Fifty Years of Shannon Theory

Sergio Verdú, *Fellow, IEEE*

***Abstract*— A brief chronicle is given of the historical development of the central problems in the theory of fundamental limits of data compression and reliable communication.**

***Index Terms*— Channel capacity, data compression, entropy, history of Information Theory, reliable communication, source coding.**

CLAUDE Shannon's "A mathematical theory of communication" [1] published in July and October of 1948 is the Magna Carta of the information age. Shannon's discovery of the fundamental laws of data compression and transmission marks the birth of Information Theory. A unifying theory with profound intersections with Probability, Statistics, Computer Science, and other fields, Information Theory continues to set the stage for the development of communications, data storage and processing, and other information technologies.

This overview paper gives a brief tour of some of the main achievements in Information Theory. It confines itself to those disciplines directly spawned from [1]—now commonly referred to as Shannon theory.

Section I frames the revolutionary nature of "A mathematical theory of communication," in the context of the rudimentary understanding of the central problems of communication theory available at the time of its publication.

Section II is devoted to lossless data compression: the amount of information present in a source and the algorithms developed to achieve the optimal compression efficiency predicted by the theory.

Section III considers channel capacity: the rate at which reliable information can be transmitted through a noisy channel.

Section IV gives an overview of lossy data compression: the fundamental tradeoff of information rate and reproduction fidelity.

The paper concludes with a list of selected points of tangency of Information Theory with other fields.

I. Before 1948

The major communication systems existing in 1948 were

- Telegraph (Morse, 1830's);
- Telephone (Bell, 1876);
- Wireless Telegraph (Marconi, 1887);
- AM Radio (early 1900's);
- Single-Sideband Modulation (Carson, 1922);
- Television (1925–1927);
- Teletype (1931);
- Frequency Modulation (Armstrong, 1936);
- Pulse-Code Modulation (PCM) (Reeves, 1937–1939);
- Vocoder (Dudley, 1939);
- Spread Spectrum (1940's).

In those systems we find some of the ingredients that would be key to the inception of information theory: a) the Morse code gave an efficient way to encode information taking into account the frequency of the symbols to be encoded; b) systems such as FM, PCM, and spread spectrum illustrated that transmitted bandwidth is just another degree of freedom available to the engineer in the quest for more reliable communication; c) PCM was the first digital communication system used to transmit analog continuous-time signals; d) at the expense of reduced fidelity, the bandwidth used by the Vocoder [2] was less than the message bandwidth.

In 1924, H. Nyquist [3] argued that the transmission rate is proportional to the logarithm of the number of signal levels in a unit duration. Furthermore, he posed the question of how much improvement in telegraphy transmission rate could be achieved by replacing the Morse code by an "optimum" code.

K. Küpfmüller [4] (1924), H. Nyquist [5] (1928), and V. Kotel'nikov [6] (1933) studied the maximum telegraph signaling speed sustainable by bandlimited linear systems. Unbeknownst to those authors, E. Whittaker [7] (1915) and J. Whittaker [8] (1929) had found how to interpolate losslessly the sampled values of bandlimited functions. D. Gabor [9] (1946) realized the importance of the duration–bandwidth product and proposed a time–frequency uncertainty principle.

R. Hartley's 1928 paper [10] uses terms such as "rate of communication," "intersymbol interference," and "capacity of a system to transmit information." He summarizes his main accomplishment as

> the point of view developed is useful in that it provides a ready means of checking whether or not claims made for the transmission possibilities of a complicated system lie within the range of physical possibility.

Intersymbol interference and basic observations with RLC circuits lead Hartley to conclude that the capacity is proportional to the bandwidth of the channel. But before being able to speak of "capacity," Hartley recognizes the need to introduce a "quantitative measure of information." He uses the letter H to denote the amount of information associated with n selections and states that

$$H = n \log s$$

where s is the number of symbols available in each selection. The principle that "information" is the outcome of a selection among a finite number of possibilities is firmly established in [10].

Manuscript received June 9, 1998.

The author is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Publisher Item Identifier S 0018-9448(98)06315-9.

The aforementioned papers by Nyquist and Hartley had not quantified the effects of noise, nor had they modeled sources of information probabilistically. Much of the credit for importing random processes into the toolbox of the 1940's communications engineer is due to N. Wiener [11][1] and to S. Rice [12].

Probabilistic modeling of information sources has in fact a very long history as a result of its usefulness in cryptography. As early as 1380 and 1658, tables of frequencies of letters and pairs of letters, respectively, had been compiled for the purpose of decrypting secret messages [13].[2] At the conclusion of his WWII work on cryptography, Shannon prepared a classified report [14][3] where he included several of the notions (including entropy and the phrase "information theory") pioneered in [1] (cf. [16]). However, Shannon had started his work on information theory (and, in particular, on probabilistic modeling of information sources) well before his involvement with cryptography.[4] Having read Hartley's paper [10] in his undergraduate days, Shannon, as a twenty-two-year-old graduate student at MIT, came up with a ground-breaking abstraction of the communication process subject to a mean-square fidelity criterion [19]. After writing his landmark Master's thesis on the application of Boole's algebra to switching circuits [20] and his Ph.D. dissertation on population dynamics [21], Shannon returned to communication theory upon joining the Institute for Advanced Study at Princeton and, then, Bell Laboratories in 1941 [16].

By 1948 the need for a theory of communication encompassing the fundamental tradeoffs of transmission rate, reliability, bandwidth, and signal-to-noise ratio was recognized by various researchers. Several theories and principles were put forth in the space of a few months by A. Clavier [22], C. Earp [23], S. Goldman [24], J. Laplume [25], C. Shannon [1], W. Tuller [26], and N. Wiener [27]. One of those theories would prove to be everlasting.

II. Lossless Data Compression

A. The Birth of Data Compression

The viewpoint established by Hartley [10] and Wiener [11] is echoed by Shannon in the Introduction of [1]:

> [The] semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages.

Shannon then makes the key observation that the source of information should be modeled as a random process:

> We can think of a discrete source as generating the message, symbol by symbol. It chooses successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities is known as a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source.

Shannon recognizes that to exploit the redundancy of the source one should take into account not only the frequencies of its symbols but its memory. But before proceeding to tackle that problem, he considers a single random variable taking n values with probabilities $p_1, \cdots, p_n$ and defines its *entropy*:[5]

$$H = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

Shannon points out the similarity with Boltzmann's entropy in statistical mechanics [29] and gives an axiomatic rationale for this measure of information, as the only measure that is i) continuous in the probabilities, ii) increasing with n if the random variable is equiprobable, and iii) additive, in the sense that if the random value is the result of two choices, its entropy can be obtained by summing the entropy of the first choice and the entropy of the second choice given the first.

Much more important than the axiomatic justification of entropy are the fundamental theorems that it satisfies. Shannon goes on to consider memoryless sources, and proves the following result using the law of large numbers:

Shannon's Theorem 3 [1]: Given any $\epsilon > 0$ and $\delta > 0$, we can find N_0 such that the sequences of any length $N \geq N_0$ fall into two classes

1) A set whose total probability is less than ϵ.
2) The remainder, all of whose members have probabilities $[p]$ satisfying the inequality

$$\left| H - \frac{\log p^{-1}}{N} \right| < \delta. \tag{2}$$

Shannon refers to the second class as the "typical sequences." They are characterized by probabilities that decrease exponentially with blocklength, $p = a^{-N}$, with $a \approx 2^H$. Shannon's Theorem 3 states that the set of atypical sequences has vanishing probability. The relevance of this result to data compression is that for the purposes of coding we can treat the typical sequences as roughly equiprobable while disregarding the atypical sequences. The resulting code maps source strings of length N to strings of length slightly larger than HN. The decoder can recover the original source string with probability at least $1-\epsilon$. Thus the rate of H encoded bits

[1] Originally a WWII classified report acknowledged in [1] to have influenced Shannon's thinking.

[2] Even higher order statistics had been envisioned. Jonathan Swift's *Gulliver's Travels* (1726) describes a machine by which "the most ignorant person may write in philosophy, poetry and politics." The machine selects words at random based on "the strictest computation of the general proportion between the numbers of particles, nouns and verbs."

[3] Later declassified and superseded by [1] and [15].

[4] According to interviews with Claude Shannon recorded in [16]–[18].

[5] Full and sole credit is due to Shannon for the introduction of entropy in information theory. Wiener never worked with entropy; instead, he introduced, apparently at J. von Neumann's suggestion and independently of Shannon, the differential entropy [27] which he used in the context of Gaussian random variables. A distant relative of the differential entropy dating back to 1934 is Fisher's information [28], which gives a fundamental limit on the achievable mean-square error of parametric estimation.

per source symbol is *achievable* provided we are willing to tolerate a nonzero probability of failing to recover the original sequence. By increasing the blocklength, and thus the delay and complexity of encoding and decoding operations, we can make that probability as small as desired.

But, is that the best we can do? Shannon's Theorem 3 does not address that question, since it only suggests a suboptimal code. (The optimal code of rate R simply disregards all but the 2^{NR} most probable sequences of length N.) Shannon finds the answer in Theorem 4: as long as we require probability of error strictly less than 1, asymptotically, we cannot encode at rates below the entropy. This statement is commonly known as the strong *converse* source coding theorem. The converse (or weak converse) source coding theorem asserts that error probability cannot vanish if the compression rate is below the entropy.

The foregoing discussion was circumscribed to *fixed-length* codes (fixed-length source strings mapped to fixed-length encoded strings). Shannon also notices that by allowing encoded sequences of *variable* length, it is possible to actually achieve zero error probability without increasing the *average* encoding rate. For example, this can be accomplished by representing the typical sequences of length N with sequences of length roughly equal to HN, and leaving all the other sequences uncompressed—a prefix bit indicating whether the encoded sequence is typical. Many other possibilities arise in variable-length data compression. Shannon gives the example of a memoryless source whose symbol probabilities are powers of $1/2$. In this special case, it is easy to find a code that encodes the ith symbol with a string of $-\log p_i$ bits. Much less obvious is what to do with arbitrary distributions. Shannon describes an "arithmetic process," discovered contemporaneously and independently by R. Fano, that assigns to each symbol the appropriately truncated binary expansion of the cumulative distribution function evaluated at the symbol. The average rate of that scheme is not optimal but is only slightly above the entropy.

B. The Asymptotic Equipartition Property

For memoryless sources, Shannon's Theorem 3 is equivalent to the weak law of large numbers for independent and identically distributed random variables taking a finite number of positive values. Because of its relevance to data compression, it is natural to investigate whether Theorem 3 applies to sources with memory. This requires replacing the entropy of an individual random variable by the *entropy rate*, namely, the limit of the entropy of an N-block divided by N. Shannon [1] shows that the entropy rate of a stationary process is equal to the limiting conditional entropy of a single source symbol given the past symbols. Having made the case that the statistics of natural language can be approximated arbitrarily well by Markov chains of increasing order,[6] Shannon [1] notices that Theorem 3 (and, thus, the achievability part of the source coding theorem) applies to stationary Markov chain sources. In 1953, a step-by-step proof of the generalization of Shannon's Theorem 3 to Markov chains was given by A. Khinchin in the first Russian article on information theory [31].

In 1953, B. McMillan [32] used the statistical-mechanics phrase "asymptotic equipartition property" (AEP) to describe the typicality property of Shannon's Theorem 3: the set of atypical sequences has vanishing probability. Moreover, McMillan showed a fundamental generalization of Shannon's Theorem 3 which is commonly referred to as the Shannon–McMillan theorem: the asymptotic equipartition property is satisfied by every *stationary ergodic* process with a finite alphabet. Unlike memoryless sources, for which the AEP is equivalent to the weak law of large numbers, showing that the AEP is satisfied for stationary ergodic sources requires a nontrivial use of the ergodic theorem. While the fundamental importance of ergodic theory to information theory was made evident by McMillan in 1953, the key role that entropy plays in ergodic theory was revealed by A. Kolmogorov [33] in 1958 and would eventually culminate in D. Ornstein's 1970 proof [34] of one of the pillars of modern ergodic theory: the isomorphy theorem.[7]

Shannon's Theorem 3 states that the normalized log-probability of the source string converges in probability as its length goes to infinity. Although this is enough for most lossless source coding theorems of interest, almost-sure convergence also holds as shown in [38] and (with a simpler proof) in [39]. Generalizations of the Shannon–McMillan theorem to continuous-valued random processes and to other functionals of interest in information theory have been accomplished in [40]–[45].

Sources that are either nonstationary or nonergodic need not satisfy Theorem 3[8]; that is, some sources require less than the entropy rate to be encoded, some require more. It is shown in [47] that the AEP is not only sufficient but necessary for the validity of the source coding theorem (in the general setting of finite-alphabet sources with nonzero entropy). Furthermore, [47] shows that the AEP is equivalent to the simpler statement in which the absolute value in (2) is removed.

C. Fixed-to-Variable Source Coding

As studied by Shannon, and used earlier in telegraphy, fixed-to-variable codes map individual information symbols (or, in general, fixed-length words of symbols) to unequal-length strings—with shorter strings assigned to the more likely symbols. In 1948, Shannon had left open two major problems in fixed-to-variable source coding: 1) the construction of a minimum average-length code, and 2) the converse variable-length source coding theorem.

The variable-length source code that minimizes average length was obtained by D. Huffman [48], as an outgrowth of a homework problem assigned in R. Fano's MIT information theory class [49]. The practicality of the Huffman code has withstood the test of time with a myriad applications ranging from facsimile [50] to high-definition television [51].

[6] A view challenged in [30] by N. Chomsky, the father of modern linguistics.

[7] Tutorials on the interplay between information theory and ergodic theory can be found in [35]–[37].

[8] General coding theorems for nonstationary/nonergodic sources can be found in [46].

No formula is known for the minimum average length in terms of the distribution of the source. In [1], Shannon showed that the minimum average length does not exceed the entropy plus one bit,[9] but he did not give a lower bound.

Before Huffman, another MIT student, L. Kraft, had attacked the construction of minimum redundancy codes unsuccessfully. However, in his 1949 Master's thesis [54], Kraft gave a basic condition (known as the Kraft inequality) that must be satisfied by the codeword lengths of a prefix code (i.e., a code where no codeword is the prefix of another).[10] Seven years later, and apparently unaware of Kraft's thesis, McMillan [56] showed that that condition must hold not just for prefix codes but for any uniquely decodable code. (A particularly simple proof was given in [57].) It is immediate to show (McMillan [56] attributes this observation to J. L. Doob) that the average length of any code that satisfies the Kraft inequality cannot be less than the source entropy. This, in turn, implies the converse variable-length source coding theorem, which had already been proven by Khinchin [31] using a method based on Shannon's Theorem 3.

The optimality of the Huffman code must have seemed at the time to leave little room for further work in fixed-to-variable source coding.[11] That, however, proved not to be the case, because of two major difficulties: 1) the distribution of the source may not be known[12] when the code is designed (Section II-E), and 2) although the Huffman algorithm need not operate symbol by symbol, its complexity grows very rapidly with the length of the source block.[13] The incentive for encoding blocks of source symbols stems from two important classes of sources for which symbol-by-symbol encoding may be decidedly suboptimal: sources with memory and binary (or other small alphabet) sources. Both difficulties encountered by the Huffman code also apply to the Shannon–Fano code mentioned in Section II-A. The second shortcoming is circumvented by the arithmetic coding method of J. Rissanen [60] (generalized in [61] and [62] and popularized in [63]), whose philosophy is related to that of the Shannon–Fano code.[14] The use of arithmetic coding is now widespread in the data-compression industry (and, in particular, in image and video applications [69]). Much of the success of arithmetic coding is due to its rational exploitation of source memory by using the conditional probability of the next symbol to be encoded given the observed past.

[9] Tighter distribution-dependent bounds are known [52], [53].

[10] Kraft [54] credits the derivation of the inequality to R. M. Redheffer, who would later coauthor the well-known undergraduate text [55].

[11] Minimum average-length source-coding problems have been solved with additional constraints such as unequal symbol lengths, infinite alphabets, lexicographic ordering of encoded strings, maximum codeword length, etc. See [58] for a recent survey.

[12] As a result of its emphasis on asymptotic stationary settings, Shannon theory has not been engulfed in the Bayesian/non-Bayesian schism that has plagued the field of statistics.

[13] For most Markov sources the minimum average length per letter approaches the entropy rate hyperbolically in the blocklength [59].

[14] The Shannon–Fano code is frequently referred to as the Shannon–Fano–Elias code, and the arithmetic coding methods described in [64] and [65] are attributed to P. Elias therein. Those attributions are unfounded [66]. In addition to [1], other contributions relevant to the development of modern arithmetic coding are [67] and [68].

D. Variable-to-Fixed Source Coding

So far we have considered data-compression methods whereby fixed-size blocks of source symbols are encoded into either variable-length or fixed-length strings. The variable-to-fixed source coding approach is advantageous whenever block formatting of encoded data is required. The key notion here is that of parsing (i.e., inserting commas) the source sequence into consecutive variable-length phrases. In variable-to-fixed source coding, those phrases belong to a predetermined fixed-size dictionary. Given the size of the dictionary, the Tunstall algorithm [70] selects its entries optimally under the condition that no phrase is the prefix of another and that every source sequence has a prefix in the dictionary. For memoryless sources, the Tunstall algorithm maximizes the expected length of the parsed phrases. Further results on the behavior of the Tunstall algorithm for memoryless sources have been obtained in [71] and [72]. For Markov sources, optimal variable-to-fixed codes have been found in [73] and [74].

Variable-to-fixed codes have been shown to have certain performance advantages over fixed-to-variable codes [75], [76].

Although variable-to-variable source coding has not received as much attention as the other techniques (cf. [77]), it encompasses the popular technique of runlength encoding [78], already anticipated by Shannon [1], [79], as well as several of the universal coding techniques discussed in the next subsection.

E. Universal Source Coding

A. Kolmogorov [80] coined the term "universal" to refer to data-compression algorithms that do not know *a priori* the distribution of the source. Since exact statistical knowledge of the source is the exception rather than the rule, universal source coding is of great practical interest.

If we apply a lossless data-compression algorithm tuned to one source to a different source we still recover the message error-free but with degraded compression efficiency. For memoryless sources, the increase in rate for compressing assuming distribution Q when the true source distribution is P is equal to the divergence[15] of P with respect to Q for both fixed-to-variable [81] and variable-to-fixed [82] coding. If the uncertainty on the source distribution can be modeled by a class of distributions, it was shown by B. Fitingof in [83] and by L. Davisson in [84] that for some uncertainty classes there is no asymptotic loss of compression efficiency if we use a source code tuned to the "center of gravity" of the uncertainty set. Constructive methods for various restricted classes of sources (such as memoryless and Markov) have been proposed by R. Krichevsky and V. Trofimov [59] and by T. Tjalkens and F. Willems [85].

In universal source coding, the encoder can exploit the fact that it observes the source output and, thus, can "learn" the source distribution and adapt to it. The same is true for the decoder because its output is a lossless reconstruction of the source sequence. Adaptive Huffman coding was initially considered in [86] and [52], and modified in [87] and [88]. For

[15] cf. Section III-G.

large-alphabet sources, lower encoding/decoding complexity can be achieved by the adaptive fixed-to-variable source codes of B. Ryabko [89], [90].[16] Showing experimental promise, the nonprobabilistic sorting method of [93] preprocesses sources with memory so that universal codes for memoryless sources achieve good compression efficiency.

Suppose now that we adopt a parametric description of the source uncertainty, say a family of distributions indexed by a string of parameters. In practice, it is useful to consider uncertainty classes that include distributions described by different numbers of parameters (e.g., Markov chains of various orders). We could envision a two-step universal compression procedure: first, using the source sequence, we estimate the unknown parameter string and describe it to the decoder; second, we compress the source sequence using a code tuned to the source distribution with the estimated parameters. What criterion do we adopt in order to estimate the source model? The choice of estimation criterion presents us with a tradeoff: the more finely we estimate the distribution (i.e., the more complex the model) the more efficiently we can compress the source, but also the longer it takes to describe the parameter string to the decoder. Rissanen [94] showed that there are fundamental reasons to choose the *minimum description length* (MDL) criterion for model selection. According to the MDL principle, the parameter string is chosen to minimize the compressed sequence length plus $\frac{m}{2}\log N$ if N is the length of the source sequence and m is the length of the parameter string. The relevance of the information-theoretic MDL principle transcends data compression and is now established as a major approach in statistical inference [95].

The most widely used universal source-coding method is the algorithm introduced by A. Lempel and J. Ziv in slightly different versions in 1976–1978 [96]–[98]. Unlike the methods mentioned so far in this subsection, the Lempel–Ziv algorithm is not based on approximating or estimating the source distribution. Like variable-to-fixed source coding, Lempel–Ziv coding is based on parsing the source sequence. The simple Lempel–Ziv parsing rule (the next phrase is the shortest phrase not seen previously) can be encoded and decoded very easily.[17] Remarkably, the Lempel–Ziv algorithm encodes any stationary ergodic source at its entropy rate as shown by Ziv [100] and Wyner–Ziv [101], [102]. The analysis of the statistical properties of the Lempel–Ziv algorithm has proven to be a fertile research ground [98], [103]–[108].

Despite its optimality and simplicity, the Lempel–Ziv algorithm is not the end of the story in universal source coding. Prior knowledge of general structural properties of the source can be exploited to give better transient (i.e., nonasymptotic) compression efficiency.[18] So far, the most fruitful effort in this direction has its roots in the finite-memory "context-tree" model introduced by Rissanen [109] and has led to the universal optimal method of F. Willems, Y. Starkov, and T. Tjalkens [110]. The method of [110] is devoted to the universal estimation of the conditional probability of the next symbol given the past, which is then fed to a standard arithmetic encoder.[19] The coding rate of the method of [110] achieves the optimum speed of approach to the entropy rate (established in [94]).

Compression of memoryless sources with countably-infinite alphabets and unknown distributions has many practical applications. Several methods for universal encoding of the integers have been proposed in [112]–[115].

Germane to universal source coding is the topic of entropy estimation pioneered by Shannon [1], [116] in the framework of English texts. The empirical estimation of the entropy of natural language is surveyed in [117] and [118]. An obvious approach to entropy estimation is to apply a universal data compressor and observe the rate at which bits are generated at the output. Representative references of the state-of-the-art in entropy estimation [119]–[121], [108] illustrate the recent interest in string-matching approaches.

Nonprobabilistic measures of the compressibility of individual data strings can be defined as the length of the shortest compression achievable by a given class of compression algorithms. The methods and results are crucially dependent on the class of data compressors allowed. J. Ziv and A. Lempel [100], [98] considered the class of finite-state machines, among which the Lempel–Ziv is asymptotically optimal for all sequences. In the mid-1960's, A. Kolmogorov [80], [122], G. Chaitin [123], and R. Solomonoff [124] considered the class of compressors that output a binary program for a universal Turing machine. The resulting measure, which suffers from the shortcoming of being noncomputable, is called *Kolmogorov complexity* or *algorithmic complexity* and its methods of study lie in recursive function theory rather than Shannon theory. However, for some random sources, the expected Kolmogorov complexity rate converges to the entropy rate [101], [125].

F. Separate Compression of Correlated Sources

In the post-Shannon era, one of the most important advances in the theory of fundamental limits of data compression was achieved by D. Slepian and J. Wolf in [126]. Consider two information sources compressed by separate individual encoders that do not have access to the output of the other source. Noiseless separate decompression of the encoded streams requires that the coding rates be equal to the individual entropies. If joint decoding were allowed would it be possible to improve compression efficiency? In particular, can the sum of the rates be strictly smaller than the sum of the individual entropies? Let us assume that the sources are dependent (the answer is obviously negative otherwise) and, thus, the sum of their entropies is strictly larger than their joint entropy. Had we allowed joint source encoding, the answer would be affirmative as the required rate-sum would be equal to the joint entropy. Slepian and Wolf's surprising result was that this conclusion holds even with separate encoding. Shortly afterwards, T. Cover [127] introduced the powerful technique

[16] Rediscovered in [91] and [92].

[17] Practical issues on the implementation of the Lempel–Ziv algorithm are addressed in [99].

[18] Reference [77] gives a survey of the interplay between delay and redundancy for universal source coding with various knowledge of the statistics of the source.

[19] The connections between universal source coding and universal prediction are surveyed in [111].

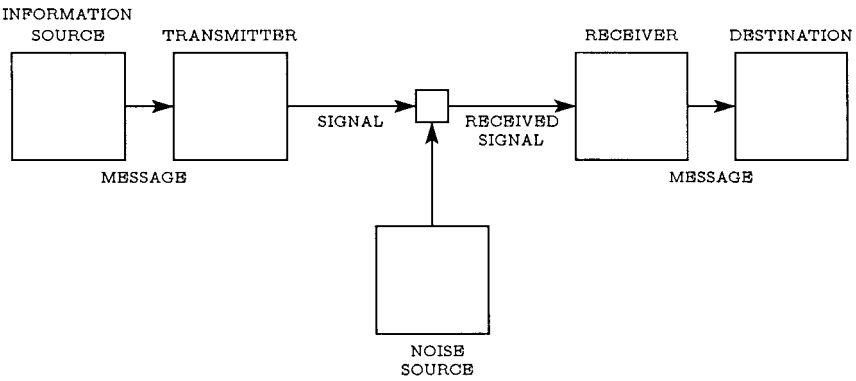

Fig. 1. Figure 1 of [1].

of random binning to generalize the Slepian–Wolf result to jointly stationary/ergodic sources.

Despite the existence of potential applications, the conceptual importance of Slepian–Wolf coding has not been mirrored in practical data compression. Not much progress on constructive Slepian–Wolf schemes has been achieved beyond the connection with error-correcting channel codes revealed in [128]. Interestingly, channel coding presents another avenue for potential applications of Slepian–Wolf coding as shown in [129].[20]

The combinatorial aspects of zero-error source coding [130] are particularly interesting in the context of separate compression of correlated sources or the more canonical setting of compression with decoder side-information. Pioneering contributions in this direction were made by H. Witsenhausen [131] and R. Ahlswede [132]. Inspired by the distributed-computing applications envisioned in the late 1970's, *interactive compression* models allow several rounds of communication between encoders so as to compute a function dependent on their individual observations [133]. The efficient exchange of remote edits to a common file is a typical application. Achievability and converse results have been obtained by A. Orlitsky *et al.* in [134]–[137].

III. Reliable Communication

A. The Birth of Channel Capacity

After fifty years, it is not easy to fully grasp the revolutionary nature of Shannon's abstraction of the fundamental problem of communication depicted in [1, Fig. 1] (Fig. 1). Shannon completes the picture he initiated nine years earlier [19] by introducing a new concept: the "channel," which accounts for any deterministic or random transformation corrupting the transmitted signal. The function of the transmitter is to add "redundancy:"

> The redundancy must be introduced to combat the particular noise structure involved ... a delay is generally required to approach the ideal encoding. It now has the additional function of allowing a large sample of noise to affect the signal before any judgment is made at the receiving point as to the original message.

[20] For example, thanks to Slepian–Wolf coding, the branch from "receiver" to "observer" in [1, Fig. 8] is redundant in order to achieve capacity.

In a world where modulation was generally thought of as an instantaneous process and no error-correcting codes had been invented[21] Shannon's formulation of the problem of reliable communication was a stroke of genius.

Shannon first develops his results on reliable communication within the context of discrete memoryless channels. He defines

> the channel capacity by
>
> $$C = \operatorname{Max}\left(H(x) - H_y(x)\right) \tag{3}$$
>
> where the maximum is with respect to all possible information sources used as input to the channel

and claims

> It is possible to send information at the rate C through the channel *with as small a frequency of errors or equivocation as desired* by proper encoding. This statement is not true for any rate greater than C.

Denoting by $N(T, q)$ the maximum codebook size of duration T and error probability q, Shannon gives the stronger statement.

Shannon's Theorem 12 [1]:

$$\lim_{T\to\infty} \frac{\log N(T,q)}{T} = C \tag{4}$$

where C is the channel capacity, provided that q does not equal 0 or 1.

Shannon justifies the achievability part of this statement succinctly and intuitively, introducing the celebrated technique of random encoding. Error probability is averaged with respect to the codebook choice and shown to vanish asymptotically with T if the transmission rate is lower than C. Shannon notices that this argument leads to the conclusion that not only does there exist a capacity-achieving code, but in fact almost all codes are good. However,

> no explicit description of a series of approximation[s] to the ideal has been found. Probably this is no accident but is related to the difficulty of giving an explicit construction for a good approximation to a random sequence.

[21] With the possible exception of the Hamming $(7, 4)$ code quoted in [1]; see also [138].

Although all the foregoing claims would eventually be shown to hold (Section III-D), Shannon [1] does not prove, even informally, the converse part of Theorem 12 (i.e., $\leq$ in (4)), nor the weaker statement that the error probability of codes with rate above capacity cannot vanish.[22]

B. Mutual Information

In addition to the difference between unconditional and conditional entropy maximized in (3), [1, Part IV] introduces its counterpart for continuous random variables.

$$\iint P(x,y)\log\frac{P(x,y)}{P(x)P(y)}\,dx\,dy.$$

Although no name is given in [1] for this quantity, Shannon realizes that it is "one of the central definitions in communication theory." By 1956, Shannon [140] refers to it as "mutual information," using a terminology attributed to Fano [141], which is commonly accepted today.

Shannon gives the data-processing inequality in [1], and a general approach to define mutual information encompassing the discrete and continuous definitions as special cases. This approach would be followed up by I. Gelfand, A. Yaglom, and A. Kolmogorov in [142]–[144], and by M. Pinsker (with special emphasis on Gaussian random variables and processes) in [145]. Early on, the Russian school realized the importance of the random variable whose expectation is mutual information (i.e., the log-likelihood ratio between joint and marginal-product distributions) and dubbed it *information density*.

The optimization problem posed in Shannon's formula was initially tackled in [146] and [147]. An iterative algorithm for computing the capacity of arbitrary discrete memoryless channels was given independently by S. Arimoto [148] and R. Blahut [149] in 1972.

C. Gaussian Channels

Undoubtedly, the channel that has yielded the biggest successes in information theory is the Gaussian channel.

Shannon [1] formulates the continuous-time ideal strictly bandlimited white Gaussian channel and uses the sampling theorem to show the equivalence to a discrete-time channel sampled at twice the bandwidth.[23] As a formal analog of the entropy of a discrete random variable, Shannon defines the differential entropy of a continuous random variable, and shows that it is maximized, subject to a variance constraint, by the Gaussian distribution. Taking the difference between the output differential entropy and the noise differential entropy, Shannon goes on to obtain his famous formula for the capacity of power-constrained white Gaussian channels with flat transfer function[24]

$$C = W\log\left(\frac{P+N}{N}\right) \tag{5}$$

where W is the channel bandwidth, P is the transmitted power, and N is the noise power within the channel band. Taking the limit of (5) as the bandwidth grows without bound [154], a fundamental conclusion is reached: the minimum energy necessary to transmit one bit of information lies 1.6 dB below the noise one-sided spectral density. Expressions similar to (5) were put forth in 1948 by Wiener [27] and Tuller [26] without information-theoretic justification: Wiener simply takes the difference between Gaussian differential entropies and Tuller arrives at the formula arguing that "if N is the rms amplitude of the noise mixed with the signal, there are $1 + S/N$ significant values of signal that may be determined." By the mid-1950's, the explicit construction of channel codes with rates approaching (5) became the "holy grail" of information theory [155].

The geometric view of time-domain functions as points in a finite-dimensional space, now prevalent in communications engineering, was championed by Shannon [150] in 1949, where he justified the achievability of the rate in (5) using a sphere-hardening reasoning. Since any strictly bandlimited signal has infinite duration, the rate of information of any finite codebook of bandlimited waveforms is, strictly speaking, equal to zero. A rigorous derivation of (5) requires a careful definition of "almost-strict" bandwidth or duration (cf. [156]), and a heavy dose of functional analysis. This was accomplished by A. Wyner in [157] using the fundamental results of [158].

Referring to (5), Shannon [1] asserts that

> to approximate this limiting rate of transmission the transmitted signals must approximate in statistical properties a white noise.

A proof that capacity-achieving codes must have empirical distributions that maximize input–output mutual information was given in [159].

The generalization of (5) to dispersive/nonwhite Gaussian channels is given by the "water-filling" formula obtained by Shannon in [150] using the differential entropy of a Gaussian stationary process found in [1]. Rigorous justifications [160]–[162] of the water-filling formula for dispersive/nonwhite Gaussian channels usually appeal to the Toeplitz distribution theorem [163], [164], although it can be circumvented, as shown in [165].

Shannon [1] also studies white Gaussian channels subject to amplitude constraints, rather than power constraints. Not only does he give bounds but he notices that for low signal-to-noise ratios the capacity is essentially given by (5). A closed-form expression is not known to exist but algorithms for the computation of amplitude-constrained capacity were given in [166], and in [167] for the practically important quadrature-modulation channel. Gaussian channel capacity has been found for models that incorporate structural constraints at the receiver

[22] Alluding to [1], Shannon says in [139]: "I was confident I was correct, not only in an intuitive way but in a rigorous way. I knew exactly what I was doing, and it all came out exactly right."

[23] In contrast to the 1920's papers by Nyquist and Küpfmüller (Section I), Shannon's crisp statement [1] and proof [150] of the sampling theorem were instrumental in popularizing this result in engineering. Thus there is indeed some justification for the term "Shannon sampling theorem" [151], [152].

[24] Four months after the publication of [1], M. Golay [153] refers to (5) as "the now classical expression for the information reception capacity of a channel."

(e.g., quantized observations) and at the transmitter (e.g., specific modulation formats). Those results have underlied much of the development of modem technology since the early 1960's [168], [169].

The capacity of additive non-Gaussian channels is also considered in [1]. For a fixed noise power, Gaussian noise is shown to be least favorable.[25] Moreover, Shannon gives an upper bound on capacity which depends on the "non-Gaussianness" of the noise distribution through its entropy-power, i.e., the variance of a Gaussian distribution with identical entropy. An equivalent bound can be given in terms of the divergence between the actual noise distribution and a Gaussian distribution with the same variance [173]. Shannon [1] states the deceptively simple "entropy-power inequality" proved in [174]: the entropy-power of the sum of two independent random variables is lower-bounded by the sum of the individual entropy-powers. The behavior of the capacity of additive non-Gaussian noise power-constrained channels was investigated by V. Prelov for various asymptotic scenarios [175]–[178].

The development of the results on the capacity of additive Gaussian noise channels where the transmitted signals are subject to fading is surveyed in [179].

D. The Channel Coding Theorem

The major result left unproven in [1] was the converse channel coding theorem. This follows directly from Fano's inequality [180]—a fundamental result in information theory which gives a lower bound on the error probability of equiprobable hypotheses tests.

Actually, Fano's inequality leads to a more general result: if the messages to be transmitted through the channel are not equiprobable but are generated by a source with entropy $H > C$, then reliable communication (i.e., vanishing error probability) is impossible. The achievability counterpart of this *source-channel coding* setting states that if the source entropy is below channel capacity, then there exist codes that achieve vanishing error probability. This follows easily by separating the source- and channel-coding operations. At the encoder, the source is compressed to a rate equal to its entropy and fed to a channel encoder which assumes equiprobable messages; at the receiver front-end a source-independent channel decoder selects a codeword, which is then fed to a source decompressor independent of the channel statistics. Because of this structure, the source-channel coding theorem is also known as the separation theorem.[26]

The validity of the strong converse in Shannon's Theorem 12 was established by J. Wolfowitz in 1957 [182] for binary memoryless channels.

A. Feinstein [183] gave the first step-by-step proof of the achievability part of the channel coding theorem[27] in 1954. Along with a deterministic greedy method to choose the codebook, Feinstein used a suboptimal decoder that selects the message whose information density with the received signal exceeds a given threshold; an error is declared if no such message exists, and if more than one such message exists, then the message with lowest index is chosen. A combinatorial variation of Feinstein's proof was proposed by J. Wolfowitz [182] for discrete memoryless channels. This marked the first use of empirical distributions (types) in Shannon theory. Fully developed by I. Csiszár and J. Körner in [185], the method of types (surveyed in [186]), has been influential in the evolution of the Shannon theory of discrete memoryless channels and sources.

Arguably the most natural proof of the direct coding theorem follows by formalizing the intuitive argument put forth by Shannon [1] that evaluates the average probability of the ensemble of all codes. In this case, the decoder is slightly different from Feinstein's: if more than one message satisfies the threshold test, then an error is declared.[28]

Other proofs of the direct coding theorem were given by Shannon in [188] and R. Gallager in [189] by evaluating the average performance of random encoding achieved by the maximum-likelihood decoder. Popularized in [161], Gallager's simple bounding method has found widespread use.

Aside from Shannon's treatment of the nonwhite Gaussian channel [150], the first works that dealt with the capacity of channels with memory did not appear until the late 1950's: [190], [188], [191], [182], and [192]. The capacity of most channels with memory is given by the limit of maximal normalized mutual informations. R. Dobrushin [193] showed that such a limiting expression holds for a wide class of channels exhibiting a certain type of ergodic behavior. If the capacity of memoryless channels is not always computable in closed form, the limit present in the formula for channels with memory poses yet another hurdle for explicit computation, which is, nevertheless, surmountable in cases such as the stationary Gaussian channel (cf. Section III-C) and timing channels [194], [195]. A general capacity formula that does not hinge on ergodicity or stationarity restrictions was obtained in [196] by introducing a new way of proving the converse coding theorem. The approach of [196] shows a dual of Feinstein's result: the average error probability of any code is essentially lower-bounded by the cumulative distribution function of the input–output information density evaluated at the code rate.[29] In 1957, Shannon had given a precursor of this lower bound in the largely overlooked [188, Theorem 2].

[25] In fact, the game between signal and noise distribution with fixed variance and input–output mutual information as payoff has a saddle point achieved by Gaussian distributions. Information-theoretic games [170]–[172] are of interest in jamming problems, for example.

[26] Other than for nonstationary sources/channels the separation theorem holds in wide generality [181].

[27] Unlike other approaches, Feinstein's proof leads to a stronger notion of reliability where error probability is measured taking the worst case over all codewords—in lieu of the average. A proof of Feinstein's result in abstract measurable spaces was given in [184].

[28] The so-called "typical-sequences" proof follows this approach except that the decoder contains a superfluous (upper) threshold. Further unnecessary complication results from considering typicality with respect to individual and joint entropies rather than mutual information [187], [168], [101].

[29] Called the *information spectrum* in [46], the distribution function of the information density replaces its average (mutual information) as the fundamental information measure when dealing with nonstationary/nonergodic channels [197].

For channels with cost constraints (e.g., power limitations), capacity is given by the maximal mutual information over all input random variables that satisfy a corresponding average cost constraint. A simpler formula exists for the minimum cost per transmitted bit [198], which in the Gaussian case reduces to the -1.6 dB result quoted in Section III-C.

In a departure from the conventional discrete-time model where there as many output symbols as input symbols, R. Dobrushin [199] and S. Stambler [200] showed coding theorems for channels subject to random deletions and insertions of symbols at instants unknown to the decoder.

E. Constrained Sequences

Under the heading "Capacity of the discrete noiseless channel," [1] poses a nonprobabilistic problem quite different from those discussed in data compression/transmission. Although there is no evidence that Shannon had in mind recording applications when formulating this problem, this discipline has found a wealth of practical applications in magnetic and optical recording [201]. In those applications, certain sequences of bits are forbidden. For example, it may be required that the transmitted sequence contain at least d and at most k 0's amid 1's, or that the sequence satisfy certain spectral properties. Shannon found the fundamental asymptotic limits on the amount of information that can be encoded per symbol, when the allowable sequences are defined by a finite-state machine. In the last fifty years, considerable progress has been achieved in the design of constrained encoders that approach Shannon's bounds (cf. [201]).

F. Zero-Error Channel Capacity

Any text typed by anyone other than an infallible typist will have a nonzero probability of being erroneous, with the probability going to unity as the length increases. An information theorist can make this probability go to zero by handing the typist a redundant text derived from the original using a code that takes into account the statistics of the typist's mistakes.[30]

Imagine now that a certain typist makes mistakes but only by occasionally mistyping a neighboring letter in the keyboard—(t may become r or g, but not u). This seemingly ordinary typist opens a whole new world of information-theoretic problems. We can now encode/decode texts perfectly. For example, the typist could be given texts drawn exclusively from the alphabet $\{\mathtt{b}, \mathtt{i}, \mathtt{t}, \mathtt{s}\}$. The probability of error does not just go to zero asymptotically, it *is* zero. The rate at which information can be encoded infallibly is called the zero-error capacity. This measure no longer depends on the probabilities with which mistakes are made—all the information relevant to finding the zero-error capacity can be summarized in a graph in which pairs of letters mistakable for each other are connected by an edge. Exceedingly difficult and radically different from the nonzero error setting, the zero-error capacity problem was formulated by Shannon [140] in 1956. Once again, Shannon created single-handedly a new research field, this time within combinatorial graph theory.

Most channels of practical interest have zero zero-error capacity. Among channels with positive zero-error capacity, the most difficult channel that has been solved corresponds to a circular typewriter with five keys. Shannon [140] showed that the zero-error capacity of this channel is no less than half of that achievable by an infallible typist ($\log_2 5 = 2.32$ bits per keystroke). In 1979, L. Lovász [202] showed that Shannon's lower bound is in fact the zero-error capacity. In 1997, N. Alon [203] disproved the conjecture in [140] that the zero-error capacity of independent channels operating in parallel is the sum of the zero-error capacities.

A tutorial survey on the results and combinatorial challenges of zero-error information theory is given in [130].

G. Error Exponent

A school of Shannon theorists has pursued a refinement of the channel coding theorem that studies the behavior of the error probability as a function of the blocklength instead of just focusing attention on the channel capacity. Rice [204] and Feinstein [205] observed the exponential decrease of error probability as a function of blocklength in Gaussian and discrete memoryless channels, respectively. The exponent of the minimum achievable probability of error is a function of the rate, which Shannon [206] christened as the *reliability function*. Upper and lower bounds on the reliability function (which coincide for all but low rates) were found in [207] for binary-symmetric channels; in [208] for symmetric discrete memoryless channels; and in [188], [209] for general discrete memoryless channels. The behavior of the reliability function of erasure channels was found in [210].

A new approach to upper-bounding the error probability averaged with respect to the choice of the encoder was found by Gallager in [189]. In addition to the proof of the direct coding theorem (Section III-D), this result led to a general lower bound on the reliability function. Lower bounds on error probability leading to improved upper bounds on the reliability function were obtained in [211] and [212].

Further improvements in bounding the reliability function were shown in [213]–[216]. The power of the method of types is illustrated by the proofs of bounds on the reliability function given in [185] and [217]. To this date, the reliability function of discrete memoryless channels (including the binary-symmetric channel) is not known for all rates.

Other than the application of Chernoff's bound, the information-theoretic work on the reliability function evolved independent of the large-deviations work initiated in the statistical literature in the 1950's. The important role played in the error exponent problem by the divergence measure introduced in statistics by S. Kullback and R. Leibler [218][31] was made evident by R. Blahut in [214]. A more fundamental information measure than either entropy or mutual information, divergence had been surprisingly slow in emerging to its rightful position in information theory, until it was popularized in the texts [185] and [168]. The vacuum left

[30] Alas, information theory has not made any inroads in this particular information technology, and typists continue to be rated by raw words per minute rather than by Shannon capacity.

[31] Earlier, A. Wald had used the nonnegativity of divergence in [219].

by this historical neglect has led to the overrated role played by the differential entropy measure in information theory.

Originating from the analysis of sequential decoders for convolutional codes [220], [221] and related to the reliability function, the *cutoff rate* is a measure of the "noisiness" of the channel, which has received much attention in the development of information and coding theory. Progress in coding theory has refuted the notion (e.g., [222]) that transmitting above cutoff rate requires unwieldy decoding complexity. While there are appealing heuristics on the different behavior of codes below and above cutoff rate (e.g. [168], [169]) the view that cutoff rate is a key measure of the channel transmission capabilities is not supported by the operational characterization that has been discovered so far [223].

Some applications, such as concatenated coding and transmission of sources with residual redundancy, have spurred work on a variant of Shannon's channel coding setup whereby the decoder outputs not one but a fixed-size list of messages. The problem of list decoding was introduced by P. Elias [224]. Capacity and error exponents have been studied in [225]–[228]. Zero-error list decoding (Section III-F) was investigated in [229] and [230].

H. Channels with Feedback

The first (and most widely used) feedback model in information theory was introduced by Shannon in [140]. It considers an encoder that, before sending the ith symbol, knows without error the $(i-1)$th symbol received by the decoder. Shannon [140] shows that even this kind of ideal feedback fails to increase the capacity of the discrete memoryless channel. Because of the lack of channel memory, not only is feedback useless to predict the future behavior of the channel, but it is futile for the encoder to try to compensate for previous channel behavior, as far as channel capacity is concerned. However, feedback does increase the reliability function [231]. Moreover, a number of constructive schemes [232]–[235] made evident that the availability of feedback may simplify the coding and decoding operations. Elias and Shannon [140] showed that the zero-error capacity of discrete memoryless channels could indeed increase with feedback.

Shannon [140] anticipated that feedback would help to increase the capacity of channels with memory, at a time when the capacity of channels with memory had not yet been tackled. The ability of feedback to increase the capacity of channels with memory was studied by a number of authors in the late 1960's: [236]–[238]. In particular, P. Ebert [238] and M. Pinsker[32] independently showed that feedback may increase the capacity of a Gaussian channel by at most a factor of two—a factor that was shown to be the best possible in [239]. The additive upper bound of [240] shows that the increase afforded by feedback cannot exceed half a bit-per-channel use.

I. Channels with Unknown Parameters

In parallel with the setting discussed in Section II-E, the channel description available to the encoder/decoder may be incomplete. Suppose that the actual channel conditional (output given input) distributions are known to belong to an uncertainty class. Depending on whether the number of parameters describing the uncertainty class remains fixed or grows with blocklength we have a *compound channel* or an *arbitrarily varying channel*, respectively. These models are relevant to practical communications systems subject to jamming, time-varying conditions, etc., and have received much attention in the Shannon theory literature [241].

The objective of the encoder/decoder is to guarantee reliable communication regardless of the actual channel in effect. This leads to a minimax problem where the probability of error is maximized over the uncertainty class and minimized over the choice of encoder/decoder.

The compound discrete memoryless channel capacity problem was posed and solved in 1959 by D. Blackwell, L. Breiman, and A. Thomasian [242] and by R. Dobrushin [243]. A year later, Wolfowitz [244] proved the strong converse (for maximal error probability) using the method of [182]. The formula for compound channel capacity is similar to (3) except that for every input distribution, the mutual information is minimized with respect to the channels in the uncertainty class, thus yielding a capacity that is less than or equal to the worst capacity of the channels in the uncertainty set.[33] The capacity of compound channels with memory was investigated in [245] and [246].

Arbitrarily varying channels[34] were introduced in [248]. The capacity was found by Ahlswede and Wolfowitz [249] in the binary output case under the pessimistic maximal error probability criterion, in which the "jammer" is allowed to know the codeword sent by the communicator. A partial generalization of the solution in [249] was obtained by Csiszár and Körner in [250]. However, a full solution of the discrete memoryless case remains elusive. In contrast, if error probability is averaged with respect to the choice of codewords, the arbitrarily varying channel capacity has been progressively solved in a series of papers [251]–[254]. Ahlswede showed in [252] that if the average-error-probability capacity is nonzero, then it does not increase further if the error probability is averaged over the choice of codebooks, i.e., if the "jammer" does not know which code is used by the communicator.

The capacity of the memoryless Gaussian arbitrarily varying channel is known both when the jammer knows the codebook [255] and when it does not [256]. In either case, the effect of the power-constrained jammer is equivalent to an additional source of Gaussian noise, except that the capacity is equal to zero if the jammer knows the codebook and has as much power as the transmitter.

Recent references on the capabilities of list decoding for arbitrarily varying channels can be found in [228] and [257].

If the receiver has incomplete knowledge of the channel or its complexity is constrained, it is of interest to investigate the capacity degradation suffered when the decoder is not maximum-likelihood. If the encoder does know both the channel distribution and the suboptimal decoding rule, then it can partially compensate for the mismatch at the receiver.

[32] Unpublished.

[33] Equality holds in special cases such as when the uncertainty is the crossover probability of a binary-symmetric channel.

[34] The term "arbitrarily varying" was coined in [247].

Recent results on the capacity achievable with mismatched decoding have been obtained in [258]–[262].

Channel uncertainty at the receiver need not result in loss of capacity. For example, known training sequences can be used to probe the channel. Alternatively, *universal decoding* operates in a "blind mode" and attains the same asymptotic performance as a maximum-likelihood rule tuned to the channel law. Universal decoders have been found for various uncertainty models; foremost among them are the maximum empirical mutual information decoder introduced by V. Goppa in [263] and further studied in [185], the Lempel–Ziv-based decoder introduced by J. Ziv in [264] and further studied in [265], the independence decoding rule of I. Csiszár and P. Narayan [266], and the merging decoder introduced by M. Feder and A. Lapidoth in [267].

J. Multiuser Channels

1) Two-Way Channels: Published in 1961, Shannon's last single-authored technical contribution [268] marks the foundation of the discipline of multiuser information theory. Undoubtedly inspired by telephony, [268] is devoted to the *two-way* channel subject to mutual interference between the signals transmitted in opposite directions. A new dimension arises: the tradeoff between the transmission speeds at each terminal, e.g., maximum speed in one direction is feasible when nothing is transmitted in the other direction. Thus the transmission capabilities of the two-way channel are not described by a single number (capacity) as in the conventional one-way channel but by a two-dimensional "capacity region" that specifies the set of achievable rate pairs. Shannon [268] gave a limiting expression for the capacity region of the discrete memoryless two-way channel. Unfortunately, it is not yet known how to explicitly evaluate that expression even in "toy" examples. Of more immediate use were the inner and outer bounds found in [268], and later improved in [269]–[273].

2) Multiaccess Channels: Shannon concludes [268] with

> In another paper we will discuss the case of a channel with two or more terminals having inputs only and one terminal with an output only, a case for which a complete and simple solution of the capacity region has been found.

In the terminology of [268], "inputs" and "output" are to be understood as "inputs to the channel" and "output from the channel." Thus the channel Shannon had in mind was what we now refer to as the *multiple-access channel*: several transmitters sending information to one receiver.

Multiple-access communication dates back to the systems invented in the 1870's by Thomas Edison and Alexander Graham Bell to transmit simultaneous telegraphic messages through a single wire. Time-division and frequency-division multiplexing methods were already well-known at the time of the inception of information theory. Code-division multiple access (CDMA) had also been suggested as one of the possible applications of the spread-spectrum modulation technology that sprung up from World War II. In fact, one of the early proponents of CDMA was Shannon himself [16].

Shannon wrote no further papers on multiple-access channels and it is not known what solution he had found for the multiple-access capacity region. But in a short span of time in the early 1970's several independent contributions [274]–[278] found various characterizations of the capacity region of the two-user discrete memoryless multiple-access channel. Most useful among those is the expression found by H. Liao [276] and R. Ahlswede [278] for the capacity region as the convex hull of a union of pentagons. Shortly after, Wyner [128] and Cover [279] showed (using the suboptimal successive cancellation decoder) that the memoryless Gaussian multiple-access channel admits a very simple capacity region: the pentagon defined by the single-user capacities of the channels with powers equal to the individual powers and to the sum of the powers.[35] The generalization of the capacity region to (non-Gaussian) memoryless multiple-access channels subject to power constraints did not take place until [282] (cf. [198]). The proof of the achievability part of the multiple-access coding theorem is most easily carried out by using the formalization of Shannon's approach discussed in Section III-D.

In spite (or, maybe, because) of the simplicity of these models, they lead to lessons pertinent to practical multiuser communication systems; for example, in many instances, orthogonal multiplexing strategies (such as time- or frequency-division multiplexing) incur a penalty in capacity. Thus letting transmitted signals interfere with each other (in a controlled way) increases capacity provided that the receiver takes into account the multiaccess interference.

Noiseless feedback can increase the capacity of memoryless multiple-access channels as shown in [283] and [284]. However, the capacity region with feedback is not yet known except in special cases such as the Gaussian multiple-access channel [285]. The upper bounds on the capacity of single-user nonwhite Gaussian channels with feedback (Section III-H) have been generalized to multiple-access channels in [286]–[288].

The capacity region of multiple-access channels with memory was given in [289]. The counterpart of the water-filling formula for the dispersive Gaussian channel was found explicitly in [290] for the two-user case and an algorithm for its computation for an arbitrary number of users was given in [291]. The practical issue of transmitter asynchronism was tackled in [292] and [293] at the frame level, and in [294] at the symbol level.

The error exponents of multiple-access channels were investigated in [295]–[297].

When the message sources are correlated it is interesting to consider the problem of joint source-channel multiuser encoding. This has been done in, among others, [298] and [299], where it is shown that the separation principle of single-user source-channel coding does not hold in the multiuser setting.

3) Interference Channels: In contrast to the multiple-access setting in which the receiver is interested in decoding the information sent by all the users, suppose now that we

[35] Multiaccess error-control codes derived from single-user codes have been proposed for the Gaussian multiple-access channel in [280] and for the discrete multiple-access channel in [281].

have as many receivers as transmitters and each receiver is interested in decoding only one of the sources. Think, for example, of telephone channels subject to crosstalk. We could take a multiple-access approach to this problem and use codes that ensure that each receiver can reliably decode the information sent by all the transmitters. However, higher rates are possible if we take advantage of the fact that each receiver requires reliable decoding of only one of the transmitters. In spite of many efforts surveyed in [300] and exemplified by [301]–[304], the capacity region of even the simplest two-user memoryless Gaussian interference channel remains an open problem. One of the practical lessons revealed by the study of the interference channel is the equivalence of powerful interference and no interference [301], [305]: unlike background noise, the known structure of a powerful interfering signal makes it feasible to recover it at the receiver with very high reliability and then subtract it from the received signal.

4) Broadcast Channels: In [306], Cover introduced the dual of the multiaccess channel: one sender that transmits one signal simultaneously to several receivers. If the same information is to be transmitted to each receiver, then the model reduces to a single-user (compound) channel. Otherwise, the problem becomes quite interesting and challenging. For example, in television broadcasting we may want receivers within the coverage area of a station to receive high-definition signals, while more distant (lower signal-to-noise ratio) receivers would be content to receive low-definition television. By superposition of the encoded streams it is possible to trade off the rate of information sent to different types of receivers. Although a general solution for the capacity region of the broadcast channel is not yet known, considerable progress (surveyed in [307]) has been made in exploring the fundamental limits of various classes of memoryless broadcast channels. On the practical side, superposition coding is gaining increasing attention for broadcast applications [308], [309] and other applications that require unequal error protection [310].

For certain nonergodic single-user channels, maximizing average transmission rate makes more sense than the overly conservative coding strategy that guarantees reliability in the worst case channel conditions. Those situations are another promising application of the broadcast channel approach [306], [311].

5) Wiretap Channels: The methods of multiuser information theory have been successfully applied to channels subject to eavesdropping. The basic model was introduced by Wyner [312] and generalized in [313]. The Shannon-theoretic limits of secret sharing by public discussion have been investigated by U. Maurer [314] and by Ahlswede and Csiszár [315].

K. Other Roles of Channel Capacity

Channel capacity has proven to be the key quantity, not only in reliable information transmission, but in a number of other problems.

1) Information Radius: Any parametrized family of distributions $\{P_{Y|\theta}, \theta \in \Theta\}$ can be viewed as a "channel" from the "input" space Θ to the output space where Y is defined. The maximal input–output mutual information is a measure of the dissimilarity ("information radius") of the family of distributions. More precisely, the maximal mutual information is the saddle point of a game whose payoff is the divergence measure and which is maximized over the family of distributions and minimized by a distribution that acts as the center of gravity [161], [185], [316].

2) Minimax Redundancy in Universal Lossless Coding: Consider a game between a source encoder and a source selector whose payoff is the difference between the expected codelength and the source entropy. This is a special case of the game in the previous paragraph. Thus its saddle point is the capacity of the parametrized family of source distributions [317]–[320].

3) Identification: R. Ahlswede and G. Dueck [321] introduced the following seemingly innocuous variation of Shannon's channel-coding setting. Suppose that the recipient of the message is only interested in knowing whether a certain preselected message is the true message.[36] Let us assume that the encoder and decoder ignore which message was preselected by the recipient; for, otherwise, the setting would become a standard hypothesis-testing problem. The situation is similar to the familiar one except that the decoder is free to declare a list of several messages to be simultaneously "true." The recipient simply checks whether the message of interest is in the list or not. Erroneous information is delivered whenever the preselected message is in the list but is not the true message, or if the preselected message is the true message but is not in the list. How many messages can be transmitted while guaranteeing vanishing probability of erroneous information? The surprising answer is that the number of messages grows doubly exponentially with the number of channel uses. Moreover, the second-order exponent is equal to the channel capacity. This result was shown in [321] (achievability) and [46] (converse).

4) System Simulation: Random processes with prescribed distributions can be generated by a deterministic algorithm driven by a source of random bits (independent flips of a fair coin). A key quantity that quantifies the "complexity" of the generated random process is the minimal rate of the source of bits necessary to accomplish the task. The *resolvability* of a system is defined as the minimal randomness required to generate any desired input so that the output distributions are approximated with arbitrary accuracy. Under fairly general conditions, [46] showed that the resolvability of a system is equal to its Shannon capacity.

IV. Lossy Data Compression

Quantization (or analog-to-digital conversion) saw its first practical applications with PCM in the 1930's (Section I) and its evolution is chronicled elsewhere in this issue [322]. The Shannon-theoretic discipline of rate-distortion theory deals with the fundamental limits of lossy data compression in the asymptotic regime of long observations. Constructive methods

[36] For example, the message may be header information in a communication network and the recipient is only interested in determining whether it is the addressee.

and their relationship to the development of information-theoretic data compression limits are reviewed in [323] and [324].

A. The Birth of Rate-Distortion Theory

As we mentioned, in his 1939 letter to V. Bush [19], Shannon had come up with an abstraction of the problem of waveform transmission using a mean-square fidelity criterion. The closing chapter in [1] "Part V: The rate for a continuous source" returns to source coding but now with a basic new ingredient:

> Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a given tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way.

Shannon then considers an arbitrary fidelity (or distortion) criterion and states that the minimum rate at which information can be encoded within a certain tolerance is the minimum mutual information between the source and any other random variable that satisfies the average distortion constraint. Shannon also states the source/channel separation theorem with a fidelity criterion (reproduction with distortion d is possible if $R(d) < C$, and impossible if $R(d) > C$). Shannon gives a quick intuitive argument along the lines used to prove the achievability part of the channel coding theorem and accepts the converse part as a straightforward consequence of the definitions.[37]

It is not until 1959 that Shannon, already at MIT, returns to the fundamental limits of lossy data compression [325], and refers to the function he defined in [1] as the "rate-distortion function $R(d)$." He proves the rate distortion theorem for discrete memoryless sources (using the random coding approach) and evaluates the rate-distortion function in several interesting special cases.

B. Evaluation of Rate-Distortion Functions

In [1], Shannon solves the optimization problem posed by the formula for the rate-distortion function in the case of a Gaussian bandlimited continuous-time random process under the mean-square error criterion. He shows that the rate is equal to the bandwidth times the logarithm of the signal-to-reconstruction-error ratio.[38] Dealing with the discrete-time counterpart, Shannon [325] shows that the Gaussian rate-distortion function is equal to the positive part of one-half of the logarithm of the signal-to-reconstruction-error ratio. This means that every additional bit of encoded information results in an increase of 6 dB in fidelity.

But prior to 1959, the rate-distortion function had attracted the attention of Kolmogorov (and his disciples) who called it the ϵ-entropy [327].[39] The dual to Shannon's water-filling formula for channel capacity (Section III-C) is the "flooding" formula[40] for the rate-distortion function of nonwhite Gaussian processes. It was originally given by Kolmogorov [327], with refined derivations due to Pinsker [330], [331], and B. Tsybakov [332].

When applied to Gaussian sources (with mean-square-error fidelity) and Gaussian channels (with power constraints), the separation theorem leads to particularly interesting conclusions. If the source and the channel have identical bandwidth and their spectra are flat, then the optimum encoding/decoding operations consist of simple instantaneous attenuation/amplification (or single-sideband modulation if frequency translation is required) [333], [334]. If the channel has more bandwidth than the source, then the achievable signal-to-noise ratio (in decibels) is equal to that achievable in the identical-bandwidth case times the ratio of channel-to-source bandwidth. To achieve this limit, nontrivial encoding/decoding is necessary; however, the original analog signal can still be sent uncoded through a portion of channel bandwidth without loss of optimality [335].

A very important rate-distortion function, found by Shannon [325], is that of a binary memoryless source with bit-error-rate fidelity. Ordinarily, the communication engineer specifies a certain tolerable end-to-end bit-error rate. This reliability measure is less stringent than the block-error probability used in the development of channel capacity. According to the separation theorem and Shannon's binary rate-distortion function, if the desired bit-error rate is ϵ, then the maximum transmission rate is equal to channel capacity times the factor

$$(1 + \epsilon \log \epsilon + (1 - \epsilon) \log (1 - \epsilon))^{-1}.$$

Contemporaneously with Shannon [325], Erokhin [336] found the rate-distortion function (for low distortion) of equiprobable discrete sources under bit-error-rate fidelity. Further work on the rate-distortion function in the low-distortion regime was reported by Y. Linkov [337], [338] and by F. Jelinek [339].

Iterative algorithms for the computation of the rate-distortion function of discrete sources have been proposed in [149] and [340].

A number of other sources/fidelity criteria have been shown to admit explicit rate-distortion functions: the Wiener process [341], the Poisson process and other continuous-time Markov processes [342], binary Markov chains with bit-error-rate fidelity [343], and various sources with absolute error criterion [344]–[346]. The rate-distortion function of random fields was studied in [347] and [348].

The Shannon lower bound [325] on the rate-distortion function for difference-distortion measures has played a prominent role in rate-distortion theory (cf. [323]). Other lower bounds can be constructed using Gallager's technique [161]. A formula for the minimum distortion achievable per encoded bit was found in [198].

[37] Such expediency is not far off the mark in contrast to the channel-coding problem with reliability measured by block error probability.

[38] In 1948, Shannon authored (with B. Oliver and J. Pierce) a tutorial [326] on the bandwidth–fidelity tradeoff in PCM.

[39] In addition to Shannon's version with an average distortion constraint, Kolmogorov [327] considered a maximum distortion constraint (cf. [328]). A nonprobabilistic gauge of the size of subsets of metric spaces is also called ϵ-entropy [329].

[40] Usually referred to as "reverse water-filling."

C. Coding Theorems

In addition to proving the lossy source coding theorem for memoryless sources, Shannon [325] sketched an approach to deal with sources with memory. A substantial class of sources was encompassed by R. Dobrushin [193] in 1963, proving a general version of the source–channel separation theorem with distortion. Several less ambitious generalizations (but with more explicit scope) were carried out in the West in the subsequent decade (cf. [323] and [322]).

By the mid-1970's, a shift in focus away from ergodic/stationary sources was spearheaded by L. Davisson, R. Gray, J. Kieffer, D. Neuhoff, D. Ornstein, and their coworkers (see [35, Part V][41]) who studied sliding-block and variable-length encoding methods in addition to the traditional fixed-length block-encoding approach used by Shannon [325]. A 1993 survey of rate-distortion coding theorems can be found in [351].

D. Universal Lossy Data Compression

Spurred by its practical importance and by the existence results [352]–[357] proved in the 1970's, the quest for universal lossy data compression algorithms that attain the rate-distortion function has attracted the efforts of many an information theorist during the 1990's. The notable advances in this topic are exemplified by [358]–[370]. In contrast to universal lossless data compression, we cannot yet state that a fully constructive optimum algorithm has been found. Moreover, while objective distortion measures may serve as useful design guidelines, the ultimate judges of the performance of most lossy data compression algorithms are the eye and the ear.

E. Multiterminal Lossy Data Compression

Consider a digital-to-analog converter operating on the compressed version of the "left" audio source and having access to the uncompressed "right" audio source. How much improvement in compression efficiency can we expect due to the auxiliary information? If the analog-to-digital converter has access to the uncompressed "right" source then the problem is fairly easy to solve using standard rate-distortion theory. Otherwise, we face a counterpart of the problem of decoding with side-information (Section II-F) in the lossy setting,[42] which was solved by Wyner and Ziv [371], [372]. In contrast to the almost-lossless setting of the Slepian–Wolf problem, in this case the absence of side-information at the encoder does incur a loss of efficiency in general. Applications and generalizations of the Wyner–Ziv rate-distortion problem have been considered in [373]–[376], [335], and [377].

The *multiple-descriptions* problem is another multiterminal lossy source-coding problem that has received much attention in the last two decades.[43] The practical relevance of this setting stems from communications systems with diversity: for increased reliability, several channels are set up to connect transmitter and receiver. If those individual channels are prone to outage, we may consider sending the same compressed version of the source through each channel in parallel. However, such a strategy is wasteful because the receiver could get a lower distortion version of the source whenever more than one channel is operational. By appropriate choice of codes it is possible to trade off the rates and distortions achievable for every subset of operational channels. Some of the more salient advances that have been reported in the two-channel case can be found in [378]–[383].

A close relative of multiple descriptions coding is the *successive refinement* problem. Sometimes, the decoder is required to provide a preliminary coarse rendition of the source before proceeding to obtain a finer version after receiving additional encoded data (e.g., a Web browser downloading an image). To that end, it would be wasteful to use codes for which the preliminary encoded data is of no use for the decompression of the higher definition version. In fact, certain sources have the property that no penalty in rate is incurred by requiring the decoding of a preliminary coarse version. The successive refinement problem was introduced by V. Koshélev [384] and by W. Equitz and T. Cover [385], and solved in more generality by B. Rimoldi in [386].

Other multiterminal lossy source-coding problems have been studied by T. Berger and coworkers [387]–[389].

V. Information Theory and Other Fields

To conclude, we offer some pointers to the interactions of Information Theory with various other scientific disciplines.

1) *Probability*

- Central Limit Theorem [390]
- Large Deviations [391]–[393]
- Random Processes and Divergence [394]
- Measure Concentration [395], [396]
- Queueing Theory [194], [397]

2) *Statistical Inference* [398], [399]

- Minimum Description Length [95]
- Hypothesis Testing [168]
- Decentralized Hypothesis Testing [400]
- Parameter Estimation [401]
- Density Estimation [402]
- Minimax Nonparametric Estimation [403], [404]
- Spectral Estimation [405]
- Bayesian Statistics [406]
- Inverse Problems [407]
- Prediction of Discrete Time-Series [111]
- Pattern Recognition and Learning [408]
- Neural Networks [409], [410]
- Speech Recognition [411]

[41] Also [349] and [350] for more recent references.

[42] A recent trend in high-fidelity audio recording is to carry out analog-to-digital conversion at the microphone. The left and right digital-to-analog converters could cooperate to lower the required rate of recorded information even if the analog-to-digital converters had no access to each other's sources. The fundamental limit for this symmetrical setup is unknown.

[43] See [378] for an account of the early history of the results on multiple descriptions.

3) *Computer Science*

- Algorithmic Complexity [412], [413]
- Data Structures: Retrieval and Hashing [59]
- Cryptology [15], [414], [314]
- Computational Complexity [415], [439]
- Quantum Computing [416]
- Random Number Generation [417]–[419]

4) *Mathematics*

- Ergodic Theory and Dynamical Systems [420], [37]
- Combinatorics and Graph Theory [130]
- Inequalities and Convex Analysis [421], [422]
- Harmonic Analysis [324]
- Differential Geometry [423], [424]
- Stochastic Combinatorial Search [425]
- Number Theory [426]
- Systems Control [427], [428]

5) *Physics* [429]

- Thermodynamics [430]
- Physics of Computation [431]
- Statistical Mechanics [432]
- Quantum Information Theory [433]
- Chaos [434]

6) *Economics*

- Portfolio Theory [101], [440]
- Econometrics [428]

7) *Biology*

- Molecular Biology [435]
- Sensory processing [436], [437]

8) *Chemistry* [438]

ACKNOWLEDGMENT

This paper has benefited from comments and suggestions by J. Abrahams, A. Barron, E. Biglieri, R. Blahut, I. Csiszár, D. Forney, A. Lapidoth, N. Merhav, A. McKellips, P. Narayan, A. Orlitsky, V. Prelov, E. Telatar, F. Willems, B. Yu, and K. Zeger.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July–Oct. 1948.
[2] H. Dudley, "The vocoder," *Bell Labs. Rec.*, vol. 18, pp. 122–126, Dec. 1939.
[3] H. Nyquist, "Certain factors affecting telegraph speed," *Bell Syst. Tech. J.*, vol. 3, pp. 324–352, Apr. 1924.
[4] K. Küpfmüller, "Uber Einschwingvorgange in Wellenfiltern," *Elek. Nachrichtentech.*, vol. 1, pp. 141–152, Nov. 1924.
[5] H. Nyquist, "Certain topics in telegraph transmission theory," *AIEE Trans.*, vol. 47, pp. 617–644, Apr. 1928.
[6] V. A. Kotel'nikov, "On the transmission capacity of "ether" and wire in electrocommunications," *Izd. Red. Upr. Svyazi RKKA* (Moscow, USSR) (material for the first all-union conference on questions of communications), vol. 44, 1933.
[7] E. T. Whittaker, "On the functions which are represented by the expansion of interpolating theory," in *Proc. Roy. Soc. Edinburgh*, vol. 35, pp. 181–194, 1915.
[8] J. M. Whittaker, "The Fourier theory of the cardinal functions," *Proc. Math. Soc. Edinburgh*, vol. 1, pp. 169–176, 1929.
[9] D. Gabor, "Theory of communication," *J. Inst. Elec. Eng.*, vol. 93, pp. 429–457, Sept. 1946.
[10] R. V. L. Hartley, "Transmission of information," *Bell Syst. Tech. J.*, vol. 7, pp. 535–563, July 1928.
[11] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York: Wiley, 1949.
[12] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 23–24, pp. 282–332 and 46–156, July 1944 and Jan. 1945.
[13] F. Pratt, *Secret and Urgent*. Blue Ribbon Books, 1939.
[14] C. E. Shannon, "A mathematical theory of cryptography," Tech. Rep. MM 45-110-02, Bell Labs. Tech. Memo., Sept. 1, 1945.
[15] ———, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, pp. 656–715, Oct. 1949.
[16] R. Price, "A conversation with Claude Shannon," *IEEE Commun. Mag.*, vol. 22, pp. 123–126, May 1984.
[17] J. R. Pierce, "The early days of information theory," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 3–8, Jan. 1973.
[18] F. W. Hagemeyer, "Die Entstehung von Informationskonzepten in der Nachrichtentechnik: eine Fallstudie zur Theoriebildung in der Technik in Industrie- und Kriegsforschung," Ph.D. dissertation, Free University of Berlin, Berlin, Germany, 1979.
[19] C. E. Shannon, "Letter to Vannevar Bush Feb. 16, 1939," in *Claude Elwood Shannon Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, pp. 455–456, 1993.
[20] ———, "A symbolic analysis of relay and switching circuits," M.S. thesis, MIT, Cambridge, MA, 1938.
[21] ———, "An algebra for theoretical genetics," Ph.D. dissertation, MIT, Cambridge, MA, Apr. 15, 1940.
[22] A. G. Clavier, "Evaluation of transmission efficiency according to Hartley's expression of information content," *Elec. Commun.: ITT Tech. J.*, vol. 25, pp. 414–420, June 1948.
[23] C. W. Earp, "Relationship between rate of transmission of information, frequency bandwidth, and signal-to-noise ratio," *Elec. Commun.: ITT Tech. J.*, vol. 25, pp. 178–195, June 1948.
[24] S. Goldman, "Some fundamental considerations concerning noise reduction and range in radar and communication," *Proc. Inst. Elec. Eng.*, vol. 36, pp. 584–594, 1948,
[25] J. Laplume, "Sur le nombre de signaux discernables en présence de bruit erratique dans un système de transmission à bande passante limitée," *Comp. Rend. Acad. Sci. Paris*, pp. 1348–, 1948.
[26] W. G. Tuller, "Theoretical limitations on the rate of transmission of information," Ph.D. dissertation, MIT, Cambridge, MA, June 1948, published in *Proc. IRE*, pp. 468–478, May 1949.
[27] N. Wiener, *Cybernetics, Chaper III: Time Series, Information and Communication*. New York: Wiley, 1948.
[28] R. A. Fisher, "Probability, likelihood and quantity of information in the logic of uncertain inference," *Proc. Roy. Soc. London, A*, vol. 146, pp. 1–8, 1934.
[29] L. Boltzmann, "Beziehung zwischen dem Zweiten Hauptsatze der Mechanischen Waermertheorie und der Wahrscheilichkeitsrechnung Respektive den Saetzen uber das Waermegleichgwicht," *Wien. Ber.*, vol. 76, pp. 373–435, 1877.
[30] N. Chomsky, "Three models for the description of language," *IEEE Trans. Inform. Theory*, vol. IT-2, pp. 113–124, Sept. 1956.
[31] A. I. Khinchin, "The entropy concept in probability theory," *Usp. Mat. Nauk.*, vol. 8, pp. 3–20, 1953, English translation in *Mathematical Foundations of Information Theory*. New York: Dover, 1957.
[32] B. McMillan, "The basic theorems of information theory," *Ann. Math. Statist.*, vol. 24, pp. 196–219, June 1953.
[33] A. N. Kolmogorov, "A new metric invariant of transitive dynamical systems and automorphism in Lebesgue spaces," *Dokl. Akad. Nauk. SSSR*, vol. 119, pp. 861–864, 1958.
[34] D. S. Ornstein, "Bernoulli shifts with the same entropy are isomorphic," *Adv. Math.*, vol. 4, pp. 337–352, 1970.
[35] R. M. Gray and L. D. Davisson, Eds., *Ergodic and Information Theory*. Stroudsbourg, PA: Dowden, Hutchinson & Ross, 1977.
[36] I. Csiszár, "Information theory and ergodic theory," *Probl. Contr. Inform. Theory*, vol. 16, pp. 3–27, 1987.
[37] P. Shields, "The interactions between ergodic theory and information theory," this issue, pp. 2079–2093.
[38] L. Breiman, "The individual ergodic theorems of information theory," *Ann. Math. Statist.*, vol. 28, pp. 809–811, 1957.
[39] P. Algoet and T. M. Cover, "A sandwich proof of the Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 16, pp. 899–909, 1988.
[40] S. C. Moy, "Generalizations of the Shannon–McMillan theorem," *Pacific J. Math.*, vol. 11, pp. 705–714, 1961.
[41] K. Marton, "Information and information stability of ergodic sources," *Probl. Inform. Transm.*, vol. 8, pp. 179–183, 1972.

[42] J. C. Kieffer, "A simple proof of the Moy–Perez generalization of the Shannon–McMillan theorem," *Pacific J. Math.*, vol. 351, pp. 203–206, 1974.

[43] D. S. Ornstein and B. Weiss, "The Shannon–McMillan–Breiman theorem for amenable groups," *Israel J. Math.*, vol. 39, pp. 53–60, 1983.

[44] A. R. Barron, "The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 13, pp. 1292–1303, 1985.

[45] S. Orey, "On the Shannon-Perez-Moy theorem," *Contemp. Math.*, vol. 41, pp. 319–327, 1985.

[46] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.

[47] S. Verdú and T. S. Han, "The role of the asymptotic equipartition property in noiseless source coding," *IEEE Trans. Inform. Theory*, vol. 43, pp. 847–857, May 1997.

[48] D. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, vol. 40, pp. 1098–1101, Sept. 1952.

[49] I. S. Reed, "1982 Claude Shannon lecture: Application of transforms to coding and related topics," *IEEE Inform. Theory Newslett.*, pp. 4–7, Dec. 1982.

[50] R. Hunter and A. Robinson, "International digital facsimile coding standards," *Proc. Inst. Elec. Radio Eng.*, vol. 68, pp. 854–867, July 1980.

[51] K. Challapali, X. Lebegue, J. S. Lim, W. Paik, R. Girons, E. Petajan, V. Sathe, P. Snopko, and J. Zdepski, "The grand alliance system for US HDTV," *Proc. IEEE* , vol. 83, pp. 158–174, Feb. 1995.

[52] R. G. Gallager, "Variations on a theme by Huffman," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 668–674, Nov. 1978.

[53] R. M. Capocelli and A. DeSantis, "Variations on a theme by Gallager," in *Image and Text Compression*, J. A. Storer, Ed. Boston, NA: Kluwer, pp. 181–213, 1992.

[54] L. Kraft, "A device for quantizing, grouping and coding amplitude modulated pulses," M.S. Thesis, Dept. Elec. Eng., MIT, Cambridge, MA, 1949.

[55] I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, 2d ed. New York: McGraw-Hill, 1966.

[56] B. McMillan, "Two inequalities implied by unique decipherability," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 115–116, Dec. 1956.

[57] J. Karush, "A simple proof of an inequality of McMillan," *IEEE Trans. Inform. Theory*, vol. IT-7, pp. 118, Apr. 1961.

[58] J. Abrahams, "Code and parse trees for lossless source encoding," in *Proc. Compression and Complexity of Sequences 1997*, B. Carpentieri, A. De Santis, U. Vaccaro, and J. Storer, Eds. Los Alamitos, CA: IEEE Comp. Soc., 1998, pp. 145–171.

[59] R. E. Krichevsky, *Universal Compression and Retrieval*. Dordrecht, The Netherlands: Kluwer, 1994.

[60] J. Rissanen, "Generalized Kraft inequality and arithmetic coding," *IBM J. Res. Devel.*, vol. 20, pp. 198–203, 1976.

[61] R. Pasco, "Source coding algorithms for fast data compression," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1976.

[62] J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM J. Res. Develop.*, vol. 23, pp. 149–162, Mar. 1979.

[63] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Commun. Assoc. Comp. Mach.*, vol. 30, pp. 520–540, June 1987.

[64] N. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1963.

[65] F. Jelinek, *Probabilistic Information Theory*. New York: McGraw-Hill, 1968.

[66] P. Elias, personal communication, Apr. 22, 1998.

[67] E. Gilbert and E. Moore, "Variable-length binary encodings," *Bell Syst. Tech. J.*, vol. 38, pp. 933–967, 1959.

[68] T. M. Cover, "Enumerative source coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 73–77, 1973.

[69] R. Aravind, G. Cash, D. Duttweiler, H. Hang, B. Haskell, and A. Puri, "Image and video coding standards," *ATT Tech. J.*, vol. 72, pp. 67–89, Jan.–Feb. 1993.

[70] B. P. Tunstall, "Synthesis of noiseless compression codes," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, Sept. 1967.

[71] G. L. Khodak, "The estimation of redundancy for coding the messages generated by a Bernoulli source," *Probl. Inform. Transm.*, vol. 8, pp. 28–32, 1972.

[72] F. Jelinek and K. Schneider, "On variable-length to block coding," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 765–774, Nov. 1972.

[73] T. J. Tjalkens and F. M. J. Willems, "Variable to fixed-length codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 246–257, Mar. 1987.

[74] S. Savari and R. G. Gallager, "Generalized Tunstall codes for sources with memory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 658–668, Mar. 1997.

[75] J. Ziv, "Variable-to-fixed length codes are better than fixed-to-variable length codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. 36, pp. 861–863, July 1990.

[76] N. Merhav and D. L. Neuhoff, "Variable-to-fixed length codes provide better large deviations performance than fixed-to-variable length codes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 135–140, Jan. 1992.

[77] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–206, Mar. 1981.

[78] S. W. Golomb, "Run length encodings," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 399–401, July 1966.

[79] C. E. Shannon, "Efficient coding of a binary source with one very infrequent symbol," Bell Labs. Memo. Jan. 29, 1954, in *Claude Elwood Shannon Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993, pp. 455–456.

[80] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transm.*, vol. 1, pp. 1–7, 1965.

[81] E. Gilbert, "Codes based on inaccurate source probabilities," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 304–314, May 1971.

[82] F. Fabris, A. Sgarro, and R. Pauletti, "Tunstall adaptive coding and miscoding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2167–2180, Nov. 1996.

[83] B. M. Fitingof, "Optimal encoding with unknown and variable message statistics," *Probl. Inform. Transm.*, vol. 2, pp. 3–11, 1966.

[84] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[85] T. J. Tjalkens and F. M. J. Willems, "A universal variable-to-fixed length source code based on Lawrence's algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 247–253, Mar. 1992.

[86] N. Faller, "An adaptive system for data compression," in *7th Asilomar Conf. on Circuits, Systems, and Computing*, 1973, pp. 593–597.

[87] D. E. Knuth, "Dynamic Huffman coding," *J. Algorithms*, vol. 1985, pp. 163–180, 1985.

[88] J. S. Vitter, "Dynamic Huffman coding," *ACM Trans. Math. Software*, vol. 15, pp. 158–167, June 1989.

[89] B. Ryabko, "A fast on-line adaptive code," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1400–1404, July 1992.

[90] ———, "Data compression by means of a book stack," *Probl. Inform. Transm.*, vol. 16, pp. 265–269, 1980.

[91] J. Bentley, D. Sleator, R. Tarjan, and V. Wei, "Locally adaptive data compression scheme," *Commun. Assoc. Comp. Mach.*, vol. 29, pp. 320–330, 1986.

[92] P. Elias, "Interval and recency rank source coding: Two on-line adaptive variable-length schemes," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 3–10, Jan. 1987.

[93] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Tech. Rep. 124, Digital Systems Res. Ctr., Palo Alto, CA, May 10, 1994.

[94] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[95] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," this issue, pp. 2743–2760.

[96] A. Lempel and J. Ziv, "On the complexity of an individual sequence," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75–81, Jan. 1976.

[97] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 337–343, May 1977.

[98] ———, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[99] T. A. Welch, "A technique for high performance data compression," *Computer*, vol. 17, pp. 8–19, June 1984.

[100] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 405–412, July 1978.

[101] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[102] A. D. Wyner and J. Ziv, "The sliding-window Lempel-Ziv algorithm is asymptotically optimal," *Proc. IEEE*, vol. 82, pp. 872–877, June 1994.

[103] F. M. J. Willems, "Universal data compression and repetition times," *IEEE Trans. Inform. Theory*, vol. 35, pp. 54–58, Jan. 1989.

[104] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1258, Nov. 1989.

[105] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 78–83, Jan. 1993.

[106] W. Szpankowski, "Asymptotic properties of data compression and suffix trees," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1647–1659, Sept. 1993.

[107] P. Jacquet and W. Szpankowski, "Asymptotic behavior of the Lempel–Ziv parsing scheme and digital search trees," *Theor. Comp. Sci.*, vol. 144, pp. 161–197, June 1995.

[108] A. D. Wyner, J. Ziv, and A. J. Wyner, "On the role of pattern matching in information theory," this issue, pp. 2045–2056.
[109] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656–663, Sept. 1983.
[110] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 653–664, May 1995.
[111] N. Merhav and M. Feder, "Universal prediction," this issue, pp. 2124–2147.
[112] V. I. Levenshtein, "On the redundancy and delay of decodable coding of natural numbers," *Probl. Cybern.*, vol. 20, pp. 173–179, 1968.
[113] P. Elias, "Universal codeword sets and representation of the integers," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 194–203, Mar. 1975.
[114] Q. F. Stout, "Improved prefix encodings of the natural numbers," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 607–609, Sept. 1980.
[115] R. Ahlswede, T. S. Han, and K. Kobayashi, "Universal coding of integers and unbounded search trees," *IEEE Trans. Inform. Theory*, vol. 43, pp. 669–682, Mar. 1997.
[116] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 47–51, Jan. 1951.
[117] T. M. Cover and R. C. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 413–421, July 1978.
[118] L. Levitin and Z. Reingold, "Entropy of natural languages—Theory and practice," *Chaos, Solitons and Fractals*, vol. 4, pp. 709–743, May 1994.
[119] P. Grassberger, "Estimating the information content of symbol sequences and efficient codes," *IEEE Trans. Inform. Theory*, vol. 35, pp. 669–675, May 1989.
[120] P. Hall and S. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, Mar. 1993.
[121] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1319–1327, May 1998.
[122] A. N. Kolmogorov, "Logical basis for information theory and probability theory," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 662–664, 1968.
[123] G. J. Chaitin, "On the length of programs for computing binary sequences," *J. Assoc. Comput. Mach.*, vol. 13, pp. 547–569, 1966.
[124] R. J. Solomonoff, "A formal theory of inductive inference," *Inform. Contr.*, vol. 7, pp. 1–22, 224–254, 1964.
[125] T. M. Cover, P. Gacs, and R. M. Gray, "Kolmogorov's contributions to information theory and algorithmic complexity," *Ann. Probab.*, vol. 17, pp. 840–865, July 1989.
[126] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, 1973.
[127] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 226–228, Mar. 1975.
[128] A. D. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 2–9, Jan. 1974.
[129] S. Shamai (Shitz) and S. Verdú, "Capacity of channels with side information," *Euro. Trans. Telecommun.*, vol. 6, no. 5, pp. 587–600, Sept.–Oct. 1995.
[130] J. Körner and A. Orlitsky, "Zero-error information theory," this issue, pp. 2207–2229.
[131] H. S. Witsenhausen, "The zero-error side information problem and chromatic numbers," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 592–593, Sept. 1976.
[132] R. Ahlswede, "Coloring hypergraphs: A new approach to multiuser source coding," *J. Comb. Inform. Syst. Sci.*, vol. 4, pp. 76–115, 1979; Part II, vol. 5, pp. 220–268, 1980.
[133] A. C. Yao, "Some complexity questions related to distributive computing," in *Proc. 11th Assoc. Comp. Mach. Symp. Theory of Computing*, 1979, pp. 209–213.
[134] A. Orlitsky, "Worst-case interactive communication. I: Two messages are almost optimal," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1534–1547, Sept. 1990.
[135] ———, "Average case interactive communication," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1534–1547, Sept. 1992.
[136] M. Naor, A. Orlitsky, and P. Shor, "Three results on interactive communication," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1608–1615, Sept. 1993.
[137] R. Ahlswede, N. Cai, and Z. Zhang, "On interactive communication," *IEEE Trans. Inform. Theory*, vol. 43, pp. 22–37, Jan. 1997.
[138] A. R. Calderbank, "The art of signaling: Fifty years of coding theory," this issue, pp. 2561–2595..
[139] A. Liversidge, "Profile of Claude Shannon," *Omni* magazine, Aug. 1987, in *Claude Elwood Shannon Collected Papers*, N. J. A. Sloane and A. D. Wyner, Eds. Piscataway, NJ: IEEE Press, 1993, pp. xix–xxxiii.
[140] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 112–124, Sept. 1956.
[141] J. G. Kreer, "A question of terminology," *IEEE Trans. Inform. Theory*, vol. IT-3, pp. 208, Sept. 1957.
[142] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom, "On the general definition of mutual information," *Repts. Acad. Sci. USSR*, vol. 3, pp. 745–748, 1956.
[143] I. M. Gelfand and A. M. Yaglom, "On the computation of the mutual information between a pair of random functions," *Adv. Math. Sci.*, vol. 12, pp. 3–52, 1957.
[144] I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom, "Mutual information and entropy for continuous distributions," in *Proc. 3rd All Union Math. Congr.*, 1958, vol. 3, pp. 521–531.
[145] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day, 1964, originally published in Russian in 1960.
[146] S. Muroga, "On the capacity of a discrete channel," *J. Phys. Soc. Japan*, vol. 8, pp. 484–494, 1953.
[147] C. E. Shannon, "Some geometrical results in channel capacity," *Verband Deut. Elektrotechnik. Fachber.*, vol. 19, pp. 13–15, 1956.
[148] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, Jan. 1972.
[149] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.
[150] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.
[151] A. J. Jerri, "The Shannon sampling theorem—Its various extensions and applications: A tutorial review," *Proc. IEEE*, vol. 65, pp. 1565–1596, Nov. 1977.
[152] R. J. Marks, *Introduction to Shannon Sampling and Interpolation Theory*. New York: Springer, 1991.
[153] M. J. E. Golay, "Note on the theoretical efficiency of information reception with PPM," *Proc. IRE*, vol. 37, p. 1031, Sept. 1949.
[154] C. E. Shannon, "General treatment of the problem of coding," *IRE Trans. Inform. Theory*, vol. PGIT-1, pp. 102–104, Feb. 1953.
[155] D. Slepian, "Information theory in the fifties" (Invited Paper), *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 145–147, Mar. 1973.
[156] D. Slepian, "On bandwidth," *Proc. IEEE*, vol. 64, pp. 292–300, Mar. 1976.
[157] A. D. Wyner, "The capacity of the band-limited Gaussian channel," *Bell Syst. Tech. J.*, vol. 45, pp. 359–371, Mar. 1966.
[158] H. J. Landau, D. Slepian, and H. O. Pollack, "Prolate spheroidal wave functions, Fourier analysis and uncertainty—III: The dimension of the space of essentially time- and band-limited signals," *Bell Syst. Tech. J.*, vol. 41, pp. 1295–1336, July 1962.
[159] S. Shamai (Shitz) and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. Inform. Theory*, vol. 43, pp. 836–846, May 1997.
[160] J. L. Holsinger, "Digital communication over fixed time-continuous channels with memory with special application to telephone channels," Tech. Rep. 430, MIT Res. Lab. Electron., Cambridge, MA, 1964.
[161] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[162] B. S. Tsybakov, "Capacity of a discrete-time Gaussian channel with a filter," *Probl. Inform. Transm.*, vol. 6, pp. 253–256, July–Sept. 1970.
[163] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. New York: Chelsea, 1958.
[164] R. M. Gray, "On the asymptotic eigenvalue distribution of toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, Nov. 1972.
[165] W. Hirt and J. L. Massey, "Capacity of the discrete-time Gaussian channel with intersymbol interference," *IEEE Trans. Inform. Theory*, vol. 34, pp. 380–388, May 1988.
[166] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inform. Contr.*, vol. 18, pp. 203–219, 1971.
[167] S. Shamai (Shitz) and I. Bar-David, "The capacity of average and peak-power-limited quadrature Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1060–1071, July 1995.
[168] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
[169] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear Gaussian channels," this issue, pp. 2384–2415.
[170] N. M. Blachman, "Communication as a game," in *IRE Wescon Rec.*, 1957, vol. 2, pp. 61–66.
[171] J. M. Borden, D. M. Mason, and R. J. McEliece, "Some information theoretic saddlepoints," *SIAM J. Contr. Optimiz.*, vol. 23, pp. 129–143, Jan. 1985.

[172] S. Shamai (Shitz) and S. Verdú, "Worst-case power constrained noise for binary-input channels," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1494–1511, Sept. 1992.

[173] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Inform. Contr.*, vol. 37, pp. 34–39, 1978.

[174] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inform. Contr.*, vol. 2, pp. 101–112, 1959.

[175] V. Prelov, "Asymptotic behavior of a continuous channel with small additive noise," *Probl. Inform. Transm.*, vol. 4, no. 2, pp. 31–37, 1968.

[176] ———, "Asymptotic behavior of the capacity of a continuous channel with large noise," *Probl. Inform. Transm.*, vol. 6 no. 2, pp. 122–135, 1970.

[177] ———, "Communication channel capacity with almost Gaussian noise," *Theory Probab. Its Applications*, vol. 33, no. 2, pp. 405–422, 1989.

[178] M. Pinsker, V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1877–1888, Nov. 1995.

[179] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," this issue, pp. 2619–2692..

[180] R. M. Fano, "Class notes for course 6.574: Transmission of information," MIT, Cambridge, MA, 1952.

[181] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. Inform. Theory*, vol. 41, pp. 44–54, Jan. 1995.

[182] J. Wolfowitz, "The coding of messages subject to chance errors," *Illinois J. Math.*, vol. 1, pp. 591–606, Dec. 1957.

[183] A. Feinstein, "A new basic theorem of information theory," *IRE Trans. Inform. Theory*, vol. PGIT-4, pp. 2–22, 1954.

[184] T. T. Kadota, "Generalization of Feinstein's fundamental lemma," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 791–792, Nov. 1970.

[185] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[186] I. Csiszár, "The method of types," this issue, pp. 2505–2523.

[187] A. El-Gamal and T. M. Cover, "Multiple user information theory," *Proc. IEEE*, vol. 68, pp. 1466–1483, Dec. 1980.

[188] C. E. Shannon, "Certain results in coding theory for noisy channels," *Inform. Contr.*, vol. 1, pp. 6–25, Sept. 1957.

[189] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.

[190] A. I. Khinchin, "On the fundamental theorems of information theory," *Usp. Mat. Nauk.*, vol. 11, pp. 17–75, 1956, English translation in *Mathematical Foundations of Information Theory*. New York: Dover, 1957.

[191] I. P. Tsaregradsky, "On the capacity of a stationary channel with finite memory," *Theory Probab. Its Applications*, vol. 3, pp. 84–96, 1958.

[192] A. Feinstein, "On the coding theorem and its converse for finite-memory channels," *Inform. Contr.*, vol. 2, pp. 25–44, 1959.

[193] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Amer. Math. Soc. Transl.*, vol. 33, pp. 323–438, 1963.

[194] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inform. Theory*, vol. 42, pp. 4–18, Jan. 1996.

[195] A. S. Bedekar and M. Azizoğlu, "The information-theoretic capacity of discrete-time queues," *IEEE Trans. Inform. Theory*, vol. 44, pp. 446–461, Mar. 1998.

[196] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.

[197] T. S. Han, *Information Spectrum Methods in Information Theory*. Tokyo, Japan: Baifukan, 1998, in Japanese.

[198] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.

[199] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Probl. Inform. Transm.*, vol. 3, pp. 11–26, 1967.

[200] S. Z. Stambler, "Memoryless channels with synchronization errors—General case," *Probl. Inform. Transm.*, vol. 6, no. 3, pp. 43–49, 1970.

[201] K. A. S. Immink, P. Siegel, and J. K. Wolf, "Codes for digital recorders," this issue, pp. 2260–2299.

[202] L. Lovász, "On the Shannon capacity of a graph," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 1–7, Jan. 1979.

[203] N. Alon, "The Shannon capacity of a union," *Combinatorica*, to be published.

[204] S. O. Rice, "Communication in the presence of noise—Probability of error for two encoding schemes," *Bell Syst. Tech. J.*, pp. 60–93, Jan. 1950.

[205] A. Feinstein, "Error bounds in noisy channels with memory," *IRE Trans. Inform. Theory*, vol. PGIT-1, pp. 13–14, Sept. 1955.

[206] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, May 1959.

[207] P. Elias, "Coding for noisy channels," in *IRE Conv. Rec.*, Mar. 1955, vol. 4, pp. 37–46.

[208] R. L. Dobrushin, "Asymptotic bounds on error probability for transmission over DMC with symmetric transition probabilities," *Theory Probab. Applicat.*, vol. 7, pp. 283–311, 1962.

[209] R. M. Fano, *Transmission of Information*. New York: Wiley, 1961.

[210] R. L. Dobrushin, "Optimal binary codes for low rates of information transmission," *Theory of Probab. Applicat.*, vol. 7, pp. 208–213, 1962.

[211] C. E. Shannon, R. G. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels, I," *Inform. Contr.*, vol. 10, pp. 65–103, 1967.

[212] ———, "Lower bounds to error probability for coding on discrete memoryless channels, II," *Inform. Contr.*, vol. 10, pp. 522–552, 1967.

[213] E. A. Haroutunian, "Bounds on the exponent of the error probability for a semicontinuous memoryless channel," *Probl. Inform. Transm.*, vol. 4, pp. 37–48, 1968.

[214] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 405–417, July 1974.

[215] R. J. McEliece and J. K. Omura, "An improved upper bound on the block coding error exponent for binary-input discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 611–613, Sept. 1977.

[216] S. Litsyn, "New upper bounds on error exponents," Tech. Rep. EE-S-98-01, Tel. Aviv Univ., Ramat-Aviv., Israel, 1998.

[217] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–11, Jan. 1981.

[218] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[219] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20, pp. 595–601, 1949.

[220] I. M. Jacobs and E. R. Berlekamp, "A lower bound to the distribution of computation for sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 167–174, Apr. 1967.

[221] E. Arikan, "An upper bound to the cutoff rate of sequential decoding," *IEEE Trans. Inform. Theory*, vol. 34, pp. 55–63, Jan. 1988.

[222] J. L. Massey, "Coding and modulation in digital communications," in *1974 Zurich Sem. Digital Communications*, 1974, pp. E2(1)–E2(4).

[223] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inform. Theory*, vol. 41, pp. 26–34, Jan. 1995.

[224] P. Elias, "List decoding for noisy channels," in *IRE WESCON Conv. Rec.*, 1957, vol. 2, pp. 94–104.

[225] G. D. Forney, "Exponential error bounds for erasure list, and decision feedback schemes," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 206–220, Mar. 1968.

[226] R. Ahlswede, "Channel capacities for list codes," *J. Appl. Probab.*, vol. 10, pp. 824–836, 1973.

[227] P. Elias, "Error correcting codes for list decoding," *IEEE Trans. Inform. Theory*, vol. 37, pp. 5–12, Jan. 1991.

[228] V. Blinovsky, "List decoding," *Discr. Math.*, vol. 106, pp. 45–51, Sept. 1992.

[229] P. Elias, "Zero error capacity under list decoding," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1070–1074, Sept. 1988.

[230] İ. E. Telatar, "Zero error list capacities of discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1977–1982, Nov. 1997.

[231] K. S. Zigangirov, "Upper bounds on the error probability for channels with feedback," *Probl. Inform. Transm.*, vol. 6, no. 2, pp. 87–92, 1970.

[232] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 183–189, Apr. 1966.

[233] M. S. Pinsker, "Error probability for block transmission on a Gaussian memoryless channel with feedback," *Probl. Inform. Transm.*, vol. 4, no. 4, pp. 3–19, 1968.

[234] M. S. Pinsker and A. Dyachkov, "Optimal linear transmission through a memoryless Gaussian channel with full feedback," *Probl. Inform. Transm.*, vol. 7, pp. 123–129, 1971.

[235] R. S. Liptser, "Optimal encoding and decoding for transmission of Gaussian Markov signal over a noiseless feedback channel," *Probl. Inform. Transm.*, vol. 10, no. 4, pp. 3–15, 1974.

[236] I. A. Ovseevich, "Capacity of a random channel with feedback and the matching of a source to such a channel," *Probl. Inform. Transm.*, vol. 4, pp. 52–59, 1968.

[237] M. S. Pinsker and R. L. Dobrushin, "Memory increases capacity," *Probl. Inform. Transm.*, vol. 5, pp. 94–95, Jan. 1969.

[238] P. M. Ebert, "The capacity of the Gaussian channel with feedback," *Bell Syst. Tech. J.*, vol. 49, pp. 1705–1712, Oct. 1970.

[239] S. Ihara, *Information Theory for Continuous Systems*. Singapore: World Scientific, 1993.

[240] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inform. Theory*, vol. 35, pp. 37–43, Jan. 1989.

[241] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," this issue, pp. 2148–2176.
[242] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Statist.*, vol. 30, pp. 1229–1241, Dec. 1959.
[243] R. L. Dobrushin, "Optimum information transmission through a channel with unknown parameters," *Radio Eng. Electron.*, vol. 4, pp. 1–8, 1959.
[244] J. Wolfowitz, "Simultaneous channels," *Arch. Rational Mech. Anal.*, vol. 4, pp. 371–386, 1960.
[245] W. L. Root and P. P. Varaiya, "Capacity of classes of Gaussian channels," *SIAM J. Appl. Math*, vol. 16, pp. 1350–1393, Nov. 1968.
[246] A. Lapidoth and İ. E. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Inform. Theory*, vol. 44, pp. 973–983, May 1998.
[247] J. Kiefer and J. Wolfowitz, "Channels with arbitrarily varying channel probability functions," *Inform. Contr.*, vol. 5, pp. 44–54, 1962.
[248] D. Blackwell, L. Breiman, and A. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.
[249] R. Ahlswede and J. Wolfowitz, "The capacity of a channel with arbitrarily varying cpf's and binary output alphabet," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 15, pp. 186–194, 1970.
[250] I. Csiszár and J. Körner, "On the capacity of the arbitrarily varying channel for maximum probability of error," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 57, pp. 87–101, 1981.
[251] R. L. Dobrushin and S. Z. Stambler, "Coding theorems for classes of arbitrarily varying discrete memoryless channels," *Probl. Inform. Transm.*, vol. 11, no. 2, pp. 3–22, 1975.
[252] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, vol. 44, pp. 159–175, 1968.
[253] T. Ericson, "Exponential error bounds for random codes in the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 31, pp. 42–48, Jan. 1985.
[254] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Capacity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, Mar. 1988.
[255] ———, "Capacity of the Gaussian arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 34, pp. 18–26, Jan. 1991.
[256] B. Hughes and P. Narayan, "Gaussian arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 267–284, Mar. 1987.
[257] B. L. Hughes, "The smallest list for the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. 43, pp. 803–815, May 1997.
[258] V. B. Balakirsky, "Coding theorem for discrete memoryless channels with given decision rules," in *Proc. 1st French–Soviet Work. Algebraic Coding*, July 1991, pp. 142–150.
[259] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.
[260] I. Csiszár and P. Narayan, "Channel decoding for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, pp. 35–43, Jan. 1995.
[261] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1439–1452, Sept. 1996.
[262] ———, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1528, Sept. 1996.
[263] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.
[264] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453–460, July 1985.
[265] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1746–1755, Sept. 1998.
[266] I. Csiszár and P. Narayan, "Capacity and decoding rules for classes of arbitrarily varying channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 752–769, July 1989.
[267] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1726–1745, Sept. 1998.
[268] C. E. Shannon, "Two-way communication channels," in *Proc. 4th. Berkeley Symp. Mathematical Statistics and Probability* (June 20–July 30, 1960), J. Neyman, Ed. Berkeley, CA: Univ. Calif. Press, 1961, vol. 1, pp. 611–644.
[269] G. Dueck, "The capacity region of the two-way channel can exceed the inner bound," *Inform. Contr.*, vol. 40, pp. 258–266, 1979.
[270] J. P. M. Schalkwijk, "The binary multiplying channel—A coding scheme that operates beyond Shannon's inner bound," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 107–110, Jan. 1982.
[271] ———, "On an extension of an achievable rate region for the binary multiplying channel," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 445–448, May 1983.
[272] Z. Zhang, T. Berger, and J. P. M. Schwalkwijk, "New outer bounds to capacity regions of two-way channels," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 383–386, 1986.
[273] A. P. Hekstra and F. M. J. Willems, "Dependence balance bounds for single-output two-way channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 44–53, 1989.
[274] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Information Theory*, 1971, pp. 103–135.
[275] E. C. van der Meulen, "The discrete memoryless channel with two senders and one receiver," in *Proc 2nd Int. Symp. Information Theory*, 1971, pp. 103–135.
[276] H. Liao, "Multiple access channels," Ph.D. dissertation, Univ. Hawaii, Honolulu, HI, 1972.
[277] H. Liao, "A coding theorem for multiple-access communications," in *1972 Int. Symp. Information Theory*, 1972.
[278] R. Ahlswede, "The capacity region of a channel with two senders and two receivers," *Ann. Probab.*, vol. 2, pp. 805–814, Oct. 1974.
[279] T. M. Cover, "Some advances in broadcast channels," *Adv. Commun. Syst.*, vol. 4, pp. 229–260, 1975.
[280] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 364–375, Mar. 1996.
[281] A. Grant, B. Rimoldi, R. Urbanke, and P. Whiting, "Rate-splitting multiple access for discrete memoryless channels," *IEEE Trans. Inform. Theory*, to be published.
[282] R. G. Gallager, "Power limited channels: Coding, multiaccess, and spread spectrum," in *1988 Conf. Information Science and Systems*, Mar. 1988. Full version published as Rep. LIDS-P-1714, Nov. 1987.
[283] N. T. Gaarder and J. K. Wolf, "The capacity region of a multiple-access discrete memoryless channel can increase with feedback," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 100–102, Jan. 1975.
[284] T. M. Cover and S. K. Leung, "A rate region for multiple access channels with feedback," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 292–298, May 1981.
[285] L. H. Ozarow, "The capacity of the white Gaussian multiple access channel with feedback," *IEEE Trans Information Theory*, vol. IT-30, pp. 623–629, July 1984.
[286] J. A. Thomas, "Feedback can at most double Gaussian multiple access channel capacity," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 711–716, Sept. 1987.
[287] S. Pombra and T. Cover, "Nonwhite Gaussian multiple access channels with feedback," *IEEE Trans. Inform. Theory*, vol. 40, pp. 885–892, May 1994.
[288] E. Ordentlich, "On the factor of two bound for Gaussian multiple-access channels with feedback," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2231–2235, Nov. 1996.
[289] S. Verdú, "Multiple-access channels with memory with and without frame-synchronism," *IEEE Trans. Inform. Theory*, vol. 35, pp. 605–619, May 1989.
[290] R. S. Cheng and S. Verdú, "Gaussian multiple-access channels with intersymbol interference: Capacity region and multiuser water-filling," *IEEE Trans. Inform. Theory*, vol. 39, pp. 773–785, May 1993.
[291] D. N. C. Tse and S. V. Hanly, "Multi-access fading channels: Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inform. Theory*, Nov. 1998, to be published.
[292] G. S. Poltyrev, "Coding in an asynchronous multiple-access channel," *Probl. Inform. Transm.*, vol. 19, pp. 12–21, July–Sept. 1983.
[293] J. Y. N Hui and P. A. Humblet, "The capacity region of the totally asynchronous multiple-access channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 207–216, Mar. 1985.
[294] S. Verdú, "The capacity region of the symbol-asynchronous Gaussian multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 35, pp. 733–751, July 1989.
[295] R. G. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 124–142, Mar. 1985.
[296] J. Pokorny and H. M. Wallmeier, "Random coding bound and codes produced by permutations for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 31, pp. 741–750, Nov. 1985.
[297] Y. S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 376–386, Mar. 1996.
[298] D. Slepian and J. K. Wolf, "A coding theorem for multiple-access channels with correlated sources," *Bell Syst. Tech. J.*, vol. 52, pp. 1037–1076, 1973.
[299] T. M. Cover, A. E. Gamal, and M. Salehi, "Multiple-access channels with arbitrarily correlated sources," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 648–657, Nov. 1980.

[300] E. C. van der Meulen, "Some reflections on the interference channel," in *Communications and Cryptography: Two Sides of One Tapestry*, R. E. Blahut, D. J. Costello, U. Maurer, and T. Mittelholzer, Eds. Boston, MA: Kluwer, 1994.

[301] A. B. Carleial, "A case where interference does not reduce capacity," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 569–570, Sept. 1975.

[302] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 49–60, Jan. 1981.

[303] M. H. M. Costa, "On the Gaussian interference channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 607–615, Sept. 1985.

[304] M. H. M. Costa and A. E. Gamal, "The capacity region of the discrete memoryless interference channel with strong interference," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 710–711, Sept. 1987.

[305] H. Sato, "The capacity of the Gaussian interference channel under strong interference," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 786–788, Nov. 1981.

[306] T. M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 2–4, Jan. 1972.

[307] T. M. Cover, "Comments on broadcast channels," this issue, pp. 2524–2530.

[308] M. Sablatash, "Transmission of all-digital advanced television-state-of-the-art and future directions," *IEEE Trans. Broadcast.*, vol. 40, pp. 102–121, June 1994.

[309] K. Ramchandran, A. Ortega, K. Uz, and M. Vetterli, "Multiresolution broadcast for digital HDTV using joint source channel coding," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 6–23, Jan. 1993.

[310] A. R. Calderbank and N. Seshadri, "Multilevel codes for unequal error protection," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1234–1248, July 1993.

[311] S. Shamai (Shitz), "A broadcast strategy for the Gaussian slowly fading channel," in *Proc. 1997 IEEE Int. Symp. Information Theory* (Ulm, Germany, July 1997), p. 150.

[312] A. D. Wyner, "The wiretap channel," *Bell Syst. Tech. J.*, vol. 54, pp. 1355–1387, 1975.

[313] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 339–348, May 1978.

[314] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inform. Theory*, vol. 39, pp. 733–742, May 1993.

[315] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography—Part I: Secret sharing," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 1121–1132, July 1993.

[316] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, July 1997.

[317] B. Ryabko, "Encoding a source with unknown but ordered probabilities," *Probl. Inform. Transm.*, vol. 15, pp. 134–138, 1979.

[318] R. G. Gallager, "Source coding with side information and universal coding," Tech. Rep. LIDS-P-937, Lab. Inform. Decision Syst., MIT, Cambridge, MA, 1979.

[319] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, 1980.

[320] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 714–722, May 1995.

[321] R. Ahlswede and G. Dueck, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 15–29, Jan. 1989.

[322] R. M. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, this issue, pp. 2325–2383.

[323] T. Berger and J. Gibson, "Lossy source coding," *IEEE Trans. Inform. Theory*, this issue, pp. 2693–2723.

[324] D. L. Donoho, I. Daubechies, R. A. DeVore, and M. Vetterli, "Data compression and harmonic analysis," this issue, pp. 2435–2476.

[325] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*,Mar. 1959, pp. 142–163.

[326] B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PCM," *Proc. IRE*, vol. 36, pp. 1324–1331, 1948.

[327] A. N. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IEEE Trans. Inform. Theory*, vol. IT-2, pp. 102–108, Sept. 1956.

[328] E. Posner and E. Rodemich, "Epsilon entropy and data compression," *Ann. Math. Statist.*, vol. 42, pp. 2079–2125, 1971.

[329] A. N. Kolmogorov and V. M. Tichomirov, "ϵ-entropy and ϵ-capacity of sets in metric spaces," *Usp. Math. Nauk.*, vol. 14, pp. 3–86, 1959.

[330] M. S. Pinsker, "Calculation of the rate of message generation by a stationary random process and the capacity of a stationary channel," *Dokl. Akad. Nauk SSSR*, vol. 111, pp. 753–766, 1956.

[331] ———, "Gaussian sources," *Probl. Inform. Transm.*, vol. 14, pp. 59–100, 1963.

[332] B. S. Tsybakov, "Epsilon-entropy of a vector message," *Probl. Inform. Transm.*, vol. 5, pp. 96–97, 1969.

[333] T. J. Goblick, "Theoretical limitations on the transmission of data from analog sources," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 558–567, Oct. 1965.

[334] J. Ziv, "The behavior of analog communication systems," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 587–594, Sept. 1970.

[335] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 564–579, Mar. 1998.

[336] V. D. Erokhin, "The ϵ-entropy of a discrete random object," *Theory Probab. Appl.*, vol. 3, pp. 103–107, 1958.

[337] Y. N. Linkov, "Epsilon-entropy of random variables when epsilon is small," *Probl. Inform. Transm.*, vol. 1, no. 2, pp. 18–26, 1965.

[338] ———, "Epsilon-entropy of continuous random process with discrete phase space," *Probl. Inform. Transm.*, vol. 7, no. 2, pp. 16–25, 1971.

[339] F. Jelinek, "Evaluation of distortion rate functions for low distortions," *Proc. IEEE*, vol. 55, pp. 2067–2068, 1967.

[340] K. Rose, "Mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1939–1952, Nov. 1994.

[341] T. Berger, "Information rates of Wiener processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 134–139, Mar. 1970.

[342] S. Verdú, "The exponential distribution in information theory," *Probl. Inform. Transm.*, vol. 32, pp. 86–95, Jan.–Mar. 1996.

[343] R. M. Gray, "Information rates of autoregressive processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 412–421, July 1970.

[344] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[345] H. H. Tan and K. Yao, "Evaluation of rate-distortion functions for a class of independent identically distributed sources under an absolute magnitude criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 59–64, Jan. 1975.

[346] K. Yao and H. H. Tan, "Absolute error rate-distortion functions for sources with constrained magnitudes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 499–503, July 1978.

[347] B. Hajek and T. Berger, "A decomposition theorem for binary Markov random fields," *Ann. Probab.*, vol. 15, pp. 1112–1125, 1987.

[348] L. A. Bassalygo and R. L. Dobrushin, "Rate-distortion function of the Gibbs field," *Probl. Inform. Transm.*, vol. 23, no. 1, pp. 3–15, 1987.

[349] M. Effros, P. A. Chou, and R. M. Gray, "Variable-rate source coding theorems for stationary nonergodic sources," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1920–1925, Nov. 1994.

[350] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, pp. 63–86, Jan. 1996.

[351] J. C. Kieffer, "A survey of the theory of source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1473–1490, Sept. 1993.

[352] D. J. Sakrison, "The rate of a class of random processes," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 10–16, Jan. 1970.

[353] J. Ziv, "Coding of sources with unknown statistics—Part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 389–394, May 1972.

[354] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 511–523, 1975.

[355] D. L. Neuhoff and P. C. Shields, "Fixed-rate universal codes for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 360–367, 1978.

[356] J. Ziv, "Distortion-rate theory for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 137–143, Jan. 1980.

[357] R. Garcia-Muñoz and D. L. Neuhoff, "Strong universal source coding subject to a rate-distortion constraint," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 285–295, Mar. 1982.

[358] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Ann. Probab.*, vol. 18, pp. 441–452, 1990.

[359] Y. Steinberg and M. Gutman, "An algorithm for source coding subject to a fidelity criterion based on string matching," *IEEE Trans. Inform. Theory*, vol. 39, pp. 877–886, 1993.

[360] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Trans. Inform. Theory*, vol. 39, pp. 813–820, May 1993.

[361] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universallossy source coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1728–1740, Nov. 1994.

[362] ———, "Fixed-rate universal source coding and rates of convergence for memoryless sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 665–676,

May 1995.
[363] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1109–1138, July 1996.
[364] Z. Zhang and E. Yang, "An on-line universal lossy data compression algorithm via continuous codebook refinement—II: Optimality for phi-mixing source models," *IEEE Trans. Inform. Theory*, vol. 42, pp. 822–836, May 1996.
[365] Z. Zhang and V. K. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement—Part I: Basic results," *IEEE Trans. Inform. Theory*, vol. 42, pp. 803–821, May 1996.
[366] J. C. Kieffer and E. H. Yang, "Sequential codes, lossless compression of individual sequences, and Kolmogorov complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 29–39, Jan. 1996.
[367] E. H. Yang and J. C. Kieffer, "Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 42, pp. 239–245, Jan. 1996.
[368] E. H. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1465–1476, Sept. 1997.
[369] T. Łuczak and W. Szpankowski, "A suboptimal lossy date compression based on approximate pattern matching," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1439–1451, Sept. 1997.
[370] D. L. Neuhoff and P. C. Shields, "Simplistic universal coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 778–781, Mar. 1998.
[371] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.
[372] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, 1978.
[373] T. Berger, K. B. Housewright, J. K. Omura, S. Tung, and J. Wolfowitz, "An upper bound to the rate distortion function for source coding with partial side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 664–666, 1979.
[374] C. Heegard and T. Berger, "Rate-distortion when side information may be absent," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 727–734, Nov. 1985.
[375] T. Berger and R. W. Yeung, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inform. Theory*, vol. 35, pp. 228–236, Jan. 1989.
[376] R. Zamir, "The rate loss in the Wyner-Ziv problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 2073–2084, Nov. 1996.
[377] T. Linder, R. Zamir, and K. Zeger, "On source coding with side information dependent distortion measures," *IEEE Trans. Inform. Theory*, submitted for publication.
[378] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 502–521, July 1987.
[379] L. H. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, pp. 1909–1921, Dec. 1980.
[380] T. M. Cover and A. E. Gamal, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 851–857, Nov. 1982.
[381] R. Ahlswede, "The rate-distortion region for multiple descriptions without excess rate," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 721–726, Nov. 1985.
[382] V. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 821–834, May 1993.
[383] Z. Zhang and T. Berger, "Multiple description source coding with no excess marginal rate," *IEEE Trans. Inform. Theory*, vol. 41, pp. 349–357, Mar. 1995.
[384] V. Koshélev, "Estimation of mean error for a discrete successive approximation scheme," *Probl. Inform. Transm.*, vol. 17, pp. 20–33, July–Sept. 1981.
[385] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, pp. 269–274, Mar. 1991.
[386] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol. 40, pp. 253–259, Jan. 1994.
[387] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources and partially separated encoders," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 828–840, Nov. 1982.
[388] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem," *IEEE Trans. Inform. Theory*, vol. 42, pp. 887–903, May 1996.
[389] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1549–1561, Sept. 1997.
[390] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, pp. 336–342, 1986.
[391] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, Feb. 1975.
[392] ———, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, pp. 768–793, Aug. 1984.
[393] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones and Bartlett, 1993.
[394] S. Kullback, J. C. Keegel, and J. H. Kullback, *Topics in Statistical Information Theory*. Berlin, Germay: Springer, 1987.
[395] K. Marton, "Bounding $\bar{d}$-distance by information divergence: A method to prove concentration inequalities," *Ann. Probab.*, vol. 24, pp. 857–866, 1996.
[396] A. Dembo, "Information inequalities and concentration of measure," *Ann. Probab.*, vol. 25, pp. 927–939, 1997.
[397] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummated union," this issue, pp. 2416–2434.
[398] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.
[399] I. Vajda, *Theory of Statistical Inference and Information*. Dordrecht, The Netherlands: Kluwer, 1989.
[400] S. I. Amari and T. S. Han, "Statistical inference under multiterminal data compression," this issue, pp. 2300–2324.
[401] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 386–391, May 1969.
[402] L. Devroye, *A Course in Density Estimation*. Boston, MA: Birkhauser, 1987.
[403] L. Bassalygo, S. Gelfand, G. Golubev, R. Dobrushin, V. Prelov, Y. Sinai, R. Khasminskii, and A. Yaglom, "Review of scientific achievements of M. S. Pinsker," *Probl. Inform. Transm.*, vol. 32, pp. 3–14, 1996.
[404] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, to be published.
[405] B. S. Choi and T. M. Cover, "An information-theoretic proof of Burg's maximum entropy spectrum," *Proc. IEEE*, vol. 72, pp. 1094–1095, 1984.
[406] A. R. Barron, "Information-theoretic characterization of Bayes performance and choice of priors in parametric and nonparametric problems," in *Bayesian Statistics 6: Proc. 6th Valencia Int. Meet.*, June 1998.
[407] I. Csiszár, "Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, pp. 2032–2066, Dec. 1991.
[408] S. Kulkarni, G. Lugosi, and S. Venkatesh, "Learning pattern classification—A survey," this issue, pp. 2178–2206..
[409] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, pp. 115–133, 1994.
[410] Y. S. Abumostafa, "Information theory, complexity and neural networks," *IEEE Commun. Mag.*, vol. 27, pp. 25–30, Nov. 1989.
[411] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1998.
[412] G. J. Chaitin, *Algorithmic Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1987.
[413] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Berlin, Germany: Springer, 1993.
[414] H. Yamamoto, "Information theory in cryptology," *IEICE Trans. Commun., Electron., Inform. Syst.*, vol. 74, pp. 2456–2464, Sept. 1991.
[415] N. Pippenger, "Information theory and the complexity of Boolean functions," *Math. Syst. Theory*, vol. 10, pp. 129–167, 1977.
[416] A. Steane, "Quantum computing," *Repts. Progr. Phys.*, vol. 61, pp. 117–173, Feb. 1998.
[417] P. Elias, "The efficient construction of an unbiased random sequence," *Ann. Math. Statist.*, vol. 43, pp. 865–870, 1972.
[418] D. E. Knuth and A. C. Yao, "The complexity of random number generation," in *Algorithms and Complexity: Recent Results and New Directions*, J. F. Traub, Ed. New York: Academic, 1976.
[419] K. Visweswariah, S. Kulkarni, and S. Verdú, "Source codes as random number generators," *IEEE Trans. Inform. Theory*, vol. 44, pp. 462–471, Mar. 1998.
[420] D. Lind and B. Marcus, *Symbolic Dynamics and Coding*. Cambridge, U.K.: Cambridge Univ. Press, 1995.
[421] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1501–1518, Nov. 1991.
[422] H. S. Witsenhausen, "Some aspects of convexity useful in information theory," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 265–271, May 1980.
[423] S. I. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford, U.K,: Oxford Univ. Press, 1999.
[424] L. L. Campbell, "The relation between information theory and the differential geometry approach to statistics," *Inform. Sci.*, vol. 35, pp. 199–210, June 1985.

[425] R. Ahlswede and I. Wegener, *Search Problems*. New York: Wiley–Interscience, 1987.
[426] S. W. Golomb, "Probability, information theory and prime number theory," *Discr. Math.*, vol. 106, pp. 219–229, Sept. 1992.
[427] O. Hijab, *Stabilization of Control Systems*. New York: Springer-Verlag, 1987.
[428] J. K. Sengupta, *Econometrics of Information and Efficiency*. Dordrecht, The Netherlands: Kluwer, 1993.
[429] W. T. Grandy, "Resource letter ITP-1: Information theory in physics," *Amer. J. Phys.*, vol. 16, pp. 466–476, June 1997.
[430] H. S. Leff and A. F. Rex, Eds., *Maxwell's Demon: Entropy, Information, Computing*. Princeton, NJ: Princeton Univ. Press, 1990.
[431] R. Landauer, "Information is physical," *Phys. Today*, vol. 44, pp. 23–29, May 1991.
[432] N. Sourlas, "Statistical mechanics and error-correcting codes," in *From Statistical Physics to Statistical Inference and Back*, P. Grassberger and J. P. Nadal, Eds. Dordrecht, The Netherlands: Kluwer, 1994, pp. 195–204.
[433] C. H. Bennett and P. Shor, "Quantum information theory," this issue, pp. 2724–2742.
[434] S. Hayes, C. Grebogi, and E. Ott, "Communicating with chaos," *Phys. Rev. Lett.*, vol. 70, no. 20, pp. 3031–3034, May 17, 1993.
[435] H. P. Yockey, *Information Theory and Molecular Biology*. New York: Cambridge Univ. Press, 1992.
[436] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network Comput. Neural Syst.*, vol. 3, pp. 213–251, May 1992.
[437] R. Linsker, "Sensory processing and information theory," in *From Statistical Physics to Statistical Inference and Back*, P. Grassberger and J. P. Nadal, Eds. Dordrecht, The Netherlands: Kluwer, 1994, pp. 237–248.
[438] K. Eckschlager, *Information Theory in Analytical Chemistry*. New York: Wiley, 1994.
[439] C. Papadimitriou, "Information theory and computational complexity: The expanding interface," *IEEE Inform. Theory Newslett.* (Special Golden Jubilee Issue), pp. 12–13, Summer 1998.
[440] T. M. Cover, "Shannon and investment," *IEEE Inform. Theory Newslett.* (Special Golden Jubilee Issue), pp. 10–11, Summer 1998.