equiv a\,\partial_p(b) + \partial_p(a)\,b \pmod{p}$. Compute $H_5(2, 4)$ from the definition and from the truncated logarithm. Compute $H_3(2, 2)$ and $H_3(1, 1, 2)$, and check the chain rule for $(1, 1, 2) = (2, 2) \circ \big((2, 2), (1)\big)$ in $\mathbb{F}_3$.

Chapter 4

Divergence and the Elementary Inequalities

Khinchin's second paper opens with a chapter of inequalities: two generalizations of Shannon's inequality $H_A(B) \le H(B)$ and three lemmas taken from Feinstein. They are the working tools of the later chapters. All of them follow from one fact, the positivity of relative entropy, together with the chain rule; and the form in which they are used, the data-processing inequality, says that composition of Markov kernels is non-expansive for a divergence. That statement is an enrichment of the category $\mathsf{FinStoch}$, and this chapter is organized around it.

4.1 Relative entropy

Definition 4.1. For distributions p, q on a finite set X, the ***relative entropy***, or Kullback–Leibler divergence, is

$$D(p\,\|\,q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)},$$

with the conventions $0 \log \frac{0}{q} = 0$ and $D(p\,\|\,q) = \infty$ if $p(x) > 0 = q(x)$ for some x.

Proposition 4.2 (Gibbs' inequality). [classical] *$D(p\,\|\,q) \ge 0$, with equality iff $p = q$.*

Proof. $\ln t \le t - 1$ with equality only at $t = 1$, so $-\sum_x p(x) \ln \frac{q(x)}{p(x)} \ge \sum_x (p(x) - q(x)) \ge 0$, the sums being over the support of p. □

For a state p on X and a kernel $f \colon X \to Y$ write $p \rtimes f$ for the joint distribution $p(x) f(y \mid x)$ on $X \times Y$.

Proposition 4.3 (Chain rule for relative entropy). [classical] *For joint distributions r, r' on $X \times Y$ with first marginals p, p',*

$$D(r\,\|\,r') = D(p\,\|\,p') + \sum_{x} p(x)\, D(r(\cdot \mid x)\,\|\,r'(\cdot \mid x)).$$

Theorem 4.4 (Data processing). [classical] *For every kernel $f\colon X \to Y$ and states p, p' on X,*

$$D(f \circ p \,\|\, f \circ p') \leq D(p \,\|\, p').$$

Proof. Apply Proposition 4.3 to $r = p \rtimes f$ and $r' = p' \rtimes f$ in two ways. Conditioning on X, the conditional distributions agree, so $D(r \,\|\, r') = D(p \,\|\, p')$. Conditioning on Y, $D(r \,\|\, r') = D(f \circ p \,\|\, f \circ p') + \sum_y (f \circ p)(y)\, D(r(\,\cdot \mid y) \,\|\, r'(\,\cdot \mid y))$, and the last sum is nonnegative by Proposition 4.2. □

The proof shows what the inequality is made of. The chain rule is an identity, and we saw in Chapter 1 that for entropy it is functoriality. Positivity is an order relation. Data processing is the two together.

4.2 Composition is short

Relative entropy between states extends to parallel morphisms. For $f, g\colon X \to Y$ in FinStoch put

$$D(f \,\|\, g) = \max_{x \in X} D(f(\,\cdot \mid x) \,\|\, g(\,\cdot \mid x)).$$

Theorem 4.5. [literature; [Per24]] *With this divergence on its hom-sets,* FinStoch *is a Markov category enriched in divergence spaces: for all composable morphisms,*

$$D(h \circ f \,\|\, h \circ g) \leq D(f \,\|\, g), \qquad D(f \circ k \,\|\, g \circ k) \leq D(f \,\|\, g),$$
$$D(f \otimes f' \,\|\, g \otimes g') = D(f \,\|\, g) + D(f' \,\|\, g').$$

Proof. The first inequality is Theorem 4.4 applied to each input. The second follows from the joint convexity of relative entropy, since $(f \circ k)(\,\cdot \mid w)$ is a mixture of the $f(\,\cdot \mid x)$ with weights $k(x \mid w)$. The third is additivity of relative entropy on product distributions. □

A ***divergence space*** is a set with a function D to $[0, \infty]$ vanishing on the diagonal; no symmetry or triangle inequality is asked. The theorem says that composition and tensoring are non-expansive maps of such spaces. The same holds for other divergences, for instance the Rényi divergences and the total variation distance $\mathrm{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|$, which is a metric. This is the pattern announced at the end of Chapter 1: an equation for deterministic morphisms, here $f \circ p = f \circ p'$ following from $p = p'$, has a quantitative form for all morphisms.

4.3 Entropy as distance from determinism

A state p on X is deterministic when $\mathrm{copy}_X \circ p = p \otimes p$: drawing once and copying is the same as drawing twice. For a general state the two sides differ, and the difference can be measured.

Proposition 4.6. [literature; [Per24]] *For a state p on X and a joint state r on $X \times Y$ with marginals p, q,*

$$H(p) = D(\mathrm{copy}_X \circ p \parallel p \otimes p), \qquad I(X;Y) = D(r \parallel p \otimes q).$$

With total variation in place of relative entropy the first expression becomes $1 - \sum_x p(x)^2$.

Proof. $\mathrm{copy}_X \circ p$ is concentrated on the diagonal with weights $p(x)$, and $\sum_x p(x) \log \frac{p(x)}{p(x)^2} = H(p)$. The second formula is $\sum r \log r - \sum r \log p - \sum r \log q$. For the last statement, $\frac{1}{2}\left[\sum_x (p(x) - p(x)^2) + \sum_{x \neq x'} p(x)p(x')\right] = 1 - \sum_x p(x)^2$. □

So entropy is the mutual information of a variable with itself, and it is defined by the Markov category structure together with the divergence. The quantity $1 - \sum_x p(x)^2$ is the Gini–Simpson index of ecology. Entropy and mutual information need no separate axiomatization once the enrichment is given; what needs a choice is the divergence.

4.4 Shannon's inequality and its two generalizations

By Propositions 4.2 and 4.6, $I(X;Y) \geq 0$, which is Shannon's inequality $H_A(B) \leq H(B)$.

Khinchin's first generalization is contained in the proof we gave of Proposition 1.3(iv). That proof applied Jensen's inequality for each value of the second variable separately, and therefore proves more than the summed statement.

Lemma 4.7 (Khinchin's Lemma 1.1)**.** [classical; [Khi56]] *Let r be a joint distribution on $X \times Y$ with marginals p, q. For every subset $S \subseteq Y$,*

$$-\sum_x p(x) \sum_{y \in S} r(y \mid x) \log r(y \mid x) \;\leq\; -\sum_{y \in S} q(y) \log q(y).$$

The second generalization is the conditional form.

Lemma 4.8 (Khinchin's Lemma 1.2)**.** [classical; [Khi56]] *For three schemes, $H_{AB}(C) \leq H_B(C)$. Equivalently, the **conditional mutual information** $I(A;C \mid B) = H_B(C) - H_{AB}(C)$ is nonnegative.*

Proof. $I(A;C \mid B) = \sum_b P(b)\, I(A;C)_{P|_b}$, an average of mutual informations. □

In the language of Chapter 3 these inequalities are statements about the function $X \mapsto H[X](P)$ on the monoid of observables, for a fixed law P. Recall that XY is the join of X and Y in the refinement order; write $X \wedge Y$ for the meet, the finest observable refined by both.

Proposition 4.9. [classical; cf. [Fuj78]] *For every law P the function $H[\,\cdot\,](P)$ on $\Pi(\Omega)$ is*

(i) monotone: $H[X] \geq H[Y]$ if X refines Y;

(ii) submodular: $H[XY] + H[X \wedge Y] \leq H[X] + H[Y]$;

(iii) and satisfies $H[XYZ] + H[Y] \leq H[XY] + H[YZ]$ for all X, Y, Z.

Proof. (i) is Theorem 1.7. (iii) is Lemma 4.8, since $H_{XY}(Z) = H[XYZ] - H[XY]$. For (ii), $X \wedge Y$ is a function of X and of Y, so with $W = X \wedge Y$ we have $H[XW] = H[X]$ and $H[WY] = H[Y]$, and (iii) for X, W, Y reads $H[XY] + H[W] \leq H[X] + H[Y]$. □

The difference between the two sides of (ii) is $I(X;Y) - H[X \wedge Y]$. The meet is the part of the information that X and Y share as an observable; the mutual information can be much larger, and is positive for two correlated bits whose meet is trivial (Exercise 4.4).

4.5 Feinstein's three inequalities

The remaining lemmas of Khinchin's first chapter concern decoding. Let r be a joint distribution on $X \times Y$ with marginals p, q; think of x as sent and y as received. The ***maximum a posteriori*** rule decodes y as an x maximizing $r(x,y)$, and its probability of error is

$$P_e = 1 - \sum_y \max_x r(x,y).$$

Lemma 4.10 (Khinchin's Lemmas 2.1–2.3)**.** [classical; [Fei54, Khi56]]

(i) Let $Z \subseteq X \times Y$ with $r(Z) > 1 - \delta_1$ and $U_0 \subseteq X$ with $p(U_0) > 1 - \delta_2$. For $\alpha > 0$ let U_1 be the set of $x \in U_0$ with $r(\{y : (x,y) \notin Z\} \mid x) \leq \alpha$. Then $p(U_1) > 1 - \delta_2 - \delta_1/\alpha$.

(ii) If there are pairwise disjoint sets $\Delta_x \subseteq Y$ with $r(\Delta_x \mid x) > 1 - \varepsilon$ for every x, then $P_e < \varepsilon$.

(iii) (Fano's inequality.) If $|X| = n > 1$, then $H(X \mid Y) \leq h(P_e) + P_e \log(n-1)$.

Proof. (i) $\sum_x p(x)\, r(\{y : (x,y) \notin Z\} \mid x) = 1 - r(Z) < \delta_1$, so by Markov's inequality the set of x where the conditional probability exceeds α has p-measure less than δ_1/α. (ii) The rule that decodes $y \in \Delta_x$ as x is correct with probability $\sum_x p(x)\, r(\Delta_x \mid x) > 1 - \varepsilon$, and no rule does better than the maximum a posteriori rule, because $\sum_y r(d(y), y) \le \sum_y \max_x r(x,y)$ for every map d. (iii) Given y, the posterior $r(\,\cdot \mid y)$ has largest entry $1 - e_y$, where e_y is the conditional probability of error. By Exercise 1.3, $H(r(\,\cdot \mid y)) \le h(e_y) + e_y \log(n-1)$. Average over y and use the concavity of h. □

Part (ii) says that a decoder is a map $d\colon Y \to X$ and that among all maps the pointwise maximizer of the Bayesian inverse is best. Part (iii) is the quantitative form of a triviality. If a kernel $f\colon X \to Y$ has a deterministic left inverse, $d \circ f = \mathrm{id}_X$, then x is a function of y and $H(X \mid Y) = 0$. Fano's inequality says that if $d \circ f$ is within ε of the identity, in the sense that $d(y) \neq x$ with probability at most ε, then the equivocation is at most $h(\varepsilon) + \varepsilon \log(n-1)$. Approximate retractions lose approximately nothing. Chapter 10 is built on this notion.

4.6 Relative entropy as a functor

Entropy is a functor on FinProb (Theorem 1.7) and is characterized by that (Theorem 2.6). Relative entropy has the same status on a slightly larger category. Let FinStat have finite probability spaces as objects; a morphism $(X,p) \to (Y,q)$ is a pair (f,s) of a measure-preserving map $f\colon X \to Y$ and a kernel $s\colon Y \to X$ with $f \circ s = \mathrm{id}_Y$, thought of as a hypothesis about how x is distributed given $f(x)$. Composition is $(g,t) \circ (f,s) = (g \circ f, s \circ t)$.

Theorem 4.11 (Baez–Fritz). [literature; [BF14]] *The assignment $(f,s) \mapsto D(p \,\|\, s \circ q)$ is a functor* FinStat $\to [0,\infty]$. *It is lower semicontinuous and convex-linear, and it vanishes when s is the Bayesian inverse of f. Every functor with these properties is a constant multiple of it.*

Functoriality here is Proposition 4.3. The quantity $D(p \,\|\, s \circ q)$ is the information gained on learning that the true distribution is p, when one's prior was built from the correct marginal q and the hypothesis s.

4.7 Beyond the elementary inequalities

Every inequality proved so far is a consequence of Proposition 4.9, that is, of monotonicity and submodularity. Such consequences are called Shannon-type. They are all there is for up to three variables.

Theorem 4.12. [classical; [ZY98, Mat07]] *For four or more random variables there are linear inequalities among their joint entropies that hold for all distributions and are not consequences of monotonicity and submodularity; there are infinitely many independent ones, and the closure of the set of entropy vectors is a convex cone that is not polyhedral.*

Khinchin's elementary inequalities are thus the first layer of a structure that is not finitely describable. Neither the cohomological account of Chapter 3 nor the enriched one of this chapter explains the further layers. We record this as a limit of what is understood, not of what was needed: the coding theorems use only the first layer.

Notes

Relative entropy is due to Kullback and Leibler; for the classical theory see [CT06, CK11]. The enriched-category view of data processing and Proposition 4.6 are due to Perrone [Per24]. The generalized Shannon inequalities and Feinstein's lemmas are [Khi56, Ch. I], after [Fei54]; Fano's inequality is in [Fan61]. Submodularity of entropy and the polymatroid point of view go back to Fujishige [Fuj78]; non-Shannon inequalities were discovered by Zhang and Yeung [ZY98], and the non-polyhedrality is due to Matúš [Mat07]. See [Yeu08] for the whole subject. The category $\mathsf{FinStat}$ and Theorem 4.11 are from [BF14].

Exercises

Exercise 4.1. For $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\frac{1}{4}, \frac{3}{4})$ compute $D(p \,\|\, q)$ and $D(q \,\|\, p)$. Let f be the binary symmetric channel that flips its input with probability 0.1. Compute $D(f \circ p \,\|\, f \circ q)$ and compare.

Exercise 4.2. Prove Proposition 4.3, and show that equality holds in Theorem 4.4 iff the two Bayesian inverses agree: $r(\,\cdot \mid y) = r'(\,\cdot \mid y)$ for all y in the support of $f \circ p$.

Exercise 4.3. Compute $\mathrm{TV}(\mathrm{copy} \circ p,\ p \otimes p)$ and $D(\mathrm{copy} \circ p \,\|\, p \otimes p)$ for $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. For which states on n points is each largest?

Exercise 4.4. Let X, Y be bits with joint distribution $r(0,0) = r(1,1) = 0.4$, $r(0,1) = r(1,0) = 0.1$. Compute $I(X;Y)$ and show that $X \wedge Y$ is trivial. Then give an example in which $I(X;Y) = H[X \wedge Y] > 0$.

Exercise 4.5. Let X be uniform on n points and let $Y = X$ with probability $1 - \varepsilon$, and otherwise uniform on the other $n - 1$ points, with $\varepsilon < 1 - \frac{1}{n}$. Show that equality holds in Fano's inequality.

Part II

Sources

Chapter 5

Sources

A source, for Khinchin, is a probability measure μ on the space $A^{\mathbb{Z}}$ of doubly infinite sequences of letters from a finite alphabet A; it is stationary when μ is invariant under the shift. This chapter sets up the same objects in a Markov category. The sequence space is an infinite tensor product, a source is a state on it, and stationary sources and the maps between them are the states and morphisms of a category of dynamical systems. We then define the entropy rate, compute it for Markov chains as Khinchin does in his first paper, and read finite-state sources as coalgebras.

5.1 Sequence spaces

Let BorelStoch be the category of standard Borel spaces and Markov kernels; it is a Markov category containing FinStoch, and its deterministic morphisms are the measurable maps (Appendix A). For a finite alphabet A the space $A^{\mathbb{Z}}$ carries the product σ-algebra $\mathcal{F}_A$, generated by the ***cylinders***, the sets of sequences with prescribed letters at finitely many places. For a finite set $F \subset \mathbb{Z}$ let $\pi_F \colon A^{\mathbb{Z}} \to A^F$ be the projection.

Theorem 5.1 (Kolmogorov extension). [literature; [FR20]] *In* BorelStoch *the object $A^{\mathbb{Z}}$, with the projections π_F, is the limit of the diagram of finite powers A^F and marginalization maps, and this limit is preserved by tensoring with any object. Explicitly, for every Y the kernels $Y \to A^{\mathbb{Z}}$ correspond bijectively to families of kernels $k_F \colon Y \to A^F$ that are compatible with marginalization.*

Such a limit is called a ***Kolmogorov product***, and $A^{\mathbb{Z}} = \bigotimes_{n\in\mathbb{Z}} A$ is the infinite tensor power of A. For $Y = 1$ the theorem is the classical statement that a measure on $A^{\mathbb{Z}}$ is the same as a consistent family of finite-dimensional distributions. The Kolmogorov product is a categorical product only in the subcategory of deterministic morphisms. A kernel into $A^{\mathbb{Z}}$ is not determined by its components into the single factors A, and this failure will matter below.

5.2 Stationary sources

The ***shift*** $T \colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$, $(Tx)_n = x_{n+1}$, is a deterministic isomorphism. A ***source*** $[A, \mu]$ is a state $\mu \colon 1 \to A^{\mathbb{Z}}$; it is ***stationary*** if $T \circ \mu = \mu$. By Theorem 5.1

a stationary source is the same as a sequence of distributions μ_n on A^n, $n \geq 1$, the laws of the blocks $x_1 \dots x_n$, such that both the first n and the last n letters of a block distributed by μ_{n+1} are distributed by μ_n.

Definition 5.2. Dyn is the category whose objects are pairs (X, T_X) of a standard Borel space and a measurable automorphism, and whose morphisms $(X, T_X) \to (Y, T_Y)$ are the Markov kernels $k\colon X \to Y$ that are ***equivariant***: $k \circ T_X = T_Y \circ k$.

Proposition 5.3. [literature; [Fri20], §7] Dyn *is a Markov category, with tensor product, copy and delete maps inherited from* BorelStoch. *Its unit object is the one-point system, and:*

(i) a state $1 \to (A^{\mathbb{Z}}, T)$ *is a stationary source;*

(ii) a deterministic morphism $(A^{\mathbb{Z}}, T) \to (B^{\mathbb{Z}}, T)$ *is a measurable map commuting with the shifts, a* ***stationary code****;*

(iii) a general morphism $(A^{\mathbb{Z}}, T) \to (B^{\mathbb{Z}}, T)$ *is a stationary channel in the sense of Chapter 9.*

The only point to check is that the copy maps are equivariant, and they are because T is deterministic. The proposition puts Khinchin's three kinds of object in one place. His remark that a code may be regarded as a channel is the inclusion of the deterministic morphisms among all morphisms.

5.3 The entropy rate

For a stationary source μ write $H_n = H(\mu_n)$ for the entropy of a block of length n, and $H_0 = 0$.

Proposition 5.4. [classical; [Sha48, Khi56]] *Let μ be stationary.*

(i) $H_{m+n} \leq H_m + H_n$.

(ii) The increments $c_n = H_{n+1} - H_n = H(x_0 \mid x_{-n}, \dots, x_{-1})$ *are nonincreasing in n.*

(iii) The limits $\lim_n H_n/n$ *and* $\lim_n c_n$ *exist and are equal, and equal* $\inf_n H_n/n$.

Proof. (i) is Shannon's inequality for the two halves of a block, whose laws are μ_m and μ_n by stationarity. (ii) is Lemma 4.8: conditioning on one more letter of the past does not increase the entropy; stationarity identifies the conditional entropies with the increments. (iii) $H_n = \sum_{k<n} c_k$ is a sum of a nonincreasing nonnegative sequence, so H_n/n converges to $\lim c_k$; the infimum is Fekete's lemma for the subadditive sequence H_n. □

Definition 5.5. The common value $H(\mu)$ of the limits in Proposition 5.4 is the ***entropy*** of the source, in bits per letter. It satisfies $0 \le H(\mu) \le \log|A|$.

For the source that emits independent letters with distribution p, the state $p^{\otimes\mathbb{Z}}$, the block entropies are $H_n = nH(p)$ and the entropy of the source is $H(p)$. The second expression in (iii) has a meaning that the first hides: $H(\mu)$ is the uncertainty of the next letter given the entire past. We make this precise in Chapter 7.

5.4 Stationary codes and cofreeness

A stationary code $\phi\colon A^{\mathbb{Z}} \to B^{\mathbb{Z}}$ is determined by its coordinate at time zero, $\phi_0 = \pi_0 \circ \phi\colon A^{\mathbb{Z}} \to B$, since $\phi(x)_n = \phi_0(T^n x)$. This is a universal property.

Proposition 5.6. [reformulation; cf. [Hed69]] *In the category of sets, of measurable spaces, or of compact Hausdorff spaces, let (X, T_X) be an object with an automorphism. Then composition with π_0 is a bijection between equivariant maps $(X, T_X) \to (B^{\mathbb{Z}}, T)$ and maps $X \to B$. That is, $(B^{\mathbb{Z}}, T)$ is the cofree system on B. Consequently:*

(i) stationary codes $A^{\mathbb{Z}} \to B^{\mathbb{Z}}$ correspond to measurable functions $A^{\mathbb{Z}} \to B$;

*(ii) continuous equivariant maps $A^{\mathbb{Z}} \to B^{\mathbb{Z}}$ correspond to continuous functions $A^{\mathbb{Z}} \to B$, which by compactness depend on finitely many coordinates: they are the **sliding block codes** $\phi(x)_n = \Phi(x_{n-r}, \dots, x_{n+r})$ (the Curtis–Hedlund–Lyndon theorem).*

In BorelStoch *the corresponding statement is false.*

Proof. Given $g\colon X \to B$ define $\phi(x)_n = g(T_X^n x)$. This is the unique equivariant map with $\pi_0 \circ \phi = g$, and it is measurable or continuous when g is, because $B^{\mathbb{Z}}$ is the product in those categories. A continuous function from the compact space $A^{\mathbb{Z}}$ to the finite discrete space B is locally constant, and finitely many cylinders on which it is constant cover $A^{\mathbb{Z}}$. For the last statement let $X = 1$: an equivariant kernel $1 \to B^{\mathbb{Z}}$ is a stationary source, a kernel $1 \to B$ is a distribution on B, and many stationary sources have the same one-letter marginal. □

The failure in the stochastic case is not a technicality. An equivariant kernel $\nu\colon A^{\mathbb{Z}} \to B^{\mathbb{Z}}$ consists of its time-zero marginal $A^{\mathbb{Z}} \to B$ and, in addition, of the dependence among the output letters for a fixed input. The first part is what Khinchin's notion of the memory of a channel controls. The second is not controlled by it, and we shall see in Chapter 9 that this is exactly where his treatment of ergodicity has a gap.

5.5 Markov chains

Let $Q = (q_{ab})$ be a stochastic matrix on A and π a stationary distribution, $\pi Q = \pi$. The ***stationary Markov source*** has block laws $\mu_n(x_1 \dots x_n) = \pi(x_1)\, q_{x_1 x_2} \cdots q_{x_{n-1} x_n}$.

Proposition 5.7. [classical; [Sha48]; [Khi53], §3] *The entropy of the stationary Markov source is*

$$H(\mu) = \sum_a \pi(a)\, H(q_{a\,\cdot}) = -\sum_{a,b} \pi(a)\, q_{ab} \log q_{ab}.$$

Proof. By the Markov property $c_n = H(x_0 \mid x_{-1})$ for all $n \geq 1$. □

Example 5.8. Let $A = \{0,1\}$ with $q_{01} = 0.1$ and $q_{10} = 0.3$. Then $\pi = (0.75, 0.25)$ and $H(\mu) = 0.75\, h(0.1) + 0.25\, h(0.3) \approx 0.572$ bits per letter. A source emitting independent letters with the same marginal has entropy $h(0.25) \approx 0.811$. The difference is what the memory of the chain is worth to someone predicting it.

5.6 Finite-state sources as coalgebras

A Markov chain is the simplest machine that emits a sequence. The general finite machine has a finite set S of internal states and, at each step, emits a letter and moves to a new state, the pair being drawn from a distribution that depends on the current state. That is a morphism

$$c \colon S \longrightarrow A \times S \quad \text{in } \mathsf{FinStoch},$$

a coalgebra for the functor $A \times -$. With an initial distribution on S it generates a process, called a hidden Markov source or a function of a Markov chain.

Theorem 5.9. [literature; [KK13]; cf. [HJS07]] *In* BorelStoch *the space $A^{\mathbb{N}}$ of one-sided sequences, with the isomorphism $A^{\mathbb{N}} \cong A \times A^{\mathbb{N}}$ splitting off the first letter, is the final coalgebra of the functor $A \times -$. For a coalgebra $c \colon S \to A \times S$ the unique coalgebra morphism $S \to A^{\mathbb{N}}$ assigns to each state the law of the sequence emitted from it.*

Uniqueness is the extension theorem again: a coalgebra morphism is determined on cylinders by induction on their length. If the initial distribution σ is invariant for the state process, the emitted one-sided process is stationary and extends uniquely to a stationary source on $A^{\mathbb{Z}}$.

The entropy of such a source is not given by a formula in terms of c and σ in general, but there is a case where it is. Call c ***unifilar*** if the next state is a function of the current state and the emitted letter, $c(a, s' \mid s) = e_s(a)\,[s' = \tau(s,a)]$.

Proposition 5.10. [classical; [Sha48]] *For a unifilar coalgebra with invariant state distribution σ, $H(\mu) = \sum_s \sigma(s)\, H(e_s)$.*

Proof. Let s_0 be the initial state. Given s_0 and $x_0, \dots, x_{n-1}$ the state s_n is determined, so $H(x_n \mid x_0, \dots, x_{n-1}, s_0) = \sum_s \sigma(s) H(e_s)$ for every n. Hence $H_n \geq H(x_0 \dots x_{n-1} \mid s_0) = n \sum_s \sigma(s) H(e_s)$ and $H_n \leq H(s_0) + H(x_0 \dots x_{n-1} \mid s_0)$. Divide by n. □

A general coalgebra can be made unifilar at the cost of enlarging the state space. Replace the state by the ***belief***, the conditional distribution $b \in \Delta(S)$ of the current state given the letters emitted so far. The belief emits a letter with distribution $e_b = \sum_s b(s)\, c(\,\cdot\,, S \mid s)$ and is then updated by Bayes' rule to a new belief $\tau(b, a)$ that is a function of b and the letter. This is the determinization of the coalgebra in the sense of the generalized powerset construction [SBBR13], carried out for the distribution monad instead of the powerset.

Theorem 5.11 (Blackwell)**.** [reformulation; cf. [Bla57]] *Let β be a stationary distribution of the belief process on $\Delta(S)$ arising from the invariant initial distribution σ. Then*

$$H(\mu) = \int_{\Delta(S)} H(e_b)\, d\beta(b),$$

which is the formula of Proposition 5.10 for the determinized coalgebra. For every $n \geq 1$,

$$H(x_n \mid x_0, \dots, x_{n-1}, s_0) \;\leq\; H(\mu) \;\leq\; H(x_n \mid x_0, \dots, x_{n-1}),$$

and both bounds converge to $H(\mu)$.

The measure β can be singular and is not computable in closed form in general; that is the sense in which the entropy of a hidden Markov source has no formula. The bounds are effective.

Example 5.12. Let the state follow the chain of Example 5.8 and let the emitted letter be the state passed through a binary symmetric channel with flip probability 0.2. The lower and upper bounds of Theorem 5.11 are 0.9033 and 0.9055 at $n = 4$ and agree to four places, at 0.9052, from $n = 8$ on. Hiding the state behind noise has raised the entropy from 0.572 to 0.905 bits per letter.

5.7 Entropy as an invariant

Let $\phi\colon (A^{\mathbb{Z}}, \mu) \to (B^{\mathbb{Z}}, \eta)$ be a stationary code with $\phi \circ \mu = \eta$; one says that η is a ***factor*** of μ.

Theorem 5.13 (Kolmogorov–Sinai). [classical; [Kol58, Sin59]] *If η is a factor of μ then $H(\eta) \le H(\mu)$. In particular isomorphic stationary sources have the same entropy.*

Proof for sliding block codes. If ϕ has window radius r, a block of n output letters is a function of a block of $n+2r$ input letters, so $H_n(\eta) \le H_{n+2r}(\mu)$ by Theorem 1.7. Divide by n. The general case follows by approximating a measurable ϕ_0 by functions of finitely many coordinates. □

So the entropy rate is a functor from stationary sources and stationary codes to the poset $([0,\infty), \ge)$. It is the infinite-time counterpart of Theorem 1.7, with one difference: for finite schemes the functor recorded the amount lost, a number attached to each morphism, and here only the inequality survives. It does not extend to a monotone function along stochastic morphisms. A noisy channel can raise the entropy of what passes through it, as Example 5.12 shows. The quantity that behaves well for all morphisms of Dyn is the information rate of Chapter 9.

Kolmogorov introduced Theorem 5.13 into ergodic theory in 1958, taking the definition from Shannon, in order to show that the independent sources with distributions $(\frac12,\frac12)$ and $(\frac13,\frac13,\frac13)$ are not isomorphic. Khinchin's papers, which appeared in the years just before, were among the routes by which Shannon's ideas reached Soviet probabilists.

Notes

The categorical treatment of infinite products is due to Fritz and Rischel [FR20]; categories of diagrams in a Markov category are in [Fri20, §7]. For the classical theory of sources see [Gra11, Bil65]. The Curtis–Hedlund–Lyndon theorem is [Hed69]; see [LM95]. Coalgebraic trace semantics for probabilistic systems was developed in [HJS07, KK13], and determinization in [SBBR13]. Blackwell's theorem is [Bla57]; the bounds are in [CT06, §4.5].

Exercises

Exercise 5.1. For the Markov source with $q_{01} = a$ and $q_{10} = b$, show that $H(\mu) = \frac{b\,h(a) + a\,h(b)}{a+b}$, and find the maximum over a, b. When is $H(\mu)$ equal to the entropy of the one-letter marginal?

Exercise 5.2. Let μ be a stationary source on $\{0,1\}$ and let ϕ be the sliding block code $\phi(x)_n = x_n + x_{n+1} \bmod 2$. Show that $H(\phi \circ \mu) \le H(\mu)$ directly, and that equality holds when μ is the source of independent fair bits. Is ϕ an isomorphism in that case?

Exercise 5.3. Let $N \geq 1$ and let $\mathbb{Z}/N$ act on $B^{\mathbb{Z}/N}$ by rotation. Show that for every $\mathbb{Z}/N$-set X, equivariant maps $X \to B^{\mathbb{Z}/N}$ correspond to maps $X \to B$, and count the equivariant maps $A^{\mathbb{Z}/N} \to B^{\mathbb{Z}/N}$.

Exercise 5.4. Let $c\colon S \to A \times S$ be a coalgebra in which the next state is a function $s' = \tau(a)$ of the emitted letter alone. Show that it generates a Markov source on A, find the transition matrix, and compute the entropy. Show that every stationary Markov source arises in this way with $S = A$.

Exercise 5.5. Prove the upper and lower bounds of Theorem 5.11, and show that the lower bound is nondecreasing and the upper bound nonincreasing in n.

Chapter 6

Ergodic Sources

The coding theorems are about what happens along a single long sequence, and a statement about one sequence can reflect the source only if the source does not secretly consist of several. Ergodicity is the hypothesis that it does not. This chapter states it in three equivalent ways: by invariant sets, as Khinchin does; as extremality in the convex set of stationary sources; and as determinism on the invariant σ-algebra, in the formulation of Moss and Perrone. It then shows that the entropy of a source is an affine function, which is why the ergodic case is the only one that needs to be treated.

6.1 Invariant sets

Let $[A, \mu]$ be a stationary source. A set $S \in \mathcal{F}_A$ is ***invariant*** if $T^{-1}S = S$. The invariant sets form a σ-algebra $\mathcal{I}$. A measurable function is $\mathcal{I}$-measurable iff it is invariant, $f \circ T = f$.

Definition 6.1. A stationary source is ***ergodic*** if every invariant set has measure 0 or 1; equivalently, if every invariant function is almost surely constant.

Proposition 6.2. [classical] *The source $p^{\otimes \mathbb{Z}}$ of independent letters is ergodic. A stationary Markov source with irreducible transition matrix is ergodic.*

Proof. Let S be invariant and $\varepsilon > 0$. Choose a set C depending on the coordinates in $[-m, m]$ with $\mu(S \triangle C) < \varepsilon$. For $n > 2m$ the set $T^{-n}C$ depends on a disjoint window, so for independent letters $\mu(C \cap T^{-n}C) = \mu(C)^2$. Since $S = S \cap T^{-n}S$, we get $|\mu(S) - \mu(S)^2| < 4\varepsilon$, and $\mu(S) \in \{0, 1\}$. For a Markov source the same argument works with $\mu(C \cap T^{-n}C)$ replaced by its Cesàro averages, which converge to $\mu(C)^2$ when the matrix is irreducible. □

A mixture $\lambda\mu' + (1-\lambda)\mu''$ of two different stationary sources, $0 < \lambda < 1$, is stationary and never ergodic (Corollary 6.5).

Theorem 6.3 (Birkhoff). [classical; [Bir31]] *Let μ be stationary and $f \in L^1(\mu)$. The averages $A_n f = \frac{1}{n}\sum_{k=0}^{n-1} f \circ T^k$ converge almost surely and in L^1 to the conditional expectation $\mathbb{E}[f \mid \mathcal{I}]$. If μ is ergodic the limit is the constant $\int f \, d\mu$.*

Khinchin quotes this theorem and uses the almost sure statement. For McMillan's theorem in the form he proves, convergence in L^1 is enough, and that part has a short proof by Hilbert space geometry which we give in Chapter 7.

6.2 The simplex of stationary sources

Let $M_T(A)$ be the set of stationary sources with alphabet A, with the topology of convergence of the probabilities of cylinders.

Theorem 6.4 (Ergodic decomposition). [classical] *$M_T(A)$ is a compact convex metrizable set, and a Choquet simplex. Its extreme points are the ergodic sources. Every stationary μ has a unique representation*

$$\mu = \int \mu_e \, d\lambda(e)$$

with λ a probability measure on the ergodic sources. Concretely, the regular conditional probability μ_x of μ given $\mathcal{I}$ is an ergodic source for almost every x, and $\mu = \int \mu_x \, d\mu(x)$.

Corollary 6.5. [classical] *Two different ergodic sources are mutually singular, and a proper mixture of two different stationary sources is not ergodic.*

Proof. If $\mu \neq \mu'$ are ergodic there is a cylinder C with $\mu(C) \neq \mu'(C)$. By Theorem 6.3 the invariant set on which the frequency of visits to C converges to $\mu(C)$ has μ-measure 1 and μ'-measure 0. Under a proper mixture it has measure strictly between 0 and 1. □

6.3 Ergodicity as determinism

The classical definition speaks of sets of measure zero or one. In a Markov category there is a name for a state that assigns every event probability zero or one: it is a deterministic state. The catch is that an ergodic source is certainly not deterministic as a state on $A^{\mathbb{Z}}$. It is deterministic on a quotient.

Work in the Markov category Stoch of all measurable spaces and Markov kernels. Let G be a group or monoid acting on an object X by deterministic morphisms; for sources, $G = \mathbb{Z}$ acts by the shift. Call a morphism $f\colon X \to Y$ ***invariant*** if $f \circ g = f$ for all $g \in G$. Let X_{inv} be the set X with the σ-algebra of invariant sets and $q\colon X \to X_{\mathrm{inv}}$ the map that is the identity on points.

Theorem 6.6 (Moss–Perrone). [literature; [MP23]; see also [EP23]]

(i) *$q\colon X \to X_{\mathrm{inv}}$ is the universal invariant morphism out of X: it is the colimit, in* Stoch*, of the diagram given by the action.*

(ii) An invariant state p on X is ergodic iff $q \circ p$ is a deterministic state.

(iii) In a Markov category with conditionals in which the colimit in (i) *exists and satisfies a compatibility condition, every invariant state p is a mixture of ergodic ones: the conditional $k\colon X_{\mathrm{inv}} \to X$ of p given q satisfies $p = k \circ (q \circ p)$, and k is, almost surely with respect to $q \circ p$, an ergodic invariant state.*

The proof of (iii) *is a short string-diagram argument that uses no limiting procedure and does not depend on the cardinality of G.*

Part (ii) is the classical definition read in Stoch, where the deterministic states are exactly the $\{0,1\}$-valued measures. Part (iii) recovers Theorem 6.4 for standard Borel X. What the formulation adds is a statement of where the decomposition comes from. It is the disintegration of p over the invariant quotient, and the ergodic components are the fibres. The space X_{inv} is usually not standard Borel, which is one reason to work in Stoch and not in BorelStoch here.

The same quotient organizes the ergodic theorem. On $L^2(\mu)$ the composite of q with its Bayesian inverse is the conditional expectation onto invariant functions, and Theorem 6.3 says that time averages converge to it. We return to this in Chapter 7.

6.4 Entropy is affine

The entropy of a finite scheme is a strictly concave function of the distribution. The entropy of a source is not.

Theorem 6.7. [classical; [Jac59, Par61]] *The function $H\colon M_T(A) \to [0, \log|A|]$ is affine and upper semicontinuous, and*

$$H(\mu) = \int H(\mu_e)\, d\lambda(e)$$

for the ergodic decomposition of μ.

Proof. Let $\mu = \lambda\mu' + (1-\lambda)\mu''$. The block law μ_n is the same mixture of μ'_n and μ''_n, and by concavity of H and Proposition 1.10,

$$\lambda H_n(\mu') + (1-\lambda) H_n(\mu'') \;\le\; H_n(\mu) \;\le\; \lambda H_n(\mu') + (1-\lambda) H_n(\mu'') + h(\lambda).$$

Divide by n and let $n \to \infty$. For semicontinuity, $H = \inf_n H_n/n$ by Proposition 5.4, and each H_n is continuous, depending on finitely many cylinder probabilities. An affine upper semicontinuous function on a Choquet simplex satisfies the barycentric formula. □

The term $h(\lambda)$ is the entropy of the choice between the two components. It is paid once, and its contribution per letter vanishes. This is the first instance of something that recurs: passing to rates turns a concave function into an affine one, and more generally removes whatever is not extensive.

Corollary 6.8. [classical] *For an affine upper semicontinuous function on $M_T(A)$, the supremum over all stationary sources is attained and equals the supremum over ergodic sources.*

Khinchin defines the capacity of a channel as a supremum over ergodic sources and calls it the ergodic capacity, leaving open whether a supremum over all stationary sources would be larger. The corollary reduces that question to showing that the rate of transmission is affine and upper semicontinuous in the source (Chapter 9).

6.5 An example

Example 6.9. Let μ be the equal mixture of two sources of independent bits, one with $\mathbb{P}(1) = 0.1$ and one with $\mathbb{P}(1) = 0.5$. It is stationary and not ergodic: the limiting frequency of ones is an invariant function taking two values. By Theorem 6.7,

$$H(\mu) = \tfrac{1}{2}\big(h(0.1) + 1\big) \approx 0.7345.$$

The normalized block entropies H_n/n are 0.8106, 0.7445, 0.7355 for $n = 10, 100, 1000$; the excess over the limit is at most $1/n$. But a typical long block does not have probability near $2^{-0.7345\,n}$. With probability $\frac{1}{2}$ it comes from the first component and has probability near $2^{-0.469\,n}$, and with probability $\frac{1}{2}$ it comes from the second and has probability near 2^{-n}. The entropy is the average of two rates, and no long block exhibits it.

The example shows what McMillan's theorem must assume. For a nonergodic source the quantity $-\frac{1}{n}\log \mu_n(x_1 \dots x_n)$ still converges, but to a random variable, the entropy of the ergodic component that produced x.

Notes

For ergodic theory see [Wal82, Pet83]; for its use in information theory, [Gra09, Gra11, Bil65]. Khinchin's treatment is [Khi56, §4]. The categorical ergodic decomposition is due to Moss and Perrone [MP23], with further development by Ensarguet and Perrone [EP23]; the synthetic zero–one laws of [FR20] give another route to Proposition 6.2. Affinity of the entropy rate and its integral representation are due to Jacobs [Jac59]; the corresponding statements for the rate of transmission are due to Parthasarathy [Par61] and Breiman [Bre60].

Exercises

Exercise 6.1. Let Q be a stochastic matrix with exactly two closed irreducible classes and no transient states. Show that the stationary Markov sources with matrix Q form a segment, that its endpoints are ergodic, and that the entropy is an affine function on it.

Exercise 6.2. Show that a stationary source is ergodic iff for all cylinders C, C', $\frac{1}{n}\sum_{k=0}^{n-1} \mu(C \cap T^{-k}C') \to \mu(C)\mu(C')$.

Exercise 6.3. Prove the two inequalities for $H_n(\mu)$ in the proof of Theorem 6.7, and show by an example that the upper one can be an equality for every n.

Exercise 6.4. For the source of Example 6.9 show that $-\frac{1}{n}\log \mu_n(x_1 \dots x_n)$ converges almost surely to a random variable taking the values $h(0.1)$ and 1 with probability $\frac{1}{2}$ each.

Exercise 6.5. Let μ give probability $\frac{1}{2}$ to each of the two sequences of alternating bits. Show that μ is ergodic for T and not for T^2, and that its entropy is zero.

Chapter 7

Martingales

To prove McMillan's theorem Khinchin needs to know that the conditional probability of the next letter given the last n letters converges as $n \to \infty$. He obtains this from Doob's martingale convergence theorem, to which he devotes a section. This chapter looks at the same convergence from the side of the category of probability spaces. Conditioning on a σ-algebra is an idempotent morphism, a filtration is an increasing sequence of idempotents, and the convergence of a martingale in mean is the convergence of that sequence to its supremum. The same picture contains the mean ergodic theorem. What it does not contain is almost sure convergence, and we say so.

7.1 The setting

Let $(X, \mathcal{F}, p)$ be a probability space. A ***filtration*** is an increasing sequence of sub-σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$, and $\mathcal{F}_\infty$ is the σ-algebra they generate. A ***martingale*** is a sequence of integrable functions M_n, with M_n measurable for $\mathcal{F}_n$ and $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$. It is ***closed*** if $M_n = \mathbb{E}[f \mid \mathcal{F}_n]$ for a single integrable f.

For a stationary source and a letter a, let $\mathcal{F}_n$ be generated by $x_{-n}, \ldots, x_{-1}$. Then

$$M_n^a(x) = \mu(x_0 = a \mid x_{-n}, \ldots, x_{-1}) = \mathbb{E}\big[1_{\{x_0 = a\}} \mid \mathcal{F}_n\big]$$

is a closed martingale with values in $[0, 1]$. This is the object Khinchin studies.

7.2 Probability spaces and kernels

Definition 7.1. Krn is the category whose objects are standard Borel probability spaces (X, p) and whose morphisms $(X, p) \to (Y, q)$ are Markov kernels k with $k \circ p = q$, two kernels being identified when they agree p-almost surely.

A morphism of Krn is the same as a ***coupling*** of p and q, the joint law $p \rtimes k$ on $X \times Y$. Reading the coupling from the other side gives the Bayesian inverse $k^\dagger \colon (Y, q) \to (X, p)$.

Proposition 7.2. [literature; [CJ19, Fri20, DDGK16]] *Bayesian inversion makes* Krn *a dagger category:* $k^{\dagger\dagger} = k$, $(l \circ k)^\dagger = k^\dagger \circ l^\dagger$, $\mathrm{id}^\dagger = \mathrm{id}$. *For each* $r \in [1, \infty]$ *there is a functor*

$L^r\colon \mathsf{Krn}^{\mathrm{op}} \to \mathsf{Ban}$ to Banach spaces and contractions, sending (X,p) to $L^r(p)$ and k to $(k^*g)(x) = \int g(y)\,k(dy \mid x)$. For $r=2$ it lands in Hilbert spaces and preserves daggers: $(k^\dagger)^* = (k^*)^\dagger$.

A measure-preserving map $f\colon (X,p) \to (Y,q)$ is a morphism of Krn, and $f^*g = g \circ f$ is an isometric embedding. Since f is deterministic, $f \circ f^\dagger = \mathrm{id}_Y$. The other composite

$$e_f = f^\dagger \circ f\colon (X,p) \to (X,p)$$

is therefore idempotent and self-adjoint, $e_f \circ e_f = e_f = e_f^\dagger$. As a kernel it does this: given x, forget everything except $f(x)$, and draw a new point from the conditional distribution given $f(x)$.

Proposition 7.3 (Conditional expectations are idempotents). [literature; [PVB24]]

(i) Every sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ arises, up to null sets, from a measure-preserving map f, and so defines a dagger idempotent $e_{\mathcal{G}} = e_f$ on (X,p) that splits through a deterministic map.

(ii) L^r sends $e_{\mathcal{G}}$ to the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{G}]$; for $r=2$ this is the orthogonal projection onto $L^2(\mathcal{G})$.

(iii) $\mathcal{G} \subseteq \mathcal{G}'$ iff $e_{\mathcal{G}} \circ e_{\mathcal{G}'} = e_{\mathcal{G}}$, and then also $e_{\mathcal{G}'} \circ e_{\mathcal{G}} = e_{\mathcal{G}}$.

Part (iii) is the tower property $\mathbb{E}[\mathbb{E}[f \mid \mathcal{G}'] \mid \mathcal{G}] = \mathbb{E}[f \mid \mathcal{G}]$. It defines a partial order on the idempotents, and a filtration is an increasing sequence $e_1 \le e_2 \le \cdots$ for it. A closed martingale is the image of one function under the sequence of projections $L^r(e_n)$.

7.3 Convergence in mean

Theorem 7.4. [classical] *Let $P_1 \le P_2 \le \cdots$ be an increasing sequence of orthogonal projections on a Hilbert space. Then $P_n v \to P_\infty v$ for every vector v, where P_∞ is the projection onto the closure of the union of the ranges. A decreasing sequence converges in the same sense to the projection onto the intersection of the ranges.*

Proof. If v lies in the range of some P_m then $P_n v = v$ for $n \ge m$. If v is orthogonal to all the ranges then $P_n v = 0$. Such vectors span a dense subspace and $\|P_n\| \le 1$. For a decreasing sequence apply this to $1 - P_n$. □

Corollary 7.5 (Lévy's theorems, in mean). [classical] *Let $1 \le r < \infty$ and $f \in L^r(p)$. If $\mathcal{F}_n$ increases to $\mathcal{F}_\infty$ then $\mathbb{E}[f \mid \mathcal{F}_n] \to \mathbb{E}[f \mid \mathcal{F}_\infty]$ in L^r. If $\mathcal{F}_n$ decreases to $\mathcal{F}_{-\infty} = \bigcap_n \mathcal{F}_n$ then $\mathbb{E}[f \mid \mathcal{F}_n] \to \mathbb{E}[f \mid \mathcal{F}_{-\infty}]$ in L^r.*

Proof. For $r = 2$ this is Theorem 7.4 with $P_n = \mathbb{E}[\,\cdot \mid \mathcal{F}_n]$, since $\bigcup_n L^2(\mathcal{F}_n)$ is dense in $L^2(\mathcal{F}_\infty)$. For bounded f the L^2 convergence together with the uniform bound $\|f\|_\infty$ gives L^r convergence for every finite r. Bounded functions are dense in L^r and conditional expectations are contractions of L^r. □

So the convergence of closed martingales in mean takes a few lines once conditional expectation is seen as a projection. Perrone and Van Belle show that the same argument can be run one level up, in the category itself.

Theorem 7.6 (Perrone–Van Belle). [literature; [PVB24]] *The categories* Krn *and* Hilb *are enriched in topological spaces in such a way that*

(i) every monotone net of dagger idempotents converges, in the topology of the hom-space, to its supremum if it is increasing and to its infimum if it is decreasing;

(ii) the functors L^r, $1 \le r < \infty$, are enriched, that is, continuous on hom-spaces.

Consequently martingales and backward martingales indexed by arbitrary directed sets converge in L^r, and the same holds for vector-valued martingales with any enriched functor in place of L^r.

The content of (i) for Krn is that the kernels e_n themselves converge: the conditional distributions given $\mathcal{F}_n$ converge to the conditional distribution given $\mathcal{F}_\infty$, in a topology that makes integration against bounded functions continuous in L^1. A martingale is one function watched through this convergence. For the source, the kernels are "the law of x_0 given the last n letters", and the limit is the law of x_0 given the infinite past.

7.4 Almost sure convergence

Theorem 7.7 (Doob). [classical; [Doo53]] *A martingale with $\sup_n \mathbb{E}|M_n| < \infty$ converges almost surely. In particular, in both cases of Corollary 7.5 the convergence holds almost surely as well.*

This is the theorem Khinchin quotes. Its proof needs an inequality that controls all times at once, either Doob's upcrossing inequality or the maximal inequality $\mathbb{P}(\sup_n |M_n| > t) \le t^{-1} \sup_n \mathbb{E}|M_n|$. Nothing in Theorem 7.4 or Theorem 7.6 supplies it. This is the usual division of labour in analysis: convergence in norm comes from density and uniform boundedness, almost everywhere convergence from a maximal inequality. The categorical methods available at present reach the first and not the second. For the theorem we are aiming at this does not matter, because McMillan's theorem as Khinchin states it is a statement about convergence in mean, and we shall prove it in Chapter 8 without Theorem 7.7.

7.5 Limits against colimits

A filtration gives two diagrams of Banach spaces: the inclusions $L^r(\mathcal{F}_n) \hookrightarrow L^r(\mathcal{F}_{n+1})$, and the conditional expectations $L^r(\mathcal{F}_{n+1}) \twoheadrightarrow L^r(\mathcal{F}_n)$ in the opposite direction. Work in the category of Banach spaces and contractions, where a limit of a sequence consists of the bounded compatible families.

Proposition 7.8. [observation] *Let $\mathcal{F}_n$ increase to $\mathcal{F}_\infty$.*

(i) For $1 \le r < \infty$, $L^r(\mathcal{F}_\infty)$ is the colimit of the inclusions.

(ii) The limit of the conditional expectations is the space of L^r-bounded martingales, with norm $\sup_n \|M_n\|_r$.

(iii) For $1 < r < \infty$ the map $f \mapsto (\mathbb{E}[f \mid \mathcal{F}_n])_n$ from the colimit to the limit is an isometric isomorphism.

(iv) For $r = 1$ it is isometric and in general not surjective. For the filtration generated by a sequence of fair coin tosses, the limit is the space of all finite signed measures on the sequence space, and the image of the colimit is the subspace of measures absolutely continuous with respect to the coin-tossing measure.

Proof. (i) is the density used in Corollary 7.5. (ii) is the definition of a martingale. (iii) is the classical fact that for $r > 1$ an L^r-bounded martingale is closed, with $\|f\|_r = \sup_n \|\mathbb{E}[f \mid \mathcal{F}_n]\|_r$. (iv) On a finite σ-algebra an integrable function is the density of a signed measure, and a martingale is a compatible family of signed measures on the algebras of cylinders with bounded total variation, which extends uniquely to a signed measure on the sequence space. Closed martingales correspond to measures with a density. □

For $r = 2$ the inclusions and the conditional expectations are adjoint to each other, and (iii) is the statement that for a sequence of embedding–projection pairs the colimit of the embeddings is the limit of the projections. The same coincidence, for a different kind of object, is what makes recursive domain equations solvable [SP82]. For $r = 1$ it fails, and the failure is the Lebesgue decomposition: the martingales that do not come from a function on $\mathcal{F}_\infty$ are the singular measures. The likelihood ratio of a biased coin against a fair one is the standard example (Exercise 7.4).

7.6 The mean ergodic theorem

Theorem 7.9 (von Neumann). [classical; [vN32]] *Let U be an isometry of a Hilbert space and P the orthogonal projection onto the space $\mathrm{Fix}(U)$ of invariant vectors. Then $\frac{1}{n}\sum_{k=0}^{n-1} U^k v \to Pv$ for every v.*

Proof. First $\mathrm{Fix}(U) = \mathrm{Fix}(U^*)$: if $Uv = v$ then $\|U^*v - v\|^2 = \|U^*v\|^2 - 2\,\mathrm{Re}\langle v, Uv\rangle + \|v\|^2 \leq 0$, and symmetrically. Hence the orthogonal complement of $\mathrm{Fix}(U)$ is the closure of the range of $1 - U$. On $\mathrm{Fix}(U)$ the averages are constant. On a vector $w - Uw$ they equal $(w - U^n w)/n$, which tends to zero. The averages are contractions, so the convergence extends to the closure. □

Corollary 7.10. [classical] *Let μ be a stationary source and $f \in L^1(\mu)$. Then $\frac{1}{n}\sum_{k<n} f \circ T^k \to \mathbb{E}[f \mid \mathcal{I}]$ in L^1.*

Proof. $Uf = f \circ T$ is an isometry of $L^2(\mu)$ whose invariant vectors are the $\mathcal{I}$-measurable functions. For $f \in L^2$ apply Theorem 7.9; L^2 is dense in L^1 and the averages are contractions of L^1. □

In the language of this chapter the limit is again an idempotent. Let $q\colon X \to X_{\mathrm{inv}}$ be the invariant quotient of Theorem 6.6. Then $e_{\mathcal{I}} = q^\dagger \circ q$ is the dagger idempotent of the invariant σ-algebra, and the corollary says that the Cesàro averages of the powers of T converge to it under L^1.

We now have two convergence statements of the same form. Along a filtration, an increasing sequence of idempotents converges to its supremum. Along time, the averages of an automorphism converge to the idempotent of its invariants. The first is an instance of Theorem 7.6. The second is not, since the averages are not idempotents and do not form a monotone net; whether it has a proof internal to an enriched dagger Markov category is a question we return to in Chapter 13.

Notes

Khinchin's section on martingales is [Khi56, §6], after Doob [Doo53]. The dagger structure on categories of probability spaces is discussed in [CJ19, Fri20, DDGK16]. The treatment of conditional expectations as idempotents and Theorem 7.6 are due to Perrone and Van Belle [PVB24]; see also Van Belle [VB23]. The proof of the mean ergodic theorem given here is F. Riesz's.

Exercises

Exercise 7.1. Let $X = \{1, \dots, 6\}$ with the uniform distribution and let $\mathcal{G}$ be generated by the partition $\{1,2\} \mid \{3,4,5,6\}$. Write $e_{\mathcal{G}}$ as a stochastic matrix, check that it is idempotent and symmetric, and compute its action on $f(x) = x$.

Exercise 7.2. Prove Proposition 7.3(iii). Give two σ-algebras on a three-point space whose idempotents do not commute.

Exercise 7.3. On $[0,1)$ with Lebesgue measure let $\mathcal{F}_n$ be generated by the dyadic intervals of length 2^{-n} and let $f(x) = x$. Compute $\mathbb{E}[f \mid \mathcal{F}_n]$ and show that $\|f - \mathbb{E}[f \mid \mathcal{F}_n]\|_2 = 2^{-n}/\sqrt{12}$.

Exercise 7.4. Let $x_1, x_2, \dots$ be fair coin tosses and $M_n = \prod_{k \le n} 2\theta^{x_k}(1-\theta)^{1-x_k}$ with $\theta = 0.8$. Show that M_n is a martingale with $\mathbb{E}M_n = 1$ and that $\mathbb{E}\sqrt{M_n} = \rho^n$ with $\rho = \frac{1}{2}(\sqrt{1.6} + \sqrt{0.4}) \approx 0.949$. Deduce that $M_n \to 0$ in probability and that M_n is not closed.

Exercise 7.5. Let U be the permutation matrix of $\sigma = (1\,2\,3)(4\,5)$ acting on functions on $\{1, \dots, 5\}$. Compute $\frac{1}{6}\sum_{k<6} U^k$ and identify it with a conditional expectation.

Chapter 8

McMillan's Theorem

McMillan's theorem says that for an ergodic source almost all long blocks have about the same probability, 2^{-nH}. Shannon proved it for independent letters and for Markov chains, where it is the law of large numbers, and Khinchin's first paper follows him. The general case is due to McMillan, and proving it carefully is the main task of the second chapter of Khinchin's second paper, which takes some thirty pages and rests on the theorems of Birkhoff and Doob. With the preparation of the last two chapters the proof is short, and it shows what the theorem depends on: the chain rule, two convergence statements about idempotents, and one elementary estimate.

8.1 The statement

For a stationary source μ and $n \geq 1$ let

$$i_n(x) = -\log \mu_n(x_0 \dots x_{n-1})$$

be the ***information*** of the first n letters. Its expectation is the block entropy H_n.

Definition 8.1 (Khinchin)**.** A source has the E ***property*** if $\frac{1}{n} i_n \to H(\mu)$ in probability: for all $\varepsilon, \delta > 0$ and all large n, the blocks C of length n with $|-\frac{1}{n}\log \mu_n(C) - H(\mu)| \leq \varepsilon$ have total probability more than $1-\delta$.

Theorem 8.2 (McMillan)**.** [classical; [McM53, Khi56]] *For an ergodic source, $\frac{1}{n} i_n \to H(\mu)$ in $L^1(\mu)$. In particular every ergodic source has the E property. For a stationary source, $\frac{1}{n} i_n$ converges in L^1 to the entropy $H(\mu_x)$ of the ergodic component of x.*

8.2 Information as a sum over time

Let $\mathcal{F}_k$ be generated by $x_{-k}, \dots, x_{-1}$, with $\mathcal{F}_0$ trivial and $\mathcal{F}_\infty$ generated by the whole past. For $0 \leq k \leq \infty$ put

$$g_k(x) = -\log \mu(x_0 \mid \mathcal{F}_k)(x),$$

the information of the present letter given the last k letters. In the notation of Chapter 7, $g_k = -\log M_k^{x_0}$.

Lemma 8.3. [classical] *$i_n = \sum_{k=0}^{n-1} g_k \circ T^k$, and $\mathbb{E} g_k = c_k$, the increment of Proposition 5.4.*

Proof. $\mu_n(x_0 \dots x_{n-1}) = \prod_{k<n} \mu(x_k \mid x_0, \dots, x_{k-1})$, and by stationarity the k-th factor is $\mu(x_0 \mid \mathcal{F}_k)$ evaluated at $T^k x$. □

This is the chain rule of Chapter 1 spread along the time axis. If the summands were one function g composed with the powers of T, the ergodic theorem would finish the proof. They are not, but they converge.

Lemma 8.4. [classical] *For every $0 \le k \le \infty$ and $t \ge 0$,*

$$\mu\{g_k > t\} \le |A|\, 2^{-t}.$$

Consequently $\sup_k \mathbb{E}[g_k^2] < \infty$, and $\mu(x_0 \mid \mathcal{F}_\infty) > 0$ almost surely.

Proof. For a letter a and $s = 2^{-t}$,

$$\mu\{x_0 = a,\ M_k^a \le s\} = \mathbb{E}\big[1_{\{M_k^a \le s\}}\, \mathbb{E}[1_{\{x_0 = a\}} \mid \mathcal{F}_k]\big] = \mathbb{E}\big[1_{\{M_k^a \le s\}} M_k^a\big] \le s.$$

Sum over a. The second moment is $\int_0^\infty 2t\, \mu\{g_k > t\}\, dt$. □

Lemma 8.5. [classical] *$g_k \to g_\infty$ in L^1, and $\mathbb{E} g_\infty = H(\mu)$.*

Proof. By Corollary 7.5, $M_k^a \to M_\infty^a$ in L^1 for each letter, so $M_k^{x_0} \to M_\infty^{x_0}$ in probability. The limit is almost surely positive by Lemma 8.4 and $-\log$ is continuous on $(0, 1]$, so $g_k \to g_\infty$ in probability. A sequence bounded in L^2 is uniformly integrable, so the convergence is in L^1. Then $\mathbb{E} g_\infty = \lim \mathbb{E} g_k = \lim c_k = H(\mu)$. □

8.3 Proof of the theorem

Proof of Theorem 8.2. Since T preserves μ, Lemma 8.3 gives

$$\Big\| \frac{1}{n} i_n - \frac{1}{n} \sum_{k<n} g_\infty \circ T^k \Big\|_1 \le \frac{1}{n} \sum_{k<n} \| g_k - g_\infty \|_1,$$

which tends to zero by Lemma 8.5. By the mean ergodic theorem (Corollary 7.10), $\frac{1}{n} \sum_{k<n} g_\infty \circ T^k \to \mathbb{E}[g_\infty \mid \mathcal{I}]$ in L^1. If μ is ergodic this is the constant $\mathbb{E} g_\infty = H(\mu)$. In general it is an invariant function whose integral over each ergodic component is the entropy of that component. □

The argument used four things. The chain rule put i_n in the form of Lemma 8.3. Lévy's theorem in mean, which is the convergence of an increasing sequence of idempotents (Theorem 7.4), gave the convergence of

the conditional laws. The mean ergodic theorem, which is the convergence of time averages to the invariant idempotent (Theorem 7.9), gave the limit. Lemma 8.4 supplied the uniform integrability needed to pass from conditional probabilities to their logarithms. Doob's almost sure theorem and Birkhoff's were not used. This organization of the proof is certainly known to specialists; we set it out because it isolates how little of the measure theory in Khinchin's proof is specific to the problem.

Remark 8.6. The limit can be written in one line. Let e_{past} be the idempotent of the σ-algebra generated by the past and $e_{\mathcal{I}}$ that of the invariant sets. The first determines the conditional law of x_0 given the past, hence the function g_∞; the second is applied to it:

$$\lim_n \tfrac{1}{n} i_n = L^1(e_{\mathcal{I}})\, g_\infty .$$

The entropy of an ergodic source is the time average of the information of the present given the past.

8.4 The almost sure theorem

Theorem 8.7 (Shannon–McMillan–Breiman). [classical; [Bre57, Chu61, AC88]] *For an ergodic source, $\frac{1}{n} i_n \to H(\mu)$ almost surely. The same holds for countable alphabets when $H_1 < \infty$.*

Breiman's proof combines the almost sure forms of the two convergence theorems with a maximal inequality for $\sup_k g_k$, of which Lemma 8.4 is the distributional part; Algoet and Cover later gave a proof by sandwiching i_n between Markov approximations. The theorem extends to actions of amenable groups [OW83, Lin01]. For the coding theorems convergence in probability is all that is needed.

8.5 Typical sets and the counting theorem

Fix $\varepsilon > 0$ and let T_n^ε be the set of blocks C of length n with $2^{-n(H+\varepsilon)} \le \mu_n(C) \le 2^{-n(H-\varepsilon)}$, where $H = H(\mu)$.

Corollary 8.8. [classical] *For an ergodic source, $\mu_n(T_n^\varepsilon) \to 1$; $|T_n^\varepsilon| \le 2^{n(H+\varepsilon)}$ for all n; and $|T_n^\varepsilon| \ge (1-\delta)\, 2^{n(H-\varepsilon)}$ as soon as $\mu_n(T_n^\varepsilon) \ge 1-\delta$.*

For $0 < \lambda < 1$ let $N_n(\lambda)$ be the least number of blocks of length n whose total probability is at least λ.

Theorem 8.9 (Khinchin's Theorem 3). [classical; [Sha48, Khi53]] *For an ergodic source and every $\lambda \in (0,1)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log N_n(\lambda) = H(\mu).$$

Proof. For large n the typical set has probability at least λ, so $N_n(\lambda) \le 2^{n(H+\varepsilon)}$. Conversely let S have probability at least λ. For large n, $\mu_n(S \cap T_n^\varepsilon) \ge \lambda/2$, and each block in T_n^ε has probability at most $2^{-n(H-\varepsilon)}$, so $|S| \ge \frac{\lambda}{2}\, 2^{n(H-\varepsilon)}$. $\square$

The limit does not depend on λ. That is the remarkable part of the statement, and it is special to the ergodic case.

Example 8.10. For independent bits with $\mathbb{P}(1) = 0.2$ the entropy is $h(0.2) \approx 0.7219$. The typical set with $\varepsilon = 0.05$ has probability 0.468 at $n = 100$ and 0.952 at $n = 1000$. The rates $\frac{1}{n}\log N_n(\lambda)$ are:

	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.9$
$n = 100$	0.571	0.684	0.779
$n = 1000$	0.683	0.717	0.748

The spread across λ shrinks like $1/\sqrt{n}$ (Exercise 8.5). For the Markov source of Example 5.8, $\mathbb{E}\left|\frac{1}{n} i_n - H\right|$ is 0.447, 0.216, 0.109 at $n = 4, 16, 64$.

Example 8.11. For the non-ergodic source of Example 6.9 the limit depends on λ. At $n = 1000$ the rate $\frac{1}{n}\log N_n(\lambda)$ is 0.463 for $\lambda = 0.25$ and 0.999 for $\lambda = 0.75$; the limits are $h(0.1) \approx 0.469$ and 1 (Exercise 8.4). The entropy 0.7345 of the source is not the answer to any counting question.

So for a stationary source that is not ergodic, the entropy rate is not an operational quantity. The number of bits per letter needed to describe all but a fraction ε of the blocks is a quantile of the distribution of $H(\mu_x)$, and to make ε arbitrarily small one needs the essential supremum. This is the starting point of the information-spectrum method [HV93, Han03]. Khinchin's restriction to ergodic sources is therefore not a convenience. His Theorem 4, on the best compression of a text by codes that may fail on a set of small probability, is false without it. For codes of variable length judged by their expected length the entropy rate remains the answer for every stationary source (Chapter 11).

8.6 The theorem as a statement about retractions

Definition 8.12. Let (X, p) be a finite probability space and $0 < \varepsilon < 1$. An (M, ε)-***compression*** of (X, p) is a pair of maps $e\colon X \to [M]$ and $d\colon [M] \to X$, where $[M] = \{1, \dots, M\}$, such that $d(e(x)) = x$ with probability at least $1 - \varepsilon$. Let $M_*(p, \varepsilon)$ be the least M for which one exists.

When $\varepsilon = 0$ and p has full support this says that X is a retract of $[M]$, which happens iff $|X| \le M$. In general the set where $d \circ e$ is the identity has at most M points, and any M points can be made such a set, so $M_*(p, \varepsilon) = N(1 - \varepsilon)$ in the notation above.

Proposition 8.13. [reformulation; cf. Theorem 8.9] *For an ergodic source and every $\varepsilon \in (0,1)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log M_*(\mu_n, \varepsilon) = H(\mu).$$

Up to ε, the space of blocks of length n is a retract of a set with 2^{nH} points and of no smaller one. Entropy is the exponential growth rate of the approximate cardinality of A^n as seen by μ. In Chapter 10 the dual notion, a finite set that is an approximate retract of a channel, will define capacity, and the two will be composed in Chapter 11.

McMillan's theorem says more than Proposition 8.13, since it controls the probabilities of the typical blocks from both sides.

Proposition 8.14. [classical] *Let μ be ergodic and $\varepsilon > 0$. For large n there is a set $S \subseteq A^n$ with $\mu_n(S) \geq 1 - \varepsilon$ such that, for the uniform distribution u_S on S and all $C \in S$,*

$$\left| \frac{1}{n} \log \frac{\mu_n(C)}{u_S(C)} \right| \leq 2\varepsilon + \frac{1}{n}.$$

Thus μ_n is close to a uniform distribution, but only on the exponential scale: the ratio of probabilities is $2^{o(n)}$, not close to 1, and μ_n is in general far from u_S in total variation. Gromov has made this scale into a framework [Gro13]. He considers the category of finite probability spaces and measure-preserving maps, defines an asymptotic equivalence between sequences of such spaces that ignores sets of small measure and subexponential factors, and reads the law of large numbers as the statement that the powers $P^{\otimes n}$ are asymptotically equivalent to the uniform spaces on $2^{nH(P)}$ points. Entropy is then the isomorphism from the semigroup of asymptotic classes, under tensor product, to $[0,\infty)$. McMillan's theorem extends the first half of this from powers of one space to the block laws of any ergodic source.

Notes

McMillan's theorem is [McM53]; Khinchin's proof is [Khi56, §§5–8]. The almost sure version is due to Breiman [Bre57], with the countable case by Chung [Chu61]; the sandwich proof is [AC88]. Textbook accounts are in [Bil65, Gra11, CT06]. The counting theorem and its use for coding go back to Shannon [Sha48]. The second-order behaviour in Exercise 8.5 is due to Strassen. For the non-ergodic case and the information spectrum see [Han03].

Exercises

Exercise 8.1. For independent letters with distribution p, derive the E property from Chebyshev's inequality, with the bound $\mathbb{P}(|\frac{1}{n} i_n - H| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2)$

where σ^2 is the variance of $-\log p(x_0)$. Compute σ^2 for bits with $\mathbb{P}(1) = 0.2$.

Exercise 8.2. For a stationary Markov source show that $g_k = g_1$ for all $k \geq 1$, and deduce McMillan's theorem from Corollary 7.10 alone.

Exercise 8.3. Prove Corollary 8.8 and Proposition 8.14.

Exercise 8.4. For the source of Example 6.9 show that $\frac{1}{n} \log N_n(\lambda) \to h(0.1)$ for $\lambda < \frac{1}{2}$ and $\to 1$ for $\lambda > \frac{1}{2}$.

Exercise 8.5. For independent bits with $\mathbb{P}(1) = p < \frac{1}{2}$, show that $N_n(\lambda)$ is attained by a set of sequences with the fewest ones, and use the central limit theorem to show that

$$\frac{1}{n} \log N_n(\lambda) = h(p) + \frac{\sigma \Phi^{-1}(\lambda)}{\sqrt{n}} + o\Big(\frac{1}{\sqrt{n}}\Big), \qquad \sigma = \sqrt{p(1-p)}\, \log \frac{1-p}{p},$$

where Φ is the standard normal distribution function. Compare with the table in Example 8.10.

Part III

Channels and Coding

Chapter 9

Channels

A channel, for Khinchin, is a family of probability measures ν_x on output sequences, one for each input sequence x. He imposes three conditions on it, stationarity, absence of anticipation and finite memory, connects it to a source, and proves that the resulting compound source is stationary and, when the source is ergodic, ergodic. In this chapter a channel is a morphism of the category Dyn of Chapter 5. The first of Khinchin's theorems becomes the statement that morphisms compose. The second is false as he states it, as was noticed within a few years, and the categorical formulation shows where the missing hypothesis has to go. We then study capacity as a function on morphisms.

9.1 Khinchin's definitions

Definition 9.1 ([Khi56], §9)**.** A ***channel*** $[A, \nu, B]$ with input alphabet A and output alphabet B is a Markov kernel $\nu\colon A^{\mathbb{Z}} \to B^{\mathbb{Z}}$. It is

- ***stationary*** if $\nu \circ T = T \circ \nu$, that is, $\nu_{Tx}(TS) = \nu_x(S)$;
- ***without anticipation*** if for every n the law of $(\dots, y_{n-1}, y_n)$ under ν_x depends only on $(\dots, x_{n-1}, x_n)$;
- of ***finite memory*** m if for all $s \le t$ the law of $(y_s, \dots, y_t)$ under ν_x depends only on $(x_{s-m}, \dots, x_t)$.

A ***memoryless*** channel is given by a stochastic matrix $W\colon A \to B$ and $\nu_x = \bigotimes_n W(\,\cdot \mid x_n)$: the output letters are independent given the input, and the n-th depends on x_n alone. It has all three properties with $m = 0$.

9.2 Connecting a channel to a source

By Proposition 5.3, a stationary channel is a morphism $(A^{\mathbb{Z}}, T) \to (B^{\mathbb{Z}}, T)$ of Dyn and a stationary source is a state.

Definition 9.2. The ***compound source*** of a source $[A, \mu]$ and a channel $[A, \nu, B]$ is the state

$$\omega = (\mathrm{id} \otimes \nu) \circ \mathrm{copy} \circ \mu \quad \text{on } A^{\mathbb{Z}} \times B^{\mathbb{Z}} = (A \times B)^{\mathbb{Z}},$$

that is, $\omega(M \times N) = \int_M \nu_x(N)\, d\mu(x)$. Its second marginal $\eta = \nu \circ \mu$ is the ***output source***.

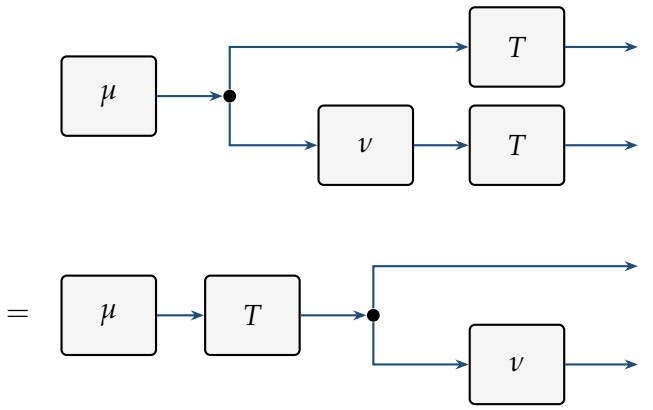


Figure 9.1: The compound source and the proof that it is stationary. The shift slides through ν by equivariance and through the copy node because it is deterministic; it is then absorbed by μ.

Theorem 9.3 (Khinchin). [reformulation; cf. [Khi56], §10] *If the source and the channel are stationary, so are the compound source and the output source.*

Proof. copy and ν are morphisms of Dyn and μ is a state, so ω is a state of Dyn. Written out (Figure 9.1),

$$\begin{aligned}(T \otimes T) \circ (\mathrm{id} \otimes \nu) \circ \mathrm{copy} \circ \mu &= (\mathrm{id} \otimes \nu) \circ (T \otimes T) \circ \mathrm{copy} \circ \mu \\ &= (\mathrm{id} \otimes \nu) \circ \mathrm{copy} \circ T \circ \mu = \omega,\end{aligned}$$

using in turn the equivariance of ν, the determinism of T, and the invariance of μ. □

The same reasoning shows that channels compose: from $[A, \nu, B]$ and $[B, \nu', C]$ one gets $[A, \nu' \circ \nu, C]$, stationary if both are. A stationary code is a deterministic channel, and coding a source before sending it through a channel is composition. Khinchin uses all of this in his last chapter, where a code is treated as a channel so that the previous theorems apply to it.

9.3 Anticipation and naturality

To see what absence of anticipation means, take one-sided time. The space $A^{\mathbb{N}}$ of sequences $x_1 x_2 \ldots$ is the limit of the tower $\cdots \to A^{n+1} \to A^n \to \cdots$ of maps forgetting the last letter.

Proposition 9.4. [reformulation] *A kernel $\nu\colon A^{\mathbb{N}} \to B^{\mathbb{N}}$ is without anticipation iff it is induced by a family of kernels $\nu_n\colon A^n \to B^n$ in* FinStoch *that commutes with the maps of the two towers, a natural transformation. The correspondence is bijective. A memoryless channel is the family $\nu_n = W^{\otimes n}$.*

Proof. If ν is without anticipation, the law of $y_1 \dots y_n$ is a function of $x_1 \dots x_n$, which defines ν_n, and marginalizing y_{n+1} gives naturality. Conversely a natural family composed with the projections $A^{\mathbb{N}} \to A^n$ is a compatible family of kernels $A^{\mathbb{N}} \to B^n$, which by Theorem 5.1 is a kernel into $B^{\mathbb{N}}$. □

Natural transformations between towers of sets are the causal functions on streams, and they are the morphisms between stream objects in the topos of trees, the presheaf category on the ordinal ω [BMSS12]. The proposition is the stochastic version: channels without anticipation are natural transformations of towers in FinStoch. Among them the memoryless ones are those that are also compatible with concatenation, $\nu_{m+n} = \nu_m \otimes \nu_n$. Memory of length m lies in between.

9.4 Ergodicity: a gap and its repair

Khinchin's second theorem about the compound source reads: if the source is ergodic and the channel is stationary, without anticipation and of finite memory, then the compound source is ergodic. He needs it in order to apply McMillan's theorem to the compound source.

Example 9.5. Let $\nu_x = \frac{1}{2}\delta_{\mathbf{0}} + \frac{1}{2}\delta_{\mathbf{1}}$ for every x, where $\mathbf{0}$ and $\mathbf{1}$ are the constant sequences. This channel ignores its input and emits one random bit forever. It is stationary. It is without anticipation and has memory 0 in the sense of Definition 9.1, since the law of any output block does not depend on x at all. For any source, $\omega = \mu \otimes \nu$, and the set $A^{\mathbb{Z}} \times \{\mathbf{0}\}$ is invariant with probability $\frac{1}{2}$. The compound source is not ergodic.

So the theorem fails under the stated hypotheses. Breiman noted that Khinchin's definition of finite memory is not restrictive enough for his results, and Takano and Feinstein had by then added the missing condition [Bre60, Tak58, Fei59].

Definition 9.6. A channel has ***finite output memory*** m if, for every x, the blocks $(y_s, \dots, y_t)$ and $(y_{s'}, \dots, y_{t'})$ are independent under ν_x whenever $s' - t > m$.

Theorem 9.7. [classical; [Tak58, Fei59, Adl61]] *Let the channel be stationary, of finite memory in the sense of Definition 9.1 and of finite output memory. If the source is ergodic, the compound source and the output source are ergodic.*

Proof. Let $E = C \times D$ and $E' = C' \times D'$ be products of cylinders in $A^{\mathbb{Z}}$ and $B^{\mathbb{Z}}$. By finite memory $f(x) = 1_C(x)\, \nu_x(D)$ depends on finitely many coordinates of x, and likewise f'. For large k the output blocks defining D and $T^{-k}D'$ are more than m apart, so by finite output memory and stationarity

$$\omega(E \cap T^{-k}E') = \int f(x)\, f'(T^k x)\, d\mu(x).$$

By the mean ergodic theorem the Cesàro averages of the right side converge to $\int f\, d\mu \int f'\, d\mu = \omega(E)\, \omega(E')$, and this criterion on cylinders is equivalent to ergodicity (Exercise 6.2). □

The categorical picture explains the shape of the gap. By Proposition 5.6, an equivariant map into $B^{\mathbb{Z}}$ is determined by its time-zero coordinate, but an equivariant kernel is not. A stationary channel consists of the conditional laws of finite output blocks given the input, and of the way these fit together across time. Khinchin's memory condition bounds the first, the dependence on the input. Ergodicity of the composite is a statement about the second, and Example 9.5 is a channel that is as trivial as possible in the first respect and as bad as possible in the second. In Dyn the ergodic states are not preserved by arbitrary morphisms; Theorem 9.7 describes a class of morphisms that do preserve them. Deterministic morphisms always do, since a factor of an ergodic source is ergodic.

9.5 Rate and capacity

Definition 9.8 ([Khi56], §10)**.** For a stationary source and channel, the ***rate of transmission*** is

$$R(\mu, \nu) = H(\mu) + H(\eta) - H(\omega) = \lim_{n \to \infty} \frac{1}{n}\, I(x_1 \dots x_n;\, y_1 \dots y_n).$$

The ***ergodic capacity*** C of the channel is the supremum of $R(\mu, \nu)$ over ergodic sources μ.

Since $\mu \mapsto \eta$ and $\mu \mapsto \omega$ are affine, Theorem 6.7 gives at once:

Proposition 9.9. [classical; [Par61]] *For every stationary channel, $R(\,\cdot\,, \nu)$ is an affine function on the simplex of stationary sources.*

Theorem 9.10. [classical; [Bre60, Par61]] *For a stationary channel of finite memory and finite output memory, $R(\,\cdot\,, \nu)$ is upper semicontinuous. Hence the supremum over all stationary sources equals the ergodic capacity, and it is attained by an ergodic source.*

The second statement follows from the first by Corollary 6.8. It answers the question Khinchin leaves open in his conclusion, where he remarks that capacity is usually defined carelessly and that his own definition may give a smaller number than a supremum over all stationary sources. It does not.

9.6 Memoryless channels

For a stochastic matrix $W\colon A \to B$ and a distribution p on A, write $I(p, W)$ for the mutual information of the joint distribution $p \rtimes W$.

Theorem 9.11. [classical; [Sha48]] *The capacity of the memoryless channel with matrix W is*

$$C(W) = \max_p I(p, W),$$

attained by the source of independent letters with a maximizing distribution.

Proof. Given the input block, the output letters are independent, so $I(x_1^n; y_1^n) = H(y_1^n) - \sum_k H(y_k \mid x_k) \le \sum_k I(x_k; y_k) \le n \max_p I(p, W)$. Independent inputs give equality. □

Example 9.12. The binary symmetric channel with flip probability δ has $C = 1 - h(\delta)$; for $\delta = 0.1$ this is 0.531. The binary erasure channel, which replaces its input by an erasure symbol with probability e, has $C = 1 - e$. The Z-channel, which sends 0 to 0 and flips 1 to 0 with probability δ, has $C = \log\bigl(1 + (1-\delta)\,\delta^{\delta/(1-\delta)}\bigr)$, which is 0.763 for $\delta = 0.1$.

9.7 Capacity as a function on morphisms

Restricted to memoryless channels, capacity is a function on the morphisms of FinStoch.

Proposition 9.13. [classical]

(i) $C(\mathrm{id}_X) = \log|X|$, *and* $C(W) = 0$ *iff W factors through the one-point set.*

(ii) $C(V \circ W) \le \min\{C(V), C(W)\}$.

(iii) $C(W \otimes W') = C(W) + C(W')$.

(iv) If W factors through a set with r elements, then $C(W) \le \log r$.

Proof. (ii) For $X \to Y \to Z$ the data-processing inequality gives $I(X;Z) \le \min\{I(X;Y), I(Y;Z)\}$. (iii) Product inputs give $\ge$; conversely for the product channel $I(XX';YY') = H(YY') - H(Y \mid X) - H(Y' \mid X') \le I(X;Y) + I(X';Y')$. (iv) follows from (i) and (ii). □

These are the properties of the logarithm of the rank of a matrix: it is $\log n$ on the identity of an n-dimensional space, it does not increase under composition, it is additive under tensor product, and it is bounded by the logarithm of the dimension of anything the matrix factors through. Capacity is a rank for stochastic matrices that is stable under small perturbations. The

next chapter makes this exact: capacity is the growth rate of the largest identity onto which the tensor powers of W retract, up to ε.

Example 9.14 (A channel with memory)**.** Let $A = B = \{0,1\}$ and $y_n = x_n \oplus x_{n-1} \oplus z_n$, where the z_n are independent bits equal to 1 with probability δ. This channel is stationary and without anticipation, has memory 1, and has output memory 0, since the outputs are independent given the input. It is the composite of the sliding block code $\phi(x)_n = x_n \oplus x_{n-1}$ with the memoryless binary symmetric channel, so its capacity is at most $1 - h(\delta)$ by the inequality above. Equality holds: ϕ maps independent fair bits to independent fair bits, so the source of independent fair bits has rate of transmission $1 - h(\delta)$. The memory costs nothing here because ϕ is invertible up to one bit.

For channels with memory (ii) persists in the form $R(\mu, \nu' \circ \nu) \leq \min\{R(\mu,\nu), R(\nu \circ \mu, \nu')\}$, by the same inequality applied to blocks.

9.8 Joinings and the backward channel

The compound source ω is a stationary measure on $A^{\mathbb{Z}} \times B^{\mathbb{Z}}$ with marginals μ and η, a ***joining*** of the two sources in the sense of ergodic theory [Gla03]. Conversely, by disintegration, every joining arises from a stationary channel, unique up to μ-null sets. So the morphisms from $(A^{\mathbb{Z}}, \mu)$ to $(B^{\mathbb{Z}}, \eta)$ in the equivariant version of the category Krn of Chapter 7 are the joinings, composition is the relatively independent joining over the middle factor, and the dagger reverses a joining. The reverse of ν with respect to μ is the ***backward channel*** $\nu^\dagger \colon B^{\mathbb{Z}} \to A^{\mathbb{Z}}$, which gives the conditional law of the input given the output.

The rate is symmetric, $R(\mu,\nu) = R(\eta,\nu^\dagger)$. The two conditional entropy rates are the noise $H(\omega) - H(\mu)$ of the channel and the ***equivocation*** $H(\omega) - H(\eta) = H(\mu) - R(\mu,\nu)$, which is the noise of the backward channel. Shannon's second theorem, in Khinchin's formulation, is a statement about the equivocation.

Notes

Khinchin's definitions and the two theorems on the compound source are [Khi56, Ch. III]. The insufficiency of his finite-memory condition and its repair are discussed in [Tak58, Fei59, Bre60]; ergodic properties of channels with infinite memory are treated by Adler [Adl61], and the modern theory is in Gray's book [Gra11], where the compound source is called the hookup. The equality of ergodic and stationary capacity is due to Parthasarathy [Par61] and Breiman [Bre60]; see also [Ned57, Jac59]. For joinings see [Gla03].

Exercises

Exercise 9.1. Derive the capacities of the erasure channel and the Z-channel in Example 9.12.

Exercise 9.2. Show that the composite of binary symmetric channels with flip probabilities δ_1, δ_2 is binary symmetric with flip probability $\delta_1(1-\delta_2) + \delta_2(1-\delta_1)$. Compute the capacity for $\delta_1 = \delta_2 = 0.1$ and compare with Proposition 9.13(ii).

Exercise 9.3. For the channel of Example 9.5 and any stationary source, compute $H(\eta)$, $H(\omega)$ and $R(\mu, \nu)$.

Exercise 9.4. Prove Proposition 9.13(iii) in full, and show that $C(W \oplus W') = \log\bigl(2^{C(W)} + 2^{C(W')}\bigr)$ for the channel $W \oplus W'$ whose input and output alphabets are the disjoint unions.

Exercise 9.5. Show that the deterministic natural families $A^n \to B^n$ are the causal functions $y_n = f_n(x_1, \ldots, x_n)$, and that the kernel $y_n = x_{n+1}$ on one-sided sequences is not induced by a natural family.

Chapter 10

Feinstein's Lemma

Shannon's proof of his theorem for noisy channels chose codewords at random and averaged. Feinstein gave the first proof that was accepted as rigorous, by a different route: he showed directly that a channel admits many input words whose output distributions are almost disjointly supported. Khinchin calls this Feinstein's fundamental lemma, proves it for stationary channels with finite memory, and derives Shannon's theorems from it. This chapter proves the lemma in its modern one-shot form, which takes a page, recovers Khinchin's version from it with McMillan's theorem, and then reads the statement as one about retractions.

10.1 Codes

Definition 10.1. Let $W\colon X \to Y$ be a morphism of FinStoch and $0 \le \varepsilon < 1$. An (M, ε)***-code*** for W consists of codewords $x_1, \dots, x_M \in X$ and pairwise disjoint decoding sets $D_1, \dots, D_M \subseteq Y$ with

$$W(D_j \mid x_j) \ge 1 - \varepsilon \qquad \text{for all } j.$$

$M^*(W, \varepsilon)$ is the largest M for which one exists.

Khinchin calls the codewords a *distinguishable group* of sequences. Note that the requirement is on each codeword separately; this is the maximal-error criterion, and it does not refer to any distribution on messages.

Fix a distribution P on X and let $Q = W \circ P$ be the output distribution. The ***information density*** is

$$i(x; y) = \log \frac{W(y \mid x)}{Q(y)},$$

a random variable on $X \times Y$ under the joint law $P \rtimes W$, with mean $I(P, W)$.

10.2 The one-shot lemma

Theorem 10.2 (Feinstein). [classical; [Fei54]; in this form see [VH94, PW25]] *For every distribution P on X, every integer $M \ge 1$ and every $\gamma > 0$, there is an (M, ε)-code for W with*

$$\varepsilon = \mathbb{P}\big[i(X; Y) \le \log M + \gamma\big] + 2^{-\gamma},$$

provided this number is less than 1.

Proof. For $x \in X$ let $E_x = \{y : i(x;y) > \log M + \gamma\}$, the set where $W(y \mid x) > M 2^{\gamma} Q(y)$. Summing this inequality over E_x gives

$$Q(E_x) \le \frac{2^{-\gamma}}{M}. \tag{10.1}$$

Choose codewords one at a time. Given $x_1, \dots, x_k$ and $D_1, \dots, D_k$, pick any x_{k+1} with

$$W\big(E_{x_{k+1}} \setminus (D_1 \cup \cdots \cup D_k) \,\big|\, x_{k+1}\big) \ge 1 - \varepsilon$$

and let $D_{k+1} = E_{x_{k+1}} \setminus (D_1 \cup \cdots \cup D_k)$. Each new set is nonempty and disjoint from the earlier ones, so the procedure stops, say after N steps, with an (N, ε)-code. Let D be the union of its decoding sets.

At that point every $x \in X$ satisfies $W(E_x \setminus D \mid x) < 1 - \varepsilon$. Averaging over P,

$$\mathbb{P}\big[i(X;Y) > \log M + \gamma,\ Y \notin D\big] < 1 - \varepsilon,$$

and therefore, by (10.1) applied to each codeword,

$$\mathbb{P}\big[i(X;Y) > \log M + \gamma\big] < 1 - \varepsilon + Q(D) \le 1 - \varepsilon + \frac{N}{M}\, 2^{-\gamma}.$$

Rearranged, $\varepsilon < \mathbb{P}[i \le \log M + \gamma] + \frac{N}{M} 2^{-\gamma}$. Comparing with the definition of ε gives $N > M$. Keep the first M codewords. □

The proof is a greedy construction and one counting estimate. The estimate (10.1) says that a decoding set on which the likelihood ratio is large has small probability under the output distribution, so that many of them fit. The construction uses nothing about the structure of X, Y or W.

10.3 Memoryless channels

Theorem 10.3. [classical; [Fei54]] *Let W be a stochastic matrix with capacity C. For all $\delta, \varepsilon > 0$ and all large n there is a $(2^{\lceil n(C-\delta)\rceil}, \varepsilon)$-code for $W^{\otimes n}$.*

Proof. Let p attain the capacity and take $P = p^{\otimes n}$. The information density of n uses is a sum of n independent copies of the one-letter density, with mean C. By the weak law of large numbers $\mathbb{P}[\frac{1}{n} i \le C - \frac{\delta}{2}] \to 0$. Apply Theorem 10.2 with $\log M = \lceil n(C - \delta) \rceil$ and $\gamma = n\delta/2 - 1$. □

Example 10.4. For the binary symmetric channel with flip probability 0.11 the capacity is 0.500. With uniform inputs, rate 0.25 and $\gamma = n/8$, the error guaranteed by Theorem 10.2 is 0.032 at $n = 200$, 3×10^{-5} at $n = 1000$, and 2×10^{-19} at $n = 5000$. At short lengths the bound is weak (Exercise 10.2).

10.4 Channels with memory

Let ν be a stationary channel without anticipation and of finite memory m. Then the law of the output block $y_1 \dots y_n$ depends only on the input letters $x_{1-m}, \dots, x_n$, and this defines a morphism

$$W_n \colon A^{n+m} \longrightarrow B^n$$

of FinStoch. The first m input letters serve to put the channel into a known condition.

Theorem 10.5 (Feinstein's fundamental lemma)**.** [classical; [Khi56], Ch. IV; with the hypotheses of Theorem 9.7] *Let ν be stationary, without anticipation, of finite memory m and of finite output memory, with ergodic capacity C. For all $\delta, \varepsilon > 0$ and all large n there is a $(2^{\lceil n(C-\delta)\rceil}, \varepsilon)$-code for W_n.*

Proof. Choose an ergodic source μ with $R(\mu,\nu) > C - \delta/4$ and let P be its block law on A^{n+m}. The information density of W_n is

$$i_n = -\log \mu(x_{1-m} \dots x_n) - \log \eta(y_1 \dots y_n) + \log \omega(x_{1-m} \dots x_n,\, y_1 \dots y_n).$$

The sources μ, η and ω are ergodic by Theorem 9.7, so by Theorem 8.2 the three terms divided by n converge in probability to $H(\mu)$, $H(\eta)$ and $-H(\omega)$; the m extra letters do not affect the limits. Hence $\frac{1}{n} i_n \to R(\mu,\nu)$ in probability, and Theorem 10.2 applies as in the memoryless case. □

This is the one place in the whole development where the ergodicity of the compound source is needed, and it is the reason the gap discussed in Chapter 9 matters. Khinchin's own proof does not pass through a one-shot statement. It builds the distinguishable sequences directly inside the typical sets supplied by McMillan's theorem, using the inequalities of Lemmas 4.7 and 4.10, and takes ten pages. The idea is the same.

The proof shows what is really required of the channel and the source: that the normalized information density converge in probability. A pair with this property is called ***information stable***, after Dobrushin [Dob59]. For an arbitrary sequence of channels W_n the capacity is the supremum, over input processes, of the largest number that the normalized information density exceeds with probability tending to one [VH94]; stationarity, ergodicity and memory conditions are sufficient conditions for this number to be a limit of mutual informations.

10.5 Codes as retractions

A code is a pair of maps $e\colon [M] \to X$ and $d\colon Y \to [M]$: e lists the codewords, and d sends D_j to j. The condition of Definition 10.1 says that every diagonal entry of the stochastic matrix

$$[M] \xrightarrow{e} X \xrightarrow{W} Y \xrightarrow{d} [M]$$

is at least $1-\varepsilon$. For $\varepsilon = 0$ this says $d \circ W \circ e = \mathrm{id}$: the identity of $[M]$ is a retract of W. In general the composite is within ε of the identity, in the total variation distance of Theorem 4.5, uniformly over inputs.

Definition 10.6. The *ε-rank* of W is $M^*(W,\varepsilon)$, the size of the largest finite set whose identity is an ε-retract of W.

Theorem 10.7 (Capacity is the asymptotic ε-rank)**.** [classical; [Fei54, Wol57]] *For a stochastic matrix W and every $\varepsilon \in (0,1)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log M^*(W^{\otimes n},\varepsilon) = C(W).$$

Proof. The lower bound is Theorem 10.3. For the upper bound in the limit $\varepsilon \to 0$, let the message J be uniform on $[M]$, sent as $e(J)$ and decoded as $\hat{J}$. The probability of error is at most ε, so by Fano's inequality (Lemma 4.10) and data processing,

$$\log M = I(J;\hat{J}) + H(J \mid \hat{J}) \le I(X^n;Y^n) + 1 + \varepsilon \log M \le nC + 1 + \varepsilon \log M.$$

Thus $\frac{1}{n}\log M^*(W^{\otimes n},\varepsilon) \le (C + \frac{1}{n})/(1-\varepsilon)$. That the limit is C for every fixed $\varepsilon < 1$, and not merely as $\varepsilon \to 0$, is Wolfowitz's strong converse. □

Set this beside Proposition 8.13. For an ergodic source μ and a memoryless channel W,

$$H(\mu) = \lim_n \tfrac{1}{n} \log \min\{M : \ (A^n, \mu_n) \text{ is an } \varepsilon\text{-retract of } [M]\},$$
$$C(W) = \lim_n \tfrac{1}{n} \log \max\{M : \ [M] \text{ is an } \varepsilon\text{-retract of } W^{\otimes n}\},$$

and neither limit depends on $\varepsilon \in (0,1)$. Entropy is how large a set is needed to hold the source; capacity is how large a set the channel can hold. Both are defined by the structure of an enriched monoidal category alone: tensor powers, composition, and the distance to an identity.

For a linear map the largest identity that is a retract of it and the smallest identity through which it factors have the same dimension, the rank. For stochastic matrices the two exact notions are different and neither is well behaved under tensor product; the first is the zero-error capacity of Chapter 12. The tolerance ε, together with the passage to the limit, is what produces a single additive invariant.

10.6 What the lemma is

It should be said plainly what the categorical language has and has not done here. It has given the statement a form, a retraction up to ε, in which it composes with other statements of the same form; the next chapter uses this. It has not shortened the proof of Theorem 10.2, which is a counting argument about sets, and it does not construct the code, which is found by exhaustive search. Feinstein's lemma and McMillan's theorem are the two places where the theory touches the particular, and only the second has so far been dissolved into general statements.

Notes

Feinstein's lemma is [Fei54]; his book [Fei58] and Khinchin's Chapter IV [Khi56] extend it to channels with memory. The one-shot form with the information density is from the information-spectrum literature [VH94, Han03]; see [PW25] for a textbook treatment and for comparison with the random-coding bounds. The strong converse is due to Wolfowitz [Wol57, Wol78]. Information stability was introduced by Dobrushin [Dob59].

Exercises

Exercise 10.1. Prove (10.1).

Exercise 10.2. For the binary symmetric channel with flip probability δ and uniform inputs on $\{0,1\}^n$, show that $i(x;y) = n + d\log\delta + (n-d)\log(1-\delta)$ with d the Hamming distance, so that the sets E_x are Hamming balls. For $n = 7$, $\delta = 0.01$ and $M = 16$, find the best error guaranteed by Theorem 10.2 and compare it with the error probability of the $(7,4)$ Hamming code.

Exercise 10.3. Compute $M^*(W,\varepsilon)$ for the identity channel on a set of k elements and for a channel that ignores its input, for all $\varepsilon \in [0,1)$.

Exercise 10.4. Show that allowing a stochastic encoder $e\colon [M] \to X$ in Definition 10.1 does not increase $M^*(W,\varepsilon)$. Give an example in which a stochastic decoder achieves a smaller maximal error than every deterministic decoder for the same codewords.

Exercise 10.5. Let W be the binary erasure channel with erasure probability e. Show that a set of M distinct codewords of length n, decoded by returning the unique codeword consistent with the unerased positions when there is one, has maximal error at most ε provided every codeword is, with probability $1-\varepsilon$, distinguished from all others by the unerased positions. Deduce that linear codes of rate below $1-e$ exist with small error, by counting.

Chapter 11

Shannon's Theorems

Everything is now in place. McMillan's theorem makes a long block from an ergodic source an approximate retract of a set of 2^{nH} points. Feinstein's lemma makes a set of 2^{nC} points an approximate retract of a long block of channel uses. If $H < C$ the first set fits inside the second, and the composite of the two retractions transmits the source through the channel. This is Shannon's theorem as Khinchin proves it in his last chapter. We present it in that form, add the converse, which Khinchin does not treat, and observe that the two halves of the theorem are of different kinds. The chapter ends with polar codes, where the martingale of Chapter 7 returns in a new role.

11.1 Coding without noise

A ***prefix code*** for a finite set X is an injective map c from X to finite binary strings such that no codeword is an initial segment of another. Write $\ell(x)$ for the length of $c(x)$.

Theorem 11.1 (Kraft's inequality and the noiseless coding theorem). [classical; [Kra49, Sha48]] *The lengths of a prefix code satisfy $\sum_x 2^{-\ell(x)} \le 1$, and every assignment of lengths satisfying this inequality is realized by a prefix code. For a distribution p on X the least expected length $L^*(p)$ of a prefix code satisfies*

$$H(p) \le L^*(p) < H(p) + 1.$$

Proof. A codeword of length ℓ determines a cylinder of measure $2^{-\ell}$ in the space of infinite binary sequences with the fair coin measure, and the prefix condition says that the cylinders are disjoint. For the lower bound let $K = \sum_x 2^{-\ell(x)} \le 1$ and $q(x) = 2^{-\ell(x)}/K$; then $\sum_x p(x)\ell(x) - H(p) = D(p \,\|\, q) - \log K \ge 0$ by Proposition 4.2. For the upper bound take $\ell(x) = \lceil -\log p(x) \rceil$. □

Applied to blocks of length n from a stationary source, the theorem gives codes whose expected length per letter lies between H_n/n and $H_n/n + 1/n$, and both tend to $H(\mu)$. So for codes of variable length judged by their expected length the entropy of the source is the optimal rate, for every stationary source. Ergodicity is needed for the stronger statement that the length is close to nH

with high probability, which is the form in which Khinchin states the result (his Theorem 4) and the form needed for transmission over a channel.

Example 11.2. For $p = (0.7, 0.2, 0.1)$, $H(p) \approx 1.157$. The optimal prefix code for single letters has expected length 1.3. For blocks of 2, 3, 4 and 6 letters the optimal expected lengths per letter are 1.165, 1.175, 1.164 and 1.162. The convergence to H is not monotone.

The proof identifies an optimal code with an embedding of the source into the fair-coin source: the encoded stream is, as nearly as integer lengths allow, a sequence of independent fair bits. In Chapter 12 this becomes a theorem of ergodic theory.

11.2 Codes and channels

Khinchin defines a code as a map $\theta \mapsto x(\theta)$ from sequences over a source alphabet A_0 to sequences over the channel alphabet A, and remarks that it can be regarded as a channel $[A_0, \rho, A]$ without noise. Connecting the source to the code and the code to the channel is the composite

$$1 \xrightarrow{\mu} A_0^{\mathbb{Z}} \xrightarrow{\rho} A^{\mathbb{Z}} \xrightarrow{\nu} B^{\mathbb{Z}}$$

of Chapter 9. The codes he uses are block codes, which map each block of n source letters to a block of $n + m$ channel letters, where m is the memory of the channel. Such a map commutes with the appropriate powers of the two shifts, not with the shifts themselves, and the sources it produces are stationary only for the block shift. This is harmless; one either works with the block shift throughout, as Khinchin does, or averages over the phase.

11.3 The first theorem

Lemma 11.3 (Retractions compose). [observation] *Let (e_s, d_s) be an (M, ε_1)-compression of a finite probability space (X, p) in the sense of Definition 8.12, and (e_c, d_c) an (M, ε_2)-code for $W\colon U \to V$ in the sense of Definition 10.1. Then for x drawn from p, sent as $e_c(e_s(x))$ through W and decoded by $d_s \circ d_c$, the result equals x with probability at least $1 - \varepsilon_1 - \varepsilon_2$.*

Proof. With probability at least $1 - \varepsilon_1$ the point x satisfies $d_s(e_s(x)) = x$. Whatever the message $j = e_s(x)$ is, the channel code returns it with probability at least $1 - \varepsilon_2$. □

The second step is where the maximal-error criterion of Definition 10.1 is used. The compression may concentrate the source on a few messages, and the

channel code must work for each of them. A code that is good only on average over messages would not compose with an arbitrary source (Exercise 11.3). This is why Feinstein's formulation suits Khinchin's proof.

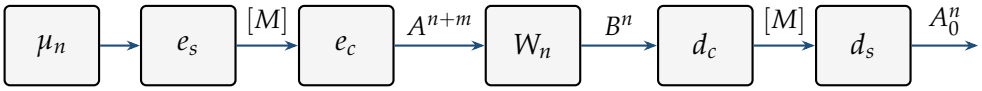


Figure 11.1: Shannon's first theorem as a composite of two approximate retractions. The middle three boxes are within ε of the identity of $[M]$ by Feinstein's lemma; the outer two are within ε of the identity on μ_n-typical blocks by McMillan's theorem.

Theorem 11.4 (Shannon's first theorem). [classical; [Sha48, Fei54, Khi56]] *Let ν be a channel as in Theorem 10.5, with ergodic capacity C, and let $[A_0, \mu]$ be an ergodic source with entropy $H_0 < C$. For every $\varepsilon > 0$ and all large n there are maps $A_0^n \to A^{n+m}$ and $B^n \to A_0^n$ such that a block of n letters from the source, encoded, transmitted and decoded, is recovered with probability at least $1 - \varepsilon$.*

Proof. Choose $\delta > 0$ with $H_0 + \delta < C - \delta$ and put $M = 2^{\lceil n(H_0+\delta)\rceil}$. By Proposition 8.13 the block law μ_n has an $(M, \varepsilon/2)$-compression for large n. By Theorem 10.5 the block channel W_n has a $(2^{\lceil n(C-\delta)\rceil}, \varepsilon/2)$-code for large n, and $2^{\lceil n(C-\delta)\rceil} \geq M$. Apply Lemma 11.3. □

11.4 The second theorem

Khinchin's second theorem says that the code can be chosen so that the rate of transmission between the source and the received signal is as close as desired to the entropy H_0 of the source, or equivalently so that the equivocation per letter is small.

Theorem 11.5 (Shannon's second theorem). [classical; [Khi56], §16] *Under the hypotheses of Theorem 11.4, for every $\varepsilon' > 0$ there is a code for which the equivocation of the source block given the received block is less than $n\varepsilon'$.*

Proof. Take the code of Theorem 11.4. The received block determines the decoded block, which differs from the source block with probability at most ε. By Fano's inequality (Lemma 4.10) the equivocation is at most $1 + \varepsilon\, n \log |A_0|$. □

In the language of Chapter 4: a morphism within ε of having a left inverse has equivocation at most $h(\varepsilon) + \varepsilon \log(\text{size})$. The second theorem is the first theorem passed through Fano's inequality, and Khinchin proves it that way.

11.5 The converse, and the two kinds of statement

Theorem 11.6 (Converse). [classical; [Fan61]] *Let W be a memoryless channel with capacity C and μ a stationary source with entropy $H_0 > C$. For every n and every pair of maps $A_0^n \to A^n$, $B^n \to A_0^n$ the probability P_e that a block is not recovered satisfies*

$$P_e \;\geq\; \frac{H_0 - C - 1/n}{\log |A_0|}.$$

Proof. Let θ be the source block and $\hat{\theta}$ the decoded block. Then $nH_0 \leq H(\theta) = I(\theta;\hat{\theta}) + H(\theta \mid \hat{\theta}) \leq nC + 1 + P_e\, n \log |A_0|$, by data processing, Theorem 9.11 and Fano's inequality. □

Compare the ingredients. The converse uses the chain rule, the data-processing inequality, the additivity of capacity over independent uses, and Fano's inequality. Each of these is a property of a function on morphisms: functoriality, monotonicity under composition, additivity under tensor, continuity at identities. The direct theorem uses McMillan's theorem and Feinstein's lemma, which assert that certain morphisms exist. The first kind of statement is what category theory handles well, and the converse can be stated for any function with those properties.

Theorem 11.7 (A universal property of capacity). [observation] *Let Φ assign a number in $[0,\infty)$ to each morphism of* FinStoch *so that*

(i) $\Phi(g \circ f) \leq \min\{\Phi(f), \Phi(g)\}$;

(ii) $\Phi(f \otimes g) \leq \Phi(f) + \Phi(g)$;

(iii) if $f\colon [M] \to [M]$ has all diagonal entries at least $1-\varepsilon$, then $\Phi(f) \geq (1-\varepsilon)\log M - 1$.

Then $\frac{1}{n}\log M^*(W^{\otimes n}, \varepsilon) \leq (\Phi(W) + \frac{1}{n})/(1-\varepsilon)$ for all W, n and ε. Capacity satisfies (i)–(iii)*, and it is the least function that does: $\Phi \geq C$ for every such Φ.*

Proof. If (e,d) is an (M,ε)-code for $W^{\otimes n}$ then by (iii), (i) and (ii), $(1-\varepsilon)\log M - 1 \leq \Phi(d \circ W^{\otimes n} \circ e) \leq \Phi(W^{\otimes n}) \leq n\Phi(W)$. Capacity satisfies (i) and (ii) by Proposition 9.13, and (iii) by Fano's inequality with a uniform input. Finally, by Theorem 10.7 the left side of the displayed bound tends to $C(W)$, so $C(W) \leq \Phi(W)/(1-\varepsilon)$ for every ε. □

The first half of the theorem, the bound, is the converse and is formal. The second half, minimality, is the coding theorem: it says that no obstruction of this general kind is sharper than mutual information, and it rests on Feinstein's lemma. Stated this way Shannon's theorem is the assertion that a certain extremal problem over monotone functions has the explicit solution $\max_p I(p, W)$.

11.6 Separation

In Figure 11.1 the composite factors through the finite set $[M]$. The source is first turned into one of M messages without reference to the channel, and the channel is used to send one of M messages without reference to the source. Theorems 11.4 and 11.6 together say that for an ergodic source and a memoryless channel nothing is lost, in the limit, by insisting on this factorization. That is the source–channel separation theorem. It is a statement about limits and about stable systems. For non-ergodic sources or channels, where the information density does not concentrate, separation can fail [VVS95], and in networks with several senders and receivers it fails in general.

11.7 Polarization: the martingale returns

Feinstein's lemma finds a code by search. For a large class of channels there is an explicit construction, due to Arıkan, whose analysis is a martingale argument.

Let $W\colon \{0,1\} \to Y$ be a channel with binary input and write $I(W)$ for its mutual information with uniform input. Take two independent uses of W and feed them $x_1 = u_1 \oplus u_2$ and $x_2 = u_2$, where u_1, u_2 are independent uniform bits. This defines two new binary-input channels:

$$W^-\colon u_1 \mapsto (y_1, y_2), \qquad W^+\colon u_2 \mapsto (y_1, y_2, u_1).$$

The first sees u_1 through noise and through the unknown u_2; the second sees u_2 twice, given u_1.

Proposition 11.8. [literature; [Ari09]] *$I(W^-) + I(W^+) = 2I(W)$ and $I(W^-) \le I(W) \le I(W^+)$, with equality iff $I(W) \in \{0,1\}$.*

Proof of the first statement. The map $(u_1, u_2) \mapsto (x_1, x_2)$ is a bijection preserving the uniform distribution, so $I(u_1u_2; y_1y_2) = 2I(W)$. By the chain rule this is $I(u_1; y_1y_2) + I(u_2; y_1y_2 \mid u_1)$, and the second term is $I(u_2; y_1y_2u_1)$ because u_1 and u_2 are independent. □

Iterating the construction n times produces 2^n channels W^b, indexed by sign sequences $b \in \{-,+\}^n$, from 2^n independent uses of W. Choose the signs by fair coin tosses and let $I_n = I(W^{b_1 \dots b_n})$. By Proposition 11.8, (I_n) is a martingale with values in $[0,1]$.

Theorem 11.9 (Arıkan). [literature; [Ari09]] *For every $\delta > 0$, as $n \to \infty$ the fraction of the 2^n channels W^b with $I(W^b) > 1 - \delta$ tends to $I(W)$, and the fraction with $I(W^b) < \delta$ tends to $1 - I(W)$.*

Proof. The increments of a martingale are orthogonal in L^2, so $\sum_{k<n} \mathbb{E}(I_{k+1} - I_k)^2 = \mathbb{E}(I_n - I_0)^2 \le 1$ and $\mathbb{E}(I_{k+1} - I_k)^2 \to 0$. On the other hand there is a function $\kappa(\delta) > 0$, not depending on the output alphabet, such that $I(W) - I(W^-) \ge \kappa(\delta)$ whenever $\delta \le I(W) \le 1 - \delta$; this follows from Mrs. Gerber's lemma [WZ73, SSZ05], and Arıkan argues instead with the Bhattacharyya parameter. By Proposition 11.8 both possible increments have this absolute value, so the probability that $I_k \in [\delta, 1-\delta]$ tends to zero. Since $\mathbb{E} I_n = I(W)$ and the middle range is asymptotically empty for every δ, the probability that $I_n > 1 - \delta$ tends to $I(W)$. □

A code is obtained by sending data bits on the indices b with $I(W^b)$ near 1, fixing the others, and decoding the bits in order; the block error probability tends to zero at any rate below $I(W)$ [Ari09, AT09]. For symmetric channels $I(W)$ is the capacity.

Example 11.10. For the erasure channel with erasure probability e both new channels are erasure channels, with probabilities $2e - e^2$ and e^2. Starting from $e = \frac{1}{2}$, the fraction of the 2^n channels with erasure probability below 0.01 is 0.234, 0.373, 0.441 and 0.473 for $n = 6, 10, 14, 18$. It tends to the capacity $\frac{1}{2}$. The mean erasure probability is $\frac{1}{2}$ at every stage.

Three things in this construction belong to the themes of this book. The transform is a string diagram built from copying, a deterministic isomorphism, and $W \otimes W$; its n-fold iterate is a tensor power of one small diagram. The conservation law is the chain rule along an isomorphism, that is, functoriality. And the proof of polarization needs only the orthogonality of the increments of a martingale, which is the statement that the projections of Chapter 7 are orthogonal projections. The result is an explicit ε-retraction of $W^{\otimes N}$ onto a set of about $2^{NI(W)}$ points. Khinchin brought the martingale into information theory to handle the source. Half a century later it turned out to construct the code.

Notes

Shannon's theorems are in [Sha48]. Khinchin's proofs are [Khi56, Ch. V]; his first paper treats noiseless coding [Khi53, §5]. Kraft's inequality is [Kra49], extended to uniquely decodable codes by McMillan [McM56]; optimal codes are due to Huffman [Huf52]. Fano's inequality and the converse are in [Fan61]; see [CT06, CK11, PW25]. The failure of separation for non-ergodic systems is studied in [VVS95]. Polar codes are due to Arıkan [Ari09]; the rate of polarization is in [AT09].

Exercises

Exercise 11.1. Find an optimal prefix code for $p = (0.4, 0.3, 0.2, 0.1)$ and compare its expected length with $H(p)$.

Exercise 11.2. Show that $L^*(p) = H(p)$ iff every $p(x)$ is a power of $\frac{1}{2}$.

Exercise 11.3. Give an example of a channel code with small error probability on average over uniformly chosen messages, and a compression of a source, whose composite fails with probability close to 1. How can the code be repaired?

Exercise 11.4. Let $r_+(W)$ be the least size of a set through which W factors in FinStoch. Show that $\Phi = \log r_+$ satisfies the hypotheses of Theorem 11.7, state the bound it gives, and give a channel with $\Phi(W) > C(W)$. Does $\Phi(W) = \log \min\{|X|, |Y|\}$ satisfy the hypotheses?

Exercise 11.5. For the erasure channel prove that W^- and W^+ are erasure channels with probabilities $2e - e^2$ and e^2. Let $f(e) = \sqrt{e(1-e)}$. Show that $\frac{1}{2}\big(f(2e-e^2) + f(e^2)\big) \leq \frac{\sqrt{3}}{2} f(e)$, and deduce Theorem 11.9 for erasure channels without any martingale theorem.

Chapter 12

Rates and Spectra

Entropy and capacity were obtained in the last chapters as limits: growth rates of the sizes of sets that are approximate retracts of, or admit approximate retractions onto, tensor powers. Two features of the definitions did the work, the tolerance ε and the passage to the limit. This chapter asks what each contributes. Without the tolerance, the limits still exist but are governed by a family of invariants and not by one number. Strassen's theory of asymptotic spectra describes such families. In the infinite setting of ergodic theory, on the other hand, the tolerance is not needed at all. None of this is in Khinchin; it is the context in which his theorems now sit.

12.1 Coding without error

Let $W\colon X \to Y$ be a stochastic matrix. Two inputs are ***confusable*** if some output has positive probability under both. The ***confusability graph*** G_W has vertex set X and an edge between confusable inputs. A $(M, 0)$-code is a set of M pairwise non-confusable inputs, so $M^*(W, 0) = \alpha(G_W)$, the independence number, and for n uses $M^*(W^{\otimes n}, 0) = \alpha(G_W^{\boxtimes n})$, where $\boxtimes$ is the strong product of graphs.

Definition 12.1 (Shannon)**.** The ***zero-error capacity*** of W is $C_0(W) = \lim_n \frac{1}{n} \log \alpha(G_W^{\boxtimes n}) = \log \Theta(G_W)$.

Example 12.2. Let $X = Y = \mathbb{Z}/5$ and let W send x to x or $x+1$ with probability $\frac{1}{2}$ each. Its capacity is $\log 5 - 1 \approx 1.322$. Its confusability graph is the pentagon C_5, with $\alpha(C_5) = 2$, so one use carries one bit without error. But $\alpha(C_5 \boxtimes C_5) = 5$: the five words $(i, 2i)$ are pairwise non-confusable. So two uses carry $\log 5$ bits, and $C_0 \geq \frac{1}{2} \log 5 \approx 1.161$.

Theorem 12.3. [classical; [Sha56, Lov79, Hae79, Alo98]] *$\Theta(C_5) = \sqrt{5}$, so the bound in Example 12.2 is the zero-error capacity. In general Θ is not multiplicative under the strong product and not additive under disjoint union: there are graphs with $\Theta(G \boxtimes H) > \Theta(G)\,\Theta(H)$ and with $\Theta(G \sqcup H) > \Theta(G) + \Theta(H)$. No algorithm is known that computes Θ.*

Compare with Proposition 9.13. Capacity is additive under tensor product, satisfies the sum formula of Exercise 9.4, and is the solution of a convex

optimization problem. Zero-error capacity has none of these properties. The exact rank of a stochastic matrix, in the sense of the largest identity that is a retract of it, does not become a homomorphism even after regularization.

12.2 Asymptotic spectra

Strassen found the right framework for regularized quantities of this kind while studying the complexity of matrix multiplication. Let S be a commutative semiring containing the natural numbers, with a preorder $\leq$ compatible with addition and multiplication, restricting to the usual order on $\mathbb{N}$, and such that every element is bounded by a natural number times any nonzero element. Define the ***rank*** $R(a) = \min\{n \in \mathbb{N} : a \leq n\}$ and the ***subrank*** $Q(a) = \max\{n \in \mathbb{N} : n \leq a\}$, and their regularizations $\tilde{R}(a) = \lim_n R(a^n)^{1/n}$ and $\tilde{Q}(a) = \lim_n Q(a^n)^{1/n}$. The ***asymptotic spectrum*** $\mathbf{X}(S)$ is the set of monotone semiring homomorphisms $S \to \mathbb{R}_{\geq 0}$.

Theorem 12.4 (Strassen). [literature; [Str88]] *$\mathbf{X}(S)$ is a compact Hausdorff space, and for every $a \in S$,*

$$\tilde{Q}(a) = \min_{\phi \in \mathbf{X}(S)} \phi(a), \qquad \tilde{R}(a) = \max_{\phi \in \mathbf{X}(S)} \phi(a).$$

Moreover $a^n \leq b^{n+o(n)}$ for large n iff $\phi(a) \leq \phi(b)$ for all $\phi \in \mathbf{X}(S)$.

Theorem 12.5 (Zuiddam). [literature; [Zui19]] *Finite graphs, with disjoint union as sum, strong product as product, and the preorder in which $G \leq H$ when there is a homomorphism from the complement of G to the complement of H, form such a semiring. Its subrank is the independence number, and therefore*

$$\Theta(G) = \min_{\phi \in \mathbf{X}(\text{graphs})} \phi(G).$$

The Lovász theta function, the fractional clique cover number and the fractional Haemers bounds are points of the spectrum.

So zero-error capacity is the minimum of a family of additive, multiplicative, monotone invariants, each of which is individually as well behaved as capacity. The family is a compact space of which only some points are known. This is the structure that the tolerance ε removes.

12.3 Sources: the Rényi family

The same phenomenon appears on the source side, in a form that can be computed completely. Recall the Rényi entropies H_α of Chapter 2, with $H_0(p) = \log|\operatorname{supp} p|$ and $H_1 = H$.

Theorem 12.6 (Exact, exponential and constant error). [classical; [CK11]] *Let p be a distribution on a finite set and consider compressions of $p^{\otimes n}$ in the sense of Definition 8.12.*

(i) With no error, $M_*(p^{\otimes n}, 0) = |\operatorname{supp} p|^n$: the rate is $H_0(p)$.

(ii) With error at most a fixed $\varepsilon \in (0,1)$, the rate is $H_1(p) = H(p)$.

(iii) With error required to be at most 2^{-nr}, a rate R strictly between $H(p)$ and $H_0(p)$ is achievable iff

$$r \le \sup_{\rho \ge 0} \rho \left(R - H_{1/(1+\rho)}(p) \right).$$

As the required exponent r goes from 0 to ∞ the relevant order $\alpha = 1/(1+\rho)$ goes from 1 to 0. Each H_α with $0 \le \alpha \le 1$ is additive under tensor product and monotone under mixing (Exercise 12.2), so each gives an obstruction of the kind in Theorem 11.7; which of them is binding depends on how small the error must be. For transformations between distributions under majorization, which is the same as the theory of bipartite pure quantum states under local operations, Jensen and Vrana showed that the asymptotic spectrum is exactly the family of power sums $p \mapsto \sum_i p_i^\alpha$ with $0 \le \alpha \le 1$, and expressed optimal rates as infima over it [JV20]; see also [Fri17, Kli07, Tur07].

The role of the tolerance can now be stated precisely.

Proposition 12.7 (Collapse). [observation] *Let μ be an ergodic source and $\varepsilon > 0$. For all large n there is a set $S \subseteq A^n$ with $\mu_n(S) \ge 1 - \varepsilon$ such that the conditioned law $\mu_n|_S$ satisfies*

$$\left| \frac{1}{n} H_\alpha(\mu_n|_S) - H(\mu) \right| \le 4\varepsilon \qquad \text{for every } \alpha \in [0, \infty].$$

Proof. Take S as in Proposition 8.14. The conditioned law has all its probabilities between $2^{-a}/|S|$ and $2^a/|S|$ with $a = n(2\varepsilon + \frac{2}{n})$. Then $\log|S| - a \le H_\infty \le H_\alpha \le H_0 = \log|S|$, and $\frac{1}{n}\log|S|$ is within $\varepsilon + \frac{1}{n}$ of $H(\mu)$. □

Before conditioning, the Rényi rates of an ergodic source are in general all different. Discarding a set of probability ε makes them equal. McMillan's theorem is the statement that the spectrum of a source, in the sense of this chapter, collapses to a point once sets of small probability may be ignored, and Shannon entropy is the value at that point. The analogue for channels is Theorem 10.7, with the strong converse: the many invariants of Theorem 12.5 are replaced by one.

12.4 Infinite time: entropy as a rank

In ergodic theory the objects are stationary sources, or more generally measure-preserving systems, and the morphisms are stationary codes. Let $B(p)$ be the source of independent letters with distribution p, a ***Bernoulli shift***.

Theorem 12.8. [classical; [Sin64, Orn70, Kri70]]

(i) (Sinai) If μ is ergodic and $H(p) \leq H(\mu)$, then $B(p)$ is a factor of μ.

(ii) (Ornstein) $B(p)$ and $B(q)$ are isomorphic iff $H(p) = H(q)$.

(iii) (Krieger) If μ is ergodic and $H(\mu) < \log k$, then μ is isomorphic to a stationary source on k letters.

For instance $B(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $B(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ both have entropy 2 and are isomorphic, as Meshalkin showed by an explicit code [Mes59].

Read these as statements about rank and subrank in the preorder of factor maps. By (i) and Theorem 5.13, the entropy of an ergodic source is the supremum of the entropies of its Bernoulli factors: a subrank, with Bernoulli shifts in the role of the finite sets $[M]$. By (iii) and Theorem 5.13 again, it is the infimum of $\log k$ over the alphabets on which the source can be represented: a rank. The two agree, with no tolerance and no regularization. Krieger's theorem is the noiseless coding theorem in this setting, and Sinai's theorem says that from any ergodic source one can extract fair bits at the rate of its entropy. The finite theorems of Chapters 8 and 11 are approximations to these, in which the infinite delay of a general stationary code is replaced by blocks of length n and the price is the tolerance.

12.5 Channels without noise

Shannon's 1948 paper begins with a channel that makes no errors but restricts which sequences may be sent. If the allowed sequences are the paths in a finite directed graph with adjacency matrix A, the number of allowed blocks of length n grows like λ^n, where λ is the largest eigenvalue of A, and the capacity is $\log \lambda$.

Theorem 12.9. [classical; [Sha48, Par64]] *For an irreducible graph, $\log \lambda$ is the maximum of $H(\mu)$ over stationary sources supported on allowed sequences, and the maximum is attained by a unique Markov source.*

In dynamics $\log \lambda$ is the topological entropy of the shift of finite type and the theorem is the variational principle for it. Here the exact count and the approximate one agree because the channel is deterministic: with no noise, ε has nothing to do.

Example 12.10. Forbid two consecutive ones. The number of allowed blocks of length n is a Fibonacci number, and the capacity is $\log \frac{1+\sqrt{5}}{2} \approx 0.694$. The maximizing source is the Markov chain that follows a 0 by a 1 with probability $\frac{3-\sqrt{5}}{2} \approx 0.382$ and always follows a 1 by a 0.

12.6 Other entropies

Several notions that carry the name entropy arise from categorical constructions of the same general shape, a growth rate attached to an endomorphism. Dimitrov, Haiden, Katzarkov and Kontsevich define the entropy of an exact endofunctor of a triangulated category as the growth rate of the complexity of the images of a generator under its powers [DHKK14]. Dikranjan and Giordano Bruno derive the topological, measure-theoretic and algebraic entropies from a single functorial scheme [DGB13]. We mention them to mark the boundary of this book: they generalize the entropy of Theorem 5.13, an invariant of a dynamical system, and not the coding theorems.

Notes

Zero-error capacity was introduced by Shannon [Sha56]; the pentagon was settled by Lovász [Lov79]. Asymptotic spectra are due to Strassen [Str88], and their application to graphs to Zuiddam [Zui19]. The resource-theoretic point of view on rates is developed in [CFS16, Fri17]. Error exponents and their expression through Rényi entropies are treated in [CK11]. For the isomorphism theory of Bernoulli shifts see [Orn70, Gla03]; for symbolic dynamics and constrained channels, [LM95].

Exercises

Exercise 12.1. Verify that $\alpha(C_5) = 2$ and that the words $(i, 2i)$, $i \in \mathbb{Z}/5$, are pairwise non-confusable for two uses of the channel of Example 12.2. Find the capacity-achieving input and confirm $C = \log 5 - 1$.

Exercise 12.2. Show that if $q = Dp$ for a doubly stochastic matrix D then $H_\alpha(q) \geq H_\alpha(p)$ for every $\alpha \geq 0$. Deduce that if $q^{\otimes m} = Dp^{\otimes n}$ for a doubly stochastic D then $nH_\alpha(p) \leq mH_\alpha(q)$ for all α.

Exercise 12.3. Prove Theorem 12.6(i). For $p = (0.9, 0.1)$ and $R = 0.7$, compute the exponent in Theorem 12.6(iii) and compare it with $D((1-\tau, \tau) \,\|\, p)$, where $h(\tau) = R$ and $0.1 < \tau < \frac{1}{2}$.

Exercise 12.4. Prove the claims of Example 12.10: count the allowed blocks, and maximize the entropy of a Markov chain on $\{0,1\}$ with $q_{11} = 0$.

Exercise 12.5. Compute the entropies of $B(\frac{1}{2}, \frac{1}{2})$, $B(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and Meshalkin's two shifts. Which of the four is a factor of which?

Part IV

Assessment

Chapter 13

What the Language Buys

Khinchin's conclusion to his second paper says that the proof he has given is long and complicated but rests on a single idea, and he restates the idea in two pages. This chapter attempts the same for the present book, and then says where its method stopped working.

13.1 The idea

There is one category. Its objects are sequence spaces with their shifts; its morphisms are equivariant Markov kernels; sources are its states, channels its morphisms, codes its deterministic morphisms. It is monoidal, its hom-sets carry a divergence for which composition is short, and, taken up to almost sure equality, it has a dagger. Shannon's theory is the study of two numerical invariants of this category. Entropy measures a state by the smallest finite set of which its long blocks are approximate retracts; capacity measures a morphism by the largest finite set that is an approximate retract of its long blocks. The coding theorem says that a state can be pushed through a morphism when the first number is smaller than the second, because retractions compose, and cannot when it is larger, because both numbers are monotone.

That such invariants exist, are finite, and are given by the formulas $-\sum p \log p$ and $\max I$ is the content of the theory. The formula for entropy is forced by the chain rule, which is a cocycle condition. The existence of the limits is McMillan's theorem and Feinstein's lemma.

13.2 What became formal

The following statements, each of which costs Khinchin a computation, came out of general principles.

Statement	Mechanism
$H(AB) = H(A) + H_A(B)$	functoriality of H on FinProb; the cocycle condition
symmetry of entropy	commutativity of the monoid of observables
uniqueness of entropy	$H^1(\Omega) \cong \mathbb{R}$, given the regularity theorem for the fundamental equation
Shannon's inequality and its generalizations	positivity of relative entropy and the chain rule
data processing	composition is short: the enrichment
the compound source is stationary	morphisms of Dyn compose
absence of anticipation	naturality between towers
ergodic decomposition	disintegration over the invariant quotient
convergence of conditional probabilities, in mean	monotone sequences of idempotents converge
McMillan's theorem, in mean	chain rule, two idempotent limits, one tail estimate
the second Shannon theorem	Fano's inequality: near-retractions lose little
the converse	any monotone subadditive normalized function bounds the rate

13.3 What did not

Five things resisted, and they are the mathematical substance of the subject.

The regularity theorem. Theorem 3.5 reduces uniqueness to the statement that measurable solutions of the fundamental equation are multiples of the binary entropy. That statement is analysis, and without a regularity hypothesis it is false. Cohomology organizes the theorem; it does not prove it.

Almost sure convergence. The theorems of Doob, Birkhoff and Breiman need maximal inequalities. The categorical convergence results currently available are in norm. For the coding theorems this is enough, which is itself worth knowing, but the gap is real.

The tail estimate. Lemma 8.4 is three lines and elementary, and the proof of McMillan's theorem cannot do without it or something like it. It is the point at which the logarithm, which is unbounded, meets a convergence theory built for bounded functions.

Feinstein's lemma. The one-shot lemma is a greedy construction and a counting bound. We gave its statement a categorical form and proved nothing

about it by categorical means. The same is true of the strong converse.

Ergodicity of the compound source. Here the categorical formulation did something useful, in that it made visible the difference between an equivariant map into a sequence space and an equivariant kernel, and so located the hypothesis missing from Khinchin's theorem. But the theorem with the hypothesis restored is proved by a mixing computation.

13.4 How much of it is higher

The title of this series promises higher categories, and the reader is owed an account. Chapters 2 and 3 use operads and the cohomology of a monoid with coefficients in a module over a presheaf of algebras; that is homological algebra in a topos, and the deformation argument of Proposition 3.8 is of the kind for which derived language exists. Everything from Chapter 4 on uses monoidal categories with one layer of extra structure on the hom-sets: a divergence, a topology, or an order. These are enriched categories and so are two-dimensional in a weak sense, but no coherence beyond dimension one was used, and no homotopy type appeared that was not a set. The approximate retractions of Chapter 10 suggest a setting in which morphisms are considered up to a controlled error and composition holds up to a larger one; we did not build it.

13.5 Khinchin, seventy years on

His definitions stand. A source is a stationary measure on a sequence space, a channel is a stationary kernel, and the theorems are about these. Three points have changed. His finite-memory condition does not imply what he uses it for; an independence condition on the noise must be added (Chapter 9). His ergodic capacity is the same as the capacity defined with all stationary sources, which he left open (Theorem 9.10). And he proves no converse, so that his capacity is shown to be achievable and not shown to be a limit; the converse is easier than anything he does prove (Theorem 11.6). His use of almost sure theorems where convergence in mean suffices is a matter of economy, not of correctness.

13.6 Open problems

Open Problem 13.1 (A synthetic McMillan theorem)**.** [programme] Formulate and prove the asymptotic equipartition property inside a Markov category with Kolmogorov products, a dagger on its category of probability spaces, and enrichments in topological spaces and in divergence spaces, from axioms

satisfied by measure-theoretic probability. The proof in Chapter 8 indicates the ingredients. What is missing is a derivation of the mean ergodic theorem in the framework of [PVB24], where the Cesàro averages are not a monotone net, and an account of how the two enrichments interact, which is where Lemma 8.4 would have to live.

Open Problem 13.2 (Maximal inequalities)**.** [programme] Find a categorical formulation of almost sure convergence in which the theorems of Doob and Birkhoff, and with them the Shannon–McMillan–Breiman theorem, have proofs comparable in length to their counterparts in mean.

Open Problem 13.3 (Higher information cohomology)**.** [programme] Compute $H^n(\Omega)$ for $n \geq 2$ for the full structure on a finite set. Decide whether the non-Shannon inequalities of Theorem 4.12, which no first-order structure explains, are reflected in it.

Open Problem 13.4 (Entropy over other bases)**.** [programme] Entropy over $\mathbb{R}$ and over $\mathbb{F}_p$ are both the defect of additivity of a derivation (Chapter 3), and both are unique solutions of the chain rule. Give a single construction, for a suitable class of rings with a derivation-like operator, of which both are instances, and identify the cohomology theory in which the resulting class lives. The appearance of the same functional equation in the theory of polylogarithms [Cat88, Kon95, EVG02] suggests where to look.

Open Problem 13.5 (Cohomology of processes)**.** [programme] Extend information cohomology to the observables of a stationary source, with the shift acting. Is the entropy rate a class in an equivariant theory, and is its affinity (Theorem 6.7) a vanishing statement?

Open Problem 13.6 (The collapse of the spectrum)**.** [programme] Define a preordered semiring of channels in which the preorder is simulation up to a tolerance, and prove that its asymptotic spectrum, in the sense of Theorem 12.4 or of a suitable extension, is the single point $W \mapsto 2^{C(W)}$. Theorem 11.7 and Proposition 12.7 are evidence that such a statement holds. The difficulty is that a preorder defined with a tolerance is not transitive without loss.

Open Problem 13.7 (Existence)**.** [programme] Find a structural proof of Feinstein's lemma: a general principle, valid in a class of enriched Markov categories, that produces approximate retractions from the concentration of an information density. The polar construction of Chapter 11 is structural but special. Is there a version of it for channels with memory in which the martingale is indexed by the tower of Proposition 9.4?

Open Problem 13.8 (Ergodic morphisms)**.** [programme] Characterize, in terms of the invariant quotients of Theorem 6.6, the morphisms ν of Dyn for which the compound source with every ergodic source is ergodic. The natural conjecture is that this holds iff the invariant quotient of the compound system is, almost surely, that of its input; Example 9.5 is a channel for which it is strictly larger.

Notes

Khinchin's own summary is the Conclusion of [Khi56]. For the programme of synthetic probability theory in Markov categories, of which several of the problems above would be instances, see [Fri20] and the works cited in Chapters 6 and 7, in particular the synthetic laws of large numbers of [FGLPS26], which are the nearest existing results to the first problem.

Appendix A

Categorical Toolkit

This appendix collects the definitions used in the text. It is a reminder, not an introduction.

A.1 Markov categories

A ***Markov category*** is a symmetric monoidal category $(\mathcal{C}, \otimes, 1)$ in which every object X carries a commutative comonoid structure

$$\mathrm{copy}_X \colon X \to X \otimes X, \qquad \mathrm{del}_X \colon X \to 1,$$

compatible with the tensor product, and in which del is natural: $\mathrm{del}_Y \circ f = \mathrm{del}_X$ for every $f \colon X \to Y$. Equivalently the unit object is terminal. Morphisms $1 \to X$ are called ***states***. A morphism f is ***deterministic*** if it also commutes with copying, $\mathrm{copy}_Y \circ f = (f \otimes f) \circ \mathrm{copy}_X$. The deterministic morphisms form a cartesian monoidal subcategory $\mathcal{C}_{\mathrm{det}}$.

Category	Objects	Morphisms
FinStoch	finite sets	stochastic matrices
BorelStoch	standard Borel spaces	Markov kernels
Stoch	measurable spaces	Markov kernels
Dyn	the same, with an automorphism	equivariant Markov kernels

In FinStoch and BorelStoch the deterministic morphisms are the (measurable) maps; in Stoch they are the kernels taking only the values 0 and 1. For a state p on X and a morphism $f \colon X \to Y$ the joint state $p \rtimes f = (\mathrm{id} \otimes f) \circ \mathrm{copy} \circ p$ on $X \otimes Y$ has marginals p and $f \circ p$.

$\mathcal{C}$ ***has conditionals*** if every morphism into a tensor product $X \otimes Y$ factors as its marginal on X followed by a morphism that produces Y from X and the input. In BorelStoch conditionals exist; this is the existence of regular conditional probabilities. A ***Bayesian inverse*** of $f \colon X \to Y$ with respect to a state p is a morphism $f^\dagger \colon Y \to X$ with $(f \circ p) \rtimes f^\dagger$ equal to $p \rtimes f$ after exchanging the factors. Two morphisms $f, g \colon X \to Y$ are equal ***p-almost surely*** if $p \rtimes f = p \rtimes g$.

An infinite tensor product $\bigotimes_{i\in I} X_i$ is a ***Kolmogorov product*** if it is the limit of the finite tensor products with their marginalization maps, the limit is preserved by tensoring with any object, and the projections to finite subproducts are deterministic [FR20]. In BorelStoch countable Kolmogorov products exist and are the usual product spaces.

For a category D and a Markov category $\mathcal{C}$, the functors $D \to \mathcal{C}_{\mathrm{det}}$, with natural transformations whose components are arbitrary morphisms of $\mathcal{C}$, form a Markov category [Fri20, §7]. Dyn is the case in which D is the group $\mathbb{Z}$ regarded as a category with one object.

A.2 Operads

A ***symmetric operad*** P in a symmetric monoidal category consists of objects $P(n)$, $n \geq 0$, of n-ary operations with actions of the symmetric groups, a unit in $P(1)$, and composition maps

$$P(m) \otimes P(k_1) \otimes \cdots \otimes P(k_m) \longrightarrow P(k_1 + \cdots + k_m)$$

that are associative, unital and equivariant. The ***partial composite*** $p \circ_i q$ inserts q into the i-th input of p and units into the others. The operad Δ of simplices has $\Delta(n) = \Delta_n$, the simplex of distributions on n points, with the composition of Chapter 1; it is an operad in topological spaces. Its algebras in sets or spaces are the convex spaces.

Given an operad P in an additive setting and a bimodule M over it, a ***derivation*** is a family of maps $d\colon P(n) \to M(n)$ with $d(p \circ_i q) = dp \circ_i q + p \circ_i dq$. In Theorem 2.7 the bimodule is $\mathbb{R}$ in every arity, with $p \circ_i r = p_i r$ and $r \circ_i q = r$.

A.3 Cohomology of a monoid

Let S be a monoid acting on an abelian group F. The complex of inhomogeneous cochains has $C^n = \{f\colon S^n \to F\}$ and coboundary

$$(\delta f)(s_1, \dots, s_{n+1}) = s_1.f(s_2, \dots, s_{n+1}) + \sum_{k=1}^{n} (-1)^k f(\dots, s_k s_{k+1}, \dots) + (-1)^{n+1} f(s_1, \dots, s_n).$$

Its cohomology is $\mathrm{Ext}^{\bullet}_{\mathbb{Z}[S]}(\mathbb{Z}, F)$. A 1-cocycle is a ***crossed homomorphism***, $f(st) = f(s) + s.f(t)$, and a 1-coboundary has the form $f(s) = s.c - c$. The complex of Definition 3.2 is the subcomplex of local cochains for the monoid of observables acting on functions of laws by conditioning. In the general setting of [Vig20] the monoid varies over a poset and the computation is carried out

for modules over a presheaf of algebras, in the topos of presheaves on that poset.

A.4 Enrichment

A category is ***enriched*** in a monoidal category $\mathcal{V}$ when its hom-sets are objects of $\mathcal{V}$ and composition is a morphism of $\mathcal{V}$. Three cases occur in the text.

Divergences. A ***divergence space*** is a set X with a function $D\colon X \times X \to [0,\infty]$ vanishing on the diagonal; a morphism is a map that does not increase D; the tensor product is the product of sets with the sum of the divergences. A category enriched in divergence spaces is one in which $D(g \circ f \parallel g' \circ f') \leq D(f \parallel f') + D(g \parallel g')$. For a Markov category one asks in addition that the tensor product of morphisms be short [Per24]. When D is symmetric and satisfies the triangle inequality this is enrichment in metric spaces in the sense of Lawvere [Law73].

Topologies. A category is enriched in topological spaces, with the canonical closed structure, when its hom-sets are spaces and composition is separately continuous. This is the enrichment used in [PVB24].

Orders. A category enriched in posets has an order on each hom-set respected by composition. The order on idempotents in Proposition 7.3 is of this kind.

A.5 Dagger categories and idempotents

A ***dagger category*** is a category with an involutive identity-on-objects functor $\dagger\colon \mathcal{C}^{\mathrm{op}} \to \mathcal{C}$. A morphism $e\colon X \to X$ is a ***dagger idempotent*** if $e \circ e = e = e^\dagger$. It ***splits*** if $e = i \circ i^\dagger$ with $i^\dagger \circ i = \mathrm{id}$; in Hilb the dagger idempotents are the orthogonal projections and all of them split. Dagger idempotents on an object are partially ordered by $e \leq e'$ iff $e \circ e' = e$. The category Krn of Definition 7.1 is a dagger category under Bayesian inversion, and $L^2\colon \mathsf{Krn}^{\mathrm{op}} \to \mathsf{Hilb}$ is a dagger functor.

A.6 Preordered semirings

A ***preordered semiring*** is a commutative semiring with a preorder such that $a \leq b$ implies $a + c \leq b + c$ and $ac \leq bc$. The conditions under which Theorem 12.4 holds are stated before it. The asymptotic preorder is $a \lesssim b$ iff $a^n \leq b^{n+o(n)}$; the theorem says that it is detected by the monotone homomorphisms to the nonnegative reals.

Appendix B

Concordance with Khinchin

The table lists the sections of the two papers in [Khi57] and where their subject is treated here. Section numbers are those of the Dover edition. The correspondence is by topic; no passage of the source is reproduced.

Khinchin	This book
The Entropy Concept in Probability Theory (1953)	
§1 Entropy of finite schemes	Chapter 1
§2 The uniqueness theorem	Chapters 2 and 3
§3 Entropy of Markov chains	Chapter 5, Proposition 5.7
§4 Fundamental theorems (Theorems 2, 3)	Chapter 8, Theorems 8.2 and 8.9
§5 Application to coding theory (Theorem 4)	Chapter 11, Theorem 11.1
On the Fundamental Theorems of Information Theory (1956)	
Ch. I, §1 Two generalizations of Shannon's inequality	Chapter 4, Lemmas 4.7 and 4.8
Ch. I, §2 Three inequalities of Feinstein	Lemma 4.10
Ch. II, §3 Source, stationarity, entropy	Chapter 5, Proposition 5.4
Ch. II, §4 Ergodic sources	Chapter 6
Ch. II, §5 The E property; McMillan's theorem	Chapter 8
Ch. II, §6 Martingales; Doob's theorem	Chapter 7
Ch. II, §7 Auxiliary propositions	Lemmas 8.4 and 8.5
Ch. II, §8 Proof of McMillan's theorem	Theorem 8.2
Ch. III, §9 Channel, noise, stationarity, anticipation, memory	Definition 9.1 and Proposition 9.4
Ch. III, §10 Connection of the channel to the source	Theorem 9.3
Ch. III, §11 The ergodic case	Example 9.5 and Theorem 9.7
Ch. IV, §§12–13 Feinstein's fundamental lemma	Chapter 10, Theorems 10.2 and 10.5

Khinchin	**This book**
Ch. V, §14 Coding	Chapter 11
Ch. V, §15 The first Shannon theorem	Theorem 11.4
Ch. V, §16 The second Shannon theorem	Theorem 11.5
Conclusion	Theorem 9.10, Chapter 13

Topics treated here that have no counterpart in the source: the functorial and operadic characterizations of entropy (Chapter 2); information cohomology and entropy modulo a prime (Chapter 3); relative entropy and the enriched structure (Chapter 4); coalgebraic sources (Chapter 5); the categorical ergodic decomposition (Chapter 6); the converse coding theorem and polar codes (Chapter 11); zero-error capacity, asymptotic spectra and the isomorphism theory of Bernoulli shifts (Chapter 12).

Appendix C

Computational Checks

The script `verify.py` that accompanies this book recomputes every numerical example in the text and in the solutions, and tests a number of structural claims on explicit or randomly generated finite instances. It uses only the Python standard library, fixes its random seed, and runs in about a second; `python3 verify.py` prints one line per check and exits with status zero if and only if all pass. Version 1.0 performs 91 checks, all of which pass under Python 3.8 or later.

A check of this kind is evidence about the examples and about the absence of slips in the finite statements. It is not a proof of any theorem, and the theorems tagged classical or literature rest on the references cited for them. What the script does establish is the following.

Ch.	What is checked
1	Examples 1.4 and 1.8. On random instances: $H(f) = H(p) - H(q)$ and functoriality (Theorem 1.7); the operadic chain rule; convex-linearity (Proposition 1.10); Khinchin's properties; Proposition 1.11.
2	The second step of Khinchin's proof on random rational distributions. Additivity, maximality and expansibility of the Rényi entropies, their limit at $\alpha = 1$, and Example 2.8. The Leibniz rule (2.1) for H and its failure for H_2.
3	On random observables and laws on a five-point space: the action axiom for $\alpha = 1, \frac{1}{2}, 2$ (Lemma 3.1); the cocycle condition for entropy (Proposition 3.3); $\delta\delta = 0$ on a generic local cochain; the Tsallis coboundary and cocycle identities and the limit $\alpha \to 1$ (Theorem 3.7 and Proposition 3.8). The fundamental equation for h and for $h(x) = x$, and, in a four-dimensional test space, that its solutions are spanned by h. The Leibniz rule for the Fermat quotient; that entropy modulo p is well defined, satisfies the chain rule, equals the truncated logarithm and satisfies the fundamental equation, for $p = 3, 5, 7, 11, 13$ (Theorem 3.10).
4	On random instances: Gibbs' inequality, the chain rule and data processing; the three statements of Theorem 4.5; Proposition 4.6; Lemma 4.7; monotonicity and submodularity on the lattice of partitions, with the meet computed explicitly (Proposition 4.9); the three inequalities of Lemma 4.10; functoriality of relative entropy on FinStat.

Ch.	What is checked
5	Examples 5.8 and 5.12, by exact enumeration of block laws; monotonicity of the two bounds of Theorem 5.11; stationarity of a coalgebraic source; that a sliding block code does not raise block entropies; the count of equivariant maps in Exercise 5.3; Proposition 5.10 on an example.
6	Example 6.9, including the bound of one bit on the excess; the criterion of Exercise 6.2 for a Markov chain and its failure for the mixture; affinity of the entropy on a segment of stationary Markov sources.
7	Idempotence, symmetry and the tower property for conditional expectations on finite spaces, and a pair that does not commute; the dyadic martingale; the likelihood-ratio martingale of Exercise 7.4, including the singularity of the limit; the mean ergodic theorem for a permutation; orthogonality of martingale increments.
8	Lemma 8.4 for the Markov and hidden Markov sources; Examples 8.10 and 8.11; the second-order prediction of Exercise 8.5.
9	Invariance of the compound source in a cyclic model (Theorem 9.3); Example 9.5; the capacities of Example 9.12 by the Blahut–Arimoto iteration against the closed forms; the four statements of Proposition 9.13 and the sum formula; symmetry of the rate under the dagger; naturality of a memoryless family (Proposition 9.4); Example 9.14.
10	The greedy construction of Theorem 10.2 on 40 random channels: whenever the guaranteed error is below 1 it produces at least M codewords with disjoint decoding sets and that maximal error. The estimate (10.1); the weak converse for the codes found; Example 10.4; exact ε-ranks of small channels by exhaustive search.
11	Kraft's inequality and the bounds of Theorem 11.1 for Huffman codes; Example 11.2; Lemma 11.3 on an explicit composite; Proposition 11.8, including the bound from Mrs. Gerber's lemma, on five channels; Example 11.10 and the orthogonality argument of Theorem 11.9.
12	Example 12.2 by exhaustive search for independent sets, and Lovász's orthonormal representation; monotonicity of all Rényi entropies under doubly stochastic maps; agreement of the two expressions for the exponent in Theorem 12.6; Proposition 12.7 for independent bits at $n = 2000$; Example 12.10; the entropies in Theorem 12.8.
D	The numerical answers to Exercises 1.1, 1.4, 1.5, 2.1, 2.4, 3.2, 3.5, 4.1, 4.3–4.5, 5.1, 5.3, 6.1, 6.4, 7.1, 7.5, 8.1, 8.5, 9.1, 10.2, 10.3, 10.5, 11.1, 11.4, 11.5, 12.2, 12.3 and 12.5.

The checks on random instances use a few dozen to a few hundred instances each. They would detect a wrong sign or a misplaced hypothesis; they say

nothing about measure-theoretic statements, which are outside the reach of a finite computation.

Appendix D

Solutions to the Exercises

Solutions are given for every exercise, numbered as in the chapters. Where an exercise asks for a computation, the numbers below are recomputed by the script of Appendix C.

Chapter 1

1.1. g has image $(\frac{3}{4}, \frac{1}{8}, \frac{1}{8})$ and one nontrivial fibre, of mass $\frac{3}{4}$ and conditional distribution $(\frac{2}{3}, \frac{1}{3})$: $H(g) = \frac{3}{4}h(\frac{2}{3}) \approx 0.689$. Then h has image $(\frac{3}{4}, \frac{1}{4})$ and one nontrivial fibre of mass $\frac{1}{4}$ and conditional distribution $(\frac{1}{2}, \frac{1}{2})$: $H(h) = 0.25$. The composite has $H(h \circ g) = H(p) - h(\frac{1}{4}) \approx 1.75 - 0.811 = 0.939$, the sum of the two.

1.2. $H(f) = H(p) - H(f_*p)$ and $H(f_*p) = 0$ iff f_*p is a point mass. If f is an isomorphism then $H(f) = 0$ since f is injective, so $H(p) = H(q)$. The converse fails: $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and (a, a, b, b, b, b) with $2a + 4b = 1$ can be given the same entropy 1.75 by the intermediate value theorem ($a = b = \frac{1}{6}$ gives $\log 6 \approx 2.58$ and $a \to \frac{1}{2}$ gives 1), and they are not isomorphic.

1.3. By Proposition 1.9, $p = (t, 1-t) \circ (u_1, p')$ where p' is the conditional distribution on the other $n-1$ points, so $H(p) = h(t) + (1-t)H(p') \leq h(t) + (1-t)\log(n-1)$ by Proposition 1.3(ii).

1.4. If $W = \sum_k \lambda_k \Pi_k$ with permutation matrices Π_k, concavity of H gives $H(Wp) \geq \sum_k \lambda_k H(\Pi_k p) = H(p)$. Conversely apply the hypothesis to the uniform state u: $H(Wu) \geq \log|X|$ forces $Wu = u$ by Proposition 1.3(ii), which says that the rows of the matrix of W (indexed by outputs) sum to one. Its columns sum to one because it is stochastic.

1.5. $H(X \mid Z) \leq H(X, Y \mid Z) = H(Y \mid Z) + H(X \mid Y, Z)$, and $H(X \mid Y, Z) = H(X \mid Y)$ because X and Z are conditionally independent given Y. If both kernels are maps, Y is a function of X and Z a function of Y, so $H(X, Y \mid Z) = H(X \mid Z)$. For strict inequality take X and Y independent fair bits and Z a point: $H(X \mid Z) = 1 < H(X \mid Y) + H(Y \mid Z) = 2$.

Chapter 2

2.1. Here $g = 6$ and $(g_1, g_2, g_3) = (3, 2, 1)$. Y has six points in blocks of sizes $3, 2, 1$, and r puts mass $\frac{1}{6}$ on each pair (i, y) with y in block i. Then $\log 6 - \frac{1}{2}\log 3 - \frac{1}{3}\log 2 \approx 2.585 - 0.792 - 0.333 = 1.459$, and $H(p) = \frac{1}{2}\cdot 1 + \frac{1}{3}\log 3 + \frac{1}{6}\log 6 \approx 0.5 + 0.528 + 0.431 = 1.459$.

2.2. By symmetry it is enough to treat $i = n$. For $k = 1$ there is nothing to prove, and $k = 2$ is recursivity. Let $q \in \Delta_k$ with $k \geq 3$ and put $s = q_{k-1} + q_k$. If $s = 0$ use expansibility, which is recursivity with $q_2 = 0$. Otherwise let $q' = (q_1, \dots, q_{k-2}, s)$. Recursivity applied to $p \circ_n q$, whose last two entries are $p_n q_{k-1}$ and $p_n q_k$, gives $H(p \circ_n q) = H(p \circ_n q') + p_n s\, H(\frac{q_{k-1}}{s}, \frac{q_k}{s})$. By induction $H(p \circ_n q') = H(p) + p_n H(q')$, and recursivity for q itself gives $H(q) = H(q') + sH(\frac{q_{k-1}}{s}, \frac{q_k}{s})$. Combine. The full composite $q \circ (p_1, \dots, p_m)$ is obtained by m partial composites, the i-th of which contributes $q_i H(p_i)$.

2.3. $\sum_{i,j}(p_i q_j)^\alpha = \sum_i p_i^\alpha \sum_j q_j^\alpha$ gives additivity. Adding a zero does not change $\sum p_i^\alpha$. For $\alpha < 1$ the function t^α is concave, so $\sum_i p_i^\alpha \leq n^{1-\alpha}$ and $H_\alpha \leq \log n$; for $\alpha > 1$ it is convex, the inequality reverses, and so does the sign of $1 - \alpha$. By l'Hôpital's rule, $\lim_{\alpha\to 1} H_\alpha(p) = -\frac{d}{d\alpha}\log \sum_i p_i^\alpha \big|_{\alpha=1} = -\sum_i p_i \log p_i$.

2.4. Continuity is clear. A map f sends p to a distribution q obtained by adding entries, and $(a+b)^2 \geq a^2 + b^2$, so $\sum q^2 \geq \sum p^2$ and $H_2(q) \leq H_2(p)$. Let $f\colon (2, u_2) \to 1$ and $f' = \mathrm{id}_1$. Then $\frac{1}{2}f \oplus \frac{1}{2}f'$ maps $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$ to $(\frac{1}{2}, \frac{1}{2})$ and has $F = \log\frac{8}{3} - 1 \approx 0.415$, whereas $\frac{1}{2}F(f) + \frac{1}{2}F(f') = 0.5$. In the proof, functoriality still gives $F(f) = I(p) - I(q)$; what fails is the identity $F(f) = \sum_i q_i I(p_i)$, that is, the chain rule.

2.5. $u_1 \circ_1 u_1 = u_1$ gives $I(u_1) = I(u_1) + I(u_1)$. Next $(1,0) \circ_1 (1,0) = (1,0,0) = (1,0) \circ_2 (1,0)$ gives $I(1,0) + 1 \cdot I(1,0) = I(1,0) + 0 \cdot I(1,0)$, so $I(1,0) = 0$. For the chain rule write $q \circ (p_1, \dots, p_m) = (\cdots((q \circ_m p_m) \circ_{m-1} p_{m-1}) \cdots) \circ_1 p_1$. Inserting at position i does not change the entries at positions before i, so the i-th step adds $q_i I(p_i)$.

Chapter 3

3.1. As in Lemma 3.1, the weight of the block $B \cap C$ in $Y_{\cdot\alpha}(Z_{\cdot\alpha} f)$ is $P(B)^\alpha \big(P(B\cap C)/P(B)\big)^\alpha = P(B\cap C)^\alpha$. And $Y_{\cdot\alpha}1 = \sum_B P(B)^\alpha$, which equals 1 for all P and Y iff $\alpha = 1$.

3.2. Write $s_Y = \sum_B P(B)^\alpha$. Then $Y_{\cdot\alpha} S_\alpha[Z] = \frac{1}{\alpha-1}\sum_B P(B)^\alpha \big(1 - \sum_C (P(B\cap C)/P(B))^\alpha\big) = \frac{1}{\alpha-1}(s_Y - s_{YZ})$, and adding $S_\alpha[Y] = \frac{1}{\alpha-1}(1 - s_Y)$ gives

$\frac{1}{\alpha-1}(1-s_{YZ})$. If Y and Z are independent, $s_{YZ}=s_Y s_Z$ and $s=1-(\alpha-1)S_\alpha$; substitute.

3.3. The observables are 1 and the atomic partition X. A 1-cochain is a pair $(f[1],f[X])$ with $f[1]$ a constant and $f[X]$ a measurable function h of $P(a)\in[0,1]$. The cocycle conditions for the four pairs reduce to $f[1]=0$ and $X.f[X]=0$, that is, $P(a)h(1)+P(b)h(0)=0$ for all P. So Z^1 is the space of measurable h with $h(0)=h(1)=0$, there are no coboundaries, and H^1 is that infinite-dimensional space.

3.4. For $h(x)=x$ both sides equal $x+y$. For h, splitting off the first outcome of $(x,y,1-x-y)$ gives $h(x)+(1-x)h(\frac{y}{1-x})$ by Proposition 1.9, and splitting off the second gives the other side. There is no contradiction, because a cocycle vanishes at deterministic laws (Step o), which forces $h(1)=0$.

3.5. $ab-a^pb^p=(a-a^p)\,b+a^p\,(b-b^p)$; divide by p and use $a^p\equiv a$. Next $H_5(2,4)=(1-32-1024)/5=-211\equiv 4$, and $\sum_{k=1}^{4}2^k/k=2+2+8\cdot 3^{-1}+4$ with $3^{-1}=2$ in $\mathbb{F}_5$, giving $9\equiv 4$. In $\mathbb{F}_3$: $H_3(2,2)=(1-8-8)/3=-5\equiv 1$ and $H_3(1,1,2)=(1-1-1-8)/3=-3\equiv 0$. The chain rule predicts $H_3(2,2)+2\cdot H_3(2,2)+2\cdot H_3(1)=1+2+0\equiv 0$.

Chapter 4

4.1. $D(p\,\|\,q)=\frac{1}{2}\log 2+\frac{1}{2}\log\frac{2}{3}\approx 0.2075$ and $D(q\,\|\,p)=\frac{1}{4}\log\frac{1}{2}+\frac{3}{4}\log\frac{3}{2}\approx 0.1887$. The channel fixes p and sends q to $(0.3,0.7)$, and $D((\frac{1}{2},\frac{1}{2})\,\|\,(0.3,0.7))=\frac{1}{2}\log\frac{5}{3}+\frac{1}{2}\log\frac{5}{7}\approx 0.1258<0.2075$.

4.2. Write $r(x,y)=p(x)r(y\mid x)$ and $r'(x,y)=p'(x)r'(y\mid x)$; then $\log\frac{r}{r'}=\log\frac{p}{p'}+\log\frac{r(y|x)}{r'(y|x)}$, and averaging against r gives the formula. In the proof of Theorem 4.4 the only inequality is the positivity of $\sum_y(f\circ p)(y)D(r(\cdot\mid y)\,\|\,r'(\cdot\mid y))$, which vanishes iff each term with $(f\circ p)(y)>0$ does.

4.3. TV $=1-\sum p^2=1-\frac{3}{8}=\frac{5}{8}$ and $D(\mathrm{copy}\circ p\,\|\,p\otimes p)=H(p)=1.5$. Both are largest at the uniform state, where they equal $1-\frac{1}{n}$ and $\log n$: the first because $\sum p^2\geq\frac{1}{n}$ by the Cauchy–Schwarz inequality, the second by Proposition 1.3(ii).

4.4. The marginals are uniform, so $I(X;Y)=1+1-H(0.4,0.4,0.1,0.1)=2-1.722=0.278$. A function of X that is also a function of Y takes the same value at $(0,0)$ and $(0,1)$, at $(0,1)$ and $(1,1)$, and so on; all four points have positive probability, so it is constant. For the second part take $Y=X$ a fair bit: $X\wedge Y=X$ and $I(X;Y)=H[X]=1$.

4.5. Given $Y = y$, the posterior of X is $1-\varepsilon$ at y and $\varepsilon/(n-1)$ elsewhere; since $\varepsilon < 1 - \frac{1}{n}$ the maximum is at y, so $P_e = \varepsilon$. By the equality case of Exercise 1.3, $H(X \mid Y = y) = h(\varepsilon) + \varepsilon \log(n-1)$ for every y.

Chapter 5

5.1. $\pi = (b, a)/(a+b)$ and Proposition 5.7 give the formula. It is a weighted average of $h(a)$ and $h(b)$, so it is at most 1, with equality iff $a = b = \frac{1}{2}$. It equals the entropy of the marginal iff the letters are independent, that is, iff the two rows of Q agree: $a + b = 1$.

5.2. An output block of length n is a function of an input block of length $n+1$, so $H_n(\phi \circ \mu) \le H_{n+1}(\mu)$; divide by n. For independent fair bits the outputs $x_n + x_{n+1}$ are again independent fair bits, so both entropies are 1. But ϕ is not an isomorphism: it identifies x with its complement, and complementation preserves μ, so ϕ is two-to-one almost everywhere. Equal entropy does not make a factor map invertible.

5.3. Given $g\colon X \to B$ put $\phi(x)_n = g(n \cdot x)$; conversely $g = \pi_0 \circ \phi$. These are inverse bijections, as in Proposition 5.6. Hence there are $|B|^{|A|^N}$ equivariant maps.

5.4. Given the past, x_{n+1} has distribution $e_{\tau(x_n)}$, which depends on x_n only. So the letters form a Markov chain with $q_{ab} = e_{\tau(a)}(b)$. The coalgebra is unifilar, and by Proposition 5.10 its entropy is $\sum_s \sigma(s) H(e_s) = \sum_a \pi(a) H(q_{a\cdot})$, where π is the stationary distribution of the letters and $\sigma = \tau_* \pi$. Conversely, given Q take $S = A$, $\tau = \mathrm{id}$ and $e_a = q_{a\cdot}$.

5.5. Upper bound: $H(\mu) = \lim c_n$ and c_n is nonincreasing (Proposition 5.4). Lower bound: for every m,

$$c_{n+m} \;\ge\; H(x_{n+m} \mid x_0, \dots, x_{n+m-1}, s_m) = H(x_{n+m} \mid x_m, \dots, x_{n+m-1}, s_m),$$

because given s_m the letters before time m are independent of those after. By stationarity the right side is $H(x_n \mid x_0, \dots, x_{n-1}, s_0)$; let $m \to \infty$. Monotonicity of the lower bound: conditioning also on s_1 and then discarding x_0, s_0, which are independent of the future given s_1, $H(x_{n+1} \mid x_0, \dots, x_n, s_0) \ge H(x_{n+1} \mid x_1, \dots, x_n, s_1)$, and the right side is the bound for n. Convergence: the gap between the bounds is $I(s_0; x_n \mid x_0, \dots, x_{n-1})$, and these terms sum to at most $H(s_0)$.

Chapter 6

6.1. A stationary distribution of Q is a convex combination $\lambda\pi' + (1-\lambda)\pi''$ of the unique stationary distributions of the two classes, and the corresponding Markov source is the same combination of the two sources supported on sequences that stay in one class. Each of those is ergodic by Proposition 6.2. The event of lying in the first class is invariant with measure λ, so interior points are not ergodic. By Proposition 5.7 the entropy is $\sum_a \pi(a) H(q_{a\cdot})$, which is linear in π.

6.2. If μ is ergodic, apply Theorem 6.3 to $f = 1_{C'}$: $A_n f \to \mu(C')$ in L^1; multiply by 1_C and integrate. Conversely the condition extends from cylinders to all measurable sets by approximation, and for an invariant S it gives $\mu(S) = \mu(S \cap T^{-k}S) \to \mu(S)^2$.

6.3. The lower inequality is concavity of H on $\Delta(A^n)$. For the upper one, let J be the label of the component: $H_n(\mu) \le H(x_1 \dots x_n, J) = H(J) + H(x_1 \dots x_n \mid J)$. Equality holds for all n when the label is a function of every block, for instance when μ' and μ'' are the two constant sequences on $\{0,1\}$: then $H_n(\mu) = h(\lambda)$ and $H_n(\mu') = H_n(\mu'') = 0$.

6.4. Write $\mu_n = \frac{1}{2}\mu'_n + \frac{1}{2}\mu''_n$. On the first component the number of ones is $0.1n + o(n)$ almost surely, so $-\frac{1}{n}\log \mu'_n(x) \to h(0.1) \approx 0.469$ and $-\frac{1}{n}\log \mu''_n(x) = 1$; the larger probability dominates the sum and the limit is $h(0.1)$. On the second component the number of ones is $\frac{1}{2}n + o(n)$, so $-\frac{1}{n}\log \mu'_n(x) \to \frac{1}{2}\log 10 + \frac{1}{2}\log\frac{10}{9} \approx 1.737$, and the limit is 1.

6.5. The two sequences are exchanged by T, so the only invariant sets are null or conull, while each of them is fixed by T^2 and is a T^2-invariant set of measure $\frac{1}{2}$. Every block has one of two values, so $H_n \le 1$ and $H(\mu) = 0$.

Chapter 7

7.1. $e_{\mathcal{G}}(y \mid x) = \frac{1}{2}$ if $x, y \in \{1,2\}$, $\frac{1}{4}$ if $x, y \in \{3,4,5,6\}$, and 0 otherwise. It is block diagonal with blocks $\frac{1}{2}J_2$ and $\frac{1}{4}J_4$, where J_k is the all-ones matrix; these are symmetric and idempotent. On $f(x) = x$ it gives the block averages: 1.5 on $\{1,2\}$ and 4.5 on $\{3,4,5,6\}$.

7.2. If $\mathcal{G} \subseteq \mathcal{G}'$, resampling given $\mathcal{G}'$ and then given $\mathcal{G}$ is resampling given $\mathcal{G}$, and in the other order the second step does not change the $\mathcal{G}$-class; so both composites are $e_{\mathcal{G}}$. Conversely if $e_{\mathcal{G}} e_{\mathcal{G}'} = e_{\mathcal{G}}$ then every $\mathcal{G}$-measurable g satisfies $\mathbb{E}[g \mid \mathcal{G}'] = g$ after applying L^2 and taking adjoints, so it is $\mathcal{G}'$-measurable. On $\{1,2,3\}$ with the uniform distribution take the partitions $\{1,2\}|\{3\}$ and

$\{1\}|\{2,3\}$. The function $f = 1_{\{1\}}$ is sent by the first projection to $(\frac12, \frac12, 0)$ and then by the second to $(\frac12, \frac14, \frac14)$; in the other order it is fixed by the second and sent to $(\frac12, \frac12, 0)$ by the first.

7.3. $\mathbb{E}[f \mid \mathcal{F}_n]$ is the midpoint of the dyadic interval containing x. On each interval f minus the midpoint is uniform on an interval of length 2^{-n} centred at 0, with variance $2^{-2n}/12$.

7.4. The factors are independent with mean $\frac12(1.6 + 0.4) = 1$, so M_n is a martingale with mean 1, and $\mathbb{E}\sqrt{M_n} = (\frac12(\sqrt{1.6} + \sqrt{0.4}))^n = \rho^n$. By Markov's inequality $\mathbb{P}(M_n > \delta) \le \rho^n/\sqrt{\delta} \to 0$. If $M_n = \mathbb{E}[f \mid \mathcal{F}_n]$ then $M_n \to f$ in L^1 by Corollary 7.5, so $f = 0$, contradicting $\mathbb{E}M_n = 1$.

7.5. σ has order 6, so $\frac16 \sum_{k<6} U^k$ averages a function over the orbits of σ: it is block diagonal with blocks $\frac13 J_3$ and $\frac12 J_2$. This is the conditional expectation, for the uniform distribution, onto the σ-algebra generated by the cycles, which is the invariant σ-algebra.

Chapter 8

8.1. $i_n = \sum_{k<n} -\log p(x_k)$ is a sum of independent identically distributed terms with mean H and variance σ^2, so $\frac1n i_n$ has variance σ^2/n. For a bit with $\mathbb{P}(1) = p$ the term takes two values differing by $\log\frac{1-p}{p}$, so $\sigma^2 = p(1-p)\log^2\frac{1-p}{p}$; at $p = 0.2$ this is $0.16 \cdot 4 = 0.64$.

8.2. By the Markov property $\mu(x_0 \mid x_{-k}, \dots, x_{-1}) = q_{x_{-1}x_0}$ for $k \ge 1$. So $i_n = g_0 + \sum_{k=1}^{n-1} g_1 \circ T^k$ with g_1 bounded on the support of μ, and $\frac1n i_n \to \mathbb{E}[g_1 \mid \mathcal{I}]$ in L^1, which is $\mathbb{E}g_1 = H(\mu)$ if the source is ergodic.

8.3. Each block of T_n^ε has probability at least $2^{-n(H+\varepsilon)}$ and the total is at most 1; each has probability at most $2^{-n(H-\varepsilon)}$ and the total is at least $1-\delta$. For Proposition 8.14 take $S = T_n^\varepsilon$ with n so large that $\mu_n(S) \ge 1-\varepsilon \ge \frac12$. Then $\frac12 2^{n(H-\varepsilon)} \le |S| \le 2^{n(H+\varepsilon)}$, and $\mu_n(C)|S|$ lies between $\frac12 2^{-2n\varepsilon}$ and $2^{2n\varepsilon}$.

8.4. Order the blocks by decreasing probability, which is by increasing number of ones. By Exercise 6.4 the blocks with about $0.1n$ ones carry probability close to $\frac12$ in total and each has probability about $\frac12 2^{-nh(0.1)}$; there are about $2^{nh(0.1)}$ of them. If $\lambda < \frac12$ they suffice. If $\lambda > \frac12$ one must add blocks from the second component, which have probability about $\frac12 2^{-n}$ each, and a fixed positive fraction of all 2^n blocks is needed.

8.5. The probability of a block decreases with its number of ones, so an optimal set consists of all blocks with fewer than k ones and some with exactly k, where $k = np + \Phi^{-1}(\lambda)\sqrt{np(1-p)} + o(\sqrt{n})$ by the central limit theorem. The

number of such blocks is $2^{nh(k/n)+O(\log n)}$, and $h(k/n) = h(p) + h'(p)(k/n - p) + O(1/n)$ with $h'(p) = \log \frac{1-p}{p}$. For $p = 0.2$, $\sigma = 0.8$; at $n = 1000$ the correction for $\lambda = 0.9$ is $0.8 \cdot 1.2816/\sqrt{1000} \approx 0.032$, predicting 0.754 against the computed 0.748, and for $\lambda = 0.1$ predicting 0.690 against 0.683. The remaining difference is the term of order $(\log n)/n$.

Chapter 9

9.1. Erasure: the output reveals whether an erasure occurred, so $H(Y) = h(e) + (1-e)H(p)$ and $H(Y \mid X) = h(e)$; $I = (1-e)H(p)$, maximal at the uniform input. Z-channel: with $q = \mathbb{P}(X = 1)$, $I = h(q(1-\delta)) - q\,h(\delta)$. Setting the derivative to zero, $(1-\delta)\log\frac{1-s}{s} = h(\delta)$ with $s = q(1-\delta)$, so $s = 1/(1 + 2^{h(\delta)/(1-\delta)})$. Substituting, $I = -\log(1-s) = \log\bigl(1 + 2^{-h(\delta)/(1-\delta)}\bigr)$, and $2^{-h(\delta)/(1-\delta)} = (1-\delta)\,\delta^{\delta/(1-\delta)}$.

9.2. The output differs from the input iff exactly one of the two channels flips. For $\delta_1 = \delta_2 = 0.1$ the flip probability is 0.18 and $C = 1 - h(0.18) \approx 0.320$, less than 0.531.

9.3. The output blocks take two values, so $H_n(\eta) = 1$ and $H(\eta) = 0$. Since $\omega = \mu \otimes \eta$, $H_n(\omega) = H_n(\mu) + 1$ and $H(\omega) = H(\mu)$. Hence $R(\mu, \nu) = 0$, as it must be for a channel that ignores its input.

9.4. For inputs (X, X') to the product channel, Y and Y' are conditionally independent given (X, X') with $H(YY' \mid XX') = H(Y \mid X) + H(Y' \mid X')$, and $H(YY') \le H(Y) + H(Y')$; so $I(XX'; YY') \le I(X;Y) + I(X';Y') \le C(W) + C(W')$. Independent optimal inputs attain the bound. For the sum channel let λ be the probability of using the first component. The output determines the component, so $I = h(\lambda) + \lambda I_1 + (1-\lambda) I_2$, maximal over the conditional inputs at $h(\lambda) + \lambda C + (1-\lambda)C'$ and then over λ at $\lambda = 2^C/(2^C + 2^{C'})$, with value $\log(2^C + 2^{C'})$.

9.5. Naturality for maps says that the first n letters of $\nu_{n+1}(x_1 \dots x_{n+1})$ are $\nu_n(x_1 \dots x_n)$, so the n-th output letter is a function f_n of $x_1, \dots, x_n$, and any such sequence of functions is natural. For $y_n = x_{n+1}$ the law of y_1 depends on x_2 and is not a function of x_1.

Chapter 10

10.1. On E_x we have $Q(y) < W(y \mid x)\, 2^{-\gamma}/M$. Summing over $y \in E_x$ gives $Q(E_x) \le W(E_x \mid x)\, 2^{-\gamma}/M \le 2^{-\gamma}/M$.

10.2. With uniform inputs the output is uniform, $Q(y) = 2^{-n}$, and $W(y \mid x) = \delta^d (1-\delta)^{n-d}$. Since $\delta < \frac{1}{2}$ the density decreases in d and E_x is a ball. For $n = 7$, $\delta = 0.01$: $i = 6.898$ at $d = 0$ and 0.269 at $d = 1$. With $\log M = 4$ and $\gamma > 0$ the threshold $4 + \gamma$ exceeds 0.269, so $E_x = \{x\}$ and one needs $\gamma < 2.898$. The guaranteed error is then $\mathbb{P}[d \geq 1] + 2^{-\gamma}$, with infimum $(1 - 0.99^7) + 2^{-2.898} \approx 0.068 + 0.134 = 0.202$. The Hamming code corrects one error and fails with probability $1 - 0.99^7 - 7(0.01)(0.99)^6 \approx 0.0020$. The lemma is sharp in rate and far from sharp at length seven.

10.3. Identity: $W(D_j \mid x_j) \geq 1 - \varepsilon > 0$ forces $x_j \in D_j$, and the sets are disjoint, so the codewords are distinct and $M^* = k$ for every $\varepsilon < 1$. Constant channel with output law Q: one needs disjoint sets of Q-measure at least $1 - \varepsilon$, so $M^* \leq \lfloor 1/(1-\varepsilon) \rfloor$, with $M^* = 1$ for $\varepsilon < \frac{1}{2}$. In both cases the answer agrees with capacity $\log k$ and 0 after taking $\frac{1}{n}\log$ of the value for the n-th power.

10.4. With a stochastic encoder the j-th diagonal entry is $\sum_x e(x \mid j)\, W(D_j \mid x) \leq \max_x W(D_j \mid x)$, so replacing $e(\cdot \mid j)$ by a maximizing x does not decrease it. For the second part let Y be a point and $M = 2$. A deterministic decoder sends the point to one message and has maximal error 1; a decoder that tosses a fair coin has maximal error $\frac{1}{2}$.

10.5. If the unerased positions distinguish x_j from every other codeword, the decoder returns x_j. For a linear code with a $k \times n$ generator matrix G and erased set S, this happens for every codeword iff the columns of G outside S have rank k, so the error probability is the same for all codewords. Choose G uniformly at random. For a fixed nonzero message v the vector vG restricted to the $n - |S|$ unerased positions is uniform, so it vanishes with probability $2^{-(n-|S|)}$, and the union bound gives failure probability at most $2^{k-n+|S|}$ given S. With probability tending to one $|S| \leq n(e+\delta)$, so the expected failure probability is at most $2^{k-n(1-e-\delta)} + o(1)$, which is small when $k/n < 1 - e - 2\delta$. Some G does at least as well as the average.

Chapter 11

11.1. Huffman's procedure merges 0.1 and 0.2, then the result with 0.3, then with 0.4, giving lengths $(1, 2, 3, 3)$ and codewords $0, 10, 110, 111$. The expected length is $0.4 + 0.6 + 0.6 + 0.3 = 1.9$ and $H(p) \approx 1.846$.

11.2. In the proof of Theorem 11.1, $L - H(p) = D(p \,\|\, q) - \log K$ with both terms nonnegative. Equality needs $K = 1$ and $p = q$, so $p(x) = 2^{-\ell(x)}$. Conversely if every $p(x)$ is a power of $\frac{1}{2}$ the lengths $-\log p(x)$ satisfy Kraft's inequality with equality.

11.3. Let the channel code have M messages, of which message 1 is always decoded wrongly and the others always correctly; its average error is $1/M$. Let the compression send a set of source blocks of probability 0.9 to message 1. The composite fails with probability at least 0.9. Repair: discard the worse half of the messages. If the average error is ε, at most half of the messages have error above 2ε, so the remaining $M/2$ form a code with maximal error 2ε, at a cost of one bit.

11.4. If f factors through $[r]$ so does every composite with f, which gives (i); if f, g factor through $[r], [s]$ then $f \otimes g$ factors through $[rs]$, which gives (ii). For (iii), if $f\colon [M] \to [M]$ has diagonal at least $1-\varepsilon$ and factors through $[r]$, then $(1-\varepsilon)\log M - 1 \le C(f) \le \log r$ by Fano's inequality and Proposition 9.13(iv). The bound is $\frac{1}{n}\log M^*(W^{\otimes n}, \varepsilon) \le (\log r_+(W) + \frac{1}{n})/(1-\varepsilon)$. For a binary symmetric channel with $0 < \delta < \frac{1}{2}$, $r_+ = 2$ and $\Phi = 1 > 1 - h(\delta)$. The function $\log\min\{|X|, |Y|\}$ fails (i): a composite $X \to 1 \to X$ has $\Phi = \log|X|$ while each factor has $\Phi = 0$.

11.5. W^-: the decoder sees y_1, y_2 and wants $u_1 = x_1 \oplus x_2$, which is known iff neither output is erased; the erasure probability is $1-(1-e)^2$. W^+: given u_1, the bit u_2 equals x_2 and equals $x_1 \oplus u_1$, so it is unknown iff both outputs are erased, with probability e^2. Next, $f(2e-e^2) = f(e)\sqrt{(2-e)(1-e)}$ and $f(e^2) = f(e)\sqrt{e(1+e)}$, and by the Cauchy–Schwarz inequality $\sqrt{(2-e)(1-e)} + \sqrt{e(1+e)} \le \sqrt{(2-e)+(1+e)}\,\sqrt{(1-e)+e} = \sqrt{3}$. Hence the average of $f(e_b)$ over the 2^n channels is at most $(\frac{\sqrt{3}}{2})^n f(e)$, and by Markov's inequality the fraction with $e_b \in [\delta, 1-\delta]$, where $f \ge \sqrt{\delta(1-\delta)}$, tends to zero geometrically. The mean of e_b is conserved, so the fraction with $e_b < \delta$ tends to $1-e$.

Chapter 12

12.1. In C_5 any three vertices contain two adjacent ones, and $\{0,2\}$ is independent. Two words $(i, 2i)$ and $(j, 2j)$ with $i \ne j$ are confusable only if $i - j \in \{0, \pm 1\}$ and $2(i-j) \in \{0, \pm 1\}$ modulo 5; for $i - j = \pm 1$ the second difference is ± 2. For the capacity, $H(Y \mid X) = 1$ for every input and the uniform input gives the uniform output, so $C = \log 5 - 1$.

12.2. For $0 < \alpha < 1$ the function $t \mapsto t^\alpha$ is concave, so $\sum_i q_i^\alpha = \sum_i (\sum_j D_{ij} p_j)^\alpha \ge \sum_{i,j} D_{ij} p_j^\alpha = \sum_j p_j^\alpha$, and $H_\alpha = \frac{1}{1-\alpha}\log\sum p^\alpha$ increases. For $\alpha > 1$ the inequality reverses and so does the sign of $1-\alpha$. The cases $\alpha = 0, 1, \infty$ follow by continuity, or directly. Then use additivity: $H_\alpha(p^{\otimes n}) = nH_\alpha(p)$.

12.3. (i) Every block in the support has positive probability, so $d \circ e$ must be the identity on all $|\operatorname{supp} p|^n$ of them, and conversely that many messages suffice.

For the numbers: $h(\tau) = 0.7$ gives $\tau \approx 0.1893$, and $D((1-\tau,\tau) \,\|\, p) \approx 0.0521$. The supremum of $\rho(0.7 - H_{1/(1+\rho)}(p))$ is attained near $\rho \approx 0.51$ and is also 0.0521. The first expression is the exponent of the probability that a block has more than τn ones, which is the event that the best set of 2^{nR} blocks misses it.

12.4. Let a_n be the number of allowed blocks of length n. An allowed block ends in 0, preceded by any allowed block of length $n-1$, or in 01, preceded by any allowed block of length $n-2$; so $a_n = a_{n-1} + a_{n-2}$ with $a_1 = 2$, $a_2 = 3$, and a_n grows like ϕ^n with $\phi = \frac{1+\sqrt{5}}{2}$. A Markov chain with $q_{01} = a$ and $q_{10} = 1$ has stationary distribution $(\frac{1}{1+a}, \frac{a}{1+a})$ and entropy $h(a)/(1+a)$. Its derivative vanishes when $(1+a)\log\frac{1-a}{a} = h(a)$, that is, $(1-a)^2 = a$, so $a = \frac{3-\sqrt{5}}{2} = \phi^{-2}$, and the maximum is $\log\phi$.

12.5. The entropies are 1, $\log 3 \approx 1.585$, 2 and 2. By Theorem 12.8 and Theorem 5.13, one Bernoulli shift is a factor of another iff its entropy is not larger. So $B(\frac{1}{2},\frac{1}{2})$ is a factor of the other three, $B(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ is a factor of Meshalkin's two, and those two are isomorphic. No other factor relations hold.

References

[AC88] P. Algoet and T. Cover. A sandwich proof of the Shannon–McMillan–Breiman theorem. *Ann. Probab.* 16 (1988).

[AD75] J. Aczél and Z. Daróczy. *On Measures of Information and Their Characterizations*. Academic Press, 1975.

[Adl61] R. Adler. Ergodic and mixing properties of infinite memory channels. *Proc. Amer. Math. Soc.* 12 (1961).

[Alo98] N. Alon. The Shannon capacity of a union. *Combinatorica* 18 (1998).

[Ari09] E. Arıkan. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inform. Theory* 55 (2009).

[AT09] E. Arıkan and E. Telatar. On the rate of channel polarization. *Proc. IEEE ISIT* (2009).

[BB15] P. Baudot and D. Bennequin. The homological nature of entropy. *Entropy* 17 (2015).

[BF14] J. Baez and T. Fritz. A Bayesian characterization of relative entropy. *Theory Appl. Categ.* 29 (2014).

[BFL11] J. Baez, T. Fritz and T. Leinster. A characterization of entropy in terms of information loss. *Entropy* 13 (2011).

[Bil65] P. Billingsley. *Ergodic Theory and Information*. Wiley, 1965.

[Bir31] G. D. Birkhoff. Proof of the ergodic theorem. *Proc. Nat. Acad. Sci. USA* 17 (1931).

[Bla57] D. Blackwell. The entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. Information Theory* (1957).

[BMSS12] L. Birkedal, R. Møgelberg, J. Schwinghammer and K. Støvring. First steps in synthetic guarded domain theory: step-indexing in the topos of trees. *Log. Methods Comput. Sci.* 8(4) (2012).

[Bra21] T.-D. Bradley. Entropy as a topological operad derivation. *Entropy* 23 (2021).

[Bre57] L. Breiman. The individual ergodic theorem of information theory. *Ann. Math. Statist.* 28 (1957); correction, 31 (1960).

[Bre60] L. Breiman. On achieving channel capacity in finite-memory channels. *Illinois J. Math.* 4 (1960).

[Cat88] J.-L. Cathelineau. Sur l'homologie de SL_2 à coefficients dans l'action adjointe. *Math. Scand.* 63 (1988).

[Cen82] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*. Amer. Math. Soc., 1982.

[CFS16] B. Coecke, T. Fritz and R. Spekkens. A mathematical theory of resources. *Inform. and Comput.* 250 (2016).

[Chu61] K. L. Chung. A note on the ergodic theorem of information theory. *Ann. Math. Statist.* 32 (1961).

[CJ19] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Structures Comput. Sci.* 29 (2019).

[CK11] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge Univ. Press, 2011.

[CT06] T. Cover and J. Thomas. *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[Dar70] Z. Daróczy. Generalized information functions. *Inform. and Control* 16 (1970).

[DDGK16] F. Dahlqvist, V. Danos, I. Garnier and O. Kammar. Bayesian inversion by ω-complete cone duality. *Proc. CONCUR 2016*.

[DGB13] D. Dikranjan and A. Giordano Bruno. Entropy in a category. *Appl. Categ. Structures* 21 (2013).

[DHKK14] G. Dimitrov, F. Haiden, L. Katzarkov and M. Kontsevich. Dynamical systems and categories. *Contemp. Math.* 621 (2014).

[Dob59] R. L. Dobrushin. General formulation of Shannon's main theorem in information theory. *Uspekhi Mat. Nauk* 14 (1959).

[Doo53] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[EP23] N. Ensarguet and P. Perrone. Categorical probability spaces, ergodic decompositions, and transitions to equilibrium. arXiv:2310.04267 (2023).

[Erd46] P. Erdős. On the distribution function of additive functions. *Ann. of Math.* 47 (1946).

[EVG02] P. Elbaz-Vincent and H. Gangl. On poly(ana)logs I. *Compositio Math.* 130 (2002).

[Fad56] D. K. Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk* 11 (1956).

[Fan61] R. Fano. *Transmission of Information*. MIT Press, 1961.

[Fei54] A. Feinstein. A new basic theorem of information theory. *IRE Trans. Inform. Theory* 4 (1954).

[Fei58] A. Feinstein. *Foundations of Information Theory*. McGraw-Hill, 1958.

[Fei59] A. Feinstein. On the coding theorem and its converse for finite-memory channels. *Inform. and Control* 2 (1959).

[FGLPS26] T. Fritz, T. Gonda, A. Lorenzin, P. Perrone and A. Shah Mohammed. Empirical measures and strong laws of large numbers in categorical probability. *Log. Methods Comput. Sci.* 22(2) (2026).

[FR20] T. Fritz and E. F. Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality* 2 (2020).

[Fri17] T. Fritz. Resource convertibility and ordered commutative monoids. *Math. Structures Comput. Sci.* 27 (2017).

[Fri20] T. Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Adv. Math.* 370 (2020).

[Fuj78] S. Fujishige. Polymatroidal dependence structure of a set of random variables. *Inform. and Control* 39 (1978).

[FuP21] J. Fullwood and A. Parzygnat. The information loss of a stochastic map. *Entropy* 23 (2021).

[Gir82] M. Giry. A categorical approach to probability theory. *Lecture Notes in Math.* 915 (1982).

[Gla03] E. Glasner. *Ergodic Theory via Joinings*. Amer. Math. Soc., 2003.

[Gra09] R. Gray. *Probability, Random Processes, and Ergodic Properties*, 2nd ed. Springer, 2009.

[Gra11] R. Gray. *Entropy and Information Theory*, 2nd ed. Springer, 2011.

[Gro13] M. Gromov. In a search for a structure, Part 1: On entropy. Preprint, 2013.

[Hae79] W. Haemers. On some problems of Lovász concerning the Shannon capacity of a graph. *IEEE Trans. Inform. Theory* 25 (1979).

[Han03] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer, 2003.

[HC67] J. Havrda and F. Charvát. Quantification method of classification processes: concept of structural *a*-entropy. *Kybernetika* 3 (1967).

[Hed69] G. Hedlund. Endomorphisms and automorphisms of the shift dynamical system. *Math. Systems Theory* 3 (1969).

[HJS07] I. Hasuo, B. Jacobs and A. Sokolova. Generic trace semantics via coinduction. *Log. Methods Comput. Sci.* 3(4) (2007).

[Hu62] Hu Kuo Ting. On the amount of information. *Theory Probab. Appl.* 7 (1962).

[Huf52] D. Huffman. A method for the construction of minimum-redundancy codes. *Proc. IRE* 40 (1952).

[HV93] T. S. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. Inform. Theory* 39 (1993).

[Jac59] K. Jacobs. Die Übertragung diskreter Informationen durch periodische und fastperiodische Kanäle. *Math. Ann.* 137 (1959).

[JV20] A. Jensen and P. Vrana. The asymptotic spectrum of LOCC transformations. *IEEE Trans. Inform. Theory* 66 (2020).

[Ken64] D. G. Kendall. Functional equations in information theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 2 (1964).

[Khi53] A. I. Khinchin. The entropy concept in probability theory. *Uspekhi Mat. Nauk* 8(3) (1953). English translation in [Khi57].

[Khi56] A. I. Khinchin. On the fundamental theorems of information theory. *Uspekhi Mat. Nauk* 11(1) (1956). English translation in [Khi57].

[Khi57] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Translated by R. A. Silverman and M. D. Friedman. Dover, 1957.

[KK13] H. Kerstan and B. König. Coalgebraic trace semantics for continuous probabilistic transition systems. *Log. Methods Comput. Sci.* 9(4) (2013).

[Kli07] M. Klimesh. Inequalities that collectively completely characterize the catalytic majorization relation. arXiv:0709.3680 (2007).

[Kol58] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk SSSR* 119 (1958).

[Kon95] M. Kontsevich. The $1\frac{1}{2}$-logarithm. Appendix to [EVG02]; circulated 1995.

[Kra49] L. G. Kraft. A device for quantizing, grouping, and coding amplitude-modulated pulses. M.S. thesis, MIT, 1949.

[Kri70] W. Krieger. On entropy and generators of measure-preserving transformations. *Trans. Amer. Math. Soc.* 149 (1970).

[Law62] F. W. Lawvere. The category of probabilistic mappings. Unpublished manuscript, 1962.

[Law73] F. W. Lawvere. Metric spaces, generalized logic, and closed categories. *Rend. Sem. Mat. Fis. Milano* 43 (1973).

[Lee64] P. M. Lee. On the axioms of information theory. *Ann. Math. Statist.* 35 (1964).

[Lei21] T. Leinster. *Entropy and Diversity: The Axiomatic Approach*. Cambridge Univ. Press, 2021.

[Lei21p] T. Leinster. Entropy modulo a prime. *Commun. Number Theory Phys.* 15 (2021).

[Lin01] E. Lindenstrauss. Pointwise theorems for amenable groups. *Invent. Math.* 146 (2001).

[LM95] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge Univ. Press, 1995.

[Lov79] L. Lovász. On the Shannon capacity of a graph. *IEEE Trans. Inform. Theory* 25 (1979).

[Mat07] F. Matúš. Infinitely many information inequalities. *Proc. IEEE ISIT* (2007).

[McM53] B. McMillan. The basic theorems of information theory. *Ann. Math. Statist.* 24 (1953).

[McM56] B. McMillan. Two inequalities implied by unique decipherability. *IRE Trans. Inform. Theory* 2 (1956).

[Mes59] L. D. Meshalkin. A case of isomorphism of Bernoulli schemes. *Dokl. Akad. Nauk SSSR* 128 (1959).

[MP23] S. Moss and P. Perrone. A category-theoretic proof of the ergodic decomposition theorem. *Ergodic Theory Dynam. Systems* 43 (2023).

[Ned57] J. Nedoma. The capacity of a discrete channel. *Trans. First Prague Conf. Information Theory* (1957).

[Orn70] D. Ornstein. Bernoulli shifts with the same entropy are isomorphic. *Adv. Math.* 4 (1970).

[OW83] D. Ornstein and B. Weiss. The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Israel J. Math.* 44 (1983).

[Par61] K. R. Parthasarathy. On the integral representation of the rate of transmission of a stationary channel. *Illinois J. Math.* 5 (1961).

[Par64] W. Parry. Intrinsic Markov chains. *Trans. Amer. Math. Soc.* 112 (1964).

[Par22] A. Parzygnat. A functorial characterization of von Neumann entropy. *Cah. Topol. Géom. Différ. Catég.* 63 (2022).

[Per24] P. Perrone. Markov categories and entropy. *IEEE Trans. Inform. Theory* 70 (2024).

[Pet83] K. Petersen. *Ergodic Theory*. Cambridge Univ. Press, 1983.

[PVB24] P. Perrone and R. Van Belle. Convergence of martingales via enriched dagger categories. arXiv:2404.15191 (2024).

[PW25] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge Univ. Press, 2025.

[Ren61] A. Rényi. On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* 1 (1961).

[SBBR13] A. Silva, F. Bonchi, M. Bonsangue and J. Rutten. Generalizing determinization from automata to coalgebras. *Log. Methods Comput. Sci.* 9(1) (2013).

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.* 27 (1948).

[Sha56] C. E. Shannon. The zero error capacity of a noisy channel. *IRE Trans. Inform. Theory* 2 (1956).

[Sin59] Ya. G. Sinai. On the concept of entropy for a dynamical system. *Dokl. Akad. Nauk SSSR* 124 (1959).

[Sin64] Ya. G. Sinai. On a weak isomorphism of transformations with invariant measure. *Mat. Sb.* 63 (1964).

[SP82] M. Smyth and G. Plotkin. The category-theoretic solution of recursive domain equations. *SIAM J. Comput.* 11 (1982).

[SSZ05] I. Sutskover, S. Shamai and J. Ziv. Extremes of information combining. *IEEE Trans. Inform. Theory* 51 (2005).

[Str88] V. Strassen. The asymptotic spectrum of tensors. *J. Reine Angew. Math.* 384 (1988).

[Tak58] K. Takano. On the basic theorems of information theory. *Ann. Inst. Statist. Math.* 9 (1958).

[Tsa88] C. Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *J. Statist. Phys.* 52 (1988).

[Turo7] S. Turgut. Catalytic transformations for bipartite pure states. *J. Phys. A* 40 (2007).

[Tve58] H. Tverberg. A new derivation of the information function. *Math. Scand.* 6 (1958).

[VB23] R. Van Belle. A categorical treatment of the Radon–Nikodym theorem and martingales. arXiv preprint (2023).

[VH94] S. Verdú and T. S. Han. A general formula for channel capacity. *IEEE Trans. Inform. Theory* 40 (1994).

[Vig19] J. P. Vigneaux. *Topology of Statistical Systems: A Cohomological Approach to Information Theory*. Thesis, Université Paris Diderot, 2019.

[Vig20] J. P. Vigneaux. Information structures and their cohomology. *Theory Appl. Categ.* 35 (2020).

[vN32] J. von Neumann. Proof of the quasi-ergodic hypothesis. *Proc. Nat. Acad. Sci. USA* 18 (1932).

[VVS95] S. Vembu, S. Verdú and Y. Steinberg. The source-channel separation theorem revisited. *IEEE Trans. Inform. Theory* 41 (1995).

[Wal82] P. Walters. *An Introduction to Ergodic Theory*. Springer, 1982.

[Wol57] J. Wolfowitz. The coding of messages subject to chance errors. *Illinois J. Math.* 1 (1957).

[Wol78] J. Wolfowitz. *Coding Theorems of Information Theory*, 3rd ed. Springer, 1978.

[WZ73] A. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications: Part I. *IEEE Trans. Inform. Theory* 19 (1973).

[Yeu08] R. Yeung. *Information Theory and Network Coding*. Springer, 2008.

[Zui19] J. Zuiddam. The asymptotic spectrum of graphs and the Shannon capacity. *Combinatorica* 39 (2019).

[ZY98] Z. Zhang and R. Yeung. On characterization of entropy function via information inequalities. *IEEE Trans. Inform. Theory* 44 (1998).