TABLE IV
GENERATORS (IN OCTAL) AND MAP OF DELETING BITS FOR BEST RATE $R = 7/8$ PUNCTURED CODES

k	Generator	Map	d_f	$W(d_f)$
2	7 5	1000001 1111110	2	8
3	15 11	1110011 1001100	2	2
4	35 23	1101100 1010011	3	49
5	63 61	1110010 1001101	4	1572
5	73 53	1001001 1110110	3	16
6[a]	163 135	1111111 1000000	4	472

[a]Partial search.

$R = 7/8, k = 5$) all the maximum d_f codes found had a very large information weight $W(d_f)$; the next largest d_f codes are, therefore, also tabulated. A small personal computer (8-MHz Intel 8086 processor) was used to obtain the listed codes, and in one case ($R = 7/8$, $k = 6$) only a partial search was performed.

The free distances of the new codes and the punctured variable-rate codes derived from specific rate 1/2 codes by Yasuda *et al.* [8] are compared in Table V. The new codes have greater free distances in five instances. The comparison indicates that not too much bit error rate performance is sacrificed when variable-rate codes are used.

TABLE V
COMPARISON OF FREE DISTANCES OF NEW CODES d_f WITH THOSE OF VARIABLE-RATE CODES d_{fv}

	$R = 4/5$		$R = 5/6$		$R = 6/7$		$R = 7/8$	
k	d_f	d_{fv}	d_f	d_{fv}	d_f	d_{fv}	d_f	d_{fv}
2	2	2	2	2	2	2	2	2
3	3	3	3	3	2	2	2	2
4	4	3	3	3	3	3	3	3
5	4	4	4	4	4	3	4	3
6	4	4	4	4	4	3	4	3

ACKNOWLEDGMENT

The author wishes to thank Dr. Tor Helleseth and Scientist Oddny. H. Brun for support and helpful discussions.

REFERENCES

[1] J. B. Cain, G. C. Clark, Jr., and J. M. Geist, "Punctured convolutional codes of rate $(n-1)/n$ and simplified maximum likelihood decoding," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 97–100, Jan. 1979.
[2] Y. Yasuda, Y. Hirata, and A. Ogawa, "Optimum soft decision for Viterbi decoding," in *Proc. 5th Int. Conf. Digital Satellite Commun.*, Genova, Italia, Mar. 1981, pp. 251–258.
[3] Y. Yasuda, Y. Hirata, K. Nakamura, and S. Otani, "Development of variable-rate Viterbi decoder and its performance characteristics," in *Proc. 6th Int. Conf. Digital Satellite Commun.*, Phoenix, AZ, Sept. 1983, pp. XII-24–XII-31.
[4] A. J. Viterbi, "Convolutional codes and their performance in communication systems," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 751–772, Oct. 1971.
[5] O. H. Brun, "Evaluation of convolutional codes by using statistical discrimination analysis," EB NERA SatCom Division Rep., Bergen, Norway, Oct. 1986.
[6] J. L. Massey and M. K. Sain, "Inverses of linear sequential machines," *IEEE Trans. Comput.*, vol. C-17, pp. 330–337, Apr. 1968.
[7] P. L. Lee, "New short constraint length, rate $1/n$ convolutional codes which minimize the required SNR for given desired bit error rates," *IEEE Trans. Commun.*, vol. COM-33, pp. 171–177, Feb. 1985.
[8] Y. Yasuda, K. Kashiki, and Y. Hirata, "High-rate punctured convolutional codes for soft decision Viterbi decoding," *IEEE Trans. Commun.*, vol. COM-32, pp. 315–319, Mar. 1984.

A Large Deviation Theory Proof of the Abstract Alphabet Source Coding Theorem

J. A. BUCKLEW, MEMBER, IEEE

***Abstract* — A modified proof of Berger's abstract alphabet source coding with a fidelity criterion theorem is provided, utilizing a variant of a result from large deviation theory due to Gärtner and Ellis. This proof does not use an asymptotic equipartition property of any kind, thus providing, in our opinion, a new and more direct approach to this key result.**

I. INTRODUCTION

Berger's proof of the abstract alphabet source coding with a fidelity criterion theorem is based upon [1, theorem 7.2.2, pp. 273–278]. This theorem asserts the existence of a code with distortion less than $D + \delta$ and with rate less than $R_1(D) + \delta$ for any $\delta > 0$ where $R_1(D)$ is the first-order approximation to the rate-distortion function $R(D)$.

At the very end of the proof of the theorem, an appeal to a generalized asymptotic equipartition property (AEP), as reported by Perez [2], is made. A counterexample to this generalized AEP has, however, been discovered. It is for this reason that Dunham [3] claims that Berger's proof is technically incorrect. He, however, does indicate that an "information stability" type of AEP result [6] shown by Pinsker *does* hold and immediately provides the needed fix.

In this correspondence we take a somewhat different line in that we modify Berger's proof in such a way that no AEP property of any type is needed. We instead invoke a result from large deviation theory that provides a new and more direct approach to this key theorem. This completes a line of research first started in [7] where we gave a proof of the source coding theorem for a less general case utilizing another large deviation theorem known as Sanov's theorem.

II. PRELIMINARIES

If c is a convex function defined on the real line $\mathscr{R}$, then the left and right derivatives $c'_+(\theta)$ and $c'_-(\theta)$ are well defined for any point θ in the domain $\mathscr{D}_c = \{\theta\colon c(\theta) < \infty\}$. The convex conjugate function of $c(\theta)$ is defined as $c^*(D) = \sup_\theta [\theta D - c(\theta)]$.

Let $c_n(\theta) = n^{-1} \log E\{\exp(\theta W_n)\}$, where $\{W_n\}$ is a sequence of nonnegative random variables. If $c(\theta) = \lim_n c_n(\theta)$ exists for all $\theta \in \mathscr{D}_c$, then $c(\theta)$ is convex and nondecreasing because each $c_n(\theta)$ is. Furthermore, $c(0) = 0$ since $c_n(0) = 0$ for all n. The following result was obtained by specializing Lemma 1.2 of Gärtner [8].

Lemma: Suppose $c(\theta) = \lim_n c_n(\theta)$ is well defined in a neighborhood of $\theta_0 \in \mathscr{R}$ and is differentiable at θ_0 with finite derivative $c'(\theta_0) = D_0$. Then for any neighborhood G of D_0 we have

$$\liminf_n n^{-1} \log P\{W_n/n \in G\} \geq -c^*(D_0).$$

Proof: See the Appendix.

Manuscript received October 23, 1986; revised January 6, 1988. This correspondence was presented at the 1987 IEEE International Symposium on Information Theory, Ann Arbor, MI, October 1987.
The author is with the Department of Electrical and Computer Engineering, University of Wisconsin—Madison, 1415 Johnson Drive, Madison, WI 53706.
IEEE Log Number 8824059.

Ellis [9] used Gärtner's exponential estimates for large deviations of random vectors in $\mathscr{R}^d$ and obtained upper bounds for arbitrary closed sets as well as lower bounds for arbitrary open sets.

III. Proof of the Source Coding Theorem

We adopt the terminology and notation of Berger [1, ch. 7]. Our modification of his proof begins with [1, eq. 7.2.31], at which point it only remains to show that

$$J_n = \int \overline{P}_n(D+\delta|x)\, d^1_\mu(x) \tag{1}$$

vanishes in the limit of large n. Here μ^n is the source probability measure restricted to n-dimensional space and $\overline{P}_n(D+\delta|x) \le [1-\nu^n(\{y\colon \rho_n(x,y)\le D+\delta\})]^{K-1}$, where ν^n is the probability measure that governs the independent selection of each of $K-1$ codewords. The total umber of codewords is K.

If $\rho(\cdot,\cdot)$ is the single letter distortion measure, we define

$$\rho_n(x,y) = n^{-1}\sum_{i=1}^{n}\rho(x_i,y_i).$$

For $\delta>0$ we can write

$$\begin{aligned}\overline{P}_n(D+\delta|x) &\le \left[1-\nu^n(\{y\colon \rho_n(x,y)\le D+\delta\})\right]^{K-1}\\ &\le \exp\left[-(K-1)\nu^n(\{y\colon \rho_n(x,y)\le D+\delta\})\right].\end{aligned} \tag{2}$$

Given a conditional probability $q^1\in Q_1(D)$ such that

$$I(Q)\equiv I(\mu^1,q^1)\le R_1(D)+\delta,$$

let q^n denote the conditional probability corresponding to n successive uses of the memoryless channel q^1, and let $\nu^n=\nu^1\times\nu^1\times\cdots\times\nu^1$, where ν^1 is the output distribution of the channel q^1 with input distribution μ^1. We note that our choice of ν^n is different from that of Berger and Dunham.

Before we proceed with Berger's theorem, we first examine the properties of the following nondecreasing convex function:

$$f(\theta) = \int \log\left[\int \exp[\theta\rho(x,y)]\, d\nu^1(y)\right] d\mu^1(x).$$

Lemma 1: The following lower bound holds for the left derivative $f'_-(0)$:

$$f'(0)\ge D^* = \int \rho(x,y)\, d\left[\nu^1(y)\cdot\mu^1(x)\right].$$

Proof: The left derivative $f'_-(0)$ is well defined since $f(0)=0$ and $f(\theta)$ is convex. Jensen's inequality yields the desired result:

$$\begin{aligned}f'_-(0) &= \lim_{\theta\uparrow 0}\frac{1}{\theta}\int\log\left[\int\exp[\theta\rho(x,y)]\,d\nu^1(y)\right]d\mu^1(x)\\ &\ge \liminf_{\theta\uparrow 0}\frac{1}{\theta}\iint\theta\rho(x,y)\,d\left[\nu^1(y)\cdot\mu^1(x)\right].\end{aligned}$$

Our objective is to construct a code with distortion less than $D+\delta$ and rate less than $R_1(D)+\delta$. We assume that $D<D^*$, since no information needs to be provided about the source and $R_1(D)=0$ if $D\ge D^*$. We further assume that $R_1(D)<\infty$. Since a "δ" of play is allowed in both the distortion and the rate, we may assume without loss of generality that $R_1(D-\epsilon)<\infty$ for some $\epsilon>0$.

Lemma 2: If $R_1(D)<\infty$, then

$$R_1(D)+\delta > I(Q)\ge \sup_{\theta\le 0}[\theta D - f(\theta)].$$

If, furthermore, $D<D^*=\int\rho(x,y)\,d[\nu^1(y)\cdot\mu^1(x)]$, then also

$$\sup_{\theta\le 0}[\theta D-f(\theta)] = \sup_{\theta}[\theta D-f(\theta)] = f^*(D).$$

Proof: A discrete alphabet version of the first assertion is given in [1, theorem 2.5.3, pp. 37–38]. The interested reader can check that the essential ideas carry through to the present setting. The second assertion follows from Lemma 1 since $f(\theta)$ is convex nondecreasing.

We now return to (1) and (2). To complete Berger's proof, we argue that the integrand of the integral defining J_n converges to zero for μ-a.e. x, so that J_n converges to zero by Lebesgue's dominated convergence theorem. Indeed, ν^n is of product form, so by the ergodic theorem,

$$\begin{aligned}&\lim_n n^{-1}\log E_{\nu^n}\left\{\exp\left[\theta\sum_{i=1}^{n}\rho(x_i,y_i)\right]\right\}\\ &\quad=\lim_n n^{-1}\sum_{i=1}^{n}\log\left[\int\exp[\theta\cdot\rho(x_i,y\}]\,d\nu^1(y)\right]\\ &\quad=\int\log\left[\int\exp[\theta\rho(x,y)]\,d\nu^1(y)\right]d\mu^1(x)=f(\theta).\end{aligned}$$

Suppose $D_0<D^*$. If $\theta_0<0$ exists such that $f'(\theta_0)=D_0$, then the large deviation lemma for the one-dimensional random variable $W_n=n\rho_n(x,y)$ and $c(\theta)=f(\theta)$ yields the lower bound

$$\liminf_n n^{-1}\log\nu^n(\{y\colon \rho(x,y)\le D_0+\delta\}) \ge -f^*(D_0) > -[R_1(D_0)+\delta].$$

We see from (2) that $\overline{P}_n(D_0+\delta)\to 0$, and hence $J_n\to 0$ as $n\to\infty$ if

$$\liminf_n n^{-1}\log K\ge R_1(D_0)+\delta.$$

This implies the existence of a code with rate less than $D_0+\delta$ and rate less than $R_1(D_0)+2\delta$.

The proof is slightly more complicated in the remaining case, when θ_0 exists such that D_0 falls between $D_-=f'_-(\theta_0)$ and $D_+=f'_+(\theta_0)$ but $D_-<D_+$. Then $f^*(D)=\sup_\theta[\theta D-f(\theta)]=[\theta_0 D-f(\theta_0)]$ is a straight line (affine) as D ranges over the interval $[D_-,D_+]$. There exists θ_- and θ_+ arbitrarily close to θ_0 such that $f'(\theta_-)$ and $f'(\theta_+)$ are arbitrarily close to $D_-=f'_-(\theta_0)$ and $D_+=f'_+(\theta_0)$, respectively, and by the previous argument there exists a code with distortion less than $D_-+\delta$ and rate less than $f'(\theta_-)+\delta$ and another code with distortion less than $D_++\delta$ and rate less than $f'(\theta_+)+\delta$. We time share these codes with time ratios $\lambda_-\ge 0$ and $\lambda_+\ge 0$ such that $\lambda_-+\lambda_+=1$ and $\lambda_-D_-+\lambda_+D_+=D_0$ and obtain a code with distortion less than $D_0+\delta$ and rate less than $\lambda_-f^*(D_-)+\lambda_+f^*(D_+)+\delta=f^*(D_0)+\delta\le R_1(D_0)+2\delta$.

IV. Discussion

The choice of codeword probability measure ν^n being of product form greatly simplified the calculation in verifying that the analog of $\lim_{n\to\infty}c_n(\theta)=c(\theta)$ holds. In the Berger/Dunham proof ν^n is not of product form. It was this fact that necessitated the use of an asymptotic equipartition property to complete the proof. In an earlier paper [7], we explicitly gave yet another proof for discrete memoryless sources and indicated how to extend it (to Polish space-valued alphabets and continuous distortion measures). This proof relied heavily on a theorem concerning the rates of convergence of empirical distributions, known as Sanov's theorem. In that paper we expressed the hope that large deviation theory can supply a proof for the abstract alphabet case. We expect that further interrelationships among the information theory theorems, large deviation theory, and Shannon–McMillan theorems will be discovered to the mutual benefit of these and other areas.

Acknowledgment

The author would like to thank Peter Ney of the University of Wisconsin, Madison and John Sadowsky of Purdue University, West Lafayette, IN, for several conversations on the interrelationships between large deviation theory and convex analysis. The author would like to thank Toby Berger of Cornell University, Ithaca, NY, for pointing out the simple proof of Lemma 1 given in the text. An anonymous reviewer should also be acknowledged for numerous suggestions resulting in the present form of this paper.

Appendix
Proof of Large Deviation Lemma

Let $Q_n(dD)$ denote the distribution of W_n/n on $\mathbf{R}$, and define the "twisted" measures

$$Q_{n,\theta}(dD) = \exp[n\theta D]\cdot Q_n(dD)/\exp[nc_n(\theta)], \quad n=1,2,\cdots.$$

As D ranges over a neighborhood $G_\epsilon = (D_0 - \epsilon, \delta_0 + \epsilon)$ of D_0, then $-\theta D \geq -\theta D_0 - |\theta|\epsilon$, and, consequently, if $G_\epsilon \subset G$ we have

$$Q_n(G) \geq Q_n(G_\epsilon) = \exp[nc_n(\theta)]\cdot\int_{G_\epsilon}\exp[-n\theta D]Q_{n,\theta}(dD)$$
$$\geq \exp[n(c_n(\theta) - \theta D_0 - |\theta|\epsilon)]\cdot Q_{n,\theta}(G_\epsilon).$$

It follows (letting $\theta = \theta_0$) that

$$\liminf_n \frac{1}{n}\log Q_n(G) \geq c(\theta_0) - \theta_0 D_0 - |\theta_0|\epsilon + \liminf_n \frac{1}{n}\log Q_{n,\theta_0}(G_\epsilon^n).$$

Note that $c(\theta_0) - \theta_0 D_0 = -c^*(D_0)$ since $c(\theta)$ is differentiable at θ_0 with finite derivative $c'(\theta_0) = D_0$. If $Q_{n,\theta_0}(G_\epsilon) \to 1$ as $n \to \infty$, the lemma follows by letting $\epsilon \to 0$.

It remains to show that $Q_{n,\theta_0}(G_\epsilon) \to 1$ as $n \to \infty$. Introduce random variables V_n such that V_n/n has distribution $Q_{n,\theta_0}(dD)$. Note also for any random variable Z and $t \geq 0$ we have $P\{Z \geq z\} \leq E\{\exp[t(Z-z)]\} = \exp[-tz]\cdot E\{\exp[tZ]\}$. Hence

$$P\{V_n/n \geq D_0 + \epsilon\} \leq E\{\exp[nt(V_n/n - D_0 - \epsilon)]\}$$
$$= \int \exp[nt(V_n/n - D_0 - \epsilon)] \cdot\exp[n\theta_0 D]Q_n(dD)/\exp[nc_n(\theta)]$$
$$= \exp[n(c_n(\theta_0 + t) - c_n(\theta_0) - t(D_0 + \epsilon))].$$

Hence

$$\limsup_n \frac{1}{n}\log P\{V_n/n \geq D_0 + \epsilon\} \leq c(\theta_0 + t) - c(\theta_0) - t(D_0 + \epsilon).$$

The right side is strictly negative for sufficiently small t since $[c(\theta_0 + t) - c(\theta_0)]/t \to D_0$ as $t \to 0$. Similarly,

$$\limsup_n \frac{1}{n}\log P\{V_n/n \leq D_0 - \epsilon\} \leq c(\theta_0 - t) - c(\theta_0) + t(D_0 - \epsilon)$$

which, again, is strictly negative for small t. The two bounds imply that $1 - Q_{n,\theta_0}(G_\epsilon)$ vanishes exponentially fast and hence $Q_{n,\theta_0}(G_\epsilon) \to 1$ as $n \to \infty$.

References

[1] T. Berger, *Rate Distortion Theory, A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] A. Perez, "Extensions of Shannon–McMillan's limit theorem to more general stochastic processes," in *Trans. 3rd Prague Conf. Information Theory, Statistical Decision Functions, and Random Processes*. Prague, Czech.: Publ. House Czech. Acad. Sci., and New York: Academic, pp. 545–574.

[3] J. Dunham, "A note on the abstract-alphabet block source coding with a fidelity criterion theorem," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 6, p. 760, Nov. 1978.

[4] J. Kieffer, "A counterexample to Perez's generalization of the Shannon-McMillan theorem," *Ann. Prob.*, vol. 1, no. 2, pp. 362–364, 1973.

[5] ——, "Correction to 'A counterexample of Perez's generalization of the Shannon-McMillan theorem,'" *Ann. Prob.*, vol. 4, no. 1, pp. 153–154, 1976.

[6] M. S. Pinsker, "Sources of messages," *Probl. Peredach. Inform.*, vol. 14, pp. 5–20, 1963 (in Russian).

[7] J. A. Bucklew, "The source coding theorem via Sanov's theorem," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 907–909, Nov. 1987.

[8] J. Gärtner, "On large deviations from the invariant measure," *Theory Probl. Appl.*, vol. 22, pp. 24–39.

[9] R. S. Ellis, "Large deviations for a general class of random vectors," *Ann. Prob.*, vol. 12, no. 1, pp. 1–12, 1984.

Finite-State Codes

FABRIZIO POLLARA, MEMBER, IEEE, ROBERT J. MCELIECE, FELLOW, IEEE, AND KHALED ABDEL-GHAFFAR

***Abstract*—We define and investigate a class of codes called *finite-state* (FS) codes. These codes, which generalize both block and convolutional codes, are defined by their encoders, which are finite-state machines with parallel inputs and outputs. We derive a family of upper bounds on the free distance of a given FS code, based on known upper bounds on the minimum distance of block codes. We then give a general construction for FS codes, based on some recent ideas of Ungerboeck, and show that in many cases the FS codes constructed in this way have a d_{free} which is the largest possible. We also discuss the issue of catastrophic error propagation (CEP) for FS codes and discover that to avoid CEP we must solve an interesting problem in graph theory, that of finding a "uniquely decodable edge labeling" of the state diagram.**

I. Introduction

We begin with a description of what we shall call a *finite-state (FS) encoder*. An (n, k, m) FS encoder is a q^m-state (time-invariant) finite-state machine (FSM) with k parallel inputs and n parallel outputs taken from a q-letter alphabet (Fig. 1).

The encoder starts from a fixed initial state. At each clock pulse, k symbols (the information symbols) are input to the encoder, and in response the encoder changes state and outputs n symbols (the code symbols). Thus if $(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots)$ is a sequence of k-symbol information blocks, then the encoder's output will be a sequence $(\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots)$ of n-symbol code blocks, which we call a *code sequence*. The set of all such code sequences is called the code generated by the FS encoder. A code generated by a (n, k, m) FS encoder will be called an (n, k, m) finite-state code. We note that if only one state exists in the encoder, the resulting $(n, k, 0)$ FS code is in fact an ordinary block code. Similarly, a linear convolutional code is just an FS code in which the finite-

Manuscript received April 9, 1987; revised December 21, 1987. This work was supported in part by the National Aeronautics and Space Administration, and by Grants from IBM and Pacific Bell.

F. Pollara is with the Jet Propulsion Laboratory, 238-420, 4800 Oak Grove Drive, Pasadena, CA 91109.

R. J. McEliece is with the Department of Electrical Engineering, 116-81, California Institute of Technology, Pasadena, CA 91125.

K. Abdel-Ghaffar was with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA. He is now with the IBM Almaden Research Center, San Jose, CA 95120.

IEEE Log Number 8824057.