

Our field has a rather remarkable history, in that it sprang almost fully fledged from the brow of one individual. In his original papers [1], Claude Shannon both posed the fundamental problems of information theory and also, to a large extent, answered them.

Shannon's good news was this: for every memoryless channel there exists a parameter C , the channel capacity, such that the probability of error $Pr(E)$ can be made arbitrarily small for any rate R less than C by use of an appropriate code and decoder. Good codes exist, and in fact a randomly chosen code will with high probability turn out to be good.

Shannon left a few loose ends, which have kept researchers occupied for nearly 50 years. In particular, his proof was nonconstructive, which left open the problem of finding specific good codes. Also, he assumed exhaustive maximum-likelihood decoding, whose complexity is proportional to the number of words in the code. It was clear that long codes would be required to approach capacity and therefore that more practical decoding methods would be needed.

I arrived at M.I.T. in the early sixties, during what can be seen in retrospect as the first golden age of information theory research. (It is notable that over half of the previous Shannon Lecturers were associated with M.I.T. during that time.) Research at M.I.T. focused on two principal problems: how does $Pr(E)$ go to zero, and how does one practically approach capacity?

The culmination of the first line of research was Gallager's elegant exponential error bounds for memoryless channels [2]. Gallager showed that the probability of error of the best block code of length N and rate R decreases exponentially with block length:

$$Pr(E) \cong \exp -NE(R),$$

where the error exponent $E(R)$ is greater than zero for all rates R less than the capacity C . A typical plot of $E(R)$ vs. R for a very noisy channel, such as a power-limited Gaussian channel, is illustrated in Figure 1. The low-rate straight-line portion of the curve can be derived by a simple union bound; it has a slope of -1 and a value of R_0 at $R = 0$.

Figure 1. Gallager's error exponent $E(R)$ for a very noisy channel.

Like Shannon, Gallager continued to assume a randomly chosen code and maximum-likelihood decoding. The decoding complexity G is then of the order of the number of codewords, $G \cong \exp NR$, and therefore $Pr(E)$ decreases only algebraically with decoding complexity:

$$Pr(E) \cong G^{-E(R)/R}.$$

It was well understood by this time that the key obstacle to practically approaching channel capacity was not the construction of specific good long codes, although the difficulties in finding asymptotically good codes were already apparent (as expressed by the contemporary folk theorem: "All codes are good, except those that we know of" [3]). Rather, it was the problem of decoding complexity.

By the early sixties at M.I.T., the decoding problem was considered to be essentially solved by the combination of:

- Long, randomly chosen ("probabilistic") convolutional codes [4];

- Sequential decoding [3].

Long convolutional codes have a dynamical structure which was at that time pictured as a branching tree structure. Sequential decoding refers to a class of exhaustive tree search techniques that are characterized by a probabilistic distribution of decoding computation. With an optimal choice of algorithm parameters, the probability of the computation exceeding a certain number G is a Pareto (algebraic) distribution:

$$Pr(\text{computation} > G) \cong G^{-\alpha(R)}.$$

The Pareto exponent $\alpha(R)$ is greater than one, and therefore the mean number of computations is bounded, when $R < R_0$ (the “computational cutoff rate,” which was then denoted as R_{comp}).

For a long enough convolutional code, decoding errors do not occur; rather, decoding failures occur when the number of computations exceeds some practical complexity limit G . Thus the probability of decoding failure still decreases only algebraically rather than exponentially with complexity G ; however, the complexity is lower and is in a much more palatable form.

In practice, sequential decoding (e.g., the Fano algorithm [5]) proved to be an efficient method of achieving essentially zero error probability on any memoryless channel at any rate $R < R_0$. There subsequently developed a quasi-religious belief that R_0 should be regarded as the “practical capacity” of a memoryless channel and that by the use of sequential decoding the practical capacity could more or less be achieved. Problem solved.

A lamentable consequence of this conclusion (which is rather ironic in view of the later history of sequential decoding) was that the M.I.T. information theory group, probably the greatest assemblage of talent that our field has ever known, began to disperse to other fields and institutions, thus bringing to an end the first golden age of information theory.

In my thesis, I took a different approach to the performance vs. complexity problem. My goal, which I regarded as essentially theoretical, was to find a class of codes and associated decoders such that the probability of error could be made to decrease exponentially at all rates less than capacity, while the decoding complexity increased only algebraically, so as to achieve an exponential tradeoff of performance vs. complexity.

The solution that I arrived at was a multilevel coding structure that I called concatenated coding, illustrated in Figure 2 [6]. In what today we would call the lowest physical layer, a relatively short random “inner code” is used with maximum-likelihood decoding to achieve a modest error probability like $Pr(E) \cong 10^{-2}$ at a rate near capacity. Then in a second layer, a long high-rate algebraic nonbinary Reed-Solomon (RS) “outer code” is used with a powerful algebraic error-correction algorithm (preferably using reliability information in error-and-erasure (E&E) or generalized-minimum-distance (GMD) decoding) to drive the error probability as low as desired, with little rate loss.

Figure 2. Concatenated coding.

I showed that by proper choice of codes, the probability of error can be made to decrease exponentially with overall code length N at all rates less than capacity:

$$Pr(E) \cong \exp -NE_c(R),$$

where the concatenation exponent $E_c(R)$ is less than $E(R)$ but nonetheless positive for all $R < C$. Meanwhile, the decoding complexity is dominated by the complexity of the algebraic RS decoder, which increases only algebraically with N .

(At that time, the complexity of decoding an RS code with minimum distance d was of the order of $O(d)$ for E&E decoding and $O(d^4)$ for GMD decoding, but soon the Berlekamp-Massey algorithm [7]-[8] reduced these to $O(d^2)$ and $O(d^3)$, respectively. Remarkably, 25 years later at the 1993 ISIT, four different authors [9]-[12], including the Shannon Lecturer (Berlekamp), independently described “one-pass” algorithms for GMD decoding of RS codes that reduce GMD decoding complexity also to $O(d^2)$.)

Another theoretical problem of that era was to develop exponential error bounds for convolutional codes analogous to Gallager’s bounds for block codes. The solution was the Viterbi-Yudkin [13]-[14] bound, which has the form

$$Pr(E) \cong \exp -\nu e(R),$$

where ν is the constraint length of the convolutional code, and the convolutional error exponent $e(R)$ is greater than zero for all rates R less than the capacity C . A typical plot of $e(R)$ vs. R for a very noisy channel is illustrated in Figure 3. The exponent is equal to $e(R) = R_0$ for all rates less than R_0 , and exceeds the block code exponent for $0 < R < C$; indeed, $e(R)/E(R)$ goes to infinity as R approaches C .

Figure 3. Convolutional error exponent $e(R)$ for a very noisy channel.

What is now called the Viterbi algorithm (VA) made its first appearance in the coding literature in [14] as a proof technique to develop this theoretical result. Using the VA (which was later shown to be a maximum-likelihood trellis search technique [15]-[16]), the decoding complexity is exponential in the constraint length, $G \cong \exp \nu R$. Thus the probability of error again is only an algebraic function of complexity,

$$Pr(E) \cong G^{-e(R)/R},$$

although the exponent $e(R)/R$ is superior to that for block codes. (Interestingly, the complexity exponent is equal to the Pareto exponent of sequential decoding [17], so the performance/complexity tradeoff of sequential decoding and of the VA are asymptotically equal.)

It is worth emphasizing that concatenated codes and the Viterbi algorithm were both developed in the course of theoretical research into issues of asymptotic performance and complexity. At the time of their development, neither was imagined to be practical. But as technology has advanced, both have turned out to be valuable standard tools in the toolkit of the communications engineer. The moral is that research motivated by good hard performance and complexity metrics tends to yield not only good papers, but also significant contributions to practice.

The superiority of $e(R)$ to $E(R)$ was used to argue that convolutional codes were inherently superior to block codes. Indeed, the relation between $e(R)$ and $E(R)$ is intimately related to the fact that for all $R < C$, an optimum block code can be constructed by properly terminating a random convolutional code. This is the basis for the graphical construction (the “inverse concatenation construction” [18]) of $e(r)$ from $E(R)$ that is schematically indicated in Figure 3.

However, this argument deserves a second look. If an optimum block code is in fact constructed as a terminated convolutional code, then it can be decoded using the VA with a complexity only of the order of $G \cong \exp \nu R$, not $\exp NR$. (Research on the currently hot topic of “trellis complexity of block codes” has revealed similar complexity reductions for general linear block codes.) Then the only performance/complexity difference between a convolutional code and a block code obtained by terminating it is the rate loss incurred by the terminating “tail.” This loss can be made as small as desired by letting the block length become very long; then there is no practical difference from a performance/complexity point of view between block and convolutional codes. (Open question: can this rate loss be avoided in another way by using short “tail-biting” block codes?)

We conclude that the dynamical structure of convolutional (or block) codes does not prevent optimal performance and permits reduced decoding complexity. Indeed on memoryless channels convolutional codes (and more recently their trellis code cousins) are used almost universally today at the lowest physical layer.

What else can be said about what kinds of codes and decoding methods can approach the performance promised by Shannon nearly 50 years ago, on the basis of our experience since then?

Codes with algebraic structure have been investigated since the earliest days of coding theory, with the goals both of explicit specification of good long codes and also of making possible efficient practical decoding algorithms.

Group structure (i.e., using symmetries in coding) certainly seems to be consistent with achieving channel capacity. Indeed, Gallager's bounds hold for codes which have a group property on appropriately symmetrical channels [19]; on a power-limited Gaussian channel with no bandwidth constraints, the highly symmetrical orthogonal, bi-orthogonal, or simplex codes can approach the Shannon limit [20]; and according to "deBuda's result," lattice codes can approach the capacity of bandwidth-limited Gaussian channels [21] (a result which Loeliger has shown is really about group codes over \mathbb{Z}_p [22]). Most of the best known trellis codes have been shown to be representable as "group codes" [23]-[25] and thus to be "geometrically uniform" [26].

However, the well-known difficulties that have been encountered in finding classes of asymptotically good algebraic block codes leave room for doubt about the asymptotic usefulness of multiplicative structure. Of course, Reed-Solomon codes [27], which are based on the fundamental theorem of algebra, are optimal and also highly useful, but they are not asymptotic. But one wonders whether the enormous amount of work that has been devoted to cyclic codes (e.g., binary BCH codes) has not been disproportionate to the practical payoff. However, in view of the exciting current work on codes from algebraic geometry, some of which are asymptotically good, it is still too early to come to any final conclusions about this question.

A very significant recent result [28]-[29], which follows easily from the chain rule of information theory, is that in principle one can approach the capacity of channels which support many bits per symbol (e.g., band-limited Gaussian channels) by multilevel codes and multistage decoding, in which coding and decoding at each level are completely decoupled [30]. The results of [29] using "turbo codes" [31] at each level show that this result is not merely theoretical.

As far as decoding methods that are useful in getting to the Shannon limit, the general approach of concatenated coding still seems to be largely valid. At the lowest physical layer, one should use near-optimum decoding of moderate-complexity codes, preferably convolutional/trellis codes; maximum-likelihood (VA) decoding, bounded-distance (Euclidean, not Hamming) decoding, or even sequential decoding are all reasonable candidates. The place for algebraic codes such as Reed-Solomon codes is at higher levels, where they can also serve for burst correction and/or error detection. Reliability information should be used wherever possible.

However, recent results on iterative decoding of "turbo codes" [31], which have achieved low error probabilities at rates well beyond R_0 , turn conventional wisdom on its head and suggest fundamentally new techniques. Beyond R_0 , there is an explosion of multiplicity of near neighbors which causes the divergence of the union bound and the computational failure of sequential decoding, and which is seen empirically in long dense codes and lattices. "Turbo codes" attack multiplicity rather than error exponent, using long interleavers to obtain an "interleaving gain" in the error coefficient, which suffices to achieve a desired finite coding gain at a non-asymptotic $Pr(E)$ [32]. Moreover, the success of the iterative decoding method used with "turbo codes" shows that nonlinear bootstrapping methods are unexpectedly effective in the challenging regime between R_0 and C .

Finally, we may mention some recent results on the kinds of equalization that are consistent with ap-

proaching the capacity of a general linear Gaussian channel, where one must contend with intersymbol interference (ISI) as well as with Gaussian noise.

Shannon showed that the optimum (capacity-achieving) transmitted spectrum may be computed by the well-known “water-pouring” method, which usually results in the transmitted power going to zero at the band edge [19]. Single-carrier modulation with linear equalization leads to excessive noise enhancement in this case. Maximum-likelihood sequence detection (not “estimation”; sorry) is known to be optimal for ISI channels [33], but is usually too complex to be feasible.

One long-known method for dealing with these problems is to use multicarrier modulation to split the channel up into many independent narrow subchannels. If the bands are sufficiently narrow, then each subchannel may be regarded as an ideal (ISI-free) Gaussian channel, and only trivial equalization is required. Achieving the capacity of each subchannel independently by appropriate bit rate and power allocation between subchannels will achieve the capacity of the aggregate channel. There has been much recent practical development of multicarrier techniques [34].

An alternative is to use single-carrier modulation with an equalization method that achieves the performance of decision-feedback equalization (DFE). “Price’s result” [35]-[36] shows that at a high signal-to-noise ratio (SNR), the gap between uncoded modulation and capacity using zero-forcing DFE is the same on any linear Gaussian channel with ISI as it is on the ideal channel with no ISI. More recently this result has been shown to hold at any SNR with minimum-mean-squared-error (MMSE) DFE [37]. Furthermore, practical methods have been found to obtain MMSE-DFE performance in combination with the coding (and shaping) gain of powerful codes by “transmitter precoding” techniques, sometimes called “DFE in the transmitter” [36]. This type of precoding is used in the new V.34 high-speed (28.8-33.6 kb/s) telephone-line modem standard. The upshot is that, with precoding, capacity can now be approached as closely on a general linear Gaussian (ISI) channel as it can be on an ideal (zero-ISI) channel [37].

In summary, at least two practical equalization methods are known that can approach the capacity of general linear Gaussian channels: multicarrier modulation, or single-carrier modulation with transmitter precoding.

Thus at least for linear Gaussian channels, we are very close to practical achievement of the performance that Shannon promised nearly 50 years ago. Is it therefore time to say “Problem solved” and move on to other things? I doubt it.

Observe how much of this progress has been achieved only recently – e.g., in trellis codes, group codes, turbo codes, algebraic geometry codes, and precoding. I doubt that it has been fully digested. In particular, we are still far from a fundamental understanding of the new code construction and decoding methods that seem to be embodied in turbo codes, nor have we fully exploited the promise of multilevel codes and multistage decoding. Moreover, every two years the boundary between “feasible” and “infeasible” advances by another factor of two.

Indeed, I believe that in 30 years we will look back with nostalgia at the current era and say: “That was a golden age.”

I would like to conclude by saying what a pleasure it has been to have this field as my professional home. It is rare good luck to be in a field in which elegant theory is motivated by practical problems and so often leads to practical advances. Indeed, as I have already said, I believe that research based on hard metrics of performance and complexity often yields the best theory as well as the best applications.

It is also rare good luck to work in such a collegial, mutually supportive field in which there is a minimum of unproductive competitiveness and a maximum of joy in shared achievement. I believe that this collegial culture springs in part from the unusual history and focus of our field, and in part from the character of the first generation of researchers, who established its ethic. I could wish nothing better for future generations

of information theorists than that this culture be preserved.

References

1. C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379-423 and 623-656, 1948.
2. R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Trans. Information Theory*, vol. IT-11, pp. 3-18, Jan. 1965.
3. J. M. Wozencraft and B. Reiffen, *Sequential Decoding*. Cambridge, MA: MIT Press, 1961.
4. P. Elias, "Coding for noisy channels," *IRE Conv. Rec.*, pt. 4, pp. 37-46, 1955.
5. R. M. Fano, "A heuristic discussion of probabilistic decoding," *IEEE Trans. Information Theory*, vol. IT-9, pp. 64-74, Apr. 1963.
6. G. D. Forney, Jr., *Concatenated Codes*, Cambridge, MA: MIT Press, 1966.
7. E. R. Berlekamp, *Algebraic Coding Theory*. New York: McGraw-Hill, 1968.
8. J. L. Massey, "Shift-register synthesis and BCH decoding," *IEEE Trans. Information Theory*, vol. IT-15, pp. 122-127, Jan. 1969.
9. E. R. Berlekamp, "Codes and games," *Proc. 1993 IEEE Intl. Symp. Inform. Theory*, San Antonio, TX, Jan. 1993.
10. R. Koetter, "A new efficient error-erasure location scheme in GMD decoding," *Proc. 1993 IEEE Intl. Symp. Inform. Theory*, San Antonio, TX, Jan. 1993.
11. S. Sakata, "Linear recurrences on 2D convex lattices and decoding of some codes from algebraic curves," *Proc. 1993 IEEE Intl. Symp. Inform. Theory*, San Antonio, TX, Jan. 1993.
12. U. Sorger, "Fast generalized-minimum-distance decoding," *Proc. 1993 IEEE Intl. Symp. Inform. Theory*, San Antonio, TX, Jan. 1993.
13. A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. IT-13, pp. 260-269, Apr. 1967.
14. H. L. Yudkin, "Channel state testing in information decoding," Sc.D. thesis, Dept. of Elec. Engg., M.I.T., 1964.
15. G. D. Forney, Jr., "Review of random tree codes," Appendix A of Final Report on Contract NAS2-3637, NASA CR73176, NASA Ames Res. Ctr., Calif., 1967.
16. G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268-278, Mar. 1973.
17. G. D. Forney, Jr., "Convolutional codes III. Sequential decoding," *Information and Control*, vol. 25, pp. 267-297, 1974.
18. G. D. Forney, Jr., "Convolutional codes II. Maximum-likelihood decoding," *Information and Control*, vol. 25, pp. 222-266, 1974.
19. R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
20. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.
21. R. de Buda, "Some optimal codes have structure," *IEEE JSAC*, pp. 893-899, Aug. 1989.
22. H.-A. Loeliger, "On existence proofs for asymptotically good Euclidean-space group codes," *Proc. DIMACS/IEEE Workshop on Coding and Quantization*, Piscataway, NJ, Oct. 1992.
23. M. D. Trott, "The algebraic structure of trellis codes," Ph.D. thesis, Dept. of Elec. Engg., Stanford U., Stanford, CA, 1992.
24. G. D. Forney, Jr. and M. D. Trott, "The dynamics of group codes: State spaces, trellis diagrams, and canonical encoders," *IEEE Trans. Information Theory*, vol. 39, pp. 1491-1513, Sept. 1993.
25. E. J. Rossin, N. T. Sindhushayana and C. D. Heegard, "Trellis group codes for the Gaussian channel," *IEEE Trans. Information Theory*, vol. 41, pp. 1217-1245, Sept. 1995.

26. G. D. Forney, Jr., "Geometrically uniform codes," *IEEE Trans. Information Theory*, vol. IT-36, pp. 1241-1260, Sept. 1991.
27. I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. SIAM*, vol. 8, pp. 300-314, 1960.
28. Y. Kofman, E. Zehavi and S. Shamai (Shitz), "Performance analysis of a multilevel coded modulation system," *IEEE Trans. Commun.*, vol. 42, pp. 299-311, Feb.-Apr. 1994.
29. J. Huber and U. Wachsmann, "Power-efficient rate design for multilevel codes with finite block length," *Proc. 1995 IEEE Intl. Symp. Inform. Theory*, Whistler, BC, Canada, Sept. 1995.
30. H. Imai and S. Hirakawa, "A new multilevel coding method using error-correcting codes," *IEEE Trans. Information Theory*, vol. IT-23, pp. 371-377, May 1977.
31. C. Berrou, A. Glavieux and P. Thitimajshima, "Near Shannon limit error-correcting coding: Turbo codes," *Proc. 1993 IEEE Int. Conf. Commun.*, pp. 1064-1070, May 1993.
32. S. Benedetto and G. Montorsi, "Design of parallel concatenated convolutional codes," *IEEE Trans. Commun.*, to appear.
33. G. D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Information Theory*, vol. IT-18, pp. 363-378, 1972.
34. J.A.C. Bingham, "Multicarrier modulation for data transmission: An idea whose time has come," *IEEE Commun. Mag.*, vol. 28, no. 4, pp. 5-14, Apr. 1990.
35. R. Price, "Nonlinearly-equalized PAM vs. capacity for noisy filter channels," *Proc. 1972 IEEE Int. Conf. Commun.*, pp. 22.12-22.17, June 1972.
36. M. V. Eyuboglu and G. D. Forney, Jr., "Combined equalization and coding using precoding," *IEEE Commun. Mag.*, vol. 29, no. 12, pp. 25-34, December 1991.
37. J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu and G. D. Forney, Jr., "MMSE decision-feedback equalizers and coding," *IEEE Trans. Commun.*, vol. 43, pp. 2582-2604, Oct. 1995.

G. David Forney, Jr. was born in New York, NY, on March 6, 1940. He received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1961, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 1963 and 1965, respectively.

In 1965, he joined the Codex Corporation, and became a Vice President and Director in 1970. From 1982 to 1986, he was Vice President and Director of Technology and Planning of the Motorola Information Systems Group, Mansfield, MA. He is currently a Vice President of the Technical Staff of Motorola, Inc. He was an Adjunct Professor at M.I.T. in 1978-80 and a Visiting Scientist in 1991 and 1994-95, and also a Visiting Scientist at Stanford University, Stanford, CA, in 1972-73 and in 1990.

Dr. Forney was Editor of the *IEEE Transactions on Information Theory* from 1970 to 1973. He was a member of the Board of Governors of the IEEE Information Theory Society during 1970-76 and 1986-94, and was President in 1992. He has been awarded the 1970 IEEE Information Theory Group Prize Paper Award, the 1972 IEEE Browder J. Thompson Memorial Prize Paper Award, the 1990 IEEE Donald G. Fink Prize Paper Award, the 1992 IEEE Edison Medal, and the 1995 IEEE Information Theory Society Claude E. Shannon Award. He was elected a Fellow of the IEEE in 1973, a member of the National Academy of Engineering (U.S.A.) in 1983, and an honorary member of the Popov Society (Russia) in 1994.