IEEE Information Theory Society Newsletter

Vol. 42, No. 4, December 1992

Editor: Nader Mehravari

1990 Shannon Lecture[1]

Thomas M. Cover [2]

Appropriately enough, Claude Shannon was asked to give the first Shannon Lecture in Ashkelon, Israel, in 1973. Perhaps noting the self referential nature of his award, Shannon chose to speak on examples of feedback. Most of the Shannon Lectures since then have contained welcome autobiographical components. I'm thankful for this tradition, because I would like to do the same.

Our field plays a unique role in communication theory, staking out as it does the extreme points of the subject, the limits of data compression and transmission. Of course there is some stated concern that it might be too theoretical and therefore irrelevant in a pragmatic society. Such attitudes will always be with us and it's good to realize that. Similar objections have been raised to asymptotic results. Maybe a good response to the relevance objections to asymptotics and to the debate on theory versus practice would be that an asymptotic limit is the first term in the Taylor series expansion at infinity. And theory is the first term in the Taylor series of practice. Anyway, there can be no doubt that information theory provides guidance and confidence in practice. But we are here not to defend the success of information theory, but to celebrate it.

There are a million reasons to do research, especially in an exciting field like information theory. In my case, I have especially admired theorems expressing the unexpected. Such theorems are like good jokes. A joke has a certain setup which leads to a comfortable point of view, which is then dashed by the punchline of the theorem. Or a joke may simply be a puzzling statement followed by a pithy and pleasing resolution. Or it could be a paradoxical juxtaposition like Mark Twain's remark about Wagner's music: "It's better than it sounds." In short, a good theorem has surprise value.

I will try to illustrate the quirky nature of mathematics and my own particular interest in eccentric results. In the interests of space, I will omit the details of some of these examples. The primary new contribution in this talk is the development of a universal portfolio which achieves universal investment goals precisely parallel to the goals achieved in universal data compression. My subjects will include the following:

- an idiosyncracy in the restoring force for the law of averages;
- a proof that in some strong sense all sporting contests are equally exciting;
- a proof that longer contests do not necessarily favor the stronger player;
- a method of deciding which of two numbers is the largest when observing only one of them;
- an information theoretic proof of Hadamard's inequality;

continued on page 3

[1] This is a synopsis of the 1990 Shannon Lecture, delivered January 1990 at the IEEE International Symposium on Information Theory, San Diego, Calif.

[2] Stanford University, Departments of Electrical Engineering and Statistics.

1990 Shannon Lecture

continued from front cover

- a proof that feedback at most doubles the capacity of an additive colored noise Gaussian channel;
- a review of the proof that the capacity of a linear threshold device is twice the number of variable weights;
- a conjecture that any feed-forward neural net has a capacity which is at most equal to twice the number of variable weights in the network;
- a proof that at least half the information in a set of labelled training patterns is contained in the nearest neighbor;
- a development of the main ideas in the degraded broadcast channel capacity region;
- a discussion of the Slepian-Wolf theorem for simultaneous data compression;
- a universal investment algorithm that outperforms the market.

Restoring force of the law of averages. We all know that there is no restoring force to the law of averages. On the other hand, the strong law of large numbers says that $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ, where $X_1, X_2, \ldots, X_n$ are independent identically distributed random variables with expected value μ.

However, it can be shown that if X has a positive probability density function, then $\Pr\{|\overline{X}_{2n} - \mu| < |\overline{X}_n - \mu|\} > \frac{1}{2}$. The punchline is that this is true not only for $\mu = EX$, but for any real number μ whatsoever. Thus the sample average of $2n$ i.i.d. samples is closer at least half the time to any real number than is the sample average of n samples.

All sporting contests are equally exciting. This seems unreasonable at first glance. Why would, for example, basketball, football and bowling be equally exciting? However, there is a strong sense in which they are. Let $p(t)$ be the probability that A wins a contest as a function of time. Thus $p(t)$ is a random process reflecting the probability that A will win the contest given all the information about the contest up through time t. It can be proved that $p(t)$ is a martingale tending to one or zero accordingly as A does or does not win the contest. If one looks at the game closely enough, it is reasonable to assume the $p(t)$ is continuous. Thus, if one looks at the level crossings of $p(t)$, say for $p(t)$ crossing points in the set $\{0, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}, 1\}$, the martingale induces a random walk. All sporting contests with continuous $p(t)$ give the same probability distribution on this random walk. So in that sense, all sporting contests are equally exciting.

Do longer contests favor the stronger player? It is intuitively obvious that longer sporting contests magnify the advantage of the stronger player. Let $P_n = \Pr\left\{\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} Y_i\right\}$ where $X_1, X_2, \ldots$ are i.i.d. $\sim F(x)$ and $Y_1, Y_2, \ldots$ are iid $\sim G(y)$. Then if $P_1 > \frac{1}{2}$ it seems that is should be true that $P_n \geq \ldots \geq P_2 \geq P_1$. However, it is easy to find distributions F and G such that P_n starts off near one, dips near zero, then comes close to one, dips down again to zero and so on. Such an example [1] can be made by letting the distribution of $X_i - Y_i$ have sparse mass points, both positive and negative, with the appropriate probabilities. Necessarily, the expected value of $X-Y$ does not exist. Thus there are games where the identity of the strongest player depends on the length of the contest.

Choose the largest number [2]. Suppose malicious nature chooses two numbers and *randomly* gives you one of the numbers. Is there a means by which you can determine whether the number x in your hand is the largest of the two numbers? At first it seems that the probability is one half that the number in your hand is the largest. That's true. It also seems that there can be no means, by inspection of this number, since the two numbers were arbitrarily chosen, for one to know which number is largest. That's untrue. Simply let $\phi(x)$ be any monotonic strictly decreasing function with range [0,1] designating the probability with which one asserts that the other number is the largest. Then, since one is more likely to switch from a small number to a large number than from a large number to a small

IEEE
Information Theory Society
Newsletter

IEEE Information Theory Society Newsletter (USPS 360-250) is published quarterly by the Information Theory Society of the Institute of Electrical and Electronics Engineers, Inc.

Headquarters: 345 East 47th Street, New York, NY 10017.

Cost is $1.00 per member per year (included in Society fee for each member of the Information Theory Society. Printed in the U.S.A. Second-class postage paid at New York, NY and additional mailing offices.

Postmaster: Send address changes to IEEE Information Theory Society Newsletter, IEEE, 445 Hoes Lane, Piscataway, NJ

number (because of the monotonicity of ϕ), the probability of selecting the larger number will be strictly greater than one half.

Hadamard's inequality. Hadamard's inequality states that the determinant of a nonnegative definite matrix is less than or equal to the product of the diagonal values. It turns out that this and many other famous matrix inequalities can be obtained most simply by information theoretic inequalities [3,4]. In this case, the relevant inequality is $H(X_1, \ldots, X_n) \leq \sum H(X_i)$. Let X be normal with mean zero and covariance matrix K. Plug this in to the entropy inequality and one determines by inspection that $\det(K) \leq \prod K_{ii}$, thus proving Hadamard's inequality.

The capacity of a channel with feedback. In addition to the great structural results of Shannon in information theory, there were a number of counterintuitive results. The first is that one could send at a positive rate (up to capacity) with probability of error tending to zero. The next is that if one allows feedback in a discrete memoryless channel, the capacity does not increase. Finally, we have the results of Pinsker and Ebert which show that feedback increases capacity by at most a factor of two for time-dependent additive Gaussian noise channels. Incidentally, the easiest proof of this result uses simple information theoretic inequalities [5].

Conjecture on neural nets. The capacity of a single neuron (a linear threshold device) is two patterns per variable weight; any number of patterns less than twice the number of variable weights is very likely to be correctly classifiable by the neuron, and a number of patterns greater than twice the number of variable weights is almost certainly not classifiable. For this reason, many researchers have advocated the use of feed-forward neural nets with an arbitrary number of neurons to form some final weighted vote. I would like to conjecture that the capacity of such a feed-forward neural net is less than twice the number of variable weights in the structure. Thus feedforward neural nets may have no more classification and learning power than a single neuron with the same number of degrees of freedom.

Nearest neighbor pattern recognition. Here is a result in pattern recognition that has some relation to information theory. Let $(X_1, \theta_1), (X_2, \theta_2) \ldots$ be i.i.d. according to some joint distribution where θ_i takes values in the set $\{1,2\}$ (the classifications) and X_i takes values in a separable metric space. Here the θ_i's are the classifications and the associated X_i's are the observations. Given a new pattern X similarly drawn, what should be the classification assigned to X?

If the underlying distribution is known, Bayes' rule can be used, achieving a probability of error R^*. It is somewhat surprising [6] that if one assigns to X the classification of the nearest neighbor to X among the labeled data set and lets the size of the labeled data set go to infinity, then the probability of error of this nearest neighbor classification algorithm is less than $2R^*(1-R^*)$. Thus at least half of the classification information in this training sample is contained in the nearest neighbor.

Broadcast channels. I was in a car with Dave Slepian and Peter Elias in Haifa, while attending a conference in the late 1960s. I was interested in schemes for sending information simultaneously to several receivers. One wants a code that is compatible with each receiver. So I asked, "Is anything known about compatible codes?" They craned their necks around and said, "What do you mean by compatible codes?" That shocked me into silence because I wasn't sure.

After some work, I found that if one receiver is a stochastically degraded version of another, one can precede the input of the channel by a fake channel and code for the overall transmission from the beginning to the end for the worst receiver, and then use the degrees of freedom at the actual input to the channel to encode some extra information for the better receiver. This leads to a determination of the capacity region for the degraded broadcast channel.

This work [11] received some credit for starting multiple user information theory, but it was Shannon [7], really, who was the first person to write a paper on multiple user information theory in 1961. Other early workers were Ahlswede and van der Meulen. Incidentally, the general broadcast channel capacity region remains unknown.

Slepian-Wolf theorem. One of the main results in multiple user information theory is the Slepian-Wolf theorem [8]. It turns out that the techniques needed to prove broadcast channel results lead to a jointly typical sequence argument which leads to a simplified proof [9] of one of the most fundamental results in multiple user theory — the Slepian-Wolf theorem. Their theorem states that one can separately data compress an X source and a Y source at respective rates $H(X)$ and $H(Y \mid X)$ and still recover both sources, X and Y, from the compressed strings.

I would now like to focus on a universal investment strategy which has goals analogous to those of universal data compression strategies. (This work [10] was published one year after this lecture.)

Universal portfolios. Information theory divides naturally into channel capacity theorems and data compression theorems. One of the successes over the last 20 years has been the advent, due to Davisson, Gray, Schalkwijck, Lempel, Ziv, Rissanen and others, of universal data compression methods. Universal data compression is robust in the sense that one doesn't need to know ahead of time the underlying distribution of the source. A typical way in which universal data compression succeeds is by first describing the empirical distribution P_n of the source (for example, the number of ones in a binary sequence) and then describing which of those sequences was observed. It typically requires about log n bits to describe the empirical distribution and then $nH(P_n)$ bits to describe the sequence which actually occurs, given this empirical distribution. Since $P_n \rightarrow P$, one has a data compression procedure which asymptotically achieves the compression rate $H(P)$.

There is a certain duality between data compression and gambling. Roughly speaking, one can say that the log of the amount of money that one wins, say, betting on a binary sequence, plus the entropy of this binary sequence is equal to a constant. Thus, the lower the entropy, the more money one makes. To achieve this maximal amount of money, one must use proportional gambling and bet on each sequence in proportion to its probability of occurring.

One wonders whether there exist universal investment algorithms for investing in the stock market which perform as well as if one had known ahead of time the empirical distribution of the daily stock performances.

There is one extra difficulty that a universal investment algorithm must face. In universal data compression, one sees the entire sequence in advance and then is allowed to describe it. However, in investment one sees the data unfold as time goes on. One cannot look into the future. One does not know ahead of time the empirical distribution of the market.

We now consider a sequential portfolio selection procedure for investing in the stock market with the goal of performing as well as if we knew the empirical distribution of future market performance. Moreover, we do not make any statistical assumption about the behavior of the market. In particular we allow for market crashes such as those occurring in 1929 and 1987. The actual sequence of stock outcomes can be chosen by a malicious nature, and we hope to do as well as if we had known the sequence (up to permutation) ahead of time.

In order to examine the performance of this so-called universal portfolio algorithm, we look at natural goals for the growth rate of wealth for arbitrary market sequences. For example, a natural goal might be to outperform the best buy-and-hold strategy, thus beating an investor who has been given a look at a newspaper dated in the future.

We propose a more ambitious goal. To motivate this goal, let us consider all constant rebalanced portfolio strategies. Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ denote a stock market vector for one investment period, where x_i is the price relative for the i^{th} stock, *i.e.*, the ratio of the closing to opening price for stock i. A portfolio $\mathbf{b} = (b_1, b_2, \ldots, b_m)$ is the proportion of the current wealth invested in each of the m stocks. Thus $S = \mathbf{b}^t\mathbf{x} = \sum_{i=1}^n b_i x_i$ is the factor by which wealth increases in one investment period using portfolio $\mathbf{b}$. We will now consider an arbitrary sequence of stock market vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Each of these vectors lies in the positive orthant of m-dimensional Euclidian space. Here x_{ij} is the price relative of stock j on day i. A constant rebalanced portfolio strategy $\mathbf{b}$ achieves wealth

$$S_n(\mathbf{b}) = \prod_{i=1}^{n} \mathbf{b}^t\mathbf{x}_i$$

Thus $S_n(\mathbf{b})$ is the amount of wealth that would accrue from rebalancing one's investments in the m stocks at the end of each day in the proportions $b_1, b_2, \ldots, b_m$. Let

$$S_n^* = \max_{\mathbf{b}} S_n(\mathbf{b})$$

denote the maximum wealth achievable on the given stock sequence maximized over all constant rebalanced portfolios. Our goal is to achieve S_n^*.

We propose the universal adaptive portfolio strategy given by

$$\hat{\mathbf{b}}_1 = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right), \quad \hat{\mathbf{b}}_{k+1} = \frac{\int \mathbf{b} S_k(\mathbf{b}) d\mathbf{b}}{\int S_k(\mathbf{b}) d\mathbf{b}}, \tag{1}$$

where the integration is over the set of $(m-1)$-dimensional portfolios

$$\mathbf{b} \geq 0, \ \sum_{i=1}^{m} b_i = 1.$$

The wealth $\hat{S}_n$ resulting from the universal portfolio is given by

$$\hat{S}_n = \prod_{k=1}^{n} \hat{\mathbf{b}}_k^t \mathbf{x}_k.$$

Thus the initial universal portfolio $\hat{\mathbf{b}}_1$ is uniform over the stocks, and the portfolio $\hat{\mathbf{b}}_k$ at time k is the performance weighted average of all portfolios $\mathbf{b}$.

Consider now m=2. We are able to show [10] that

$$\hat{S}_n = \sqrt{\frac{2\pi}{nJ_n}}\, S_n^* \qquad (2)$$

plus terms of lower order in n for *every* sequence of stock vectors. Here J_n refers to the curvature of $S_n(\mathbf{b})$ at its maximum. Thus $\hat{S}_n$ achieves, to first order in the exponent, the same growth as the target wealth S_n^*. We observe that $\frac{1}{n}\log S_n^*$ plays the same role in universal investment that the entropy H plays in universal data compression.

The main idea of the portfolio algorithm is as follows. Give an amount $d\mathbf{b}$ to each portfolio manager indexed by a constant rebalancing strategy $\mathbf{b}$. This portfolio manager will make $S_n(\mathbf{b})d\mathbf{b}$. Thus, when the performances of all the portfolio managers are pooled together at the end, an amount of wealth $\hat{S}_n = \int S_n(\mathbf{b})d\mathbf{b}$ is achieved. But the integral is approximately equal to S_n^*, because $S_n(\mathbf{b})$ is exponential in n and the arithmetic average of exponentials has an exponent given by the maximum of the exponents. The precise proof [10] uses some techniques from Laplace's method of integration.

In the calculation of tomorrow's portfolio $\hat{\mathbf{b}}_{k+1}$, we simply add together (on paper) each portfolio manager's buy and sell orders, resulting in $\hat{\mathbf{b}}_{k+1}$ as given in (1).

Notice that, unlike the universal data compression algorithms, in which one looks at a mixture of all data compression schemes indexed by the underlying distribution, we do not index the portfolio investment algorithms by the underlying distribution on the stock vector (first, there isn't any underlying distribution, and second, there are too many such distributions) but instead we index the portfolio algorithms by $\mathbf{b}$, a point in the simplex of all constant rebalanced portfolios. Thus we place a uniform distribution over all algorithms, not a uniform distribution over all underlying distributions. This also eliminates the unnecessary intermediate step of estimating the (nonexistent) distribution $F(\mathbf{x})$ of the next outcome.

We have investigated the actual performance of this algorithm on real stocks and found that $\hat{S}_n$ and S_n^* typically exponentially dominate the best buy-and-hold strategy. Thus, the motivation of universal data compression has led naturally to a universal investment strategy.

In closing, I believe that information theory shares with physics a certain coherence and beauty. The history is dramatic and powerful, but I believe that the full coherence and extent of the field is yet to be discovered. I've been proud to work with the many engineers, physicists, mathematicians, and statisticians who have made this field their intellectual home.

References

1. T. Cover. "Do Longer Games Favor the Stronger Player?" *The American Statistician*, 43(4):277-278, November, 1989.

2. T. Cover and B. Gopinath. "Pick the Largest Number," *Open Problems in Communication and Computation*, p.152, Springer-Verlag, New York, 1987.

3. A. Dembo, T. Cover and J. Thomas. "Information Theoretic Inequalities," *IEEE Transactions on Information Theory*, 37(6):1501-1518, November, 1991.

4. T. Cover and J. Thomas. *Elements of Information Theory*, Wiley & Sons, New York, 1991.

5. T. Cover and S. Pombra. "Gaussian Feedback Capacity," *IEEE Transactions on Information Theory*, 35(1):37-43, January, 1989.

6. T. Cover and P. Hart. "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, 13(1):21-27 January, 1967.

7. C. Shannon. "Two-Way Communication Channels," *Pr 4th Berkeley Symp. Math.Stat. and Prob.*, 1:611-644, 1961.

8. D. Slepian and J. Wolf. "Noiseless coding of correlated information sources," *IEEE Trans. Information Theory*, IT-19:471-480, July, 1973.

9. T. Cover. "A Proof of the Data Compression Theorem of Slepian and Wolf for Ergodic Sources," *IEEE Trans. Information Theory*, IT-21(2):226-228, March, 1975.

10. T. Cover. "Universal Portfolios," *Mathematical Finance* 1(1):1-29, January, 1991.

11. T. Cover. "Broadcast Channels," *IEEE Trans. Information Theory* IT-18(1):2-14, January, 1972.