Kybernetika

ERGODIC THEORY, ENTROPY, AND CODING PROBLEMS OF INFORMATION THEORY

ŠTEFAN ŠUJAN

ACADEMIA
PRAHA

Advances in entropy theory of measure theoretic dynamical systems during the period 1970 to 1980 are described, with particular emphasis on ideas and results relevant from the point of view of information theory. The survey is completed by a commented sample of information theoretic papers which are based on recent ideas of ergodic theory. An attempt is made to explain the results from ergodic theory in a language appropriate for an information theorist, and future perspectives of interplay between ergodic and information theories are discussed.

REFERENCES

[1] M. A. Ackoglu, A. del Junco, and M. Rahe: Finitary codes between Markov processes. Z. Wahrsch. verw. Gebiete *47* (1979), 305—314.

[2] R. L. Adler, W. Goodwyn, and B. Weiss: Equivalence of topological Markov shifts. Israel J. Math. *27* (1977), 49—63.

[3] R. L. Adler, A. G. Konheim, and M. H. McAndrew: Topological entropy. Trans. Amer. Math. Soc. *114* (1965), 309—319.

[4] R. L. Adler and B. Marcus: Topological entropy and equivalence of dynamical systems. Memoirs Amer. Math. Soc. *219* (1979).

[5] R. L. Adler and B. Weiss: Similarity of the automorphisms of the torus. Memoirs Amer. Math. Soc. *98* (1970).

[6] V. M. Alekseev: Symbolic Dynamics (in Russian). Math. Institute, AN USSR, Kiev 1976.

[7] V. M. Alekseev and M. V. Jakobson: Symbolic dynamics and hyperbolic dynamical systems (in Russian). Supplement to R. Bowen: Methods of Symbolic Dynamics (in Russian). Mir, Moskva 1979, pp. 196—240.

[8] R. B. Ash: Information Theory. J. Wiley, New York 1965.

[9] T. Berger: Rate Distortion Theory. Prentice Hall, Englewood Cliffs 1971.

[10] T. Berger: Information singular processes. IEEE Trans. Inform. Theory *IT-20* (1975), 502—511..

[11] P. Billingsley: Ergodic Theory and Information. J. Wiley, New York—London—Sydney 1965.

[12] P. Billingsley: Convergence of Probability Measures. J. Wiley, New York—London—Sydney—Toronto 1968.

[13] R. E. Blahut: Computation of channel capacity and rate-distortion functions. IEEE Trans. Inform. Theory *IT-18* (1972), 460—473.

[14] H. Blasbalg and R. van Blerkom: Message compression. IRE Trans. Space Electron. Telemech. 1962, 228—338.

[15] J. R. Blum and D. L. Hanson: On invariant probability measures I. Pacific J. Math. *10* (1960), 1125—1240.

[16] J. R. Blum and D. L. Hanson: On the isomorphism problem for Bernoulli schemes. Bull. Amer. Math. Soc. *63* (1963), 221—223.

[17] R. Bowen: Symbolic dynamics for hyperbolic systems. Amer. J. Math. *95* (1973), 429—459.

[18] R. Bowen: Topological entropy for non-compact sets. Trans. Amer. Math. Soc. *184* (1973), 413—423.

[19] R. Bowen: Smooth partitions of Anosov diffeomorphisms are weak Bernoulli. Israel J. Math. *21* (1975), 95—100.

[20] R. Bowen: Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. (Lecture Notes in Mathematics 470.) Springer-Verlag, Berlin—Heidelberg—New York 1975.

[21] R. Bowen and D. Ruelle: The ergodic theory of Axiom A flows. Invent. Mathematicae *29* (1975), 181—202.

[22] R. C. Bradley, Jr.: On the strong mixing and weak Bernoulli conditions. Z. Wahrsch. verw. Gebiete *51* (1980), 49—54.
[23] A. A. Brudno: Entropy and algorithmic complexity of trajectories of a dynamical system (in Russian). Preprint, VNIISI, Moskva 1980.
[24] E. Coven and M. Paul: Endomorphisms of irreducible shifts of finite type. Math. Systems Theory *8* (1974), 165—175.
[25] L. D. Davisson: Universal noiseless coding. IEEE Trans. Inform. Theory *IT-19* (1973), 783—795.
[26] L. D. Davisson and R. M. Gray: A simplified proof of the sliding-block source coding theorem and its universal extension. Proc. Int. Conf. on Communication, Vol. 2, pp. 34.4.1 — 34.4.5. Toronto, Canada 1978.
[27] A. del Junco and M. Rahe: Finitary codings and weak Bernoulli partitions. Proc. Amer. Math. Soc. *75* (1979), 259—364.
[28] M. Denker: Finite generators for ergodic ,measure-preserving transformations. Z. Wahrsch. verw. Gebiete *29* (1974), 45—55.
[29] M. Denker: Generators and almost topological isomorphisms. Astérisque *59* (1977), 23—35.
[30] M. Denker, C. Grillenberger, and K. Sigmund: Ergodic Theory on Compact Spaces. (Lecture Notes in Mathematics 527.) Springer-Verlag, Berlin—Heidelberg—New York 1976.
[31] M. Denker and M. Keane: Almost topological dynamical systems. Israel J. Math. *34* (1979), 139—160.
[32] M. Denker and M. Keane: Finitary codes and the law of the iterated logarithm. Z. Wahrsch. verw. Gebiete *52* (1980), 321—331.
[33] J. G. Dunham: Abstract alphabet sliding-block entropy compression coding with a fidelity criterion. Ann. Probab. *8* (1980), 1085—1092.
[34] R. Fellgett and W. Parry: Endomorphisms of a Lebesgue space II. Israel J. Math. *21* (1975), 167—172.
[35] B. M. Fitingoff: Optimal coding in case of unknown and changing message statistics (in Russian). Problemy Peredachi Informacii *2* (1966), 3—11.
[36] B. M. Fitingoff: The compression of discrete information (in Russian). Problemy Peredachi Informacii *3* (1967), 28—36.
[37] R. J. Fontana, R. M. Gray, and J. C. Kieffer: Asymptotically mean stationary channels. IEEE Trans. Inform. Theory *IT-27* (1981), 308—316.
[38] N. Friedman and D. S. Ornstein: On the isomorphism of weak Bernoulli transformations. Adv. in Math. *5* (1970), 365—394.
[39] R. G. Gallager: Information Theory and Reliable Communication. J. Wiley, New York 1968.
[40] F. R. Gantmacher: The Theory of Matrices. Vols. I and II. Chelsea, New York 1959.
[41] R. M. Gray: Sliding-block source coding. IEEE Trans. Inform. Theory *IT-21* (1975), 357—368.
[42] R. M. Gray and L. D. Davisson: The ergodic decomposition of stationary discrete random processes. IEEE Trans. Inform. Theory *IT-20* (1974), 625—636.
[43] R. M. Gray and L. D. Davisson: Source coding theorems without the ergodic assumption. IEEE Trans. Inform. Theory *IT-20* (1974), 502—516.
[44] R. M. Gray and J. C. Kieffer: Mutual information rate, distortion, and quantization in metric spaces. IEEE Trans. Inform. Theory *IT-26* (1980), 412—422.
[45] R. M. Gray and J. C. Kieffer: Asymptotically mean stationary measures. Ann. Probab. (1980), 962—973.
[46] R. M. Gray, D. L. Neuhoff, and J. K. Omura: Process definitions of distortion-rate function and source coding theorems. IEEE Trans. Inform. Theory *IT-21* (1975), 524—532.

[47] R. M. Gray, D. L. Neuhoff, and D. S. Ornstein: Non block source coding with a fidelity criterion. Ann. Probab. *3* (1975), 478—491.
[48] R. M. Gray, D. L. Neuhoff, and P. C. Shields: A generalization of Ornstein's $\bar{d}$-distance with applications to information theory. Ann. Probab. *3* (1975), 315—328.
[49] R. M. Gray and D. S. Ornstein: Sliding-block joint source/noisy channel coding theorems. IEEE Trans. Inform. Theory *IT-22* (1976), 683—690.
[50] R. M. Gray and D. S. Ornstein: Block coding for discrete stationary $\bar{d}$-continuous noisy channels. IEEE Trans. Inform. Theory *IT-25* (1979), 292—306.
[51] R. M. Gray, D. S. Ornstein, and R. L. Dobrushin: Block synchronization, sliding-block coding, invulnerable sources, and zero-error codes for discrete noisy channels. Ann. Probab. *8* (1980), 639—674.
[52] C. Grillenberger and U. Krengel: On marginal distributions and isomorphisms of stationary processes. Math. Z. *149*, (1976), 131—154.
[53] B. Hajek: Information-singularity and recoverability of random processes. IEEE Trans. Inform. Theory *IT-28* (1983), 422—429.
[54] P. R. Halmos: Measure Theory. D. Van Nostrand, Princeton N. J. 1950.
[55] P. R. Halmos: Lectures on Ergodic Theory. Chelsea, New York 1953.
[56] G. Hansel and J. P. Raoult: Ergodicité, uniformité et unique ergodicité. Indiana Univ. Math. J. *23* (1973), 221—237.
[57] R. Heim: On the algorithmic foundations of information theory. IEEE Trans. Inform. Theory *IT-25* (1979), 557—566.
[58] R. I. Jewett: The prevalence of uniquely ergodic systems. J. Math. and Mech. *19* (1970), 717—729.
[59] M. Keane: Coding problems in ergodic theory. Proc. Int. Conf. on Math. Physics. Camerino, Italy, 1974.
[60] M. Keane and M. Smorodinsky: A class of finitary codes. Israel J. Math. *26* (1977), 352—371.
[61] M. Keane and M. Smorodinsky: Bernoulli schemes of the same entropy are finitarily isomorphic. Ann. Math. *109* (1979), 397—406.
[62] M. Keane and M. Smorodinsky: The finitary isomorphism theorem for Markov shifts. Bull. (New Series) Amer. Math. Soc. *1* (1979), 436—438.
[63] A. I. Khinchine: Mathematical Foundations of Information Theory. Dover, New York 1957.
[64] J. C. Kieffer: On approximation of stationary measures by periodic and ergodic measures. Ann. Probab. *2* (1974), 530—534.
[65] J. C. Kieffer: A generalized Shannon-McMillan theorem for the action of an amenable group on a probability space. Ann. Probab. *3* (1975), 1031—1037.
[66] J. C. Kieffer: Block coding for an ergodic source relative to a zero-one valued fidelity criterion. IEEE Trans. Inform. Theory *IT-24* (1978), 432—438.
[67] J. C. Kieffer: A unified approach to weak universal source coding. IEEE Trans. Inform. Theory *IT-24* (1978), 674—682.
[68] J. C. Kieffer: On the minimum rate for strong universal block coding of a class of ergodic sources. IEEE Trans. Inform. Theory *IT-26* (1980), 693—702.
[69] J. C. Kieffer: On the transmission of Bernoulli sources over stationary channels. Ann. Probab. *8* (1980), 942—961.
[70] J. C. Kieffer: On coding a stationary process to achieve a given marginal distribution. Ann. Probab. *8* (1980), 131—141.
[71] J. C. Kieffer: Extensions of source coding theorems for block codes to sliding-block codes. IEEE Trans. Inform. Theory *IT-26* (1970), 679—692.

[72] J. C. Kieffer: Stationary coding over stationary channels. Z. Wahrsch. verw. Gebiete *56* (1981), 113—136.

[73] J. C. Kieffer: Block coding for weakly continuous channels. IEEE Trans. Inform. Theory *IT-27* (1981), 721—727.

[74] J. C. Kieffer: Perfect transmission over a discrete memoryless channel requires infinite expected coding time. J. Combin. Inform. System Sci. *5* (1980), 317—322.

[75] J. C. Kieffer: Sliding-block coding for weakly continuous channels. IEEE Trans. Inform. Theory *IT-28* (1982), 2—10.

[76] J. C. Kieffer: Characterizations of $\bar{d}$-total boundedness for classes of B sources. IEEE Trans. Inform. Theory *IT-28* (1982), 26—35.

[77] J. C. Kieffer: On obtaining a stationary process isomorphic to a given process with a desired distribution. Preprint, Univ. of Missouri at Rolla, 1982.

[78] J. C. Kieffer: Generators with prescribed marginals for nonergodic automorphisms. Lecture presented at the 9th Prague Conf. Inform. Theory, Prague, June 1982.

[79] J. C. Kieffer and M. Rahe: Selecting universal partitions in ergodic theory. Ann. Probab. *9* (1981), 705—709.

[80] A. N. Kolmogorov: A new metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces (in Russian). Doklady AN SSSR *119* (1958), 862—864.

[81] A. N. Kolmogorov: The three approaches to the definition of the concept "amount of information" (in Russian). Problemy Peredachi Informacii *5* (1965), 3—7.

[82] I. P. Kornfel'd, Ya. G. Sinai, and S. V. Fomin: Ergodic Theory (in Russian). Nauka, Moskva 1980.

[83] U. Krengel: Recent results on generators in ergodic theory. Trans. 6th Conf. Inform. Theory etc., Academia, Prague 1973, 465—482.

[84] U. Krengel: Discussion of Professor's Ornstein's paper (see [45]). Ann. Probab. *1* (1973).

[85] W. Krieger: On entropy and generators of measure-preserving transformations. Trans. Amer. Math. Soc. *119* (1970), 453—464. Erratum: ibid. *168* (1972), 519.

[86] W. Krieger: On unique ergodicity. Proc. 6th Berkeley Symp. Math. Stat. Prob., Vol. I. University of California Press, Los Angeles 1972, 327—346.

[87] N. Kryloff and N. Bogoliouboff: La théorie générale de la mesure dans son application à l'étude des systémes dynamiques de la mécanique non linéaire. Ann. Math. *38* (1937), 65—113.

[88] A. G. Kushnirenko: On metric invariants of entropy type (in Russian). Uspehi Mat. Nauk *22* (1967), 57—65.

[89] A. Leon-Garcia, L. D. Davisson, and D. L. Neuhoff: New results on coding of stationary nonergodic sources. IEEE Trans. Inform. Theory *IT-25* (1979), 137—144.

[90] K. M. Mackenthun and M. B. Pursley: Variable-rate universal block source coding subject to a fidelity criterion. IEEE Trans. Inform. Theory *IT-24* (1978), 349—360.

[91] B. Marcus: Factors and extensions of full shifts. Monatsh. Math. *88* (1979), 239—247.

[92] B. McMillan: The basic theorems of information theory. Ann. Math. Statist. *24* (1953), 196—219.

[93] L. D. Meshalkin: One particular case of isomorphism of Bernoulli schemes (in Russian). Doklady AN SSR *141* (1959), 41—44.

[94] M. Morse: Symbolic Dynamics (lecture notes). Institute for Advanced Study, Princeton 1966.

[95] D. L. Neuhoff, R. M. Gray, and L. D. Davisson: Fixed-rate universal block source coding with a fidelity criterion. IEEE Trans. Inform. Theory *IT-22* (1975), 524—532.

[96] D. L. Neuhoff and P. C. Shields: Fixed-rate universal codes for Markov sources. IEEE Trans. Inform. Theory *IT-24* (1978), 360—367.

[97] D. L. Neuhoff and P. C. Shields: Channels with almost finite memory. IEEE Trans. Inform. Theory *IT-25* (1979), 440—447.

[98] D. L. Neuhoff and P. C. Shields: Indecomposable finite state channels and primitive approximation. IEEE Trans. Inform. Theory *IT-28* (1982), 11—18.

[99] D. S. Ornstein: Bernoulli shifts with the same entropy are isomorphic. Adv. in Math. *4* (1970), 338—352.

[100] D. S. Ornstein: Factors of Bernoulli shifts are Bernoulli. Adv. in Math. *5* (1970), 349—364.

[101] D. S. Ornstein: Imbedding Bernoulli shifts in flows. Contributions to Ergodic Theory and Probability. (Lecture Notes in Mathematics 160.) Springer-Verlag, Berlin—Heidelberg —New York 1970, 178—218.

[102] D. S. Ornstein: An application of ergodic theory to probability theory. Ann. Probab. *1* (1973), 43—65.

[103] D. S. Ornstein: Ergodic Theory, Randomness, and Dynamical Systems. Yale Univ. Press, New Haven—London 1974.

[104] D. S. Ornstein and B. Weiss: Ergodic theory of amenable group actions. I: The Rohlin lemma. Bull. (New Series) Amer. Math. Soc. *2* (1980), 161—164.

[105] J. C. Oxtoby: Ergodic sets. Bull. Amer. Math. Soc. *58* (1952), 116—136.

[106] W. Parry: Intrinsic Markov chains. Trans. Amer. Math. Soc. *112* (1964), 55—66.

[107] W. Parry: Entropy and Generators in Ergodic Theory. W. A. Benjamin, New York—Amsterdam 1969.

[108] W. Parry: A finitary classification of topological Markov chains and sofic systems. Bull. London Math. Soc. *9* (1977), 86—92.

[109] W. Parry: Endomorphisms of a Lebesgue space III. Israel J. Math. *21* (1975), 167—172.

[110] W. Parry: The information cocycle and ε-bounded codes. Israel J. Math. *29* (1978), 205 to 230.

[111] W. Parry: An information obstruction to finite expected coding length. Ergodic Theory. Proceedings, Oberwolfach. (Lecture Notes in Mathematics 729.) Springer-Verlag, Berlin—Heidelberg—New York 1979, 163—168.

[112] W. Parry: Finitary isomorphisms with finite expected code-length. Bull. London Math. Soc. *11* (1979), 170 — 176.

[113] W. Parry: Topics in Ergodic Theory. (Cambridge Tracts in Mathematics 75.) Cambridge Univ. Press, Cambridge 1981.

[114] W. Parry and K. Schmidt: A note on cocycles of unitary representations. Proc. Amer. Math. Soc. *55* (1976), 185—190.

[115] W. Parry and S. Tuncel: On the classification of Markov chains by finite equivalence. Preprint. Warwick Univ., Math. Institute, March 1981.

[116] K. R. Parthasarathy: Probability Measures on Metric Spaces. Academic Press, New York 1967.

[117] M. B. Pursley and L. D. Davisson: Variable-rate coding for nonergodic sources and classes of sources subject to a fidelity constraint. IEEE Trans. Inform. Theory *IT-22* (1976), 324—337.

[118] M. B. Pursley and K. M. Mackenthun: Variable-rate coding for classes of sources with generalized alphabets. IEEE Trans. Inform. Theory *IT-23* (1977), 592—597.

[119] V. A. Rohlin: On basic concepts of measure theory (in Russian). Mat. Sbornik *67* (1949), 107—150.

[120] V.A. Rohlin: Selected problems of the metric theory of dynamical systems (in Russian). Uspehi Mat. Nauk *30* (1949), 57—128.

[121] V. A. Rohlin: On the decomposition of a dynamical system into transitive components (in Russian). Mat. Sbornik *67* (1949), 235—249.

[122] V. A. Rohlin and Ya. G. Sinai: Construction and properties of invariant measurable partitions (in Russian). Doklady AN SSSR *141* (1961), 1038—1041.
[123] D. J. Rudolph: A characterization of those processes finitarily isomorphic to a Bernoulli shift. Ergodic Theory and Dynamical Systems I. Progress in Mathematics, Vol. 10. Birkhäuser, Boston, Mass. 1981, 1—64.
[124] D. J. Sakrison: The rate distortion function of a class of sources. Inform. and Control *15* (1969), 165—195.
[125] C. E. Shannon: A mathematical theory of communication. Bell. System Techn. J. *27* (1948), 379—432, 623—656.
[126] C. E. Shannon: Coding theorems for discrete source with a fidelity criterion. IRE Nat. Conv. Rec., part 4 (1959), 142—163.
[127] C. E. Shannon: The zero-error capacity of a noisy channel. IRE Trans. *3* (1056), 8—32.
[128] P. C. Shields: The Theory of Bernoulli Shifts. Univ. of Chicago Press, Chicago 1973.
[129] P. C. Shields: Stationary coding of processes. IEEE Trans. Inform. Theory *IT-25* (1979), 283—291.
[130] P. C. Shields: Almost block independence. Z. Wahrsch. verw. Gebiete *49* (1979), 119—123.
[131] P. C. Shields and D. L. Neuhoff: Block and sliding-block source coding. IEEE Trans. Inform. Theory *IT-23* (1977), 211—215.
[132] K. Sigmund: On the prevalence of zero entropy. Israel J. Math. *10* (1971), 281—288.
[133] Ya. G. Sinai: On the notion of entropy of a dynamical system (in Russian). Doklady AN SSSR *124* (1959), 768—771.
[134] Ya. G. Sinai: On weak isomorphism of transformations with an invariant measure (in Russian). Mat. Sbornik *63* (1964), 23—42.
[135] S. Smale: Differentiable dynamical systems. Bull. Amer. Math. Soc. *73* (1967), 747—817.
[136] M. Smorodinsky: Ergodic Theory, Entropy. (Lecture Notes in Mathematics 214.) Springer-Verlag, Berlin—Heidelberg—New York 1971.
[137] M. Smorodinsky: A partition on a Bernoulli shift which is not weak Bernoulli. Math. Systems Theory *5* (1971), 201—203.
[138] Š. Šujan: Generators of an abelian group of invertible measure-preserving transformations. Monatsh. Math. *90* (1980), 68—79.
[139] Š. Šujan: Epsilon-rates, epsilon-quantiles, and group coding theorems for finitely additive information sources. Kybernetika *16* (1980), 105—119.
[140] Š. Šujan: Existence of asymptotic rate for asymptotically mean stationary sources with countable alphabets. Trans. 3rd Czechosl.-Soviet-Hung. Seminar on Information Theory. ÚTIA ČSAV, Prague 1980, 201—207.
[141] Š. Šujan: Channels with additive asymptotically mean stationary noise. Kybernetika *17* (1981), 1—15.
[142] Š. Šujan: On the capacity of asymptotically mean stationary channels. Kybernetika *17* (1981), 122—233.
[143] Š. Šujan: Continuity and quantization of channels with infinite alphabets. Kybernetika *17* (1981), 465—478.
[144] Š. Šujan: Block transmissibility and quantization. Probability and Statistical Inference (W. Grossmann et al., eds.), D. Reidel, Dordrecht—Boston—London 1982, 361—371.
[145] Š. Šujan: A local structure of stationary perfectly noiseless codes between stationary nonergodic sources. I: General considerations. Kybernetika *18* (1982), 361—376.
[146] Š. Šujan: A local structure ... II: Applications. Kybernetika *18* (1982), 465—484.
[147] Š. Šujan: Codes in ergodic theory and information: Some examples. Proc. Conf. Ergodic Theory and Related Topics, Akademie-Verlag, Berlin 1982 (to appear).
[148] Š. Šujan: Finite generators for amenable group actions (submitted).
[149] J. - P. Thouvenot: Quelques proprietes des systémes dynamiques qui se decomposent

en un produit de deux systémes dont l'un est un schema de Bernoulli. Israel J. Math. *21* (1975), 178—207.
[150] S. Tuncel: Conditional pressure and coding. Israel J. Math. *39* (1981), 101—112.
[151] P. Walters: Ergodic Theory. Introductory Lectures. (Lectures Notes in Mathematics 458.) Springer-Verlag, Berlin—Heidelberg—New York 1975.
[152] B. Weiss: The isomorphism problem in ergodic theory. Bull. Amer. Math. Soc. *78* (1972), 668—684.
[153] K. Winkelbauer: On discrete information sources. Trans. 3rd Prague Conf. Inform. Theory etc., NČSAV, Prague 1964, 765—830.
[154] K. Winkelbauer: On the asymptotic rate of nonergodic information sources. Kybernetika *6* (1970), 128—148.
[155] K. Winkelbauer: On the existence of finite generators for invertible measure-preserving transformations. Comment. Math. Univ. Carolinae *18* (1977), 789—812.
[156] J. Wolfowitz: Coding Theorems of Information Theory. 2nd ed. Springer-Verlag, New York 1964.
[157] G. M. Zaslavskij: On the isomorphism problem for stationary processes (in Russian). Teoria veroyatnostej i primen. *9* (1964), 241—298.
[158] J. Ziv: Coding of sources with unknown statistics. Part I: Probability of encoding error; Part II: Distortion relative to a fidelity criterion. IEEE Trans. Inform. Theory *IT-18* (1972), 384—394.
[159] J. Ziv: Coding theorems for individual sequences. IEEE Trans. Inform. Theory *IT-24* (1978), 405—413.
[160] J. Ziv and A. Lempel: Compression of individual sequences via variable-rate coding. IEEE Trans. Inform. Theory *IT-24* (1978), 530—536.

INTRODUCTION

The aim of this paper is to give a survey on developments in entropy methods of ergodic theory of measure theoretic dynamical systems during the period 1970 to 1980. The reason is that, unlike classical notions and results of ergodic theory (ergodicity, mixing properties, ergodic theorems, etc.), recent ideas and constructions in ergodic theory do not constitute a standard part of an information theorist's background. On the other hand, these ideas have been very fruitfully employed in information theory and, based on up-to-date state, one can expect further progress as to the interplay between ergodic and information theories.

A main problem connected with applications of ergodic theory to information theory consists of different languagues employed and different aims followed by the two theories. Our intention thus will be (a) to point out that the aims are not so different as generally judged, and (b) to explain relevant results of ergodic theory in a language appropriate for an information theorist.

Of course, as any survey also the present one is influenced by author's taste and scientific interests, however, it is hoped that the choice of contributions will be sufficiently representative. Also, readers should not expect investigations on technical details and should refer to original papers which are thoroughly indicated. On the other hand, we attempted to explain the essence of most important ideas.

PART I: PRELIMINARIES

1. Some Historical Remarks

Until Kolmogorov [80] and Sinai [133] introduced the concept of entropy into ergodic theory, the problems of ergodic theory were formulated and solved mainly within the spectral theory of induced unitary operators on L^2 spaces (see [151] for a survey). In particular, spectral invariants were not sharp enough to distinguish between such simple dynamical systems as Bernoulli schemes. Kolmogorov and Sinai showed that entropy is an isomorphism invariant, and in case of Bernoulli schemes even an easily computable one. Thus, two Bernoulli schemes with different entropies cannot be isomorphic.

Of course, that did not solve the more interesting part of the problem whether entropy is a complete invariant for Bernoulli schemes; that is, whether two Bernoulli schemes of the same entropy are isomorphic or not. The problem turned out to be of extreme complexity, nevertheless, considerable progress had been achieved in entropy theory per se. Usually, the obtained results were not coding results (an outstanding result of this type is the characterization of K-automorphisms by the property of having completely positive entropy; see [122]), but coding results appeared very early. E. g., Meshalkin [93] constructed invertible stationary codes between

some particular Bernoulli schemes (see also [16]) and Sinai [134] obtained the so-called weak isomorphism theorem, according to which two Bernoulli schemes of the same entropy are stationary codings of each other.

Real beginnings of systematic investigations on stationary codes are due to Ornstein who successfully solved the isomorphism problem for Bernoulli schemes [99]. At nearly the same time Krieger solved another longstanding problem of ergodic theory, the solution of which is now known as Krieger's finite generator theorem [85]. Smorodinsky [136] observed that Krieger's theorem also can be obtained with the aid of Ornstein's coding technique. Subsequently, both Ornstein's and Krieger's results have been improved on in many respects.

Importance of Ornstein's technique was not immediately recognized by specialists in information theory, mainly because of a very different language. Although applications of Ornstein's result to information theory were foreseen (indeed, in a hypothetical form; see [11]), it turned out that it is not the result itself but rather three main ideas of Ornstein's construction which bear relevance to coding problems of information theory:

(a) the idea of construction of a good stationary code from a good block code,
(b) the idea of getting a converging sequence of ever better stationary codes (this amounts to a method of making slight change in the structure of a good stationary code in order to get a much better one, and has no counterpart within the traditional block coding approach), and
(c) a new type of approximation arguments based on a new type of distance function, $\bar{d}$-distance, between stationary and ergodic processes.

Ornstein's theory is one of the major achievements of entropy methods in ergodic theory since the end of sixties. That is why we shall devote a good deal of our survey to problems connected with it. But first let us collect some basic concepts.

2. Basic Concepts

Let us briefly comment on abstract setting of *measure theoretic dynamical systems*. By definition, this is a quadruple $(\Omega, \mathscr{F}, \mu, T)$, where $(\Omega, \mathscr{F}, \mu)$ is a probability space and $T: \Omega \to \Omega$ is a measurable (i.e. $T^{-1}\mathscr{F} \subset \mathscr{F}$) and measure-preserving (i.e., $\mu T^{-1} = \mu$) map. We shall assume that T is *invertible* (sometimes it is called an automorphism); that is, T^{-1} is defined, measurable, and measure-preserving, too. Two dynamical systems $(\Omega, \mathscr{F}, \mu, T)$ and $(\Omega', \mathscr{F}', \mu', T')$ are said to be (mod 0) *isomorphic* if there exist sets $E \in \mathscr{F}$, $E' \in \mathscr{F}'$, and a map $\varphi : E \to E'$ such that

(i) $\mu(E) = \mu'(E') = 1$ and φ is bijective,
(ii) $T^{-1}E \subset E$, $(T')^{-1}E' \subset E'$,
(iii) if $F \subset E$, then $F \in \mathscr{F}$ if and only if $\varphi F \in \mathscr{F}'$; in this case $\mu(F) = \mu'(\varphi F)$, and
(iv) $\varphi(T\omega) = T'(\varphi\omega)$ for all $\omega \in E$.

There appear two natural problems related to the notion of isomorphism. There are classification problems dealing with the question whether two dynamical systems are isomorphic or not. Then there are representation problems which deal with the question when a dynamical system can be (isomorphically) represented by another one (which is supposed either to possess a simpler structure or to possess some additional properties).

Of course, one cannot expect significant results without imposing some additional structure on $(\Omega, \mathscr{F}, \mu, T)$. A first natural restriction is to consider only "sufficiently nice" underlying probability spaces. Usually, it is entirely sufficient, from the point of view of potential fields of applications, to consider $(\Omega, \mathscr{F}, \mu)$ as a standard Borel space, that is, to assume that Ω is a Borel subset of a complete separable metric space, and μ is defined on the σ-field $\mathscr{F}$ of all Borel subsets of Ω; cf. [116]. In ergodic theory it is commonly accepted to work with slightly less general Lebesgue spaces [119]. These are spaces isomorphic with the probability space $(I, \mathscr{L}, \lambda)$ of the unit interval I equipped with the σ-field of all Lebesgue measurable sets and the usual Lebesgue measure (the non-atomic or, continuous case) or, with a part of unit interval with Lebesgue measure plus an at most countable set of isolated points each carrying a positive mass. We do not dwell on details, for we shall deal mainly with spaces having a more specific structure.

To this end let us introduce the notion of a generator. Let $(\Omega, \mathscr{F}, \mu, T)$ be an invertible dynamical system. We call T *aperiodic* if for any $N \geqq 1$,

$$\mu\{\omega \in \Omega : T^N\omega = \omega\} = 0$$

(this makes sense in "nice" spaces as above, for in such spaces all singletons are measurable; in general, the concept of aperiodicity must be defined in a different way [83]). A countable partition $\zeta = (C_1, C_2, \ldots)$ of Ω into measurable sets is said to be a *generator* (relative to (T, μ)) if

$$\sigma\Big(\bigcup_{i \in \mathbf{Z}} T^i \sigma(\zeta)\Big) = \mathscr{F} \bmod 0\,,$$

where $\mathbf{Z} = \{\ldots, -1, 0, 1, \ldots\}$. Here $\sigma(\mathscr{E})$ stands for the σ-field generated by $\mathscr{E}$, and two σ-fields are considered as mod 0 identical if they give rise (under μ) to algebraically isomorphic measure algebras (see [54]). The proof of the following assertion can be found, e.g., in [16]:

Proposition 1. Let T be an invertible aperiodic transformation of a Lebesgue probability space $(\Omega, \mathscr{F}, \mu)$. Then there exists a generator relative to (T, μ).

Let ζ be a generator from Proposition 1. Define a map $\tau_\zeta : \Omega \to \zeta^{\mathbf{Z}}$ by the properties that $\tau_\zeta\omega = (x_i; i \in \mathbf{Z})$, where $x_i = C \in \zeta$ if and only if $T^i\omega \in C$. Let $T_\zeta : \zeta^{\mathbf{Z}} \to \zeta^{\mathbf{Z}}$ denote the shift, i.e., $(T_\zeta x)_i = x_{i+1}$, and let $\mathscr{C}^{\mathbf{Z}}$ denote the usual product σ-field on $\zeta^{\mathbf{Z}}$ induced by the σ-field $\mathscr{C} = \{D : D \subset \zeta\}$. One can easily check that

(i) τ_ζ is measurable,

(ii) τ_ζ is almost everywhere injective (in fact, it can be non-injective only on periodic points of T which have probability zero), and

(iii) $\tau_\zeta \circ T = T_\zeta \circ \tau_\zeta$ almost everywhere.

Moreover, using standard arguments based on Baire category theorem (see, e.g., Appendix A of [103]) one can deduce from the fact that ζ is a generator that τ_ζ is isomorphism between $(\Omega, \mathscr{F}, \mu, T)$ and $(\zeta^{\mathbf{Z}}, \mathscr{C}^{\mathbf{Z}}, \mu\tau_\zeta^{-1}, T_\zeta)$.

Proposition 1 suggests that it is possible to work with structures induced by measurable partitions for very general measure theoretic dynamical systems without any essential loss of information about their statistical properties.

3. Shift Spaces

Throughout the rest of the paper A will denote, unless otherwise stated, a countable discrete space so that its Borel subsets are $\mathscr{A} = \{E : E \subset A\}$. We let $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}})$ denote the measurable space of all doubly infinite sequences $x = (x_i; i \in \mathbf{Z})$ with $x_i \in A$. As well-known and easy to check, $A^{\mathbf{Z}}$ is a complete separable metric space (compact if $\|A\| < \infty$; $\|A\| = \text{card}(A)$), and $\mathscr{A}^{\mathbf{Z}}$ coincides with the σ-field of all Borel subsets of $A^{\mathbf{Z}}$. We let $T_A : A^{\mathbf{Z}} \to A^{\mathbf{Z}}$ denote the shift-transformation (see Section 2) and $X = (X_i; i \in \mathbf{Z})$ will stand for the sequence of one-dimensional projections: $X_i(x) = x_i$ for $x \in A^{\mathbf{Z}}$, $i \in \mathbf{Z}$. If μ is a T_A-invariant probability measure on $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}})$ (in symbols, $\mu \in \mathscr{M}(A)$) then the whole structure will be abbreviated as $[A, \mu]$ or $[A, \mu, X]$ and called a *stationary source*. The set A is called its alphabet. Sometimes we shall indicate that X corresponds to μ by writting that $\text{dist}(X) = \mu$. A stationary source $[A, \mu]$ is said to be *ergodic* if μ is T_A-ergodic (i.e., if T_A is an ergodic transformation of $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu)$), that is, if $\mu(E) \in \{0, 1\}$ for any event $E \in \mathscr{I}(A) = \{F \in \mathscr{A}^{\mathbf{Z}} : T_A F = F\}$. Let $\mathscr{E}(A) \subset \mathscr{M}(A)$ denote the set of all T_A-ergodic measures in $\mathscr{M}(A)$. A stationary source $[A, \mu]$ is said to be *aperiodic* if the shift T_A is aperiodic (see the preceding section). Observe that $[A, \mu]$ is aperiodic if and only if μ is non-atomic; that is, if $\mu\{x\} = 0$ for all $x \in A^{\mathbf{Z}}$.

4. Coding Structures

Let $\hat{A}$ be another countable discrete space and $\hat{\mathscr{A}} = \{\hat{E} : \hat{E} \subset \hat{A}\}$. Any measurable map $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is said to be a *code*. The code $\bar{\Phi}$ is called

— a *block code* if there is $N \in \mathbf{N} = \{1, 2, \ldots\}$ (called the order of $\bar{\Phi}$) and a map $\Phi : A^N \to \hat{A}^N$ such that

$$(\bar{\Phi}x)_1^N = \Phi(x_1^N), \quad \bar{\Phi} \circ T_A^N = T_{\hat{A}}^N \circ \bar{\Phi}$$

(here $x_M^N = (x_M, \ldots, x_N$ if $M, N \in \mathbf{Z}$, $M \leqq N$. In what follows we shall use also the notation $x^N = x_0^{N-1}$; it is often more convenient to start at time zero.);

– a *sliding-block code* if there is an $N \in \mathbf{N}$ (called the order of $\bar{\Phi}$) and a map $\Phi : A^{2N+1} \to \hat{A}$ such that

$$(\bar{\Phi}x)_0 = \Phi(x_{-N}^N), \quad \bar{\Phi} \circ T_A = T_{\hat{A}} \circ \bar{\Phi};$$

– an *infinite code* if there is a measurable map $\Phi : A^{\mathbf{Z}} \to \hat{A}$ such that

$$(\bar{\Phi}x)_i = \Phi(T_A^i x), \quad x \in A^{\mathbf{Z}}, \quad i \in \mathbf{Z}.$$

Readers acquainted with coding techniques of information theory observed that our definition of a block code differs from the usual one. The main difference is that $\hat{A}$ is allowed to be infinite. As we shall see later, however, in relevant coding problems in information theory either $\hat{A}$ will be finite or at least $\|\Phi(A^N)\| < \infty$. In both cases it is easy to see that the present definition is equivalent to the usual one in terms of code books. We shall return to this point later.

By definition, sliding-block codes as well as infinite codes are stationary in the sense that $\bar{\Phi}X$ is a stationary process if X was (in other words, we have $\mu\bar{\Phi}^{-1} \in \mathcal{M}(\hat{A})$ whenever $\mu \in \mathcal{M}(A)$), whereas a block code of order N is only N-stationary. Conversely, any stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is determined by a measurable map $\Phi : A^{\mathbf{Z}} \to \hat{A}$ as above.

If $\bar{\Phi}$ is a stationary code then

$$\zeta_{\Phi} = \{\Phi^{-1}\{\hat{a}\} : \hat{a} \in \hat{A}\}$$

is a countable measurable partition of $A^{\mathbf{Z}}$. Conversely, if ζ is a countable measurable partition of $A^{\mathbf{Z}}$ then the formula

$$(\bar{\Phi}_\zeta x)_i = C \quad \text{iff} \quad T_A^i x \in C, \quad C \in \zeta$$

defines a stationary code $\bar{\Phi}_\zeta : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ ($\hat{A} = \zeta$) such that $\zeta_{\Phi_\zeta} = \zeta$. Observe that if $\bar{\Phi}$ is a sliding-block code then ζ_Φ partitions $A^{\mathbf{Z}}$ into cylinders of some fixed length.

Thus, we can speak either about partitions or about stationary codes. Next we use this fact to define several ergodic theoretic concepts for sources. Suppose $[A, \mu, X]$ is a stationary source. A stationary source $[\hat{A}, \nu, \hat{X}]$ is said to be a *factor* of $[A, \mu, X]$ if there exists a stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ such that $\hat{X} = \bar{\Phi}X$, i.e. $\nu = \mu\bar{\Phi}^{-1}$. Call $[A, \mu, X]$ independent and identically distributed (IID) if μ is a product measure. Thus, an IID source is the same as a stationary memoryless source. Any factor of an IID source is said to be a *Bernoulli* source. It is clear that a Bernoulli source can have memory (in fact, the memory is inserted by the code). We shall see that, in general Bernoulli sources can be characterized by a sufficiently fast decrease of memory effects.

In what follows we shall use the clauses like factor, IID, Bernoulli also for the processes which correspond to sources. Thus, e.g., X is said to be Bernoulli if dist $(X) = \mu$ and $[A, \mu]$ is a Bernoulli source.

Let X be a stationary process with $\mathrm{dist}(X) = \mu$. A stationary coding $\hat{X} = \bar{\Phi}(X)$ of X is called *invertible* if $\bar{\Phi}$ is μ – a.e. injective. It follows (see, e.g., [103], Appendix A or [128]) that there is a measurable map $\Psi : A^{\mathbf{Z}} \to \hat{A}$ such that

$$\mathrm{Prob}_{\mu}\left[X_0 \neq (\bar{\Psi}(\bar{\Phi}X))_0\right] = 0 .$$

This means there is a stationary code $\bar{\Psi} : \hat{A}^{\mathbf{Z}} \to A^{\mathbf{Z}}$ by means of which we can recover X from $\bar{\Phi}X$ without any error so that we shall call an invertible code also a *perfectly noiseless* code. It is clear that it is entirely sufficient that $\bar{\Phi}$ be defined and stationary only μ-a.e. Hence, a perfectly noiseless code $\bar{\Phi}$ is but an isomorphism between the sources $[A, \mu, X]$ and $[\hat{A}, \mu\bar{\Phi}^{-1}, \bar{\Phi}X]$.

Proposition 2. Let $[A, \mu, X]$ be a stationary source. Then

(a) if $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is a stationary code such that either $\|\hat{A}\| < \infty$ or $\|\Phi(A^{\mathbf{Z}})\| < \infty$, then for any $\varepsilon > 0$ there is a sliding-block code $\bar{\Psi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ of order depending on ε such that

$$\mathrm{Prob}_{\mu}\left[(\bar{\Phi}X)_0 \neq (\bar{\Psi}X)_0\right] \leqq \varepsilon ;$$

(b) if $[A, \mu, X]$ is ergodic then so is $[\hat{A}, \mu\bar{\Phi}^{-1}, \bar{\Phi}X]$ for any stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$; and

(c) a stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is perfectly noiseless if and only if ζ_{Φ} is a generator (relative to (T_A, μ)).

Part (a) is Theorem 3.1 of [43] and follows from a simple approximation argument to the effect that any measure on $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}})$ can be approximated by values it takes on cylinder sets [54]. Part (b) is simple for sliding-block codes; for infinite codes combine the latter result with part (a). Part (c) is again the result of a standard application of the Baire category argument.

Part (c) implies that sliding-block codes cannot be invertible in general (for this will entail that $\mathscr{A}^{\mathbf{Z}}$ coincides mod 0 with the σ-field of all events which depend only on a fixed finite number of coordinates). On the other hand, there is a notion of code in certain sense lying inbetween sliding-block and infinite codes.

Let $[A, \mu, X]$ be a stationary source. A stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is said to be *finitary* if for μ-a.a. $x \in A^{\mathbf{Z}}$ there exist integers $q = q(x)$ and $r = r(x)$ with $q \leqq r$ satisfying the following condition: if $x' \in A^{\mathbf{Z}}$, $\bar{\Phi}x'$ is defined, and $x_i = x'_i$ for $q \leqq i \leqq r$ then $(\bar{\Phi}x)_0 = (\bar{\Phi}x')_0$. If $\bar{\Phi}$ is invertible and $\bar{\Phi}^{-1}$ is also finitary, we call $\bar{\Phi}$ a *finitary isomorphism* (or, a finitary perfectly noiseless code). It follows from the definition of the product topology in $A^{\mathbf{Z}}$ that a stationary code $\bar{\Phi}$ is finitary if and only if $\bar{\Phi}$ is a.e. continuous. That is why finitary isomorphisms were called almost topological by Keane [59] and Denker [29]. Since

$$(\bar{\Phi}x)_i = \Phi(T_A^i x) ; \quad x \in A^{\mathbf{Z}} , \quad i \in \mathbf{Z}$$

we can imagine a stationary code $\bar{\Phi}$ as a sequential coding. In order to determine $\hat{x} = \bar{\Phi}x$ we have to look at the coordinates $x_{-N}, \ldots, x_N$, $N = 0, 1, 2, \ldots$, until we find a sequence such that

$$\{x' \in A^{\mathbf{Z}} : x'_i = x_i,\ |i| \leqq N\} \subset \bar{\Phi}^{-1}\{\hat{x} \in \hat{A}^{\mathbf{Z}} : \hat{x}_0 = \hat{a}\}$$

for some $\hat{a} \in \hat{A}$. If $\bar{\Phi}$ is a.e. continuous then this will happen at some finite N (depending on x) for a.a. x. We then put $(\bar{\Phi}x)_0 = \hat{a}$ and determine $(\bar{\Phi}x)_i$ for $i \neq 0$ by shifting the procedure. In other words, a finitary code is a sequential coding procedure whose stopping time is finite with probability one [31].

Put, more symmetrically,

$$L(\bar{\Phi}, M; x) = \inf\{N \geqq 0 : (x' \in A^{\mathbf{Z}}) \,\&\, (x'^{N}_{-N} = x^{N}_{-N}) \text{ imply } (\bar{\Phi}x)^{M}_{-M} = (\bar{\Phi}x')^{M}_{-M}\}\,.$$

The problem of finding $(\bar{\Phi}x)_0 \in \hat{A}$ for $x \in A^{\mathbf{Z}}$ satisfying $L(\bar{\Phi}, 0; x) < \infty$ can be visualized by an infinite tree [74]. The tree consists of a zeroth order node, at least one node of order i ($i \geqq 1$) and at most one branch connecting each node of order i and each node of order $i + 1$ ($i \geqq 0$). If there is no branch from a given node to any node of the next order, we label that by a letter from $\hat{A}$. Given $x \in A^{\mathbf{Z}}$, one looks at (x_{-1}, x_0, x_1) and this determines a branch to certain first order node. If there is no branch to second order nodes, then $L(\bar{\Phi}, 0; x) = 1$, and $(\bar{\Phi}x)_0$ is the label of that node. Otherwise, one passes to $(x_{-2}, \ldots, x_2)$ and repeats the procedure.

Suppose it takes one time unit to pass over any branch. Then $L(\bar{\Phi}, 0; x)$ is the time required to code x into $(\bar{\Phi}x)_0$. The time required to obtain the symbols $(\bar{\Phi}x)_0, \ldots, \ldots, (\bar{\Phi}x)_{N-1}$ is thus

$$\sum_{j=0}^{N-1} L(\bar{\Phi}, 0; T_A^j x)\,.$$

If $[A, \mu, X]$ is ergodic and $\mathsf{E}_\mu\, L(\bar{\Phi}, 0; X) < \infty$ then

$$\lim_{N\to\infty} N^{-1} \sum_{j=0}^{N-1} L(\bar{\Phi}, 0; T_A^j x) = \mathsf{E}_\mu\, L(\bar{\Phi}, 0; X)$$

for μ-a.a. $x \in A^{\mathbf{Z}}$. Thus, a code $\bar{\Phi}$ with finite expectation $\mathsf{E}_\mu\, L(\bar{\Phi}, 0; X)$ is practical in the sense that the time required to get N successive reproduction symbols approaches infinity no faster than linearly in N. Keane and Smorodinsky [60] (see also [1]) developed a construction of such practical finitary codes between IID sources. Later, they extended their construction (see [61, 62]) in order to get invertible codes. However, the practical aspect is lost in the sense that the constructed finitary isomorphism $\bar{\Phi}$ is such that either $\bar{\Phi}$ itself or $\bar{\Phi}^{-1}$ have infinite expected code length. Unfortunately, this is not a consequence of their coding technique but rather a typical occurrence as clarified by Parry [110–112]. We shall turn to that problem later, and now we give definitions of various types of codes according to the behaviour of the length function $L(\bar{\Phi}, M; \cdot)$.

So, let $[A, \mu, X]$ and $[\hat{A}, \nu, \hat{X}]$ be two stationary sources and $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ a finitary isomorphism between them. $\bar{\Phi}$ is said to have *finite expected code length*, if, for any $M \geqq 0$, we have that

$$\mathsf{E}_{\mu} L(\bar{\Phi}, M; X) < \infty , \quad \mathsf{E}_{\nu} L(\bar{\Phi}^{-1}, M; \hat{X}) < \infty .$$

In light of the above remark we have to find some weaker but still desirable property of finitary isomorphisms. Bowen [19] introduced a weaker notion. A finitary code $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is said to be *ε-bounded* if there is a $K = K(\varepsilon)$ such that for any $M \geqq 0$, $L(\bar{\Phi}, M; X) \leqq M + K$ everywhere except a set of measure less than ε. A code $\bar{\Phi}$ is said to be *bounded* if it is ε-bounded for all $\varepsilon > 0$. Bounded and ε-bounded isomorphisms are defined in a straightforward manner. del Junco and Rahe [27] showed that bounded isomorphism is a weaker concept than an isomorphism with finite expected code length.

Surprisingly, Parry's arguments apply equally well to bounded isomorphisms and even to ε-bounded ones. Thus, typically these kinds of isomorphisms are excluded, too. On the other hand, there is a prominent example of a bounded isomorphism namely, the code designed by Adler and Weiss [5] in order to classify toral auto-, morphisms. Thus, it is equally important to know which properties are responsible for the absence of bounded isomorphisms.

Nevertheless, (non-invertible) codes with finite expected code length are also of interest for their existence implies many interesting and desirable properties of the encoded process (like central limit theorems, invariance principles, laws of the iterated logarithm, etc., see [31, 32]). We shall report on these results later.

PART II: ROHLIN'S LEMMA AND ORNSTEIN'S CODING TECHNIQUE

5. Classical Formulations of Rohlin's Lemma

A key to many constructions in ergodic theory is the fundamental Rohlin's lemma [120] which gives a simple geometric picture of actions of measure-preserving transformations (the geometric aspects are explained in [128]). The proof of Rohlin's lemma can be found, e.g., in [55].

Theorem 3. (Rohlin's lemma). Let T be an invertible aperiodic transformation of a Lebesgue space $(\Omega, \mathscr{F}, \mu)$. For any $\varepsilon > 0$ and any $N \in \mathbf{N}$ there exists a set $E \in \mathscr{F}$ such that the sets $E, TE, \ldots, T^{N-1}E$ are pairwise disjoint and

$$\mu\left(\bigcup_{j=0}^{N-1} T^j E\right) \geqq 1 - \varepsilon .$$

If T is non-ergodic we can decompose it into ergodic components $(T_\omega; \omega \in \Omega^*)$, where $\Omega^* \subset \Omega$ is an invariant event such that $\mu(\Omega^*) = 1$ for any T-invariant probability measure μ. The transformations T_ω can be defined as restrictions of T to supports of distinct ergodic components μ_ω of μ (cf. [105, 153, 42]). It is an easy consequence of the theory of regular conditional probabilities that

$$\mu\{\omega \in \Omega^* : \mu_\omega = \mu(\cdot \mid \mathscr{I}_T)(\omega)\} = 1$$

for any invariant probability measure μ on $(\Omega, \mathscr{F})$, where $\mu(\cdot \mid \mathscr{I}_T)$ stands for the conditional probability conditioned on the σ-field $\mathscr{I}_T = \{E \in \mathscr{F} : TE = E\}$ (see [116] or [42]).

For sake of brevity a set $E \in \mathscr{F}$ having the properties listed in Theorem 3 is called a (T, N, ε)-*Rohlin set*. In general, a (T, N, ε)-Rohlin set E may posses different μ_ω-measures. If this is not the case, i.e., if

$$\mu\{\omega : \mu_\omega(E) = \mu(E)\} = 1\,,$$

then E is called a *uniform* (T, N, ε)-Rohlin set [30]. That is, a uniform (T, N, ε)-Rohlin set is a (T, N, ε)-Rohlin set which is independent of the σ-field $\mathscr{I}_T$.

Theorem 4. (Uniform Rohlin's lemma [30]). Let T be an invertible aperiodic transformation of a Lebesgue space $(\Omega, \mathscr{F}, \mu)$.

(a) For each $N \in \mathbf{N}$ and each $\varepsilon > 0$ there is a uniform (T, N, ε)-Rohlin set.

(b) If $Q \in \mathscr{F}$ satisfies $\mu\{\omega \in \Omega : \mu_\omega(Q) > 1 - \delta\} = 1$ then for each $N \in \mathbf{N}$ and each $\varepsilon > 0$ there exists a uniform $(T, N, \varepsilon + \delta)$-Rohlin set $E \subset Q$.

Part (b) was used in [30] in order to extend Krieger's finite generator theorem to aperiodic non-ergodic transformations, and was the main tool to overcome problems connected with non-uniformity of convergence in Shannon-Mcmillan's theorem and the ergodic theorem. We omit details for we shall approach the extension problem in a different way.

The next version of Rohlin's lemma will be formulated for shift spaces and the natural zero-time partitions, although it is possible to prove it for arbitrary countable measurable partitions of a Lebesgue space. So suppose $[A, \mu, X]$ is a stationary source over the alphabet $A = \{a_1, a_1, \ldots\}$. Let $\gamma_A = (C(a_1)), C(a_2), \ldots)$, where

$$C(a_i) = \{x \in A^{\mathbf{Z}} : x_0 = a_i\}\,.$$

Then the partition $\bigvee_{j=0}^{N-1} T_A^{-j}\gamma_A$ partitions $A^{\mathbf{Z}}$ according to the outputs at times zero through $N-1$, i.e., its atoms are all cylinders of the form

$$C(a^1, \ldots, a^N) = \{x \in A^{\mathbf{Z}} : x_0 = a^1, \ldots, x_{N-1} = a^N\}$$

for $(a^1, \ldots, a^N) \in A^N$. The distribution $d(\bigvee_{j=0}^{N-1} T_A^{-j}\gamma_A)$ is the vector of lexicographically

ordered measures of its atoms $\bigcap_{j=0}^{N-1} T_A^{-j} C(a_{i_j}), (i_0, \ldots, i_{N-1}) \in \mathbf{N}^N$. If $E \in \mathscr{F}$, we let $\bigvee_{j=0}^{N-1} T_A^{-j} \gamma \mid E$ denote the induced partition of E. If $\mu(E) > 0$, we can define the conditional distribution $d(\bigvee_{j=0}^{N-1} T_A^{-j} \gamma_A \mid E)$ as the lexicographically ordered vector of conditional probabilities

$$\mu\left(\bigcap_{j=0}^{N-1} T_A^{-j} C(a_{i_j}) \mid E\right) = \mu(E)^{-1}\, \mu\left(E \bigcap \bigcap_{j=0}^{N-1} T_A^{-j} C(a_{i_j})\right).$$

Theorem 5. (Strong Rohlin's lemma [128]). Let $[A, \mu, X]$ be an aperiodic stationary source over a countable alphabet A. Given $\varepsilon > 0$ and $N \in \mathbf{N}$ there exists a (T_A, N, ε)-Rohlin set E such that

$$d\left(\bigvee_{j=0}^{N-1} T_A^{-j} \gamma_A\right) = d\left(\bigvee_{j=0}^{N-1} T_A^{-j} \gamma_A \mid E\right).$$

In [128] the proof is sketched for finite partitions. An extension to countable partitions (i.e., to a countable alphabet A in our formulation) is possible as discussed in [131]. Theorem 5 says that we can choose the Rohlin set E in such a way that the statistical properties of source N-tuples over E are the same as over the entire space. In Section 7 we shall deal with a more "information-theoretic" approach to Theorem 5.

6. Making Block Codes Stationary

We pause here to explain the "stationarization" method for block codes mentioned in Section 1. Note that Ornstein himself did not mention block codes at all but considered good maps from N-tuples to N-tuples obtained by a combination of Shannon-McMillan's theorem and a marriage lemma (see [103] and Chapter 9 of [128]). An application of Ornstein's idea to actual stationarization of block codes was carried over by Gray and Ornstein [49] who used this technique to prove a sliding-block joint source/channel coding theorem. Later, Gray, Ornstein, and Dobrushin [51] applied that technique in their investigation of zero-error stationary codes for a class of noisy channels. Our next assertion is quoted from that paper.

The quadruple (T_A, N, E, γ_A) with properties listed in Theorem 5 is called an *ε-gadget*. The set E is called its base and the set

$$A^{\mathbf{Z}} \setminus \bigcup_{j=0}^{N-1} T_A^j E$$

its garbage. Suppose $\hat{A}$ is a finite set and $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ a block code of order N. We define a map $\bar{\Psi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ with the aid of the measurable map $\psi : A^{\mathbf{Z}} \to \hat{A}$

determined by the properties that

$$\Psi(x) = \begin{cases} \hat{a} & \text{if } x \in T_A^i(E \cap C(x^N)) \text{ and } (\Phi(x^N))_i = \hat{a}\,; \\ a^* & \text{if } x \text{ belongs to the garbage}\,, \end{cases}$$

where a^* is a fixed reference letter in $\hat{A}$ and $C(x^N)$ stands for the cylinder set as defined in Section 5. The action of the stationary code $\overline{\Psi}$ can be described as follows. If $x \in C(a_0, \ldots, a_{N-1}) \cap E \subset C(a_0, \ldots, a_{N-1})$, then $x^N = (a_0, \ldots, a_{N-1})$, i.e., given $x \in C(a_0, \ldots, a_{N-1}) \cap E$ the next N source outputs (starting from time zero) will be $a_0, \ldots, a_{N-1}$. In other words, we "relabel" the sets $T_A^i(E \cap C(a_0 \ldots, a_{N-1}))$, $0 \leqq i \leqq \leqq N - 1$, by the symbols a_i. The map Ψ simply substitutes $\Phi(a_0, \ldots, a_{N-1})$ for $(a_0, \ldots, a_{N-1})$.

As well-known, the rate of a block code of order N is the number $R(\bar{\Phi}) = N^{-1} \cdot \log \|\Phi(A^N)\|$. If $\bar{\Phi}$ is a stationary code it is natural to define its rate as the entropy rate $h(\bar{\Phi}X)$ of the encoded process (see [11] for definitions and basic facts concerning entropy).

Proposition 6. Let $\overline{\Psi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ be a stationary code constructed as above with the aid of a block code $\bar{\Phi}$ of order N and of an ε-gadget (T_A, N, E, γ_A). Then

$$h(\mu\overline{\Psi}^{-1}) \leqq R(\bar{\Phi}) + H(N^{-1})\,,$$

where $H(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log (1 - \alpha)$.

Thus, the rate of the code remains almost unchanged. Put $\overline{X}_i(x) = \Psi(T_A^i x)$, i.e., $\overline{X} = \overline{\Psi}X$. The proof of Proposition 6 is based on the estimate

$$h(\mu\overline{\Psi}^{-1}) \leqq h(\overline{X} \mid Z) + h(Z)\,,$$

where $Z_i(x) = 1_E(T_A^i x)$ (1_E is the indicator function of E) and E is the gadget's base. If $Z_i = 1$ then $Z_{i+1} = \ldots = Z_{i+N-1} = 0$ so that, if N is large, the entropy rate $h(Z)$ must be small. Indeed, it is easy to check that $h(Z) \leqq H(\mu(E)) \leqq H(N^{-1})$ for we must have $\mu(E) \leqq N^{-1}$. The rest of the proof (not interesting for our purpose) is a combinatorial argument giving that $h(\overline{X} \mid Z) \leqq R(\bar{\Phi})$.

An important conclusion can be drawn from the proof to the effect that instead of working with a gadget we can work with an auxiliary binary stationary coding $X \to Z$. By reformulation of the stationarization method we see that the process Z can be used to indicate when to use the block code. In the next section we make that idea precise.

7. Strong Rohlin's Lemma

For simplicity suppose that A and $\hat{A}$ are finite alphabets. Let $\overline{\Psi} : A^{\mathbf{Z}} \to \{0, 1\}^{\mathbf{Z}}$ be a binary sliding-block code, say of order m. Let $\overline{X} = \overline{\Psi}X$. A block $\overline{X}_0^N = (0, 1, 1, \ldots \ldots, 1, 0)$ of length $N + 1$ is called an *N-cell*.

Theorem 7. (Strong Rohlin's lemma [131]). Let $[A, \mu, X]$ be a stationary and ergodic source. Given $\delta > 0$ and $N \in \mathbf{N}$ there is an $m \in \mathbf{N}$ and a binary sliding-block code $\overline{\Psi} : A^{\mathbf{Z}} \to \{0, 1\}^{\mathbf{Z}}$ of order m such that for the encoded process $\overline{X} = \overline{\Psi}X$ the following holds true:

(a) $N^{-1} \leqq \text{Prob}\,[\overline{X}_0 = 0] \leqq 2N^{-1}$,

(b) $\text{Prob}\,[\overline{X}_0^N \text{ is an } N\text{-cell} \mid \overline{X}_0 = 0] \geqq 1 - \delta$, and

(c) $\sum_{a \in A^N} |\text{Prob}\,[\overline{X}_0^N \text{ is an } N\text{-cell} \mid X^N = a] - \text{Prob}\,[X^N = a]| < \delta$.

Assertion (b) says that the waiting time between two consecutive zeroes in the $\overline{X}$ process is, with high probability, equal exactly to N. Assertion (c) says that the distribution of source N-tuples is almost independent of the N-cells in its binary encoding $\overline{X}$.

This suggests connections between Theorems 3 and 7. Indeed, put $E = \{x \in A^{\mathbf{Z}} : \bar{x}_0^N \text{ is an } N\text{-cell}\}$. Then $E \in \mathscr{A}^{\mathbf{Z}}$. Suppose $x \in E \cap T_A E$. Since the map $x \mapsto \bar{x}$ is a stationary coding, we have that

$$T_A \overline{x} = T_B \bar{x}\,, \quad B = \{0, 1\}\,.$$

Consequently, $\bar{x}_0^N = (0, 1, \ldots, 1, 0)$ and $\bar{x}_1^{N+1} = (0, 1, \ldots, 1, 0)$. But this is impossible so that $E \cap T_A E = \emptyset$. The same argument applies to the sets $T_A^i E$, $T_A^j E$ unless $|i - j| = N$. Using (a) and (b) of Theorem 7 it is easy to derive that $\mu(E \cup T_A E \cup \ldots \cup T_A^{N-1} E) \geqq 1 - \delta$, hence, E is a (T_A, N, δ)-Rohlin set. Consequently, Theorem 7 implies Rohlin's lemma. Furthermore, assertion (c) entails

$$\sum_{a \in A^N} |\text{Prob}\,[X^N = a] - \text{Prob}\,[X^N = a \mid E]| \leqq \delta \quad (*)$$

This is slightly weaker than the condition of exact independence in Theorem 5. On the other hand, as the binary coding $\overline{\Psi}$ in Theorem 7 is a sliding-block coding, the set E can be chosen to depend on only a finite number of coordinates. Thus, we get a particular case of Dunham's strong version of Rohlin's lemma [33]. The fact E depends on a finite number of coordinates is crucial to Dunham's proof of an abstract alphabet sliding-block entropy compression coding theorem, for it allows to control the distortion of the encoded process.

Finally, note that Theorem 7 is valid also for a countable alphabet and for aperiodic stationary and non-ergodic sources, however, the technical details are much more involved.

8. Bernoulli Processes

In Section 4 we introduced the concept of a Bernoulli source. Of course, it is desirable to have manageable criteria for deciding whether a given source is (isomorphic to) a Bernoulli source or not. It is easy to prove a general criterion which is, however, not a very practical one.

The shift-transformation T on $\{1, 2, \ldots, K\}^{\mathbf{Z}}$ is said to be a *Bernoulli shift* if there exists a probability vector $(p_1, \ldots, p_K)$ such that the process X is an IID process such that

$$\text{Prob}\,[X_0 = k] = p_k, \quad 1 \leqq k \leqq K\,.$$

In this case we denote T as $T(p_1, \ldots, p_K)$. In what follows we shall say that two transformations T and T' are isomorphic if the corresponding dynamical systems are (and this will be used only if the underlying probability spaces will be clear from the context). Shields [128, Theorem 2.1] obtained the following result:

Theorem 8. Let T be an invertible measure-preserving transformation of a Lebesgue space $(\Omega, \mathscr{F}, \mu)$. Then T is isomorphic to a Bernoulli shift $T(p_1, \ldots, p_K)$ if and only if there exists a measurable partition $\zeta = (C_1, \ldots, C_K)$ of Ω such that

(a) $d(\zeta) = (p_1, \ldots, p_K)$,
(b) ζ is a generator (relative to (T, μ)), and
(c) the sequence $(T^n\zeta;\ n \in \mathbf{N})$ is independent.

In particular, a stationary source $[A, \mu, X]$ is Bernoulli if there is an IID source $[\hat{A}, \nu, \hat{X}]$ and a perfectly noiseless code $\bar{\Phi} : \hat{A}^{\mathbf{Z}} \to A^{\mathbf{Z}}$ such that $X = \bar{\Phi}\hat{X}$, This is different from, but equivalent to, the definition given in Section 4. The equivalence follows from the fact that any factor of a Bernoulli shift is a Bernoulli process [100].

Moreover, we can dispense with finite partitions or finite alphabets. This can be made either using finite approximations to countable partitions based on Theorem 9.5 of [103] (which says that if T has finite entropy and is a union of an increasing sequence of Bernoulli transformations then T itself is Bernoulli) or by developing Ornstein's theory directly in terms of countable partitions as done by Smorodinsky [136].

On the other hand, the criterion from Theorem 8 amounts to construction of an independent generator which is an extremely difficult task even for simple sources like mixing Markov ones. Hence, other criteria for Bernoullicity are desirable. In what follows, we describe them in terms of codes.

Let $[A, \mu, X]$ and $[A, \mu', X']$ be two stationary sources over the same finite alphabet A. Let

$$d_N(x^N, x'^N) = N^{-1} \sum_{i=0}^{N-1} \delta(x_i, x'_i)$$

denote the Nth order average Hamming distance. For any $N \in \mathbf{N}$ let $\mu^N \vee \mu'^N$ denote the set of all joint probability vectors on $A^N \times A^N$ yielding μ^N and μ'^N as marginals. The number

$$\inf_{\lambda \in \mu^N \vee \mu'^N} \int d_N(x^N, x'^N)\, \mathrm{d}\lambda(x^N, x'^N) = \bar{d}_N(X, X')$$

is said to be the *Nth order $\bar{d}$-distance* between the two sources. The *$\bar{d}$-distance*

between X and X' (or, between μ and μ') is the limit

$$\bar{d}(X, X') = \lim_{N\to\infty} \sup \bar{d}_N(X, X')$$

(see [103] or [102]; in the latter paper you can find a detailed discussion on $\bar{d}$-distance as well as equivalent definitions). All characteristic properties of Bernoulli processes reflect, in some approximative manner, independence properties of IID processes.

If X is IID then X_i is independent of its "past" $X_{i-1}, X_{i-2}, \ldots$ for all $i \in \mathbf{Z}$. If, $Y = (Y_i;\ i \in \mathbf{Z})$ is a sliding-block coding of X of some finite order m then, for each i, the random variables Y_i do not depend on past coordinates at least $2m$ apart, i.e. on $Y_{i-2m}, Y_{i-2m-1}, \ldots$. If $\bar{\Phi} : A^{\mathbf{Z}} \to \bar{A}^{\mathbf{Z}}$ is an infinite code then, by Proposition 2(a), we can find a sliding-block approximation $\tilde{Y}$ to $Y = \bar{\Phi}X$ such that Prob $[Y_0 \neq \tilde{Y}_0] < \varepsilon$. Thus, one expects Y to satisfy some approximate version of the independence property enjoyed by $\tilde{Y}$. Call a stationary process $Z = (Z_i;\ i \in \mathbf{Z})$ an independent N-blocking of Y if, for each integer k,

(i) dist (Z_{Nk+1}^{Nk+N}) = dist (Y_{Nk+1}^{Nk+N}) and

(ii) the vector Z_{Nk+1}^{Nk+N} is independent of Z_i for $i \leqq Nk$.

A stationary process Y is said to be *almost block independent* (ABI) if for any $\varepsilon > 0$ there exists an $N_0 \in \mathbf{N}$ such that for any $N \geqq N_0$ we can find an independent N-blocking Z of Y so that $\bar{d}(Y, Z) < \varepsilon$ (see [129, 130]).

If X is IID then for each $N \in \mathbf{N}$ the vector X_1^N does not depend on the past $(X_i;\ i \leqq 0)$. A stationary process Y is said to be *very weak Bernoulli* (VWB) if for any $\varepsilon > 0$ there is an N such that for each $m \in \mathbf{N}$, $\bar{d}(Y, Y \mid Y_{-m}^0) < \varepsilon$ for a collection of pasts Y_{-m}^0 of total probability at least $1 - \varepsilon$ [101].

A weak Bernoulli property (stronger than VWB) was introduced by Friedman and Ornstein [38] for the purpose of showing that mixing Markov processes are Bernoulli. Indeed, as mentioned above, to find an independent generating partition is extremely difficult but, for mixing Markov chains, the natural zero-time partition is weak Bernoulli (see [128] for a detailed investigation). Furthermore, weak Bernoulli is in fact a mixing condition so that usually it is quite easy to verify (see, e.g., [19]). On the other hand, weak Bernoulli is too strong to be a characteristic property of Bernoulli processes for there is a Bernoulli source which has a factor that is not weak Bernoulli ([137], see also [22]).

The third property involves also approximation in entropy. It follows easily from the definition of $\bar{d}$-distance that closeness in $\bar{d}$ implies closeness in finite dimensional distributions and closeness in entropy (that is, entropy is $\bar{d}$-continuous [102]). If the converse is true, we say the process is finitely determined (FD). Thus, a stationary process Y is FD if, given $\varepsilon > 0$ we find $\delta > 0$ and $N \in \mathbf{N}$ such that for every ergodic process $\bar{Y}$ the conditions that

$$\sum_{a \in A^N} \left| \text{Prob}\,[Y_1^N = a] - \text{Prob}\,[\bar{Y}_1^N = a] \right| < \delta$$

and

$$|h(Y) - h(\overline{Y})| < \delta$$

imply that $\bar{d}(Y, \overline{Y}) < \varepsilon$. Originally all these properties were formulated in terms of partitions. Our formulation follows [129] (see also [76]). Now we can formulate a characterization theorem for Bernoulli sources:

Theorem 9. Let $[A, \mu, X]$ be a stationary source over a finite discrete alphabet A. Then the following assertions are equivalent: (a) X is a Bernoulli process, (b) X is ABI, (c) X is FD, and (d) X is VWB.

The proof of this theorem summarizes the basic achievements of Ornstein's theory. In the next section we shall sketch the proof of one part which makes more transparent the fundamental features of Ornstein's coding technique.

9. Ornstein's Coding Technique

We shall sketch the proof that any ABI process is Bernoulli. One possible way is to show how to encode a given ABI process in a stationary manner from a suitable IID process.

Let Y be an ABI process over a finite or countable alphabet. Let U be an IID process over the alphabet $[0, 1]$ such that U_0 is uniformly distributed. Let V be an IID process over a discrete alphabet $\hat{A}$ such that U and V are independent.

First we construct an initial coding. We claim that for any $\varepsilon > 0$ there exists a measurable map $\Psi : [0, 1]^{\mathbf{Z}} \times \hat{A}^{\mathbf{Z}} \to A$ such that $\bar{d}(Y, \overline{Y}) < \varepsilon$, where we have put $\overline{Y} = \overline{\Psi}(U, V)$ (and $\overline{\Psi}$ is the stationary code induced by ψ). The idea is similar to that used in Section 6. We use a binary encoding of V to indicate when to block code U using an appropriate block coding function $\Phi_N : [0, 1]^N \to A^N$. Let N be large and δ small. Let λ^N denote the Lebesgue measure on $[0, 1]^N$. Since U_0 is uniformly distributed and U is IID we can find Φ_N such that

$$\lambda^N[\Phi_N^{-1}(\boldsymbol{a})] = \text{Prob}\,[Y_1^N = \boldsymbol{a}], \quad \boldsymbol{a} \in A^N.$$

As in the proof of Rohlin's lemma in [131] we can show that given an ergodic source X over the alphabet A, $N \in \mathbf{N}$, and $\delta > 0$ there is a measurable map $\beta : A^{\mathbf{Z}} \to \{0, 1\}$ which gives rise to a stationary code $\bar{\beta}$ such that the encoded process $R = \bar{\beta}X$ satisfies

$$\text{Prob}\,[R_0 \text{ is in an } N\text{-cell}] > 1 - \delta.$$

We call R an (N, δ)-process. Let $R = \bar{\beta}V$ be the (N, δ)-process of V, where N and δ are as above. Next we want to block code U by means of Φ_N. Let $X = (U, V)$. For a given x, we first code to obtain $\bar{\beta}x$. Whenever $(\bar{\beta}x)_i$ is a start of an N-cell, we apply Φ_N to code x_i^{i+N-1} onto $\Phi_N(x_i^{i+N-1})$. If $(\bar{\beta}x)_i$ does not lie in an N-cell, we assign to x_i a distinguished letter. This defines a code

$$\bar{\Phi} = \bar{\Phi}_{\Phi_N, R} : [0, 1]^{\mathbf{Z}} \times \hat{A}^{\mathbf{Z}} \to A^{\mathbf{Z}}.$$

Let

$$\bar{Y} = \bar{\Phi}_{\Phi_N, R}(U, V).$$

Since U and V were independent, U and R are also independent. If δ is sufficiently small, it follows that Y and $\bar{Y}$, conditioned on R-typical sequences are $\bar{d}$-close (for details see [131] and [129]). From the way how R has been constructed we conclude that even Y and $\bar{Y}$ themselves are $\bar{d}$-close and this gives us the desired initial coding.

Having a good (ε-close) initial coding, the main point of Ornstein's technique is an idea of how to make a much better coding in exchange to only a small change in the structure of the code.

Assume $\bar{Y}$ is a sliding-block coding of $X = (U, V)$. Then there exists a $k \in \mathbf{N}$ and a map $f : [0, 1]^{2k+1} \times \hat{A}^{2k+1} \to A$ such that $\bar{Y} = \bar{f}(U, V)$. If $M > 2k + 1$, then f induces a map

$$f_M : [0, 1]^M \times \hat{A}^M \to A^{M-2k}$$

according to the formula

$$(f_M(u^M, v^M))_i = f(u_{i-k}^{i+k}, v_{i-k}^{i+k}), \quad k + 1 \leqq i \leqq M - k.$$

We can and do assume that V is an aperiodic process. Since U and V are independent aperiodic processes, they are also jointly aperiodic. Hence, if $\bar{\lambda}^M$ is the measure on $[0, 1]^M \times \hat{A}^M$ induced by (U, V), then we can find a map

$$\Psi'_M : [0, 1]^M \times \hat{A}^M \to A^M$$

such that

$$\bar{\lambda}^M[\Psi'^{-1}_M(\mathbf{a})] = \text{Prob}[Y_1^M = \mathbf{a}], \quad \mathbf{a} \in A^M.$$

Recall that by our construction $\bar{Y}$ satisfies $\bar{d}(Y, \bar{Y}) < \varepsilon$. Consequently, if M is large enough, then we will have also $\bar{d}_M(Y, \bar{Y}) < \varepsilon$. Moreover, this inequality shows that the distributions of Y_1^M and $\bar{Y}_1^M$ are close, for Ψ'_M partitions $[0, 1]^M \times \hat{A}^M$ according to the distribution of Y_1^M. It follows there is a map

$$\Psi_M : [0, 1]^M \times \hat{A}^M \to A^M$$

such that

$$(*) \qquad \begin{cases} \bar{\lambda}^M[\Psi_M^{-1}(\mathbf{a})] = \text{Prob}[\bar{Y}_1^M = \mathbf{a}], \quad \mathbf{a} \in A^M, \\ \mathsf{E}_{\bar{\lambda}^M}[\bar{d}_M(\Psi_M(U_1^M, V_1^M), \Psi'_M(U_1^M, V_1^M))] < \varepsilon. \end{cases}$$

On the other hand, the partition induced by Ψ'_M refines that one induced by Ψ_M. Consequently, if M is much larger than $2k + 1$, we can get

$$\mathsf{E}_{\bar{\lambda}^M}[\bar{d}_M(\Psi_M(U_1^M, V_1^M), f_M(U_1^M, V_1^M))] < 2\varepsilon.$$

In other words, we used the coding of (U, V) onto $\bar{Y}$ to induce a distribution of $\bar{Y}$ on M-tuples, i.e., on $[0, 1]^M \times \hat{A}^M$. The $\bar{d}$-fit of Y and $\bar{Y}$ allows to induce a distribution of Y on the same space in such a way that $Y_i = \bar{Y}_i$ with high probability.

Now we choose a new IID process W over the alphabet $\hat{A}$ so that W is independent

of each U and V. Let $\bar{R}$ denote the associated $(M, \bar{\delta}_1)$-process of W, and

$$\bar{\bar{Y}} = \bar{\Phi}_{\Psi_M, R}(U, V, W).$$

Given $\bar{\varepsilon} > 0$ we can assume M so large and $\bar{\delta}_1$ so small that $\bar{d}(\bar{\bar{Y}}, Y) < \bar{\varepsilon}$. We can choose $\hat{\varepsilon}$ so small and M so large that $\text{Prob}\left[\bar{Y}_i \neq \bar{\bar{Y}}_i\right] < 3\varepsilon$.

Thus, by change of no more than 3ε in our coding we can get from an ε-close coding a coding which is as close to Y in $\bar{d}$ sense as we please. This is the idea of construction of a converging sequence of codes. Pick mutually independent IID processes $U, V^{(1)}, V^{(2)}, \ldots$ such that U_0 is uniformly distributed over $[0, 1]$, and $V^{(1)}, V^{(2)}, \ldots$ are each over the same alphabet $\hat{A}$. We apply the above technique first to $(U, V^{(1)})$, then to $(U, V^{(1)}, V^{(2)})$, etc. This defines a sequence $(Y^{(n)}; n \in \mathbf{N})$ of codings of $\bar{U} = (U, V^{(1)}, V^{(2)}, \ldots)$ such that, as $n \to \infty$, we have

$$\bar{d}(Y, Y^{(n)}) \to 0, \quad \text{Prob}\left[Y_i \neq Y_i^{(n)}\right] \to 0.$$

Consequently, there exists a stationary coding $\tilde{Y}$ of $\bar{U}$ such that $\bar{d}(Y, \tilde{Y}) = 0$. A suitable quantization allows to express any IID source over the alphabet $[0, 1]$ in the form $\bar{U} = (U, V^{(1)}, V^{(2)}, \ldots)$.

Furthermore, it is clear that the same construction can be carried over provided one starts with only approximate versions of $(*)$. Thus, we really need only that $h(U) > h(Y)$, e.g., U can be any IID process over a countable alphabet with enough entropy [129].

PART III: FINITARY CODES

10. The Finitary Isomorphism Theorem

It is clear that the limiting coding $\bar{U} \to \tilde{Y}$ obtained via the construction sketched in the preceding section cannot be finitary. Indeed, we make a good code better in exchange to increase in length M of the block coding function. If we stopped at some fixed length M then we could not get invertible codes. Thus, one has to develop a different construction for the proof of the next assertion:

Theorem 10 [61]. Let $T(p_0, \ldots, p_{m_1-1})$ and $T(q_0, \ldots, q_{m_2-1})$ be two Bernoulli shifts with the same entropy, and let X and Y denote the corresponding IID processes. Then there exists a finitary isomorphism $\bar{\Phi}$ such that $Y = \bar{\Phi}X$.

As the construction is quite involved and, moreover, explained in detail elsewhere (see [82]), we shall give only a brief outline and devote more place to other results related to finitary coding.

Let $A = \{0, 1, \ldots, m_1 - 1\}$ and $B = \{0, 1, \ldots, m_2 - 1\}$. Using continuity of the entropy function on finite probability vectors we can assume that at least one letter of A and B has the same probability. This will be used to construct a time invariant indication of when to code. So suppose that $p_0 = q_0$. For each $x \in A^{\mathbf{Z}}$ let $N_x =$

$= \{n \in \mathbf{Z} : x_n = 0\}$; similarly for $y \in B^{\mathbf{Z}}$. Put

$$M_x^{(1)} = \{x' \in A^{\mathbf{Z}} : N_x = N_{x'}\}\ ; \quad M_y^{(2)} = \{y' \in B^{\mathbf{Z}} : N_y = N_{y'}\}\ .$$

The code we intend to construct will be defined on a subset $\overline{M}_1 \subset A^{\mathbf{Z}}$ of full measure. $\overline{M}_1$ will consist of entire sets of type $M_x^{(1)}$, and its performance will be time-invariant, i.e., $N_{\Phi x} = N_x$ and $\bar{\Phi}(M_x^{(1)}) = M_{\Phi x}^{(2)}$.

On the first step one constructs (based on the assumption that $p_0 = q_0$) a time-invariant indication of when to code. To this end let $N_1 < N_2 < \ldots$ be a sequence of positive integers (to be chosen in an appropriate way). Let us consider a configuration

$$0^{n_0} \underset{l_1}{-} 0^{n_1} \underset{l_2}{-} \ldots \underset{l_k}{-} 0^{n_k}\,(=\Sigma)$$

which consists of n_0 zeroes followed by l_1 blank spaces followed by n_1 zeroes followed by l_2 blank spaces, etc. Σ is called a skeleton of order r if $l_t \geqq 1$ $(1 \leqq t \leqq k)$, $n_t \geqq 1$ $(0 \leqq t \leqq k)$, and $n_t < N_t < \min\{n_0, n_k\}$ $(1 \leqq t \leqq k-1)$. The length of Σ is defined to be the total number of blank spaces, $l_1 + l_2 + \ldots + l_k$. This choice makes it possible to define in a consistent way a sequence of skeletons of increasing orders and lengths. In particular, there exists a canonical, so-called order decomposition of each skeleton of order r into subskeletons of order $r-1$.

On the second step one constructs the actual code as a method of filling in the blank spaces. The idea is similar to Ornstein's construction of coding functions between N-tuples. One uses the Shannon-McMillan's theorem in order to get estimates for the number and the total probability of good filler blocks (i.e., blocks which may fill in the blank spaces). Based on these estimates one can use a marriage lemma (proved in [60]) to define, by induction on order of skeletons, so called partial assignments which assign to each block filling in the blank spaces in one configuration a set of blocks which can fill in the blank spaces in the corresponding configuration for the second process. The important fact which follows from the marriage lemma is that these assignments at boundedly finite-to-one maps and one can show that, with probability one, they are even one-to-one. Using an appropriate choice of taking products of partial assignments one can join them into global assignments in a consistent way. This ensures that if a blank space was filled in on the rth step then, on the next steps, the letter filling in that blank space remains unaffected. In this way, a stationary coding is defined for almost every sequence $x \in A^{\mathbf{Z}}$ and, furthermore, the coding procedure stops at some finite order $r = r(x)$ with probability one.

It is natural to expect that an analogue of Theorem 10 should hold for irreducible multistep Markov sources, for any such source is a sliding-block coding of an IID source. A proof of this more general result which makes use of a reduction to the IID case is announced in [62]. For infinite codings of IID sources, i.e., for Bernoulli sources of non-Markov type, one cannot expect for a finitary isomorphism (see [151]).

11. Finitary Codes With Finite Expected Code Length

In this section we investigate finitary codes with finite expected code length (not invertible ones!). Following [31] let us introduce the concept of a shift dynamical system. Let A denote either a finite set, say $\{1, 2, \ldots, K\}$ or, the one-point compatification $\{1, 2, \ldots, \infty\}$ of $\mathbf{N}$. A *shift dynamical system* is the quadruple $(Y, \mathscr{B}(Y), \mu, T_A)$, where Y is a closed T_A-invariant set, $\mathscr{B}(Y)$ is the Borel subsets of Y and T_A is the shift transformation on $A^{\mathbf{Z}}$. We assume that μ is a T_A-invariant probability measure on Y or, if A is infinite, a measure such that $\mu(Y \cap (\mathbf{N} \setminus \{\infty\})^{\mathbf{Z}}) = 1$. Let f be a bounded measurable function on Y. We denote by $\mathscr{B}^k_{-k}$ the σ-field $\sigma(X_i; |i| \leqq k)$. Following [31], f is said to be *sequential* (in symbols, $f \in S(Y)$) if there exists a sequence $(f_k; k \in \mathbf{N})$ of bounded functions such that

(i) f_k is $\mathscr{B}^k_{-k}$-measurable, $k \geqq 0$;

(ii) $\lim_{k \to \infty} \int \left| f - \sum_{n=0}^{k} f_n \right| d\mu = 0$; and

(iii) $\sum_{k=0}^{\infty} k \int |f k_k| \, d\mu < \infty$.

If $f \in S(Y)$, put $\sigma(f)$ to be the infimum of the sums in (iii) over all possible sequences $(f_k; k \in \mathbf{N})$. In particular, if U is an open subset of Y, then 1_U is sequential if and only if

$$U = \bigcup_{k=0}^{\infty} C_k, \quad C_k \in \mathscr{B}^k_{-k}, \quad \text{and} \quad \sum_{k=0}^{\infty} k \, \mu(C_k) < \infty .$$

The sets C_k can be chosen so that

$$\sigma(U) = \sigma(1_U) = \sum_{k=0}^{\infty} k \, \mu(C_k) .$$

Let $\bar{\Phi}$ be a finitary code from $(Y, \mathscr{B}(Y), \mu, T_A)$ to another shift dynamical system $(Y', \mathscr{B}(Y'), \nu, T_B)$. If g is a bounded measurable function on Y', then $\bar{\Phi}$ is said to *code g sequentially* if $g \circ \bar{\Phi} \in S(Y)$. The code $\bar{\Phi}$ is said to have *finite expectation* if there is a constant K such that for any set $C \in (\mathscr{B}')^k_{-k}$, r codes 1_C sequentially, and

$$\sigma(1_C \circ \bar{\Phi}) \leqq K k \, \nu(C) .$$

Theorem 11 [31]. *$\bar{\Phi}(S(Y')) \subset S(Y)$ for any code with finite expectation* (here, $\bar{\Phi}(S(Y'))$ stands for the set of all compositions $g \circ \bar{\Phi}$, $g \in S(Y')$).

One can prove that if $(Y, \mathscr{B}(Y), \mu, T_A)$ is strongly mixing with the mixing coefficients $\alpha(k)$ satisfying

$$\sum_{k=1}^{\infty} \alpha(k) < \infty$$

then for any $g \in S(Y')$ a central limit theorem is valid in the sense that properly normalized partial sums of shifts g converge in distribution to the normal law, provided the code $\bar{\Phi} : Y \to Y'$ has finite expectation (see Theorem 24 of [31]). In

a subsequent paper [32] Denker and Keane proved also the law of the iterated logarithm and an invariance principle. These results show that finitary codes with finite expectation are very desirable from the point of view of physical applications (see pp. 158–159 of [31]). Indeed, many classical dynamical systems have been shown to be continuous factors of symbolic systems which are Bernoulli processes or Bernoulli flows (see [20, 21]). It is easy to show that continuous factor maps are isomorphisms when the systems are equipped with natural invariant probability measures – the equilibrium states.

12. Finitary Isomorphisms With Finite Expected Code Length

Let $[A, \mu, X]$ and $[B, \nu, Y]$ be two stationary sources over countable discrete alphabets. Let γ_A and γ_B denote the natural zero-time partitions of $A^{\mathbf{Z}}$ and $B^{\mathbf{Z}}$, respectively. Suppose $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ is an invertible (stationary) code with $\nu = \mu\bar{\Phi}^{-1}$. An A-cylinder is any set of the form

$$\bigcap_{i=-m}^{n} T_A^{-i} C(a_i), \quad n \geqq 0, \quad m \geqq 0$$

(see Section 5 for the symbol $C(a)$); similarly, we define B-cylinders. As already mentioned, $\bar{\Phi}$ is a finitary code if and only if the set $\bar{\Phi}^{-1} C(b)$ is mod 0 a countable union of A-cylinders for each $b \in B$, and $\bar{\Phi} C(a)$ is mod 0 a countable union of B-cylinders for any $a \in A$. The length of an A-cylinder as above is $n + m$, and the future length is defined to be m. This allows us to define the length functions for the code $\bar{\Phi}$ itself. To this end, pick a $b \in B$ and express the set $\bar{\Phi}^{-1} C(b)$ as a countable union of A-cylinders C with minimal length. If $x \in C$, then $L(\bar{\Phi}, x)$ (and $L^+(\bar{\Phi}, x)$) are defined as the length (and the future length) of C. A finitary isomorphism $\bar{\Phi}$ is said to have *finite expected code length* (*future code length*) if

$$\mathsf{E}_\mu L(\bar{\Phi}, X) < \infty \; (\mathsf{E}_\mu L^+(\bar{\Phi}, X) < \infty) .$$

The *inverse* lengths of $\bar{\Phi}$ are the lengths of $\bar{\Phi}^{-1}$. Observe that

$$\mathsf{E}_\mu L(\bar{\Phi}, X) \geqq \mathsf{E}_\mu L^+(\bar{\Phi}, X) = \sum_{n=i}^{\infty} a_n ,$$

where

$$a_n = \mu\{x \in A^{\mathbf{Z}} : L^+(\bar{\Phi}, x) \geqq n\} .$$

Parry investigated the consequences of existence of a finitary isomorphism $\bar{\Phi}$ such that both $\bar{\Phi}$ and $\bar{\Phi}^{-1}$ have finite expected code length. His idea was to derive from that assumption a cocycle-coboundary equation for the conditional information functions of the two isomorphic sources. This equation admits, in general, a richer family of invaraints (see [34, 109, 110, 19, 114]) which can be used to find examples of IID and Markov sources for which no such "practical" perfectly noiseless codes can exist (see Section 4).

Let γ_A^- denote the "past" of the partition γ_A, i.e.

$$\gamma_A^- = \bigvee_{i=0}^{\infty} T_A^{-i} \gamma_A ,$$

where, as usually, the symbol on the right-hand side is interpreted as a σ-field [107], viz.

$$\bigvee_{i=0}^{\infty} T_A^{-i} \gamma_A = \sigma\Big(\bigcup_{i=0}^{\infty} T_A^{-i}\, \sigma(\gamma_A)\Big) .$$

The *information cocycle* for a stationary source $[A, \mu]$ is defined as

$$I(\gamma_A \mid T_A^{-1}\gamma_A^-) = -\sum_{C \in \gamma_A} 1_C \log \mu(C \mid T_A^{-1}\gamma_A^-) .$$

A real-valued function of the form $f \circ T_A - f$ is called a *coboundary* (with respect to T_A). Two functions which differ by a coboundary are said to be *cohomologous*. The main result of [111, 112] is the following assertion:

Theorem 12. Let $[A, \mu]$ and $[B, \nu]$ be two stationary and ergodic sources over countable discrete alphabets A and B such that $H(\gamma_A)$ and $H(\gamma_B)$ are finite. Let $\bar{\Phi}$ from $A^{\mathbf{Z}}$ to $B^{\mathbf{Z}}$ be a finitary isomorphism such that $\bar{\Phi}$ and $\bar{\Phi}^{-1}$ each have finite expected code length. Then the information cocycles

$$I(\gamma_A \mid T_A^{-1}\gamma_A^-) \quad \text{and} \quad I(\gamma_B \mid T_B^{-1}\gamma_B^-) \circ \bar{\Phi}$$

are cohomologous.

For the sake of brevity, let us denote by I_A and I_B the information cocycles of $[A, \mu]$ and $[B, \nu]$, respectively. By Theorem 12 we have the following cocycle-coboundary equation:

$$I_A = I_B \circ \bar{\Phi} + g \circ T_A - g .$$

Our problem is how to exploit this equation. Bowen [19] proposed (in the context of bounded codes; see Section 4) a method based on the central limit theorem. He showed that the limiting distribution of the sequence

$$F_n(\zeta) = n^{-1/2}\Big[\sum_{i=0}^{n-1} I_A(\zeta) \circ T_A^i - n\, h_\mu(T_A, \zeta)\Big]$$

is independent of all measurable partitions ζ which boundedly code each other (here $h_\mu(T_A, \zeta)$ stands for the entropy of the shift T_A relative to the partition ζ; see [11] or [107]). By the central limit theorem, the limiting distribution is gaussian and hence determined by its mean (which is but the entropy $h_\mu(T_A, \zeta)$) and variance. The variance is a new invariant which can be used to prove that certain Bernoulli shifts with the same entropy are not isomorphic via a finitary code with finite expected code length and inverse code length.

More precisely, the new invariant is the number (called *information variance*

in [34])

$$\sigma^2(T_A, \zeta) = \lim_{n\to\infty} n^{-1} \int [\sum_{i=0}^{n-1} I_{T_A}(\zeta) \circ T_A^i - n\ h_\mu(T_A, \zeta)]^2\ \mathrm{d}\mu\ .$$

Parry and Schmidt [114] have shown how to compute this invariant in some simple cases. In particular, it follows that Bernoulli shifts $T(1/4, 1/4, 1/4, 1/4)$ and $T(1/2, 1/8, 1/8, 1/8, 1/8)$ (which are isomorphic by [99] and even finitarily isomorphic by [93]) cannot be finitarily isomorphic via a code $\bar{\Phi}$ such that $\bar{\Phi}$ and $\bar{\Phi}^{-1}$ each have finite expected code length. The fact that Meshalkin's code has infinite expected code length was observed by Gray [41] using a random walk representation of the encoding rules from [93].

In [110] it is shown that the weaker assumption to the effect that $\bar{\Phi}$ is only a bounded (or, merely an ε-bounded) isomorphism gives rise to the same cocycle-coboundary equation. This surprising result shows that we cannot be too optimistic as far as concerns "practical" perfectly noiseless codes.

On the other hand, the information variance is not sharp enough to distinguish between Markov sources, say, those determined by matrices

$$\begin{pmatrix} p & q \\ p & q \end{pmatrix}, \quad \begin{pmatrix} p & q \\ q & p \end{pmatrix}, \quad \begin{pmatrix} q & p \\ p & q \end{pmatrix}; \quad p \neq q\ , \quad p + q = 1\ .$$

Any two of them are finitarily isomorphic by the Keane-Smorodinsky result [62]. Parry [109] introduced another invariant of the cocycle-coboundary equation, namely the group

$$\Lambda(T_A, \zeta) = \{(a, b) \in \mathbf{R} \times \mathbf{R} : (F \circ T_A)/F = \exp\{2\pi i(a + b\ I_A(\zeta))\}\ ,$$
$$F : A^{\mathbf{Z}} \to \{z \in \mathbf{C} : |z| = 1\}\}\ .$$

As pointed out in [114], it is easy to compute the group $\Lambda(T_A, \zeta)$ in case when $\zeta = \gamma_A$ and the sources are of Markov type. In particular, no two of the above three Markov sources are isomorphic via a "practical" code.

On the other hand, there exist several positive results on existence of "practical" perfectly noiseless codes (recall the bounded isomorphisms of Adler and Weiss [5] and a recent result by Adler and Marcus [4]). Hence, it is of interest to have some general conditions explaining such positive results. Here is one of them (consult [30] or Sections 17 and 18 below for definitions):

Proposition 13. Let T_1 and T_2 be two topological Markov chains over finite alphabet A and B. Suppose T_1 and T_2 are topologically mixing and

$$h_{\mathrm{top}}(T_1) = h_{\mathrm{top}}(T_2)\ ,$$

where h_{top} stands for the topological entropy. If T_1 and T_2 are equipped with (unique) measures of maximal entropy μ_1 and μ_2 (i.e., $h(\mu_i) = h_{\mathrm{top}}(T_i)$, $i = 1, 2$) then there

exists a finitary isomorphism $\bar{\Phi}$ from $A^{\mathbf{Z}}$ to $B^{\mathbf{Z}}$ such that $\mu_2 = \mu_1\bar{\Phi}^{-1}$ and $\bar{\Phi}$ and $\bar{\Phi}^{-1}$ each have finite expected code length.

The proof is based on the fact that the measures μ_1 and μ_2 are uniformly mixing with exponential rate [106] so that $\mu_i[L(\bar{\Phi}, X) = n] \to 0$ exponentially fast as $n \to \infty$, $i = 1, 2$, Hence

$$\mathsf{E}_{\mu_i} L(\bar{\Phi}, X) = \sum_{n=1}^{\infty} n\mu_i[L(\bar{\Phi}, X) = n] < \infty\,, \quad i = 1, 2$$

(here, $\bar{\Phi}$ stands either for $\bar{\Phi}$ $(i = 1)$ or for $\bar{\Phi}^{-1}(i = 2)$). In light of negative results presented above a deeper insight into reasons responsible for this exceptionally good behaviour is desirable. A partial result is given in Theorem 2 of [110]:

Theorem 14. Let T_1 and T_2 be two Markov chains determined by finite stochastic irreducible matrices. Let T_2 be of maximal type (i.e., the matrix gives rise to the measure of maximal entropy). If T_1 and T_2 are ε-bounded isomorphic $(0 \leqq \varepsilon < \frac{1}{2})$ then T_1 must be of maximal type, too.

The reader should consult [91] or [150] for related results (we shall commet on them a little bit later).

PART IV: REPRESENTATION PROBLEMS

13. Krieger's Theorem

Representation problems seem to be the most attractive from the point of view of information theory. In fact, they result in perfectly noiseless codes of quite general processes onto processes which have certain, in advance given, properties (Section 22 will be devoted to an "information theoretic" discussion of this point).

Let us describe the first result of this type due to Krieger [85]. In light of Proposition 1 we shall consider only stationary sources over countable discrete alphabets. Let INT(u) denote the integer part of a real $u \geqq 0$.

Theorem 15 (Krieger's theorem). Let $[A, \mu, X]$ be an aperiodic, stationary and ergodic source over a countable discrete alphabet A such that $h(X)$ is finite. Then there exists a finite alphabet B with at most INT $\{\exp h(X)\} + 1$ letters, a source $[B, \nu, Y]$ over the alphabet B, and a perfectly noiseless code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $Y = \bar{\Phi}X$.

In other words, we can find a source $[B, \nu, Y]$ over a finite alphabet B which has the same statistical properties as $[A, \mu, X]$. It is clear that the problem is very close to that of noiseless source coding (nevertheless, a proper explanation of the connections requires some efforts; see Section 22).

Krieger proved Theorem 15 in two steps. On the first step, he shows only existence of some finite alphabet B and of a code $\bar{\Phi}$. However, his method yields highly over-

redundant codes so that the asserted bound is not obtained. On the second step he uses an approximation technique for approximation of ergodic sources by periodic ones (see [157]) in order to get the desired bound for $\|B\|$.

Smorodinsky [136] showed how to get Theorem 15 using Ornstein's coding technique. One uses Sinai's theorem to construct an initial coding. Then a small perturbation of the initial process is performed in order one has an ergodic process with enough entropy (for reasons explained in Section 9). The rest of the construction follows the idea of getting a converging sequence of ever better codes as described in Section 9.

14. Related Representation Problems

A large part of contemporary ergodic theory deals with smooth dynamical systems. Although this part of ergodic theory stands outside of our main interest, the methods of symbolic dynamics (which traces back to Hadamard; good surveys are [6] and [94]) may also be considered as representation techniques (see [17, 21, 7, 23]).

The main idea is as follows. Given a smooth dynamical system one can find (under some conditions concerning the local stability properties of orbits) a finite family of sets which behaves, from the point of view of dynamics, as states of a Markov chain. Usually, some transitions are forbidden so that one gets a transition matrix with some entries possibly zero. Any such matrix induces a closed invariant subset of the shift space over the alphabet consisting of the states. In this way one obtains a topological Markov chain (see Section 17). The ergodic theory of the latter (problems like existence of invariant measures, their uniqueness, ergodic properties, etc.) is well developed so that one can easily get the corresponding conclusions for the original smooth system. Of course, the construction of the above mentioned family (usually called a Markov partition) may be a difficult task.

15. Improvements on Krieger's Theorem

In this section we give several improvements of Krieger's theorem to the effect that the encoded process has some prescribed properties.

Theorem 16. Let $[A, \mu, X]$ be an aperiodic, stationary and ergodic source over a countable discrete alphabet A. If $h(X) < \infty$ and $K = \text{INT}\{\exp h(X)\} + 1$, and if $p = (p_1, \ldots, p_K)$ is a probability vector with $H(p) > h(X)$ then for any $\varepsilon > 0$ we can find a source $[B, \nu, Y]$ over an alphabet B with at most K letters such that

$$|\text{Prob}[Y_0 = k] - p_k| < \varepsilon, \quad 1 \leqq k \leqq K$$

and a perfectly noiseless code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $Y = \bar{\Phi}X$.

This result was obtained by Denker [28]. An important contribution of Denker is that he completely clarified which tools are needed to prove a generator theorem.

Indeed, his proof depends only on Shannon-McMillan's theorem and Rohlin's lemma. When compared with the original Krieger's proof, Denker does not make essential use of the linear ordering of "one-dimensional time" (a point on which Krieger's proof fails when more dimensional situations are considered). Thus, Theorem 15 can be extended to all dynamical systems for which one can prove the Shannon-McMillan's theorem and the Rohlin's lemma. At present, the most general result available is for free actions of countable amenable groups of invertible transformations of a non-atomic Lebesgue space [148]. The Shannon-McMillan's theorem was proved, even for more general dynamical systems, in [65] and Rohlin's lemma in [104].

It is easy to modify Denker's proof so as to get approximations in spirit of Theorem 16 to any prescribed n-dimensional distribution ($n \in \mathbf{N}$) with enough entropy. Next consider approximation in entropy. This problem is motivated by the following consideration (see also Section 22 below). If Y is an IID process then

$$H(Y_0) = -\sum_{a \in A} \text{Prob}\,[Y_0 = a] \log \text{Prob}\,[Y_0 = a]$$

equals the entropy rate $h(Y)$ (cf., e.g., [11]). Conversely, if Y is an ergodic process such that the difference $|h(Y) - H(Y_0)|$ is small then, using the method described in [128], pp. 52–53 (see also Lemma 4.1 in [103]) one can conclude that Y is nearly independent (i.e., ε-independent, in Ornstein's language, for some small $\varepsilon > 0$). Thus, from the point of view of redundancy removal it is desirable to have the following result:

Theorem 17. Given $\varepsilon > 0$ and B with $\|B\| \leqq \text{INT}\{\exp h(X)\} + 1$. Then there is a source $[B, \nu, Y]$ such that $|h(Y) - H(Y_0)| < \varepsilon$ and a perfectly noiseless code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $Y = \bar{\Phi}X$.

Theorem 17 is a slight generalization of the usual formulation of entropy approximation property which says that for any probability vector p with $H(p) > h(X)$ we find a process Y isomorphic to X such that $|H(Y_0) - H(p)| < \varepsilon$. Both assertions can be obtained using the method of proof employed in [136].

It is of interest to have a theorem involving directly closeness to a prescribed IID process. A natural tool for measuring closeness is the weak topology on the space $\mathscr{E}(B)$. Let $\varkappa, \lambda \in \mathscr{E}(B)$. Put

$$d_w(\varkappa, \lambda) = \sum_{n=1}^{\infty} 2^{-n} \sum_{\mathbf{b} \in B^n} |\varkappa^n(\mathbf{b}) - \lambda^n(\mathbf{b})| .$$

The metric d_w is compatible with the weak topology on $\mathscr{E}(B)$ (see [116] or [12] for details on weak topology).

Theorem 18. Given $\varepsilon > 0$ and $\|B\| \leqq \text{INT}\{\exp h(X)\} + 1$ there exists a perfectly noiseless code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $d_w(\mu\bar{\Phi}^{-1}, \text{dist}\,(\overline{Y})) < \varepsilon$, where $\overline{Y}$ is the IID equiprobable process over alphabet B. More generally, if $\overline{Y}$ is any ergodic process over a finite alphabet such that $h(\overline{Y}) > h(X)$, then the above conclusion is true.

Theorem 18 is implicitly contained already in Krieger's original paper. It can be shown, however, that it is a particular case (apart from some technical details) of a zero-error transmission theorem of Kieffer [72] (the idea of the proof is sketched in [146]).

A principal novelty of Kieffer's proof (which also follows the lines of Ornstein's technique) is that Kieffer develops converging sequences of both encoders and decoders so that, in the limit, the decoder becomes the inverse of the encoder. Thus, there is no need for a Baire category argument in order to prove invertibility of the limiting code.

As pointed out in [146] we cannot replace closeness in d_w by closeness in $\bar{d}$. The reason is that the first part of Theorem 18 would then force X to be an FD process, i.e., a Bernoulli process (see Theorem 9).

Other types of approximations involve the property that the encoded process ΦX be a factor of some reasonable type of process, e.g., a factor of a mixing Markov chain (see [86] and [30]) or a factor of a strictly ergodic process (the surprising result that this is always possible to find such a coding is known as strictly ergodic embedding; see [30, 86, 58, 56]).

However, all these results are approximations. A natural question arises whether it is possible to prove exact results. We devote the next section to that problem.

16. Kieffer's Isomorphism Theorem

Grillenberger and Krengel [52] proved a theorem on stationary coding of processes in order to achieve a given marginal distribution. Recently, Kieffer [77] obtained a theorem which unifies the Grillenberger-Krengel theorem and his previous zero-error transmission theorem [72].

Let $\mathscr{M}(B)$ denote the set of all T_B-invariant probability measures on $B^{\mathbf{Z}}$, $\|B\| < \infty$. In this section only, $\mathscr{E}(B)$ will mean the set of all *aperiodic*, stationary and ergodic measures on $B^{\mathbf{Z}}$. We say that a set $\mathscr{M} \subset \mathscr{E}(B)$ obeys *condition* (A) if the conditions that $(\mu_n;\ n \geqq 1) \subset \mathscr{E}(B)$ and $d_w(\mu_n, \mathscr{M}) \to 0$ imply that $\bar{d}(\mu_n, \mathscr{M}) \to 0$ as $n \to \infty$. In other words, if $\mu \in \mathscr{M}$ then for any $\varepsilon > 0$ we can find a $\delta > 0$ such that if $\nu \in \mathscr{E}(B)$ and $d_w(\mu, \nu,) < \delta$ then there exists a $\nu' \in \mathscr{M}$ with $\bar{d}(\nu, \nu') < \varepsilon$. Let

$$\tilde{h}(\mathscr{M}) = \sup \{h(\mu) : \mu \in \mathscr{M}\} .$$

Theorem 19 (Kieffer's isomorphism theorem). Let X be a stationary ergodic aperiodic process with a finite state space and let $\mathscr{M} \subset \mathscr{E}(B)$ be a weak G_δ subset of $\mathscr{E}(B)$ obeying the condition (A). If $h(X) < \tilde{h}(\mathscr{M})$, then X is isomorphic with a process Y with dist $(Y) \in \mathscr{M}$.

The proof follows in main lines Ornstein's technique as modified for channel coding purpose [49, 51]. First of all, one proves a synchronization lemma which ensures the existence of synchronization words. These words cannot be mistaken

for cyclic shifts of themselves and thus will be able to play the role of the auxiliary binary codding employed in Section 9, namely, to indicate when to block code.

On the second step, one determines, using a marriage lemma, a good block coding function which is then used to produce an initial stationary coding. Then one shows how to get a very good code from a good one and repeated use of this assertion yields a converging sequence of ever better codes so that, in the limit, we get the conclusion of Theorem 19.

However, the technical details differ considerably from those sketched in Section 9 and a proper understanding of the proof requires a reader familiar with some techniques of channel coding (see [72] and [75]).

The rest of this section will be devoted to Grillenberger-Krengel theorem. Let B be a finite set, $m \geqq 2$, and $\pi : B^m \to [0, 1]$ a given probability vector. Let $X = (x_i; i \in \mathbf{Z})$ be the sequence of coordinate maps $B^{\mathbf{Z}} \to B$. The vector π is said to be *invariant* if

$$\mathrm{dist}_\pi (X_1^{m-1}) = \mathrm{dist}_\pi (X_2^m) .$$

If $\mathbf{b} = (b_1, \ldots, b_m) \in B^m$, let $\pi(b_m \mid b_1, \ldots, b_{m-1}) = \pi(\mathbf{b})/\pi_{m-1}(b_1, \ldots, b_{m-1})$ if the denominator is positive, and 0 otherwise, where $\pi_{m-1} = \mathrm{dist}_\pi (X_1^{m-1})$. For $n > m$, let π_n be the probability vector on B^n defined by

$$\pi_n(b_1, \ldots, b_n) = \pi_{m-1}(b_1, \ldots, b_{m-1})\, \pi(b_m \mid b_1, \ldots, b_{m-1}) \ldots$$
$$\ldots \pi(b_n \mid b_{n-m+1}, \ldots, b_{n-1}) .$$

These relations determine a consistent family of finite-dimensional distributions π_n, $n = 1, 2, \ldots$, and we let $\hat{\pi}$ denote the unique T_B-invariant probability measure on $B^{\mathbf{Z}}$ for which $\hat{\pi}\{x \in B^{\mathbf{Z}} : x_0^{n-1} = \mathbf{b}\} = \pi_n(\mathbf{b})$, $\mathbf{b} \in B^n$, $n \in \mathbf{N}$. This $\hat{\pi}$ is an $(m-1)$-step Markov measure called the *Markov extension of* π (such extensions have been previously considered in [52, 85, 64]).

Lemma 20 [70]. The shift T_B is ergodic (mixing) with respect to $\hat{\pi}$ if and only if there exists an ergodic (mixing) stationary source $[B, \mu, X]$ such that $\mathrm{dist}_\mu (X_1, \ldots, X_m) = \pi$.

Accordingly, π itself is called ergodic or mixing. It is easy to see that if π is a mixing probability vector then there exists a unique measure $\lambda_\pi^* \in \mathcal{M} = \{\lambda \in \mathcal{E}(B) : \mathrm{dist}_\lambda (X^m) = \pi\}$ maximizing entropy, i.e. $h(\lambda_\pi^*) = \sup \{h(\lambda) : \lambda \in \mathcal{M}\}$. This measure is $(m-1)$-step Markov so that

$$h(\lambda_\pi^*) = H(\pi) - H(\pi_{m-1})$$

(see e.g. [11] or [52]). This will be our $\tilde{h}$.

Lemma 21. If π is a mixing invariant probability vector on B^m then the set $\mathcal{M} = \{\lambda \in \mathcal{E}(B) : \mathrm{dist}_\lambda (X^m) = \pi\}$ obeys the condition (A).

It should be noted that Lemma 21 is very non-trivial (cf. Theorem 2 of [70]).

At present, there are no simple methods for a direct verification of condition (A), and thus one is forced to use rather tricky coding techniques for this purpose. As a corollary to Lemma 21 one obtains:

Theorem 22 (Grillenberger-Krengel theorem). Let X be an aperiodic, stationary and ergodic process such that $h(X) < H(\pi) - H(\pi_{m-1})$, where π is a mixing invariant probability vector on B^m. Let A denote the alphabet of X. Then there exists an invertible stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $\operatorname{dist}((\bar{\Phi}X)^m) = \pi$.

The original proof of Theorem 22 in [52] as well as its simplification in [70] were based on Denker's proof of Theorem 16. Grillenberger and Krengel observed that having an ε-approximation to π it is possible to redistribute the probabilities in the encoded process in such a way that we get exact coincidence of distributions.

On the other hand, Kieffer's approach via Theorem 19 is based on Ornstein's coding technique so that we have a unique method for solving a large class of classification and representation problems.

PART V: CLASSIFICATION PROBLEMS FOR MARKOV CHAINS

17. Topological Markov Chains

Topological Markov chains naturally appear as symbolic representations of smooth dynamical systems (see Section 14). The classification problems for topological Markov chains however lead to results which are very stimulating also from the point of view of coding problems.

Let A be a finite set and σ an $\|A\| \times \|A\|$ irreducible $0-1$ matrix. Let $A(\sigma) = \{x \in A^{\mathbf{Z}} : \sigma(x_i, x_{i+1}) = 1 \text{ for all } i \in \mathbf{Z}\}$. Then $A(\sigma)$ is a closed T_A-invariant set. The restriction T_σ of T_A to $A(\sigma)$ is a homeomorphism of a compact metric space, i.e., a topological system [30] known under several names: *intrinsic Markov chain* [106], *subshift of finite type* [135] or, a *topological Markov chain* [2].

If $\boldsymbol{P}$ is a stochastic matrix such that $\boldsymbol{P}(i, j) = 0$ if and only if $\sigma(i, j) = 0$, then the T_A-invariant Markov probability measure $\mu_{\boldsymbol{P}}$ determined by $\boldsymbol{P}$ is supported by $A(\sigma)$. In what follows we shall consider only Markov probability measures supported by sets $A(\sigma)$ for irreducible $0-1$ matrices σ. In general, an irreducible topological Markov chain (i.e., one constructed from an irreducible matrix σ) can have many invariant measures. However, there is one which deserves particular attention. Let

$$\lambda_\sigma = \sup \{\lambda : \sigma v \geqq \lambda v, v = (v_1, \ldots, v_{\|A\|}), v_i > 0, \sum v_i = 1\} .$$

λ_σ is called the *Perron value* of σ; it is the largest eigenvalue in the sense that $\lambda_\sigma < |\lambda|$ for all other eigenvalues and λ_σ has multiplicity one [40]. Let $v > 0$ be a column vector and $u > 0$ a row vector associated with λ_σ. Let $\boldsymbol{P} = (p(i, j))$, where

$$p(i, j) = \sigma(i, j)\, v_j / v_i \lambda_\sigma ,$$

and $\boldsymbol{p} = (p_1, \ldots, p_{\|A\|})$, where

$$p_i = u_i v_i / \sum_i u_i v_i .$$

Then $\boldsymbol{p}\boldsymbol{P} = \boldsymbol{p}$ so that the pair $(\boldsymbol{p}, \boldsymbol{P})$ induces an invariant Markov probability measure on $A(\sigma)$ called the *Parry measure* after Parry [106] who proved its uniqueness in the following sense:

Proposition 23. If μ_σ is the Parry measure then $h(\mu_\sigma) = \log \lambda_\sigma$. Conversely, if μ is a T_σ-invariant probability measure on $(A(\sigma), A(\sigma) \cap \mathscr{A}^{\mathbf{Z}})$ such that $h(\mu) = \log \lambda_\sigma$ then $\mu = \mu_\sigma$.

A topological Markov chain $(A(\sigma), T_\sigma)$ is said to be a *finite extension* of a topological Markov chain $(B(\tau), T_\tau)$ (and $(B(\tau), T_\tau)$ a *finite factor* of $(A(\sigma), T_\sigma)$) if there exists a stationary code $\bar{\Phi} : A(\sigma) \to B(\tau)$ which is (i) boundedly finite-to-one, (ii) continuous, and (iii) surjective. Two topological Markov chains are called *finitely equivalent* if there exists a topological Markov chain which is a common finite extension of both.

Since $A(\sigma) \subset A^{\mathbf{Z}}$, $B(\tau) \subset B^{\mathbf{Z}}$, a finite factor map $\bar{\Phi} : A(\sigma) \to B(\tau)$, being continuous, admits a similar description as a finitary code. We can find integers $0 \leqq l < k$ and a map $\Phi : A^k \to B$ such that for each $i \in \mathbf{Z}$, $(\bar{\Phi}\boldsymbol{x})_{i+l} = \Phi(x_i^{i+k-1})$. Such a code is called a k-block map.

Classification problems for topological Markov chains are related with the topological entropy ([3], an alternate definition imitating the definition of Hausdorff dimension was introduced by Bowen [18]). Let T be a homeomorphism of a compact metric space Y. Let $\mathscr{U} = (U_1, \ldots, U_m)$ be an open cover of Y. Put

$$\mathscr{U}^N = \{U_{i_0 \ldots i_{N-1}} = U_{i_0} \cap T^{-1}U_{i_1} \cap \ldots \cap T^{-N+1}U_{i_{N-1}}\} .$$

Let $k(\mathscr{U}^N)$ denote the minimal cardinality of a subcover of $\mathscr{U}^N$. Then the *topological entropy* of the dynamical system (Y, T) is defined by

$$h_{\text{top}}(T) = \sup_{\mathscr{U}} H(T \mid \mathscr{U})$$

where

$$H(T \mid \mathscr{U}) = \inf_{N \geqq 1} N^{-1} \log k(\mathscr{U}^N) .$$

If $(Y, T) = (A(\sigma), T_\sigma)$, where σ is an irreducible $0-1$ matrix, then the relation

$$h_{\text{top}}(T_\sigma) = \log \lambda_\sigma$$

is a theorem due to Parry [106]. The next theorem says that finite equivalence is the right notion of "similarity" for topological Markov chains (the natural concept of topological conjugacy is too strong; see examples in [4]).

Theorem 24. [108]. Two topological Markov chains $(A(\sigma), T_\sigma)$ and $(B(\tau), T_\tau)$ are finitely equivalent if and only if they have the same topological entropy; that is, if and only if $\lambda_\sigma = \lambda_\tau$.

The proof of Theorem 24 is based on an interesting lemma due to Furstenberg (various proofs of Furstenberg lemma are given in [4, 108, 115]).

Lemma 25. (Furstenberg). Let σ and τ be irreducible nonnegative matrices with integral entries. Then $\lambda_\sigma = \lambda_\tau$ if and only if there exists a strictly positive integral matrix $\boldsymbol{U}$ with $\boldsymbol{U}\sigma = \tau\boldsymbol{U}$.

Let us recall several related results. The first one is due to Coven and Paul [24].

Proposition 26. Let $(A(\sigma), T_\sigma)$ and $(B(\tau), T_\tau)$ be two irreducible topological Markov chains with $\lambda_\sigma = \lambda_\tau$. Let $\bar{\Phi} : A(\sigma) \to B(\tau)$ be stationary and continuous. Then the following are equivalent: (a) $\bar{\Phi}$ is surjective, (b) $\bar{\Phi}$ is boundedly finite-to one, and (c) $\bar{\Phi}$ is measure-preserving relative to the Parry measures.

Proposition 26 was applied in [91] to study the particular case when λ_σ is a positive integer so that $(A(\sigma), T_\sigma)$ is related to the full shift over some finite alphabet. The following is a simple consequence of Proposition 26:

Corollary 27. Let $(A(\sigma), T_\sigma)$ and $(B(\tau), T_\tau)$ be two irreducible topological Markov chains, and let $\bar{\Phi} : A(\sigma) \to B(\tau)$ be a stationary and continuous surjection. Then $\bar{\Phi}$ is boundedly finite-to-one if and only if $\lambda_\sigma = \lambda_\tau$.

18. Stochastic Markov Chains

Recall that we always assume that the stochastic Markov chains are supported by topological Markov chains in the sense described in Section 17. So, let $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu, T_A)$ and $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu, T_B)$ be two such (stationary) Markov chains. Then $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu, T_A)$ is said to be a *finite extension* of $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu, T_B)$ (and the second one a *finite factor* of the former one) if there exists a boundedly finite-to-one continuous measure-preserving surjection $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ with $\bar{\Phi} \circ T_A = T_B \circ \bar{\Phi}$ a.e. The two Markov chains are said to be *finitely equivalent* if they have a common Markov finite extension. A seemingly weaker notion of equivalence would be the result of dropping out the Markov property of the common extension. However, the following result is true:

Proposition 28. [150]. Let $(A(\sigma), T_\sigma)$ be a topological Markov chain, and let $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu, T_B)$ be a stochastic Markov chain. Let $\bar{\Phi} : A(\sigma) \to B^{\mathbf{Z}}$ be a boundedly finite-to-one continuous surjection (onto the support of the measure ν which is supposed to be $B(\tau)$ for some τ) such that $\bar{\Phi} \circ T_\sigma = T_B \circ \bar{\Phi}$. Then there exists a unique T_σ-invariant probability measure which makes $\bar{\Phi}$ measure-preserving. If $\bar{\Phi}$ is a k-block map then this measure is k'-step Markov for some $k' \leqq k$.

A surprising result is that even finite factor maps induce the same cocycle-coboundary equation as isomorphisms with finite expected code length and inverse code length:

Theorem 29. [113]. Let $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu, T_B)$ be a finite factor of $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu, T_A)$ where the two dynamical systems are stochastic Markov chains supported by the topological ones. Let $\bar{\Phi}$ denote the corresponding code. Then there exists a continuous function g on $A^{\mathbf{Z}}$ such that

$$I_A = I_B \circ \bar{\Phi} + g \circ T_A - g \,.$$

If $\bar{\Phi}$ is a k-block map then $g(x) = g(x^k)$, i.e., g depends on at most k coordinates.

It should be noted that Theorem 29 is valid only for Markov chains as indicated. In general, a finite factor map of arbitrary measure theoretic dynamical systems need not give rise to the cocycle-coboundary equation.

Tuncel [150] introduced a new invariant of the cocycle-coboundary equation. To this end, let $\boldsymbol{P} = (\boldsymbol{P}(i,j))$ be an irreducible stochastic matrix. For each $t \in \mathbf{R}$ let $\boldsymbol{P}^t(i,j) = \boldsymbol{P}(i,j)^t$ when $\boldsymbol{P}(i,j) > 0$, and $\boldsymbol{P}^t(i,j) = 0$ otherwise. Let

$$\beta_{\boldsymbol{P}}(t) = \lambda_{\boldsymbol{P}^t} \,,$$

where $\lambda_{\boldsymbol{P}^t}$ is the Perron value of $\boldsymbol{P}^t$. This defines a function $\beta_{\boldsymbol{P}} : \mathbf{R} \to \mathbf{R}^+$. If $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu_{\boldsymbol{P}}, T_A)$ is the Markov chain determined by $\boldsymbol{P}$, we call $\beta_{\boldsymbol{P}}$ the *β-function* of $\mu_{\boldsymbol{P}}$. An explicit form of this function was obtained in [150, 115]:

Lemma 30. Let $\boldsymbol{P}$ be an irreducible stochastic $\|A\| \times \|A\|$ matrix, and let $\mu_{\boldsymbol{P}}$ denote the Markov measure on $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}})$ determined by $\boldsymbol{P}$. Let $t \in \mathbf{R}$. Then

$$\log \beta_{\boldsymbol{P}}(1-t) = \lim_{n\to\infty} n^{-1} \log \int \exp\Big(t \sum_{i=0}^{n-1} I_A \circ T_A^i\Big) \, \mathrm{d}\mu_{\boldsymbol{P}} \,.$$

Furthermore, $\beta_{\boldsymbol{P}}$ is an analytic function, $\beta_{\boldsymbol{P}}(1) = 1$,

$$\beta'_{\boldsymbol{P}}(1) = -h(\mu_{\boldsymbol{P}}) \,, \quad \text{and} \quad \beta''_{\boldsymbol{P}}(1) = \sigma^2(T_A, \gamma_A) + h(\mu_{\boldsymbol{P}})^2 \,.$$

Thus, the known invariants, the entropy and the information variance, can be derived from the β-function. A combination of Lemma 30 with Theorem 29 gives the next result:

Theorem 31. If a Markov chain $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \mu_{\boldsymbol{Q}}, T_B)$ is a finite factor of a Markov chain $(A^{\mathbf{Z}}, \mathscr{A}^{\mathbf{Z}}, \mu_{\boldsymbol{P}}, T_A)$ then $\beta_{\boldsymbol{P}} = \beta_{\boldsymbol{Q}}$. In particular, the β-function is invariant under finite equivalence within the class of all stochastic Markov chains supported by topological Markov chains. Furthermore, both entropy and information variance are invariant under finite equivalence within the indicated class.

It is conjectured that the β-function is even a complete invariant (some evidence in favour of this conjecture is gathered in [115]).

Adler and Marcus [4] employed a more restrictive notion of factors and equivalence in the sense that they required the factor maps to be not only finite-to-one but also one-to-one almost everywhere (with respect to any invariant probability measure which is ergodic and positive on all open sets; a topological formulation

of almost everywhere valied assertions is possible, see the concluding section of [4]). It is easy to show that the β-function is not complete with respect to this stronger concept of equivalence. For example, the Markov chains determined by matrices

$$\begin{pmatrix} p & q \\ p & p \end{pmatrix} \text{ and } \begin{pmatrix} p & p \\ p & p \end{pmatrix}, \quad 0 < p < \tfrac{1}{2}, \quad p + q = 1$$

are both aperiodic and have identical β-functions, $\beta(t) = p^t + q^t$. However, the group invariant introduced in Section 12 distinguishes them.

Next fix a probability vector $\boldsymbol{p} = (p(1), \ldots, p(n))$ and consider all irreducible matrices $\boldsymbol{M}$ which have exactly n non-zero entries in each row, and these entries form a permutation of the vector $\boldsymbol{p}$. Markov chains determined by matrices $\boldsymbol{M}$ and their inverses are called *Bernoulli-type shifts*. Using Lemma 30, it is easy to calculate the β-function, which is the same for all such shifts based on the same vector $\boldsymbol{p}$:

$$\beta(t) = p(1)^t + \ldots + p(n)^t .$$

Proposition 32. [115]. All Bernoulli-type shifts based on the same probability vector are finitely equivalent.

In particular, if $T(\boldsymbol{p})$ and $T(\boldsymbol{q})$ are Bernoulli shifts over the same alphabet such that $\boldsymbol{p}$ is merely a permutation of $\boldsymbol{q}$, then they are finitely equivalent. The converse is a conjecture. A slightly weaker result is available supporting that conjecture. It was proved in [150] using a variational principle for the β-function. Let $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ be an isomorphism between two stationary sources $[A, \mu]$ and $[B, \nu]$. We call $\bar{\Phi}$ *regular* if there is an integer $k \geqq 0$ such that $(\bar{\Phi}\boldsymbol{x})_0$ is determined by knowing $(x_j;\ j \leqq k)$, and the same is true for Φ^{-1}.

Proposition 33. Two Bernoulli shifts are regularly isomorphic if and only if there exists an enumeration of one alphabet giving identical one-dimensional distributions.

PART VI: UNIVERSAL CODES IN ERGODIC THEORY

19. Interpretations of the Ergodic Decomposition

Let us return to a classical result of ergodic theory which traces back to Krylov and Bogolyubov ([87], see also [121, 105, 154]). Let $[A, \mu]$ be a stationary source over a countable discrete alphabet A. A sequence $x \in A^{\mathbf{Z}}$ is said to be *regular*, in symbols $x \in R_A$, if there exists a measure $\mu_x \in \mathscr{E}(A)$ such that for any cylinder E (i.e., for any E of the form $\{x \in A^{\mathbf{Z}} : x_i = a_1, \ldots, x_{i+n-1} = a_n\}$, $i \in \mathbf{Z}$, $n \geqq 1$, $a_1, \ldots, a_n \in A$) we have

$$\mu_x(E) = \lim_{n \to \infty} n^{-1} \sum_{j=0}^{n-1} 1_E(T_A^j x) .$$

Theorem 34. (Ergodic decomposition theorem). The set R_A is invariant, measurable, and $\mu(R_A) = 1$ for any $\mu \in \mathcal{M}(A)$. A measure $\mu \in \mathcal{M}(A)$ is ergodic if and only if it satisfies $\mu\{x \in R_A : \mu_x = \mu\} = 1$. Given $\mu \in \mathcal{M}(A)$, the function $x \mapsto \mu_x(E) : R_A \to \to [0, 1]$ is μ-integrable for each $E \in \mathcal{A}^{\mathbf{Z}}$, and

$$\mu(E) = \int_{R_A} \mu_x(E)\, \mu(\mathrm{d}x)\,.$$

More generally, if f is μ-integrable, then the function $x \mapsto \mathsf{E}_{\mu_x}(f)$ is also integrable, and

$$\int f\, \mathrm{d}\mu = \int_{R_A} \left[\int f\, \mathrm{d}\mu_x\right] \mu(\mathrm{d}x)\,.$$

There are two different interpretations of this result. The first one is due to Gray and Davisson [42, 43]. Accordingly, having some $\mu \in \mathcal{M}(A)$ means the "true" source statistics is described by one of its ergodic subsources μ_x, $x \in R_A$. The measure μ itself is considered merely as a weighting prior that expresses our degree of evidence in favour of the unknown "true" source.

Or, we can simply suppose that μ itself is the "true" statistics, i.e., we say that the true source is stationary but non-ergodic [154].

Davisson [25] pointed out that this is not merely a play of words and that different interpretations suggest various formulations of the basic aims of universal source coding. Within the former interpretation, one seeks for codes which perform optimally for each member of the class of available ergodic subsources. In this way we can get the best result possible each time when a particular ergodic subsource turns out to be the "true" one. Within the second interpretation, one usually seeks for codes which perform optimally for even the "worst" component.

A related third approach deals with universal coding for classes of sources without assuming some prior measure. This part will be devoted to an explanation of these approaches within coding problems of ergodic theory.

20. Universal Codes for Classes of Sources

Let $(\Omega, \mathcal{F})$ be a standard Borel space (in applications, we shall usually work with a shift space). Let $T : \Omega \to \Omega$ be an automorphism of $(\Omega, \mathcal{F})$ and let $\mathcal{M}(T)$(and $\mathcal{E}(T)$) denote the set of all T-invariant (and ergodic) probability measures on $(\Omega, \mathcal{F})$. Fix a finite set B. Let $\mathcal{P}$ denote the set of all finite partitions ζ of Ω into sets from $\mathcal{F}$ indexed by B. Since $(\Omega, \mathcal{F})$ is standard, each $\zeta \in \mathcal{P}$ induces a measurable map $\bar{\Phi}_\zeta : \Omega \to \to B^{\mathbf{Z}}$ (see Section 2) which is stationary: $\bar{\Phi}_\zeta \circ T = T_B \circ \bar{\Phi}_\zeta$. Since the Baire category argument works equally well in standard Borel spaces, we can extend Proposition 2 (c). Let $\mu \in \mathcal{M}(T)$. Then $\zeta \in \mathcal{P}$ is a generator (relative to (T, μ)) if and only if $\bar{\Phi}_\zeta$ is an isomorphism between $(\Omega, \mathcal{F}, \mu, T)$ and $(B^{\mathbf{Z}}, \mathcal{B}^{\mathbf{Z}}, \mu\bar{\Phi}_\zeta^{-1}, T_B)$ (if $(\Omega, \mathcal{F}, \mu)$ is a Lebesgue

space then it should be more appropriate to work with a completion of the σ-field $\mathscr{B}^{\mathbf{Z}}$; however, the mod 0 properties of an isomorphism eliminate the differences between σ-fields and their completions).

Let $\mathscr{E} \subset \mathscr{E}(T)$. For each $\mu \in \mathscr{E}$, let a subclass $\mathscr{P}_\mu \subset \mathscr{P}$ be specified. A partition ζ is said to be *universal*, if $\zeta \in \cap \{\mathscr{P}_\mu : \mu \in \mathscr{E}\}$. E.g., for each $\mu \in \mathscr{E}$, $\mathscr{P}_\mu$ may be chosen so that for each $\zeta \in \mathscr{P}_\mu$, the encoded process (using the coding Φ_ζ) has some prescribed property. Then a universal partition gives rise to a code such that the encoded processes each have the specified properties, for all $\mu \in \mathscr{E}$.

Kieffer and Rahe [79] found various sufficient conditions for the existence of universal partitions. Let us introduce several notations. If $\zeta, \xi \in \mathscr{P}$ and $\mu \in \mathscr{M}(T)$ then the partition distance is defined to be the number

$$|\zeta - \xi|_\mu = \tfrac{1}{2} \sum_{b \in B} \mu(C^b \, \Delta D^b) ,$$

where $\zeta = (C^b;\ b \in B)$ and $\xi = (D^b;\ b \in B)$. We assume some fixed ordering on B so that the elements of $\mathscr{P}$ can be considered as ordered partitions, too. For each μ, let

$$\varrho_\mu(\zeta, \xi) = |\zeta - \xi|_\mu ; \quad \zeta, \xi \in \mathscr{P} .$$

On $\mathscr{E}(T)$ we define the least σ-field such that for each $E \in \mathscr{F}$, the map $\mu \mapsto \mu(E) : \mathscr{E}(T) \to [0, 1]$ is measurable. From now on we assume that $\mathscr{E}$, $\mathscr{D}$ denote measurable subsets of $\mathscr{E}(T)$ and we let $\mathscr{F}(\mathscr{E})$, $\mathscr{F}(\mathscr{D})$ denote the induced σ-fields.

Theorem 35. Let $\mathscr{E} \subset \mathscr{E}(T)$ and let $\widetilde{\mathscr{P}} \subset \mathscr{P}$ be countable. Let $\{\mathscr{P}_\mu : \mu \in \mathscr{E}\}$ be such that (a) for any $\zeta \in \mathscr{P}$, $\{\mu \in \mathscr{E} : \zeta \in \mathscr{P}_\mu\} \in \mathscr{F}(\mathscr{E})$, (b) for any $\mu \in \mathscr{E}$, $\zeta \in \mathscr{P}_\mu$, and $\xi \in \mathscr{P}$, the condition that $|\zeta - \xi|_\mu = 0$ entails $\xi \in \mathscr{P}_\mu$, and (c) for any $\mu \in \mathscr{E}$, $\widetilde{\mathscr{P}} \cap \mathscr{P}_\mu \neq \emptyset$. Then $\cap \{\mathscr{P}_\mu : \mu \in \mathscr{E}\} \neq \emptyset$.

If $\widetilde{\mathscr{P}}$ is ϱ_μ-dense in $\mathscr{P}$ for each $\mu \in \mathscr{E}$, Theorem 35 gives rise to a different sufficient condition:

(a) for each $\mu \in \mathscr{E}$, $\mathscr{P}_\mu$ is a non-empty ϱ_μ-open set,
(b) for each $\zeta \in \mathscr{P}$, $\{\mu \in \mathscr{E} : \zeta \in \mathscr{P}_\mu\} \in \mathscr{F}(\mathscr{E})$.

Now let $\mathscr{E} \subset \mathscr{E}(T)$. A function $\Phi : \mathscr{E} \times \mathscr{P} \to [0, \infty]$ is called *admissible* if for each $\zeta \in \mathscr{P}$, $\Phi(\cdot, \zeta)$ is measurable and for each $\mu \in \mathscr{E}$, $\Phi(\mu, \cdot)$ is ϱ_μ-continuous. Further, if $\mathscr{Q} \subset \mathscr{P}$ and $\mu \in \mathscr{M}(T)$, put

$$\varrho_\mu(\zeta, \mathscr{Q}) = \inf \{|\zeta - \xi|_\mu : \xi \in \mathscr{Q}\} , \quad \zeta \in \mathscr{P} .$$

The idea of admissible function is that, given μ, the zeroes of $\Phi(\mu, \cdot)$ give desired partitions provided Φ is chosen in an appropriate way (this idea appears also in [76] and [78]).

Theorem 36. Let $\mathscr{E} \subset \mathscr{E}(T)$, and let $\{\mathscr{P}_\mu : \mu \in \mathscr{E}\}$ be a family of non-empty sets $\mathscr{P}_\mu \subset \mathscr{P}$. Suppose that (a) for each $\zeta \in \mathscr{P}$, the map $\mu \mapsto \varrho_\mu(\zeta, \mathscr{P}_\mu)$ is measurable and

(b) there exists a sequence $(\Phi_n; n \geqq 1)$ of admissible functions $\Phi_n : \mathscr{E} \times \mathscr{P} \to [0, \infty)$ such that $\zeta \in \mathscr{P}_\mu$ if and only if $\inf\{\Phi_n(\mu, \zeta) : n \geqq 1\} = 0$ for each $\mu \in \mathscr{E}$ and $\xi \in \mathscr{P}$. Then $\bigcap \{\mathscr{P}_\mu : \mu \in \mathscr{E}\} \neq \emptyset$.

Condition (a) of Theorem 36 is satisfied, if, for example, the conditions (a) and (b) formulated after Theorem 35 are valid. This, combined with Theorem 36 gives sufficient conditions in the sense that $\Phi(\mu, \zeta) = 0$ for all $\mu \in \mathscr{E}$.

Theorem 37. Let $\mathscr{E} \subset \mathscr{E}(T)$ and $\Phi : \mathscr{E} \times \mathscr{P} \to [0, \infty)$ be given. Suppose that (a) there exist admissible functions $(\Phi_n; n \geqq 1)$ such that $\Phi = \inf \Phi_n$, (b) given $\mu \in \mathscr{E}$ and $\varepsilon > 0$ there is a $\delta > 0$ such that the conditions $\zeta \in \mathscr{P}$ and $\Phi(\mu, \zeta) < \delta$ imply there is a ξ for which $|\zeta - \xi|_\mu < \varepsilon$ and $\Phi(\mu, \xi) = 0$, and (c) for each $\mu \in \mathscr{E}$ there exists $\zeta \in \mathscr{P}$ with $\Phi(\mu, \zeta) = 0$. Then there exists $\zeta \in \mathscr{P}$ such that $\Phi(\mu, \zeta) = 0$ for any $\mu \in \mathscr{E}$.

The approximation property (b) is crucial here. It was used recently by Kieffer [78] in order to extend Grillenberger-Krengel theorem (see Theorem 22) to aperiodic non-ergodic sources. Actually, Kieffer constructed a function $\Phi : \mathscr{P} \to [0, \infty)$, continuous with respect to the partition metric, whose zeroes are the generators for $(T_A, \mathrm{dist}(X))$ and which has an approximation property analogous to condition (b) of Theorem 37. By conclusion of that theorem, we find a generator relative to all ergodic sources simultaneously.

Finally, we have the following criterion:

Theorem 38. Let $\mathscr{E} \subset \mathscr{E}(T)$ and let $\mathscr{P}_\mu$, $\mu \in \mathscr{E}$ be nonempty subsets of $\mathscr{P}$ such that (a) each $\mathscr{P}_\mu$ is ϱ_μ-closed and (b) for any $\zeta \in \mathscr{P}$, the map $\mu \mapsto \varrho_\mu(\zeta, \mathscr{P}_\mu)$ is measurable. Then $\bigcap \{\mathscr{P}_\mu : \mu \in \mathscr{E}\} \neq \emptyset$.

Let $\mathscr{M}^a(B)$ and $\mathscr{E}^a(B)$ denote the sets of all aperiodic elements of $\mathscr{M}(B)$ and $\mathscr{E}(B)$, respectively. The notations $\mathscr{M}^a(T)$ and $\mathscr{E}^a(T)$ have the same meaning.

Let $\mathscr{P}_\mu = \{\zeta : d_w(\mu\bar{\Phi}_\zeta^{-1}, \nu_\mu) < \varepsilon\}$, where $\nu_\mu \in \mathscr{M}(B)$. If $\mu \in \mathscr{E}^a(T)$, a coding $\bar{\Phi}$ such that the latter inequality is valid is possible by Lemma 5, p. 22 of [103]. Using the conditions (a) and (b) formulated after Theorem 35 we get our first universal coding result:

Theorem 39. Let $\mathscr{E} \subset \mathscr{E}^a(T)$ and let $\{\nu_\mu : \mu \in \mathscr{E}\} \subset \mathscr{M}(B)$ be such that for each $E \in \mathscr{B}^{\mathbf{Z}}$, the map $\mu \mapsto \nu_\mu(E)$ from $\mathscr{E}$ to $[0, 1]$ is measurable. For any $\varepsilon > 0$ we can find a partition $\zeta \in \mathscr{P}$ such that $d_w(\mu\bar{\Phi}_\zeta^{-1}, \nu_\mu) < \varepsilon$, $\mu \in \mathscr{E}$.

A simple consequence is a universal version of Rohlin's lemma:

Corollary 40. Let $\mathscr{E} = \mathscr{M}^a(T)$, $N \geqq 1$, and $\varepsilon > 0$ be given. Then there exists a partition $\zeta \in \mathscr{P}$ such that, for any $\mu \in \mathscr{E}$, $d_w(\mu\bar{\Phi}_\zeta^{-1}, \nu) < \varepsilon$, where $\nu \in \mathscr{M}(B)$ is an arbitrary fixed measure. In particular, there exists a (T, N, ε)-Rohlin set E, the same for all $\mu \in \mathscr{E}$.

Put $\Phi(\mu, \zeta) = \bar{d}(\mu\bar{\Phi}_\zeta^{-1}, \nu_\mu)$, where $\mu \in \mathscr{E} \subset \mathscr{E}^a(T)$ and ν_μ corresponds to μ as in Theorem 39. An application of Theorem 37 gives the next result.

Theorem 41. Let $\mathscr{E} \subset \mathscr{E}^a(T)$. Let $\{\nu_\mu : \mu \in \mathscr{E}\} \subset \mathscr{E}^a(B)$ be Bernoulli measures such that the map $\mu \mapsto \nu_\mu(E)$ is measurable for each $E \in \mathscr{B}^{\mathbf{Z}}$. Let $h(\nu_\mu) \leqq h_\mu(T)$, $\mu \in \mathscr{E}$. Then there is a $\zeta \in \mathscr{P}$ such that $\mu\bar{\Phi}_\zeta^{-1} = \nu_\mu$ for all $\mu \in \mathscr{E}$.

The assumption (a) of Theorem 37 is clearly satisfied. Assumption (c) reads as follows: for each $\mu \in \mathscr{E}$ there exists a partition $\zeta \in \mathscr{P}$ such that $\bar{d}(\mu\bar{\Phi}_\zeta^{-1}, \nu_\mu) = 0$. But this follows from Sinai's theorem (see [134] or [136]). The approximation property (b) follows, on account of our assumptions, from Proposition 8, p. 26 of [103]. Theorem 41 prepares the way to a universal form of Sinai's theorem:

Theorem 42. Let $\nu \in \mathscr{E}^a(B)$ be a Bernoulli measure and let $\mu \in \mathscr{M}^a(T)$ be given. Suppose there exists a probability space $(\Lambda, \mathscr{L}, \lambda)$ and a family of measures $\{\mu_\theta : \theta \in \Lambda\} \subset \mathscr{E}(T)$ such that (a) the map $\theta \mapsto \mu_\theta(E)$ is measurable for each $E \in \mathscr{F}$ and (b) $\mu(E) = \int \mu_\theta(E)\, \lambda(d\theta)$. Suppose that

$$\lambda\{\theta \in \Lambda : h_{\mu_\theta}(T) \geqq h(\nu)\} = 1 .$$

Then there exists $\zeta \in \mathscr{P}$ such that $\mu\bar{\Phi}_\zeta^{-1} = \nu$.

In other words, if $[B, \nu]$ is a Bernoulli source and $\mu \in \mathscr{M}^a(T)$ is such that all of its ergodic components have enough entropy then we can encode the system $(\Omega, \mathscr{F}, \mu, T)$ onto $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu, T_B)$. The last two assertions concern a similar extension of Ornstein's isomorphism theorem.

Theorem 43. Let $\mathscr{E} \subset \mathscr{E}^a(T)$ be a set of Bernoulli measures. Suppose for each $E \in \mathscr{B}^{\mathbf{Z}}$, the map $\mu \mapsto \nu_\mu(E) : \mathscr{E} \to [0, 1]$ is measurable, where ν_μ is a Bernoulli measure for each $\mu \in \mathscr{E}$. Let $h_\mu(T) = h(\nu_\mu)$, $\mu \in \mathscr{E}$. Then there exists a partition $\zeta \in \mathscr{P}$ such that, for any $\mu \in \mathscr{E}$, (a) $\mu\bar{\Phi}_\zeta^{-1} = \nu_\mu$ and (b) ζ is a generator (relative to (T, μ)).

In other words, there exists a universal perfectly noiseless code $\bar{\Phi}_\zeta : \Omega \to B^{\mathbf{Z}}$; that is, $\bar{\Phi}_\zeta$ is an isomorphism between $(\Omega, \mathscr{F}, \mu, T)$ and $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}}, \nu_\mu, T_B)$ for all $\mu \in \mathscr{E}$ simultaneously. The proof proceeds using again Theorem 37 for a conveniently chosen admissible function. An immediate corollary is a universal version of Ornstein's theorem:

Theorem 44. Let $\mu_1 \in \mathscr{M}(T)$ and $\mu_2 \in \mathscr{M}(T_B)$ be given. Suppose there exist probability spaces $(\Lambda_i, \mathscr{L}_i, \lambda_i)$, $i = 1, 2$, where $(\Lambda_i, \mathscr{L}_i)$ are standard Borel spaces. Let $\mathscr{E}_i = \{\mu_\theta^i : \theta \in \Lambda_i\}$ be families of Bernoulli measures, $\mathscr{E}_1 \subset \mathscr{E}^a(T)$, $\mathscr{E}_2 \subset \mathscr{E}^a(B)$. Let the map $\theta \mapsto \mu_\theta^i$ be a measurable injection and let $\mu_i(E) = \int \mu_\theta^i(E)\, \lambda_i(\mathrm{d}\theta)$. Suppose there is an injective measurable map φ from Λ_1 onto Λ_2 such that $\lambda_2 = \lambda_1\varphi^{-1}$ and

$$h_{\mu_\theta^1}(T) = h(\mu_{\varphi(\theta)}^2) ; \quad \theta \in \Lambda_1 .$$

Then there exists a $\zeta \in \mathscr{P}$ such that $\mu_1 \bar{\Phi}_\zeta^{-1} = \mu_2$ and the code $\bar{\Phi}_\zeta$ is perfectly noiseless.

21. Universal Isomorphism Theorems

An advantage of Kieffer-Rahe approach is, beside its generality, that it applies also to stationary non-invertible codes. On the other hand, it has also some drawbacks. First of all, being based on Ornstein's approximation techniques, it does not seem capable of extensions to finitary codes. Also, once having a result like Theorem 44, a natural question arises whether one can find a complete invariant for isomorphisms of mixtures of Bernoulli sources. Also here the approach described in Section 20 does not seem to give any suggestions.

In [145] we developed a completely different approach to extensions of isomorphism theorems from ergodic to the aperiodic non-ergodic case. Since the general theory, according to which a stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ is an isomorphism between two stationary aperiodic sources $[A, \mu]$ and $[B, \nu]$ if and only if it is composed of "local" isomorphisms between corresponding ergodic components, has been published recently in this journal, we omit the general theory and merely give a brief account of its applications.

Let $[A, \mu, X]$ be an aperiodic stationary source over a countable discrete alphabet A such that

$$H^*(\mu) = \operatorname{ess.sup} \{h(\mu_x) : x \in R_A[\mu]\}$$

is finite. Note that $H^*(\mu)$ is the *asymptotic rate* defined first by Winkelbauer [153, 154] as

$$H^*(\mu) = \sup_{0<\varepsilon<1} \limsup_{n\to\infty} n^{-1} \log L_n(\varepsilon, \mu),$$

where

$$L_n(\varepsilon, \mu) = \min \{\|E\| : E \subset A^n, \mu^n(E) > 1 - \varepsilon\}, \quad 0 < \varepsilon < 1.$$

It was generalized in the spirit of Kolmogorov-Sinai invariant to transformations of abstract probability spaces [155], to free actions of the group $\mathbf{Z}^d$ [138], and to actions of countable amenable groups [148] as a new invariant. Let

$$K = \operatorname{INT} \{\exp H^*(\mu)\} + 1.$$

Theorem 45. [155]. Let $[A, \mu, X]$ and K be as above. Then there exists a finite alphabet B with at most K letters and a source $[B, \nu, Y]$ such that we can find an isomorphism $\bar{\Phi} : C^{\mathbf{Z}} \to B^{\mathbf{Z}}$ with $Y = \bar{\Phi}X$.

There is an asymmetry related to "local" and "global" codes. We explain it on the theorem. As pointed out in Section 4, Theorem 45 says, equivalently, that there exists a finite generator (relative to (T_A, μ)) with $\|\zeta\| \leqq K$. As shown in [155] we then

have

$$\mu\{x \in R_A : \zeta \text{ is a generator relative to } (T_A, \mu_x)\} = 1 .$$

Clearly ζ is not the best we may wish. For example, if $x \in R_A$ is such that $h(\mu_x) \ll \ll H^*(\mu)$ then, by Krieger's theorem, there exists a generator ζ_x relative to (T_A, μ_x) with at most INT $\{\exp h(\mu_x)\} + 1 \ll K$ elements. However, the result is natural for if we take a partition ζ with $\|\zeta\| \leqq K$ then we can find, with positive probability, an ergodic component μ_x with $h(\mu_x) > \log \|\zeta\|$. Consequently, such a partition cannot be a generator "universally", i.e., for μ-almost all ergodic components of $[A, \mu]$. Furthermore, as shown in [148], one has to take the optimum generators for ergodic components in order to obtain the best bound K for the mixture.

Of course, this asymmetry does not appear in the general theory [145], where no additional restrictions are imposed upon the local and global isomorphisms.

Next, let us prepare the notations for an alternate formulation of the universal Ornstein's isomorphism theorem. Let $[A, \mu, X]$ be stationary. For each $\varepsilon \in (0, 1)$ define

$$H_\varepsilon(\mu) = \lim_{n \to \infty} n^{-1} \log L_n(\varepsilon, \mu)$$

so that

$$H^*(\mu) = \lim_{\varepsilon \to 0} H_\varepsilon(\mu)$$

(see [153, 154]). Let

$$d^A(t) = \mu\{x \in R_A : h(\mu_x) \leqq t\} , \quad t \geqq 0 ;$$

$$c^A(\delta) = \inf\{t : d^A(t) \geqq \delta\} , \qquad 0 < \delta < 1 .$$

As proved in [139] the limit defining $H_\varepsilon(\mu)$ exists if and only if $1 - \varepsilon$ is a continuity point of $c^A(\cdot)$, and in this case we have that $H_\varepsilon(\mu) = c^A(1 - \varepsilon)$. If $[B, \nu, Y]$ is another stationary source, we let d^B and c^B denote analogous quantities.

Theorem 46. Let $[A, \mu]$ and $[B, \nu]$ be two stationary aperiodic sources over countable discrete alphabets. If they are isomorphic then $d^A(t) = d^B(t)$, $t \geqq 0$. Conversely, if $d^A(t) = d^B(t)$ for all $t \geqq 0$ and the sources each have the property that almost all ergodic components are Bernoulli sources (possibly with infinite entropies in which case it is assumed that the infinite entropy components have the same total weights) then the sources $[A, \mu]$ and $[B, \nu]$ are isomorphic.

Thus, the distribution function of entropy is a complete invariant for the class of all sources whose components are Bernoulli. Extensions to finitary codes are also given in [145, 146]. Another application is an extension of Theorem 18 to the aperiodic non-ergodic case.

Theorem 47. Let $[A, \mu, X]$ be a stationary aperiodic source over a countable discrete alphabet A such $H^*(\mu)$ is finite. Let $[B, \nu, Y]$ be an ergodic process over a finite alphabet B such that $h(\nu) > H^*(\mu)$. Then there exists an invertible stationary

code $\bar{\Phi} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $d_w(\mu\bar{\Phi}_\xi^{-1}, \nu) < \varepsilon$. In particular, if $K = \text{INT}\{\exp H^*(\mu)\} + 1$ and Y is the equiprobable IID process over the alphabet B then the above conclusion is valid.

PART VII: CODING PROBLEMS OF INFORMATION THEORY

22. Basic Coding Problems

In this section we discuss the basic coding problems of information theory and motivate the choice of problems to be dealt with in subsequent sections. Also, we attempt to illustrate how ergodic theory can make precise intuitive but necessary vague formulations of several information theoretic problems. In his pioneering paper [125] Shannon formulated the basic types of coding problems in information theory:

(a) noiseless source coding,
(b) coding for sources with a fidelity criterion,
(c) channel coding, and
(d) joint source/channel coding,

and formulated fundamental (although often only heuristic) ideas as to their proofs (for problems (a), (c), and (d); as far as concerns (b) see [126]).

We start with so-called overall source coding operation which involves (a) and (b). The task of the overall source coding operation is to transmit information produced by a stationary source X across a noiseless channel with finite capacity C in such a way that the resulting reconstruction of X at the receiver approximates X as well as possible. When $h(X) > C$, a perfect transmission is excluded, and consequently the overall source coding operation splits into two distinct steps [14]. On the first step, we must carry over the process of *entropy reduction*. Its goal is to transform the given process X into its approximation $\hat{X}$ satisfying $h(\hat{X}) \leqq C$. This necessarily inserts some distortion in the reproduction process $\hat{X}$ so that the coding $X \to \hat{X}$ cannot be invertible. In order we can evaluate the distortion, we seek for a functional relationship between the entropy of $\hat{X}$ and the minimum attainable average distortion – the distortion-rate function (for historical reasons, however, the dual approach, i.e. the rate-distortion theory was prefered; see [9]).

The first part of overall source coding operation still fails to fulfil the goal of actual data compression, i.e., the goal of sending less source characters over the channel. This is the objective of the second step called *noiseless source coding* operation. Its goal is to map the process $\hat{X}$ obtained on the first step into an appropriate input process $Y = (Y_i; i \in \mathbf{Z})$ of the given noiseless channel. A noiseless channel does not produce additional errors as to the identity of $\hat{X}$. Hence, it is required that the coding $\hat{X} \to Y$ be nearly invertible. Moreover, a natural requirement is that the overall

source coding operation be efficient in the sense that the channel be used ar rates near its capacity. Formally, this means that $h(Y) \approx C = \log \|B\|$, where B is the channel alphabet. As $h(Y) \leqq \log \|B\|$ for any process Y over the alphabet B, the requirement that $h(Y) \approx \log \|B\|$ forces Y to be nearly independent and use each of $\|B\|$ letters almost equiprobably. For these reasons the second step is often called the operation of *redundancy removal* [41].

In order to keep clear why, and how, ergodic theory comes in, let us sketch the approach to redundancy removal within the frame of traditional block coding technique (see [125, 126]). A block code of order N (see Section 4) partitions source sequences $x = (x_i; i \in \mathbf{Z})$ into consecutive non-overlapping blocks of length N, and codes them individually. The individual coding function either maps N-tuples of source letters into N-tuples of reproduction letters (a fixed-rate code) or, maps the source N-tuples into variable length non-overlapping blocks of reproduction letters (a variable-rate code). Although we implicitly assumed in the above considerations that $\hat{X}$ is again a stationary process, the overall source coding operation can be described also without that restriction [41, 9, 39]. But for reason to be clear below let us consider the operation of redundancy removal separately and suppose $\hat{X}$ is a stationary process such that $h(\hat{X}) \leqq C = \log \|B\|$.

It is intuitively clear that a method aiming to produce redundancy removal has to take into account the probabilities (frequencies) of source N-tuples so that a fixed-rate code is not appropriate. Thus, consider a variable-rate code of order N which assigns to each N-tuple $\hat{x}^N = (\hat{x}_0, \ldots, \hat{x}_{N-1})$ a word composed of letters from B having the length $l(\hat{x}^N)$. The average length is

$$l(N, \hat{X}) = \sum_{\hat{x}^N} l(\hat{x}^N) \operatorname{Prob}\left[\hat{X}^N = \hat{x}^N\right].$$

A classical variable-length source coding theorem (see p. 785 of [25] or [39, 8]) implies that $l(N, \hat{X}) \geqq H(\hat{X}^N)$. The *Nth order redundancy* is defined to be the quantity

$$r_N(\hat{X}) = N^{-1}\left[l(N, \hat{X}) - H(\hat{X}^N)\right] \geqq 0.$$

Now, the best result we can expect is the existence of a sequence of block codes of orders $N = 1, 2, \ldots$ such that

$$\lim_{N \to \infty} r_N(\hat{X}) = 0.$$

However, for block codes we cannot speak about a limiting coding $\hat{X} \to Y$ which performs a complete reduction of redundancy. Thus, we have no process Y as in the above informal discussion of the overall source coding operation. Does this mean that it has been merely an aesthetic but necessarily vague description of noiseless source coding? In particular, what should the clauses "almost equiprobable" and "nearly independent" mean?

Suppose for a moment there exists a stationary coding $\hat{X} \to Y$ so that Y is again

a stationary process. We can say that Y is almost equiprobable if, for some small $\varepsilon > 0$, we have that

$$(*) \qquad \left|\text{Prob}\left[Y_0 = b\right] - \|B\|^{-1}\right| < \varepsilon\,, \quad b \in B\,.$$

We define Y to be nearly independent if, for any $n \geqq 1$, the random variable Y_n is nearly independent of the vector $(Y_0, \ldots, Y_{n-1}) = Y^n$, i.e., for some small $\varepsilon > 0$,

$$(**) \qquad \sum_{b \in B,\, \boldsymbol{b} \in B^n} \left|\text{Prob}\left[Y_n = b,\, Y^n = \boldsymbol{b}\right] - \text{Prob}\left[Y_n = b\right] \text{Prob}\left[Y^n = \boldsymbol{b}\right]\right| < \varepsilon\,.$$

Now we can make the intuitive description of noiseless source coding a rigorous fact:

Proposition 48. For every $\varepsilon > 0$ there is a $\delta > 0$ such that the assumption that

$$\left|h(Y) - \log \|B\|\right| < \delta$$

implies Y satisfies $(*)$ and $(**)$ above (i.e., Y is ε-equiprobable and ε-independent).

Thus, for stationary processes, the requirement that $h(Y) \approx C = \log \|B\|$ forces Y to be almost equiprobable and nearly independent. The proof of Proposition 48 can be found in [103, 128], and the proposition itself serves as a starting point to approximation arguments used in the proof of Ornstein's isomorphism theorem. Moreover, we can prove even a little bit more:

Proposition 49. For every $\varepsilon > 0$ there exists a $\delta > 0$ such that the inequality $\left|h(Y) - \log \|B\|\right| < \delta$ implies the existence of a stationary coding $Y \to \hat{\hat{X}}$ such that

$$\text{Prob}\left[\hat{X}_0 \neq \hat{\hat{X}}_0\right] < \varepsilon\,.$$

But this can be accepted as a reasonable formulation of the vague clause that the coding $\hat{X} \to Y$ is "nearly invertible". Since we have been completely free as to the choice of ε in the latter two assertions, a seemingly plausible hypothesis is that, using some appropriate limiting operation, it might be possible to find stationary codings $\hat{X} \to Y$ and $Y \to \hat{\hat{X}}$ such that

(a) Y is exactly independent and equiprobable, and
(b) $\text{Prob}\left[\hat{X}_0 \neq X_0\right] = 0$, i.e., the coding $\hat{X} \to Y$ is invertible.

However, ergodic theory shows us that the hypothesis is overoptimistic. Indeed, (a) forces $h(Y) = \log \|B\|$ whence (b) entails $h(\hat{X}) = \log \|B\|$ (for entropy is an isomorphism invariant [11]). Hence, in general we can expect only results weaker than (a) and (b).

First let us argue that (a) is excluded. Actually, suppose (a) is valid and $h(\hat{X}) \neq h(Y) = \log \|B\|$. Then we must have $h(\hat{X}) < h(Y)$ for $h(\hat{X}) \leqq C$ and $C = h(Y)$. On the other hand, $\hat{X} \to Y$ is a stationary coding so that $h(Y) \leqq h(\hat{X})$, a contradiction.

Consequently, the only type of assertions that remain at our disposal is as follows. There exist stationary codings $\hat{X} \to \hat{Y}$, $\hat{Y} \to \hat{\hat{X}}$ such that (b) is valid and $\hat{Y}$ is close to Y from (a). But these assertions (depending on which criterion of closeness has been chosen) are just improvements on Krieger's finite generator theorem (see Section 15). Thus, Krieger's theorem and its improvements represent solutions to the problem of noiseless source coding within the frame of stationary codes.

By these considerations we are tempted to infer that it is only the entropy mismatch $h(\hat{X}) < h(Y)$ which is responsible for violation of (a) and (b) above. Again, ergodic theory helps to clarify the situation. Indeed, suppose that $h(\hat{X}) = h(Y) = \log \|B\|$ so that Y is as in (a). Of course, this does not force $\hat{X}$ to be also independent and equiprobable for the alphabet of $\hat{X}$ can be much larger than B. However, Ornstein's isomorphism theorem tells us that there exists an invertible stationary coding $\hat{X} \to Y$ if and only if $\hat{X}$ is a Bernoulli source. This is a measure theoretic counterpart of a result by Marcus [91] who investigated connections of topological Markov chains with entropies $\log n$, $n \in \mathbf{N}$, with full shifts over n symbols. Indeed, $h(Y) = \log \|B\| = h_{\text{top}}(T_B)$.

To summarize, we have seen the close connections between generators problems and noiseless source coding. Furthermore, it became clear that it would be desirable to develop the theory of entropy reducing coding for stationary codes.

Finally, let us deal with channel coding problems. In our review only noiseless channels were investigated. The loss is not as discouraging as it might appear at glance [51]. The point is that noisy channels possess families of input sources which can be transmitted over the channel directly without first encoding them, and then exactly decoded from the channel output using a stationary code. Such sources have been called *invulnerable* in [51]. The concept of invulnerability is closely related to that of zero-error transmission over noisy channels. Thus, the problems of noiseless source coding and noiseless transmission over noisy channels exhibit not only an external similarity but, as we have already mentioned, there exists a unique method for proving both kinds of coding theorems.

23. Entropy Reducing Coding

Since the theory of source coding for sources with a fidelity criterion evolved primarily within the block coding approach [9] it is reasonable to start with a result of Kieffer [71] who showed that it is not necessary to give separate proofs to block and sliding-block versions of source coding theorems.

Let $(A, \mathscr{A})$ and $(\hat{A}, \hat{\mathscr{A}})$ be measurable spaces and $\varrho : A \times \hat{A} \to [0, \infty)$ a jointly measurable function. For each $n \in \mathbf{N}$ let

$$\varrho_n(x^n, y^n) = n^{-1} \sum_{i=0}^{n-1} \varrho(x_i, y_i) ; \quad x^n \in A^n, \quad y^n \in \hat{A}^n .$$

The family $(\varrho_n;\ n \geqq 1)$ is called a *single-letter fidelity criterion* and ϱ itself a *distortion measure*. Other kinds of fidelity criteria also can be of interest, e.g., the $0-1$ ones (see [66]; a $0-1$ fidelity criterion can be used to unify the tasks of noiseless coding and of entropy reducing coding, respectively). A map $\bar{\Phi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is said to be a block code if there exists an $N \in \mathbf{N}$, a finite set $B \subset \hat{A}^N$, and a measurable map $\Phi : A^N \to B$ such that

$$(\bar{\Phi}x)_{iN}^{iN+N-1} = \Phi(x_{iN}^{iN+N-1}); \quad x \in A^{\mathbf{Z}}, \quad i \in \mathbf{Z}$$

(this definition slightly differs from that one adopted in Section 4, the difference being a consequence of more general alphabets involved). The *rate* of the foregoing code is the number

$$R(\bar{\Phi}) = N^{-1} \log \|\Phi(A^N)\| .$$

If $\bar{\Phi}$ is used to code $\mu \in \mathscr{M}(A)$, the resulting *average distortion* is, by definition,

$$\bar{\varrho}(\bar{\Phi}, \mu) = \int_{A^{\mathbf{Z}}} \varrho_N(x^N, (\bar{\Phi}x)^N)\, \mu(\mathrm{d}x) .$$

Similarly, a map $\bar{\Psi} : A^{\mathbf{Z}} \to \hat{A}^{\mathbf{Z}}$ is called a sliding-block code if there is an $N \in \mathbf{N}$, a finite set $B \subset \hat{A}$, and a measurable map $\Psi : A^{2N+1} \to B$ such that

$$(\bar{\Psi}x)_i = \Psi(x_{i-N}^{i+N}); \quad x \in A^{\mathbf{Z}}, \quad i \in \mathbf{Z} .$$

Put $M_n(\bar{\Psi}) = \|\{(\bar{\Psi}x)_1^n : x \in A^{\mathbf{Z}}\}\|$, $n = 1, 2, \ldots$. The *rate* of $\bar{\Psi}$ is defined as the limit

$$r(\bar{\Psi}) = \lim_{n \to \infty} n^{-1} \log M_n(\bar{\Psi})$$

and the average distortion when $\bar{\Psi}$ is used to code $\mu \in \mathscr{M}(A)$ is

$$\bar{\varrho}(\bar{\Psi}, \mu) = \int_{A^{\mathbf{Z}}} \varrho(x_0, (\bar{\Psi}x)_0)\, \mu(\mathrm{d}x) .$$

Theorem 50. [71]. Let A, $\hat{A}$, and ϱ be given as above. Let $[A, \mu]$ be a stationary source. Then

(a) given a block code $\bar{\Phi}$ and an $\varepsilon > 0$, there is a sliding-block code $\bar{\Psi}$ such that

$$r(\bar{\Psi}) \leqq R(\bar{\Phi}) + \varepsilon, \quad \bar{\varrho}(\bar{\Psi}, \mu) \leqq \bar{\varrho}(\bar{\Phi}, \mu) ;$$

(b) given a sliding-block code $\bar{\Psi}$ and an $\varepsilon > 0$, there is a block code $\bar{\Phi}$ such that

$$R(\bar{\Phi}) \leqq r(\bar{\Psi}) + \varepsilon, \quad \bar{\varrho}(\bar{\Phi}, \mu) \leqq \bar{\varrho}(\bar{\Psi}, \mu) .$$

An analogue of Theorem 50 can be obtained also for variable-rate block and sliding-block codes (see [71], Theorem 2). Theorem 50 allows to extend and unify many previous results. We quote several of them which bear connections to recent ideas of ergodic theory.

Let $R > 0$. Then the *optimum performance theoretically attainable* (OPTA)

using block codes is defined by

$$\delta_b(R, \mu) = \inf \{\bar{\varrho}(\bar{\Phi}, \mu) : \bar{\Phi} \text{ block code, } R(\bar{\Phi}) \leqq R\} .$$

Similarly, one defined the OPTA $\delta_s(R, \mu)$ using sliding-block codes.

Corollary 51. Let $[A, \mu]$ be a stationary source. For each $R > 0$ we have $\delta_b(R, \mu) = \delta_s(R, \mu)$.

This result has been known previously only in certain special cases [47, 131, 26]. Shields and Neuhoff [131] considered the case when $A = \hat{A}$ is a finite alphabet. The idea of their proof is as follows. Suppose that $\overline{\Psi}$ is a sliding-block code of order m. Take $N > 2m + 1$. Use $\overline{\Psi}$ to code typical source N-blocks into $N - 2m$ blocks, and then fill in the remaining $2m$ places arbitrarily. If N is large then the number of typical N-blocks is near $\exp\{Nh\}$, where h is the entropy of the encoded process. This gives the bound $R(\bar{\Phi}) \leqq r(\overline{\Psi}) + \varepsilon$. If N is enough larger than m, then the additional distortion from filling in the $2m$ places arbitrarily will be small (in fact, if X is an ergodic source and $\hat{X} = \overline{\Psi}X$, where $\overline{\Psi}$ is a sliding-block code, then $(X, \hat{X})$ is jointly ergodic so that the empirical distortions converge with probability one to $\bar{\varrho}(\bar{\Phi}, \mu)$).

The converse is more involved. On the first step one uses Theorem 7 to construct an auxiliary binary sliding-block coding, and this is then used to indicate when to use the block code. Then one proceeds similarly to Section 6 (this gives the desired bound to rate). Using the independence property from Theorem 7 one can control also the average distortion of the resulting sliding-block code (to this end recall that Theorem 7 entails the existence of a strong Rohlin set depending on only a finite number of coordinates (see Section 7) so that we really obtain a sliding-block code of some finite order).

Other applications of Theorem 50 will be discussed later in connection with universal source coding problems.

24. Sliding-Block Source Coding and $\bar{\varrho}$-Distance

A main task of source coding theorems is to relate the OPTA to an information theoretic optimum – the distortion-rate function (DRF). This gives an operational meaning to the DRF and, at the same time, allows to calculate, at least for some classes of sources, the OPTA (as to that consult [13]). However, the proof of coding theorems even for ergodic sources is a non-trivial task, the difficulties coming from the requirement that a property like ergodicity of the partitioned process should hold (this is necessary in order one can employ a Shannon-style random coding argument to a process obtained by a block code of some finite order; see [9, 43, 48]), As discussed in detail by Gray, Neuhoff, and Omura [46], the problems

are caused mainly by a somewhat artificial mutual information constraint involved in the definition of the DRF. They developed process definitions of OPTA's and DRF's in which the mutual information rate constraint is replaced by the constraint concerning the entropy of the reproduction.

The results to be presented below are also in the spirit of [46] and follow [47]. We assume that ϱ is a non-negative distortion measure on $(A \cup \hat{A}) \times (A \cup \hat{A})$. If A is finite, ϱ may be any finite-valued function, if A is a metric space, we take ϱ to be the metric and assume that $A \cup \hat{A}$ is a complete separable metric space under ϱ. If $[A, \mu, X]$ is ergodic and $X \to \hat{X}$ is a sliding-block coding then the pair process $(X, \hat{X})$ is again ergodic so that

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \varrho(X_i, \hat{X}_i) = \mathsf{E}_\mu \varrho(X_0, \Phi(X_{-N}^N)) = \bar{\varrho}(\bar{\Phi}, \mu)$$

for any sliding-block code $\bar{\Phi}$ of order N. We indicate its order by a superscript, viz. $\bar{\Phi} = \bar{\Phi}^{(N)}$. Let

$$\delta(R, N) = \inf \{\bar{\varrho}(\bar{\Phi}^{(N)}, \mu) : h(\bar{\Phi}^{(N)}x) \leqq R\} .$$

Then

$$\inf_{N\geqq 1} \delta(R, N) = \delta_s(R, \mu) , \quad R > 0$$

(cf. Section 23 for the definition of the sliding-block OPTA). Our results will relate the OPTA to the process $\bar{\varrho}$-distance between source and constrained entropy reproduction processes. This and equivalent definitions of $\bar{\varrho}$-distance are given in [48]: if $[A, \mu]$ and $[\hat{A}, \nu]$ are two stationary processes then we define

$$\bar{\varrho}([A, \mu], [\hat{A}, \nu]) = \inf_{p\in\mu\vee\nu} \mathsf{E}_p \varrho(X_0, \hat{X}_0) ,$$

where $\mu \vee \nu$ denotes the set of all distributions of stationary pair processes $(X, \hat{X})$ such that dist $(X) = \mu$ and dist $(\hat{X}) = \nu$. Recall from [48] that if ϱ is a metric then so is $\bar{\varrho}$.

Theorem 52. Suppose A and $\hat{A}$ are finite and ϱ is an arbitrary finite-valued distortion measure. Let $[A, \mu]$ be a stationary and aperiodic source. Then

$$\delta_s(R, \mu) = \inf_{[\hat{A},\nu]:h(\nu)\leqq R} \bar{\varrho}([A, \mu], [\hat{A}, \nu]) .$$

The proof of Theorem 52 does not involve any random coding argument and is based on Rohlin's lemma. For later reference we point out the following result which can be obtained as a byproduct:

$$\delta_s(R, \mu) = \lim_{N\to\infty} \delta(R, N) = \delta^*(R, \mu) ,$$

where $\delta^*(R, \mu)$ is the OPTA using infinite codes:

$$\delta^*(R, \mu) = \inf \{\bar{\varrho}(\bar{\Phi}^{(\infty)}, \mu) : h(\bar{\Phi}^{(\infty)}X) \leqq R\} .$$

This means that sliding-block codes of finite orders perform, in the limit of orders, as a code which has at its disposal the entire source sequence. Such codes were foreseen by Krengel [84].

The proof of Theorem 52 follows easily from the following assertion.

Proposition 53. Given two finite alphabet stationary aperiodic sources $[A, \mu]$, $[\hat{A}, \nu]$, a finite-valued distortion measure ϱ, and a $\delta > 0$. There exists an $N(\delta)$ such that for any $N \geqq N(\delta)$ we can find a sliding-block code $\bar{\Phi}^{(N)}$ for $[A, \mu]$ such that

$$\bar{\varrho}(\bar{\Phi}^{(N)}, \mu) \leqq \bar{\varrho}([A, \mu], [\hat{A}, \nu]) + \delta ,$$
$$h(\bar{\Phi}^{(N)}X) \leqq h(\nu) + \delta .$$

Recall from Section 6 that a quadruple (T, N, E, ζ), where E is a strong (T, N, ε)-Rohlin set and ζ is a finite measurable partition, is sais to be an ε-gadget. Two ε-gadgets $(T. N, E, \zeta)$ and (U, N, F, ξ) are said to be isomorphic (denoted by "$\sim$") if

$$d(\bigvee_{i=0}^{N-1} T^{-i}\zeta \mid E) = d(\bigvee_{i=0}^{N-1} U^{-i}\xi \mid F) \quad \text{(cf. Section 5)}$$

Let γ_A, $\gamma_{\hat{A}}$ denote the natural zero-time partitions of $A^{\mathbf{Z}}$ and $\hat{A}^{\mathbf{Z}}$, respectively. Let $\tilde{\zeta}_0$ and $\tilde{\xi}_0$ denote the first and the second coordinate zero-time partitions of $A^{\mathbf{Z}} \times \hat{A}^{\mathbf{Z}}$. Since $[A \times \hat{A}, \mu \times \nu]$ is jointly aperiodic, we can construct by Theorem 5 a good joint gadget $(T_{A\times\hat{A}}, N, \tilde{F}, \tilde{\zeta}_0 \vee \tilde{\xi}_0)$. Also, we can construct a good source gadget (T_A, N, F, γ_A). Then the projection of the joint gadget on the first coordinate is isomorphic to the source gadget, i.e.

$$d(\bigvee_{i=0}^{N-1} T_A^{-i}\gamma_A \mid F) = d(\bigvee_{i=0}^{N-1} T_{A\times\hat{A}}^{-i}\tilde{\zeta}_0 \mid \tilde{F}) .$$

As in [128] we can find a partition η such that the isomorphism is extended to

$$(T_A, N, F, \gamma_A \vee \eta) \sim (T_{A\times\hat{A}}, N, \tilde{F}, \tilde{\zeta}_0 \vee \tilde{\xi}_0) .$$

The new gadget thus obtained is isomorphic to the second coordinate projection $(U, N, \tilde{F}, \tilde{\xi})$ of the joint gadget:

$$d(\bigvee_{i=0}^{N-1} T_A^{-i}\eta \mid F) = d(\bigvee_{i=0}^{N-1} T_{A\times\hat{A}}^{-i}\tilde{\xi}_0 \mid \tilde{F}) .$$

This implies closeness of probabilities assigned to atoms as well as closeness in entropy. This can be used to bound the distortion and the entropy of the reproduction process which is obtained as a stationary coding determined by (an extension of) η. An application of the standard approximation argument allows to find a sliding-block coding with nearly the same distortion and entropy.

Having a finite alphabet coding theorem one usually employs a quantization argument in order to extend it to more general alphabets (see, e.g., [124] and a de-

tailed explanation in [43]). We omit these results and refer an interested reader to [44] and the references therein.

We close this section by a mismatch theorem which evaluates the loss in the code's performance when the code was designed for some source $[A, \nu]$ but the true source is $[A, \mu]$.

Proposition 54. [48]. Let $[A, \mu]$ and $[A, \nu]$ be two stationary sources over a separable metric alphabet A. Then

$$|\delta_b(R, \mu) - \delta_b(R, \nu)| \leqq \bar{\varrho}([A, \mu], [A, \nu]),$$
$$|\delta_s(R, \mu) - \delta_s(R, \nu)| \leqq \bar{\varrho}([A, \mu], [A, \nu]).$$

This result is very close to universal source coding theorems and we shall comment on it in the next section.

25. Universal Source Coding

The idea of universal source coding traces back to Fitingoff [35, 36] but its modern origins were founded by Davisson [25] (for noiseless source coding) and by Gray and Davisson [43] (for source coding with a fidelity criterion). Davisson classified the notions of universality similarly to the classification of statistical decision rules, and a similar classification within source coding with a fidelity criterion was given in [95]. In this section we shall deal with two types of universal coding called weak and strong universal coding.

Let $(A, \mathscr{A})$ be an arbitrary measurable space, $\hat{A}$ a set, and $\varrho : A \times \hat{A} \to [0, \infty)$ such that $\varrho(., y) : A \to [0, \infty)$ is measurable for each $y \in \hat{A}$. We assume that $(\varrho_n; n \geqq 1)$ is a single-letter fidelity criterion determined by ϱ (see Section 23). In what follows it is more convenient to work with code books rather than with block codes. As already pointed out, a code book $B \subset \hat{A}^N$ gives rise to a block coding function $\Phi : A^N \to B$ by the rule that $\Phi(x^N) = \hat{x}^N \in B$, where

$$\varrho_N(x^N, \hat{x}^N) = \varrho_N(x^N \mid B) = \min_{y^N \in B} \varrho_N(x^N, y^N).$$

We put $R(B) = N^{-1} \log \|B\|$ and $\bar{\varrho}(B, \mu) = \mathsf{E}_\mu \varrho_N(X^N \mid B)$. Let $\mathscr{E} \subset \mathscr{E}(A)$. We suppose there is a common reference letter $y^* \in \hat{A}$ such that for any $\mu \in \mathscr{E}$, $\mathsf{E}_\mu \varrho(X_0, y^*) < \infty$.

A sequence $(B_N; N \geqq 1)$ is said to be *universal weakly minimax sequence* of codes for $\mathscr{E}$ at the rate R (or, simply, weak universal) if

(i) B_N is a finite subset of A^N, $N = 1, 2, \ldots$,
(ii) $R(B_N) < R$, $N = 1, 2, \ldots$, and
(iii) $\lim_{N \to \infty} \bar{\varrho}(B_N, \mu) = D(R, \mu)$ for each $\mu \in \mathscr{E}$,

where $R \to D(R, \mu)$ is the DRF of the source μ; see [9] or [43]. This type of coding is also called universal fixed-rate coding [95] and it was investigated by many authors

(e.g., see [158, 67, 25, 95]). In particular, various types of conditions were derived under which weak universal coding is possible. A common feature of these conditions is a *separability* property singled out as a necessary and sufficient condition by Kieffer [67]. Let us first formulate the necessary condition:

Proposition 55. A necessary condition for the existence of a universal weakly minimax sequence $(B_N; N \geqq 1)$ for $\mathscr{E}$ at every rate $R > 0$ is the following:

there exists a countable class $\mathscr{B}$ of block code books such that for any $\mu \in \mathscr{E}$ and any $\varepsilon > 0$, for each block code book B there is a $B' \in \mathscr{B}$ such that $R(B') < R(B) + \varepsilon$ and $\bar{\varrho}(B', \mu) < \bar{\varrho}(B, \mu) + \varepsilon$.

The condition is satisfied, for example, in the following cases:

(a) A, $\hat{A}$, or $\mathscr{E}$ are countable,

(b) $\hat{A}$ is a separable metric space, ϱ is bounded and such that $\varrho(a, .)$ is continuous on $\hat{A}$ for each $a \in A$,

(c) $\varrho(x, y) = f[d(x, y)]$, where d is a metric on $A \cup \hat{A}$, A or $\hat{A}$ are separable under d, and $f : [0, \infty) \to [0, \infty)$ is a non-decreasing function such that for each $a > 0$, $\lim_{x \to \infty} f(x + a)/f(x) = 1$.

All weak universal coding results of [95] are special cases of the following general result of Kieffer:

Theorem 56. Suppose $\mathscr{E}$ is separable in the sense of the condition formulated in Proposition 55. Then for any rate $R > 0$ these exists a universal weakly minimax sequence of codes $(B_N; N \geqq 1)$ for $\mathscr{E}$ at the rate R. Moreover, that sequence can be chosen so that, for any $\mu \in \mathscr{E}$,

$$\lim_{N \to \infty} \left| \mathsf{E}_\mu \varrho_N(X^N \mid B_N) - D(R, \mu) \right| = 0$$

As said at the beginning of this section, a main task of source coding is to relate the OPTA and the DRF, i.e., to prove the following relation:

$$\delta_b(R, \mu) = D(R, \mu), \quad \mu \in \mathscr{M}(A).$$

If $\mu \in \mathscr{M}(A) \setminus \mathscr{E}(A)$ then Ziv [158] observed and Gray and Davisson [43] studied in detail the surprising fact that the OPTA for a mixture of ergodic subsources is not given by the DRF calculated for that mixture, but rather by the weighted average of the DRF's of the subsources. The most general result is the following corollary to Theorem 56:

Theorem 57. Let $\mathscr{E} \subset \mathscr{E}(A)$ be separable. Let $(\Omega, \mathscr{F}, \lambda)$ be a probability space and $\{\mu_\omega : \omega \in \Omega\} \subset \mathscr{E}$ a measurable family. Let $\mu = \int \mu_\omega \, \lambda(\mathrm{d}\omega)$ satisfy $\mathsf{E}_\mu \varrho(X_0, y^*) < \infty$ for some reference letter $y^* \in \hat{A}$. Then

$$\delta_b(R, \mu) = \int_\Omega D(R, \mu_\omega) \, \lambda(\mathrm{d}\omega), \quad R > 0.$$

As in the ergodic case $\delta_b(R, \mu) = D(R, \mu)$, we can deduce from the convergence assertion in Theorem 56 also

$$(*) \qquad \lim_{N\to\infty} \bar{\varrho}(B_N, \mu) = \delta_b(R, \mu), \quad \mu \in \mathscr{E}.$$

Observe that a combination of Theorems 56 and 50 shows that weak universal (fixed-rate) block coding for $\mathscr{E}$ is possible if and only if weak universal (fixed-rate) sliding-block coding is possible. All results of Sections 23 – 25 are valid also for variable-rate codes [71]. An interested reader should start reading [71] and [67]; the methods used there unify the methods developed in a series of previous papers (see [158, 117, 118, 90, 89]).

As shown by (*), weak universal coding corresponds to pointwise convergence of average distortions $\bar{\varrho}(B_N, \mu)$ to the OPTA $\delta_b(R, \mu)$. If the convergence is uniform in $\mu \in \mathscr{E}$, we speak about *strong universal coding* (see [95, 43]). In light of Proposition 54, if $\mathscr{E}$ can be covered by a finite set of ε-balls (in the process $\bar{\varrho}$-distance) for each $\varepsilon > 0$, we get strong universal coding for $\mathscr{E}$. Thus, if ϱ is a metric distortion measure, a sufficient condition for the existence of strong universal coding of $\mathscr{E}$ at any rate $R > 0$ is that $\mathscr{E}$ be $\bar{\varrho}$-totally bounded (this is Corollary 2 of [48]; see also [95]). Before formulating general assertions about strong universal codding, two remarks are worth make.

If $R(B_N) \to R$ and if the convergence $\bar{\varrho}(B_N, \mu) \to \delta_b(R, \mu)$ is uniform in $\mu \in \mathscr{E}$, then for any $\varepsilon > 0$ and for N large enough we find a single code book B_N such that

$$|R(B_N) - R| \leqq \varepsilon,$$

$$|\bar{\varrho}(B_N, \mu) - \delta_b(R, \mu)| \leqq \varepsilon, \quad \mu \in \mathscr{E}.$$

This is the main advantage of strong universal coding, for in the weak case we only could assert existence of a sequence of code books which was good for all sources just asymptotically in the limit of increasing block length.

On the other hand, the condition of $\bar{\varrho}$-total boundedness of $\mathscr{E}$ is much more restrictive than the separability condition. This puts the following problem. If the alphabets are finite and ϱ is a finite valued metric distortion measure, then a set $\mathscr{E}$ is $\bar{\varrho}$-totally bounded if it is $\bar{d}$-totally bounded [95]. However, even classes such small as the class of all first-order binary Markov sources are not $\bar{d}$-totally bounded. Neuhoff and Shields [96] investigated the question whether strong universal coding is possible at least at sufficiently large rates. They obtained the following result:

Theorem 58. (a) Let $\mathscr{E}$ be the class of all first-order Markov sources over an alphabet with K letters. If $R \geqq \log(K - 1)$ then strong universal coding of $\mathscr{E}$ at the rate R is possible. If $K > 2$ and $0 < R < \log(K - 1)$, then strong universal coding of $\mathscr{E}$ at the rate R is impossible.

(b) For any K and n there exists a number $R^*(K, n)$ such that strong universal coding of $\mathscr{E}$ at the rate R is possible if and only if $R \geqq R^*(K, n)$ for the class $\mathscr{E}$

of all n-step Markov sources over an alphabet with K letters that are either non-ergodic or have transient states.

Part (b) puts strong limitations to classes $\mathscr{E}$. Indeed, only rates $0 < R < \log \|A\|$ are of interest. But as shown in [96], p. 365, $\lim_{n\to\infty} R^*(K, n) = \log K$.

Kieffer [68] developed general criteria for strong universal source coding which include Theorem 58 and other known results as special cases. Let A be a finite alphabet. Then the weak closure $\bar{\mathscr{E}}$ of any set $\mathscr{E} \subset \mathscr{M}(A)$ is weakly compact (as $\mathscr{M}(A)$ itself is weakly compact [116]). For each Borel set $\mathscr{E} \subset \mathscr{M}(A)$, let $\mathscr{B}(\mathscr{E})$ denote the σ-field of all Borel subsets of $\mathscr{E}$. Given $\mu \in \mathscr{M}(A)$ there exists a probability measure $\hat{\mu}$ on $(\mathscr{E}(A), \mathscr{B}(\mathscr{E}(A)))$ such that

$$\mu = \int_{\mathscr{E}(A)} \nu \hat{\mu}(\mathrm{d}\nu)$$

(this is a different but equivalent formulation of the ergodic decomposition theorem (see Theorem 34 above); cf. [42, 43, 15]). If $\mathscr{E} \in \mathscr{B}(\mathscr{E}(A))$, let $\hat{\mathscr{E}} = \{\mu \in \mathscr{M}(A) : \hat{\mu}(\mathscr{E}) = 1\}$, i.e., $\hat{\mathscr{E}}$ consists of all those measures in $\mathscr{M}(A)$ which have their ergodic components in $\mathscr{E}$. Let $\hat{A}$ be a finite set and $\varrho : A \times \hat{A} \to [0, \infty)$ an arbitrary function. Let $(\varrho_n; n \geqq 1)$ denote the corresponding single-letter fidelity criterion. Put

$$\alpha(\mu) = \mathsf{E}_\mu[\min_{\hat{a}\in\hat{A}} \varrho(X_0, \hat{a})], \quad \mu \in \mathscr{M}(A),$$

i.e., $\alpha(\mu)$ is the minimum possible distortion when coding the source $[A, \mu]$. Let $\beta(\mu)$ denote the corresponding rate, viz.

$$\beta(\mu) = R(\alpha(\mu), \mu)$$

($R(\cdot, \mu)$ is the usual RDF [9]). If $\mathscr{E} \in \mathscr{B}(\mathscr{E}(A))$, put

$$R^*(\mathscr{E}) = \begin{cases} 0 & \text{if} \quad \mathscr{E} = \bar{\mathscr{E}}; \\ \sup\limits_{\mu\in\bar{\mathscr{E}}\setminus\mathscr{E}} \operatorname{ess.sup}\limits_{\nu\in\mathscr{E}(A)[\hat{\mu}]} \beta(\nu) & \text{if} \quad \mathscr{E} \neq \bar{\mathscr{E}}. \end{cases}$$

If $A = \hat{A}$ and ϱ is the Hamming distance then $\alpha \equiv 0$, $\beta(\mu) = h(\mu)$, and thus the latter essential supremum is the asymptotic rate $H^*(\mu)$ (see Section 21). In this case

$$R^*(\mathscr{E}) = \begin{cases} 0 & \text{if} \quad \mathscr{E} = \bar{\mathscr{E}}; , \\ \sup\{H^*(\mu) : \mu \in \bar{\mathscr{E}} \setminus \mathscr{E}\} & \text{if} \quad \mathscr{E} \neq \bar{\mathscr{E}} \end{cases}$$

and $R^*(\mathscr{E}) = R^*(K, n)$ if $\mathscr{E}$ is as in Theorem 58(b). Kieffer [68] obtained the following general result:

Theorem 59. Let $\mathscr{E} \in \mathscr{B}(\mathscr{E}(A))$ be such that (a) the restriction of the entropy functional h to $\bar{\mathscr{E}}$ is d_w-continuous and (b) $\bar{\mathscr{E}} \subset \hat{\mathscr{E}}$. Then strong universal (fixed-rate) coding of $\mathscr{E}$ can be done at rates above $R^*(\mathscr{E})$, that is, for any $R \geqq R^*(\mathscr{E})$ and for any $\varepsilon > 0$, there exists a fixed-rate block code book B with $R(B) \leqq R$ and $\bar{\varrho}(B, \mu) \leqq$ $\leqq D(R, \mu) + \varepsilon$ for all $\mu \in \mathscr{E}$.

Note that in light of previous results it is possible to prove Theorem 59 also for variable-rate block codes as well as for fixed- and variable-rate sliding-block codes.

The rest of this section is devoted to results which help to understand the meaning of conditions formulated in Theorem 59. First of all, there is a "converse" to Theorem 59 which says that the bound $R \geqq R^*(\mathscr{E})$ is the best one we can get. To this end, let $\mathscr{E} \subset \mathscr{E}(A)$. A function $R : \mathscr{E} \to [0, \infty)$ is said to be an admissible rate function if for any $\varepsilon > 0$ there exists a variable-rate code $\mathscr{C}$ such that $\bar{\varrho}(\mathscr{C}, \mu) \leqq D(R(\mu), \mu) + \varepsilon$ and $\bar{r}(\mathscr{C}, \mu) \leqq R(\mu) + \varepsilon$ for all $\mu \in \mathscr{E}$ (see [71] and [68] for definitions).

Theorem 60. Suppose $\mathscr{E}$ has the properties listed in Theorem 59. Let $R^* \geqq 0$ be such that every uniformly d_w-continuous function $R : \mathscr{E} \to [R^*, \infty)$ is an admissible rate function. Then $R^* \geqq R^*(\mathscr{E})$.

Next, let us clarify the role of the continuity condition (a) in Theorem 59. A *noiseless code* can be defined as a pair (σ, n), where $n \geqq 1$ and $\sigma : A^n \to \mathbf{N}$ is a length function, i.e.,

$$\sum_{x \in A^n} 2^{-\sigma(x)} \leqq 1$$

(see [71]). We say that strong universal noiseless coding of a set $\mathscr{E} \subset \mathscr{E}(A)$ is possible if for any $\varepsilon > 0$ there exists a code (σ, n) such that for every $\mu \in \mathscr{E}$, $\bar{r}((\sigma, n), \mu) \leqq h(\mu) + \varepsilon$.

Proposition 61. Let $\mathscr{E} \subset \mathscr{M}(A)$. Then strong universal noiseless coding of $\mathscr{E}$ is possible if and only if $h \mid \bar{\mathscr{E}}$ is d_w-continuous.

There are only several results known concerning examples of $\bar{d}$-totally bounded classes of sources [95, 68, 47] but no general criteria are available. Kieffer [76] obtained a characterization of $\bar{d}$-total boundedness for classes of Bernoulli sources. It turns out that this amounts to a uniform version of the characterization theorem for Bernoulli sources (see Theorem 9). It should be clear what should mean conditions like VWB, FD, etc., uniformly for a class of sources. Let $\mathscr{D}(A)$ denote the class of all IID sources over the alphabet A. We say that $\mathscr{E}$ is a continuous stationary coding of $\mathscr{D}(A)$ if there exists a stationary code $\bar{\Phi} : A^{\mathbf{Z}} \to A^{\mathbf{Z}}$ such that

(i) the map $\mu \mapsto \mu\bar{\Phi}^{-1}$ from $\mathscr{D}(A)$ to the class of all Bernoulli sources is $\bar{d}$-continuous, and

(ii) $\{\mu\bar{\Phi}^{-1} : \mu \in \mathscr{D}(A)\} \supset \mathscr{E}$.

Theorem 62. Let $\mathscr{E}$ be a set of Bernoulli sources over a fixed finite alphabet A. Then the following are equivalent: (a) $\mathscr{E}$ is uniformly VWB, (b) $\mathscr{E}$ is uniformly FD, (c) $\mathscr{E}$ is uniformly ABI, (d) $\mathscr{E}$ is a continuous stationary coding of $\mathscr{D}(A)$, (e) there exists a weakly closed set $\mathscr{E}' \subset \mathscr{M}(A)$ such that $\mathscr{E} \subset \mathscr{E}'$ and $\mathscr{E}'$ consists of Bernoulli sources and h is d_w-continuous on $\mathscr{E}'$, and (f) $\mathscr{E}$ is $\bar{d}$-totally bounded.

Finally, let us make the following remark. Since there are only a few $\bar{d}$-totally bounded classes known, it is reasonable to ask whether some weaker distance func-

tion is still compatible with strong universal coding (for a weaker metric admits for a larger class of totally bounded sets). Kieffer [68] observed that the entropy metric

$$d_e(\mu, \nu) = d_w(\mu, \nu) + |h(\mu) - h(\nu)|$$

is compatible with strong universal coding. He proved that it is weaker than $\bar{d}$ by constructing an example of an d_e compact set which is not $\bar{d}$ compact (note that d_e and $\bar{d}$ are the same for Bernoulli sources by Theorem 9 and the definition of FD).

26. Perfect Transmission Over Noisy Channels

Let B and C be two finite sets. By definition, a *channel* $[B, \nu, C]$ is a family $\nu = (\nu_u;\ u \in B^{\mathbf{Z}})$ of probability measures ν_u on $(C^{\mathbf{Z}}, \mathscr{C}^{\mathbf{Z}})$ such that for each $F \in \mathscr{C}^{\mathbf{Z}}$ the map $u \to \nu_u(F) : B^{\mathbf{Z}} \to [0, 1]$ is measurable. A channel $[B, \nu, C]$ is called *stationary* if

$$\nu_{T_B u}(T_C F) = \nu_u(F);\quad u \in B^{\mathbf{Z}},\ F \in \mathscr{C}^{\mathbf{Z}}.$$

If λ is a probability measure on $(B^{\mathbf{Z}}, \mathscr{B}^{\mathbf{Z}})$, we let $\lambda\nu$ denote the joint input/output distribution of the channel $[B, \nu, C]$; $\lambda\nu$ is uniquely determined on $\mathscr{B}^{\mathbf{Z}} \times \mathscr{C}^{\mathbf{Z}}$ by the properties that

$$\lambda\nu(E \times F) = \int_E \nu_u(F)\, \lambda(\mathrm{d}u);\quad E \in \mathscr{B}^{\mathbf{Z}},\ F \in \mathscr{C}^{\mathbf{Z}}.$$

Observe that if $[B, \nu, C]$ is stationary then $\lambda\nu \in \mathscr{M}(B \times C)$ whenever $\lambda \in \mathscr{M}(B)$. If $\lambda \in \mathscr{E}(B)$ entails $\lambda\nu \in \mathscr{E}(B \times C)$, the stationary channel $[B, \nu, C]$ is called *ergodic*.

Before entering the problem of zero-error transmission let us make several remarks concerning the usual block coding approach to channel coding. For details, refer to [8] and [156]. Following Wolfowitz an (M, n, ε) *channel code* is a collection $\mathscr{Y} = \{(\mathbf{y}_i, G_i) : 1 \leqq i \leqq M\}$ of M distinct code words $\mathbf{y}_i \in B^n$ and M mutually disjoint decoding sets $G_i \subset C^n$ such that

$$(*) \qquad \max_{1 \leqq i \leqq M} \sup_{u \in C(\mathbf{y}_i)} \nu_u[C(C^n \setminus G_i)] \leqq \varepsilon,$$

where $C(E) = \{x : x^n \in E\}$ for a set of n-tuples E. Let $C_0(\nu)$ denote the supremum of permissible rates (the rate of the foregoing code is $R(\mathscr{Y}) = n^{-1} \log M$) i.e., $C_0(\nu)$ is the largest possible rate for which coding is possible which gives the error probability $(*)$ as close to zero as we please provided only the block length n is large enough [50].

The channel coding theorems serve the purpose of establishing that a particular number C_0 is the capacity of a given channel. To this end one proves the *positive coding theorem* (that is, for each $R < C_0$, there exists $(\mathrm{INT}\{\exp(NR)\}, n, \varepsilon_n)$ channel codes with $\varepsilon_n \to 0$ as $n \to \infty$, and hence $C_0 \leqq C_0(\nu)$) and a *weak converse* (i.e., given any sequence of $(\mathrm{INT}\{\exp(NR)\}, n, \varepsilon_n)$ channel codes, $R > C_0$ there

exists an $\varepsilon_0 > 0$ such that $\varepsilon_n \geqq \varepsilon_0$ for all n large enough, and hence $C_0 \geqq C_0(\nu)$).

The source of difficulties is the positive part while the weak converse can usually be proved for arbitrary stationary channels (this depends, of course, on the adopted concept of channel capacity, see [50]). The positive part is usually proved using a random coding argument based on Feinstein's lemma [63]. The problem with Feinstein's lemma is that it gives "good" error probability not with respect to the actual channel probabilities as required by (*), but with respect to the channel output probability induced from the artificial "capacity yielding" source. McMillan [92] was the first to recognize a kind of continuity property as responsible for the possibility to derive (*) from a similar relation for the artificial probabilities over channel output n-tuples. However, he was not able to single out the type of continuity needed for Feinstein-like arguments. Consequently, the idea of continuity was quite forgotten although implicitly used by many authors who attempted to find ever less restrictive constraints as to the channel input memory and anticipation. Gray and Ornstein [50] introduced $\bar{d}$-continuous channels and showed that all existing constraints actually imply continuity properties at least as strong as $\bar{d}$-continuity. Thus, $\bar{d}$-continuous channels are the most general ones for which a Feinstein-type approach works.

On the other hand, the question of most interest is to prove joint source/channel coding theorems, i.e., coding theorems for transmission of sources across noisy channels. For this one usually combines a channel coding theorem giving good channel codes (as in (*)) and a source coding theorem (using the code word set of a good channel code as a code book for the source). A typical result of this type is as follows:

Theorem 63. Let $[A, \mu, X]$ be an ergodic source and $[B, \nu, C]$ an ergodic $\bar{d}$-continuous channel. Let $h(X) < C(\nu)$, where $C(\nu)$ is the *Shannon capacity*, i.e., $C(\nu)$ is the supremum of information rates $I(\lambda\nu)$ over all $\lambda \in \mathscr{E}(B)$. For any $\varepsilon > 0$ and for n large enough there exist block codes $\bar{f} : A^{\mathbf{Z}} \to B^{\mathbf{Z}}$ and $\bar{g} : C^{\mathbf{Z}} \to A^{\mathbf{Z}}$ of order n such that

$$(**) \qquad \text{Prob}\,[X^n \neq (gY)^n] \leqq \varepsilon\,,$$

where Y is the channel output process corresponding to the input process $\bar{f}X$.

Kieffer in a series of papers [69, 73, 75] observed that in order to get (**) it is not necessary to have (*). Indeed, the probabilities in (**) are determined from knowing merely the joint input/output distributions so that, by paraphrasing Feinstein's approach, the continuity of the actual channel probabilities $u \mapsto \nu_u(.)$ might be replaced by continuous dependence of $\lambda\nu$ on the input distribution λ. Following Kieffer, a stationary channel $[B, \nu, C]$ is said to be *weakly continuous* if the assumptions that $\lambda_n \in \mathscr{E}(B)$, $\lambda \in \mathscr{E}(B)$, and $d_w(\lambda_n, \lambda) \to 0$ imply that $d_w(\lambda_n\nu, \lambda\nu) \to 0$ as $n \to \infty$. Any $\bar{d}$-continuous stationary channel is weakly continuous [69]. Moreover, weakly continuous channels are the most general for which we can reasonably ask for a cod-

ing theorem (see [73, 75]). We refer the reader to the latter two references for block and sliding-block channel coding theorems and devote the rest to stationary coding.

The problem of zero-error transmission using block codes was formulated already by Shannon [127] who observed that for certain channels one actually can obtain block codes (of sufficiently large rates) which result in zero error probability. On the other hand, the problem in general turned out to be very difficult and the corresponding zero-error (block coding) capacity has been calculated only in a few special cases.

However, if one no longer insists on block codes, powerful results are obtainable using Ornstein's coding technique as first observed in [51] and then, in great generality in [72].

A stationary source $[A, \mu]$ is said to be *zero-error transmissible* over a stationary channel $[B, \nu, C]$ if there exist processes X, U, V over alphabets A, B, C, and stationary codes $\bar{f}: A^{\mathbf{Z}} \to B^{\mathbf{Z}}$, $\bar{g}: C^{\mathbf{Z}} \to A^{\mathbf{Z}}$ such that $\operatorname{dist}(X) = \mu$, $U = \bar{f}X$, $\operatorname{dist}(V \mid U) = \nu$, and $X = \bar{g}V$ a.e. The formula $\operatorname{dist}(V \mid U) = \nu$ means that $\operatorname{dist}(V \mid U = u) = \nu_u$ for almost all $u \in B^{\mathbf{Z}}$.

A stationary source $[B, \lambda]$ is called *ν-invulnerable* if there exist processes U, V, and a stationary code $\bar{h}: C^{\mathbf{Z}} \to B^{\mathbf{Z}}$ such that $\operatorname{dist}(U) = \mu$, $\operatorname{dist}(V \mid U) = \nu$, and $U = \bar{h}V$ a.e.

Lemma 64. A stationary source $[A, \mu]$ is zero-error transmissible over a stationary channel $[B, \nu, C]$ if and only if it is isomorphic to a ν-invulnerable source.

As already mentioned, the zero-error transmission theorem is a particular case of Kieffer's isomorphism theorem (see Theorem 19). However, the proof of this fact is quite involved for there do not exist simple method of verification of the condition (A) formulated in Section 16.

Theorem 65. Let $[B, \nu, C]$ be an ergodic and weakly continuous channel with Shannon capacity $C(\nu)$. Let $[A, \mu]$ be an ergodic aperiodic source over a finite alphabet A. If $h(\mu) < C(\nu)$ then $[A, \mu]$ is zero-error transmissible over $[B, \nu, C]$. Conversely, if $[A, \mu]$ is zero-error transmissible then $h(\mu) \leqq C(\nu)$.

In [72] the theorem is obtained as a consequence of a more general result which shows that the conclusion of zero-error transmissibility is valid if, roughly speaking, the conditions of ergodicity and weak continuity are satisfied only locally (in the same spirit are the block and sliding-block transmission results in [73] and [75]).

Theorems 65 gives a new interpretation to the Shannon capacity and, furthermore, the usual ε-formulations of transmission theorems using codes of finite orders can be obtained as approximations to infinite zero-error codes. Unfortunately, this way of proving block and sliding-block transmission theorems is not yet possible, for the proofs of Theorem 65 in [72] and [83] presuppose knowledge of these results. Hence, it is very desirable to have a simple or, at least, a direct proof.

A natural further step is to study the structure of codes $\bar{f}$ and $\bar{g}$. For example, is

it possible to choose them as finitary codes? Unfortunately, one can expect answers rather in negative (see [74]). On the other hand, it is quite easy to extend Theorem 65 to transmission of aperiodic non-ergodic sources as done in [146].

Finally, note that Kieffer's approach to channel coding problems makes it possible to overcome the difficulty that the structure obtained by quantizations of the channel alphabets is not a channel. As a consequence, one can extend block transmission theorems to channels with alphabets which are standard Borel spaces [143, 144].

PART VIII : CONCLUSIONS

27. Open Problems and Perspectives

In the last decade it was possible to recognize a strong trend towards both-sided exchange of ideas between ergodic theory and information. In this section I tried to collect some ideas and suggestions of which seems to me of importance for the future interplay between the two topics. The formulations are sometimes vague and also express my own interests. In any case, I hope they will at least stimulate interest in this challenging part of contemporary mathematics.

1. *Algorithmic methods.* In applications of information theory we meet again and again the problems of testing statistical models. Though we often have some evidence in favour of properties like ergodicity and stationarity, it is extremely difficult to test them in a rigorous manner. Also, a dynamical system may undergo uncontrolled changes in time so that we cannot be sure that a physical observable is measured along a single typical orbit. Thus, the only information at our disposal is frequently just the individual sequence of observed outcomes.

There is already some progress in coding problems related to individual sequences instead of assuming some prior probabilistic model. The coding algorithms are then based on various complexity measures [159, 160]. Of course, the choice of complexity measure depends on which type of coding device is at our disposal. For example, there exist binary seqeunces for which the finite-state-complexity is one (i.e., the largest possible) while the normalized Kolmogorov-Solomonoff-Chaitin complexity [23] is zero (see [159]). On the other hand, this is not so serious from the point of view of data compression, for we usually are given the type of coding devices in advance.

Heim [57] formulated parts of information theory in terms of complexity measures and computable probabilities. An interesting problem is to check whether it is possible to develop an algorithmic counterpart of, say, the Ornstein's theory. Any result of this type would give an universal code for all Bernoulli sources of the same entropy. In light of presented universal results (cf. Theorems 44 and 46) this does not seem so strange.

2. *Stationary entropy compression coding.* There exist several problems connected

with that area. Dunham [33] has shown how to stationarize a block code so that both the rate and the average distortion remain almost unaffected. Is it possible to obtain also a counterpart of Ornstein's technique of making good codes much better in case when goodness is measured in terms of average distortion?

Another open problem is to relate isomorphism of sources with their behaviour from the point of view of distortion-rate theory. Because of the "averaging effect" in computations of average distortion it does not seem likely, that identical DRF's relative to one fixed distortion measure could give some results. However, what about the case when we have a "sufficiently rich" class of distortion measures (by sufficiently rich I mean a class such that any dissimilarity is recognized by at least one distortion measure while the other ones can average it out).

3. *Codes with prescribed properties.* One usually has some desirable class of sources (for example, IID ones in redundancy removal problems) and some family of codes connected with each particular coding problem. Characterization theorems are of interest. By this I mean (hopefully simple) characterization of all those sources which can be coded using a code from the given family so as to get a source belonging to the given class. One problem of this type has been recently investigated by Rudolph [123]. He introduced the concept of a "finitarily Bernoulli" process and proved that entropy is a complete finitary isomorphism invariant for that class. This gives a characterization of all those processes finitarily isomorphic to a Bernoulli shift.

A problem going in the opposite direction is to find information theoretic interpretations (similarly as was done for Ornstein's technique) to other technique which produced important results in ergodic theory (e.g., to coding techniques used to derive relative isomorphism theorems, see [150]).

A slightly more technically formulated problem: let $[A, \mu, X]$ be an IID source and let $\delta(R, \mu)$ be the OPTA using stationary codes. Thus, we get performance as close to $\delta(R, \mu)$ as we please using Bernoulli encoded processes (for any stationary coding of an IID process is Bernoulli). Is it possible to get the same conclusion using only stationary codes $\bar{\Phi}$ such that $\bar{\Phi}X$ is again IID? Note that codes having this property exist by [60]. The problem is how to control the average distortion. At present, only a weaker is known [147].

4. *Ergodic theory for channels.* It would be very desirable to have a classification of channels according to their transmissibility properties. At present, some results are available which throw light upon the structure of channels. Neuhoff and Shields [97] introduced the concept of channel $\bar{d}$-distance and obtained interesting results like characterizations of all channels which can be obtained as $\bar{d}$-limits of sequences of channels with very simple properties (like primitive or finite-state channels; see also [158]).

5. *Zero-entropy and related processes.* As shown by Sigmund [132] zero-entropy processes prevail in the sense that they form a topologically large set. However, there are also other, more realistic, reasons for investigations on zero-entropy processes.

In fact, a manual for measuring some physical quantity may produce very complex finite strings of outcomes, however, it is felt that in the limit of ever longer strings, the complexity will remain bounded due to unchanging prescription of how to get an outcome. In light of connections between entropy and complexity [23] one can expect that observations correspond to zero-entropy processes much more frequently than usually judged. In any case, as already mentioned, the methods based on complexity measures have the advantage of working universally, i.e., without the requirement of a prior statistical model and even without any assumptions concerning stationarity at all.

A well-known result is that zero entropy processes can be distinguished by the sequence entropy [88]. A complete classification is still absent, however. Also, is the sequence entropy the best we can do, or do there exist more natural invariants?

Information-theoretical considerations enable us to define processes which, being in fact random, behave from the point of view of both rate and distortion as deterministic provided the observation time is large enough – so called information singular processes introduced by Berger [10] and studied recently in [53]. Such processes might become interesting also from the physical point of view. Intuitively, they seem to describe, in information theoretic terms, motions which look like random for a long period of time but sooner or later they finally move into a steady state.

6. *AMS theory.* Processes obtained as fixed- or variable-rate block codings of stationary processes have a weaker stationarity property called asymptotically mean stationarity (AMS; see [45]). The basic concepts of AMS theory for channels have been introduced in [37] and several coding theorems have been obtained both for AMS sources and AMS channels [141, 142, 140]. On the other hand, all these results can serve just as preliminaries to a complete theory of AMS processes.

Of course, this is only a very small sample of problems which I judge as important and interesting. I attempted to present only such problems which seem to be extremely difficult or which at least seem to require development of new ideas and concepts. I think it is more stimulating to give such an account than to present a plenty of highly technical questions.

28. Acknowledgment

First of all, I should like to express my sincere thanks to the editors of the journal "Kybernetika" for their kind suggestion and encouragement to prepare a survey paper.

I profited much from a substantial help of many collegues (too many to be all mentioned here) who kept me informed about their results by sending me their new papers, often prior to publication. Especially, my thanks are due to Manfred Denker, Robert M. Gray, Michael Keane, and William Parry. Special thanks are due to John Kieffer for many stimulating discussions on relations between ergodic and information theories. On the other hand, I am fully responsible for conclusions and judgments made throughout the paper.