INFORMATION SCIENCES **42**, 239–253 (1987)

Metric Bounds on Losses in Adaptive Coding

J. DAVID FLICK

PETER SALAMON

and

BJARNE ANDRESEN*

Department of Mathematical Sciences, San Diego State University, San Diego, California 92182

ABSTRACT

A noiseless-channel coding system in which the source probabilities change continuously in time is introduced. Using a geometry defined by the second derivative matrix of the information-theoretic entropy function, a bound on the number of excess channel symbols (the redundancy) sent by a retrospective encoder during a fixed time interval is derived. Results on minimal-cost coding and their implications for a real-time coding algorithm are then explored. Underlying most of the discussion here is the assumption made on the separability of the time scales corresponding to the (high) character rate as compared with the rate of change of the source probabilities.

I. INTRODUCTION

A noiseless-channel coding system in which the source symbol probabilities vary continuously with time is examined. The source-to-channel encoder is designed to adapt to the current source probabilities with the objective of minimizing the redundancy, i.e. the expected extra number of channel symbols sent across the channel (relative to the minimum possible), in a given interval of time. By analogy with physics, we will refer to these extra bits transmitted as "dissipation." This adaptation must take place in order for dissipation to be minimized, since the optimal code at a given time is dependent on the current source probabilities.

*Permanent address: Physics Laboratory, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

52 Vanderbilt Ave., New York, NY 10017. 0020-0255/87/$03.50

We first assume that the encoder has complete foreknowledge of the distribution of source characters at each instant of the time interval $[0, \tau]$. The problem of optimal coding in the absence of information about the future and an associated algorithm follow in Section V. We assume throughout that the rate of "drift" in the source probabilities is slow, i.e., the time scale for determining the instantaneous probabilities used to generate the source symbols is much smaller than the time scale on which the source symbol probabilities change. This assumption is valid in the limit as the rate of transmission r goes to infinity.

We determine the optimal codes and code changes subject to the above assumptions in three stages. Initially, we assume that K code changes occur at given times $t_1, \ldots, t_K$ and optimize the choice of codes for each of the K time intervals. Next we optimize the choice of times $t_2, \ldots, t_K$ at which to change codes. Finally we optimize the choice of the number K of code changes. Unless we assume there is a nonzero cost of changing codes, the optimal choice of K is simply the number of symbols sent, i.e. change codes to be optimal at each instant. We therefore assume a nonzero cost per code change.

A central concept in the development of the bounds on dissipation presented here is the geometry we use on the set of states of the source, i.e., on the set of probability distributions over the source symbols. The geometry is Riemannian and uses the second-derivative matrix of the information-theoretic entropy as metric matrix. This metric was originally introduced by R. A. Fisher [2, 4, 7], and the associated distances can be interpreted as the number of distinguishable intermediate distributions [7]. This distance is therefore natural to the problem at hand, which concerns the time evolution of the source and the minimal extra redundancy required to code in the presence of such evolution.

The problem and the methodologies considered here were motivated by recent results on the minimum dissipation in thermodynamic processes in finite time [5]. In these physical processes the dissipation is measured by loss of available work or by production of entropy, rather than extra bits transmitted as in the present example. In fact a similar geometrical structure can be defined on the state space of any model in which a system shows optimizing behavior. The problem of minimizing deviations from optimality inevitably relates such deviations to the distance traversed in a geometry defined by the second derivative of the objective function. Besides the thermodynamic example cited above and the example of adaptive coding discussed in the present paper, an economic example has also been analyzed [6].

Section II will introduce notation and concepts associated with our adaptive-coding model. Section III will examine the geometry on the states of the source defined by the second derivative of the entropy function. Section IV derives a lower bound on the dissipation for the adaptive-coding model equipped with a retrospective encoder, i.e. with perfect information on how the

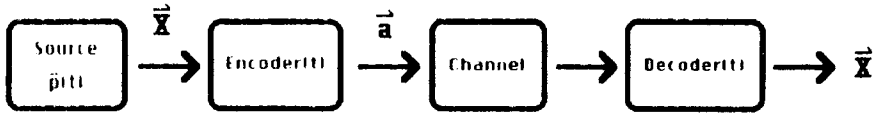

Fig. 1. The communication system.

source is to evolve probabilistically. Implications of the results of Section IV for implementing an algorithm to run in "real time" are presented in Section V. A summary is given in Section VI.

II. ADAPTIVE-CODING MODEL

Let $\mathbf{X}^T = (x_1, \ldots, x_m)$ be the set of characters of the source in a communication system (Figure 1). A character is generated at discrete times t_v (corresponding to some fixed rate r) according to the probability vector $\mathbf{p}(t_v)^T = (p_1(t_v), \ldots, p_m(t_v))$, which we will refer to as the state of the source. It is assumed that distinct trials (characters) from the source are independent. The goal of the adaptive encoder is to transmit the information represented by these characters across a noiseless channel with the channel alphabet denoted by $\mathbf{A}^T = (a_1, \ldots, a_D)$. It is well known [1] that for a fixed set of probabilities $\mathbf{p}^T = (p_1, \ldots, p_m)$ for the source symbols, the expected codeword length for a uniquely decipherable code is bounded below by

$$H_D(\mathbf{p}) = -\sum_{i=1}^{m} p_i \log_D p_i. \tag{1}$$

(Henceforth, log functions with an unspecified base will be understood to be to the base D, and we will drop the subscript D on H_D.) Furthermore, by using extensions of the source, i.e. by coding blocks of source symbols, this lower bound may be approached arbitrarily closely by increasing the length of the blocks. Thus if we assume the alphabet $\mathbf{A}$ actually represents blocks of some base set of characters which are long enough to make the lower bound a good approximation of the expected codeword length, then

$$E(\text{number of channel symbols sent in } [0, \tau])$$

$$= \sum_{v=1}^{r\tau} H(\mathbf{p}(t_v)) \approx r \int_0^{\tau} H(\mathbf{p}(t))\, dt, \tag{2}$$

where E denotes the expected value relative to the probability space of strings

of source symbols on $[0, \tau]$. Note that the integral approximation above is accurate provided the rate of source symbols, r, is large relative to the magnitude of the rate of change of $\mathbf{p}$, i.e., $r \gg |\dot{\mathbf{p}}(t)|$ for all t. If $\mathbf{p}(t)$ is changing significantly with time, Equation (2) would hold only if the code were changed to the new optimal [1] code consistent with the current probabilities at each step. If there are codewords sent where the code is not constructed for the actual $\mathbf{p}(t)$ at that time, then there will be an expected number of channel symbols sent in excess of the expression in (2). We define this quantity to be the *dissipation* in the coding process. If these extra channel symbols are used for error detection, this excess is known as redundancy. We shall explore the minimum value of this dissipation after introducing the requisite geometrical constructs.

III. GEOMETRY ON THE SET OF SOURCE STATES

It is easily verified that the second-derivative matrix of the state function $H(\mathbf{p})$ of the state $\mathbf{p}$ of the source is given by

$$D^2[H(\mathbf{p})] = \frac{-1}{\ln D}\begin{bmatrix} 1/p_1 & & & 0 \\ & 1/p_2 & & \\ & & \ddots & \\ 0 & & & 1/p_m \end{bmatrix}. \tag{3}$$

The negative of this matrix is both positive definite and symmetric and can be used to define a Riemannian metric on the set of states of the source. The length of a path $\mathbf{p}(t)$, $t \in [0, \tau]$, describing the time evolution of a source is calculated relative to the metric $-D^2[H(\mathbf{p})]$ as

$$L = \int_0^\tau \sqrt{-[\dot{\mathbf{p}}(t)]^T D^2 H(\mathbf{p}(t))\dot{\mathbf{p}}(t)}\, dt. \tag{4}$$

Note that from its very form the length of a path defined in this way is independent of parametrization. If the parameter is taken to represent time (as we have done), independence of parametrization means in effect that the length of a path is independent of how fast the path is traversed, i.e. length is a purely geometric notion.

The geometry that this length defines on the set of source states will be exploited in deriving a lower bound on the dissipation in the proposed coding system. It is rather remarkable that a metric defined on the set of source states can provide information concerning coding losses for a time-varying source. It

is no less remarkable that defining lengths of paths for thermodynamic or economic time evolutions in a similar fashion yields information about inherent "losses" in these models [5, 6]. In short, geometries of this type appear to provide a general tool useful in bounding losses in any process in which an environment leads a system which is attempting to "keep close," i.e. remain as near to equilibrium with its environment as possible.

IV. MINIMUM DISSIPATION FOR A RETROSPECTIVE ENCODER

We are now in a position to consider the central problems of this paper, namely minimal-dissipation and minimal-cost coding for an encoder blessed with complete foreknowledge of how $\mathbf{p}(t)$ is to evolve over a fixed interval of time $[0, \tau]$. As mentioned previously, we shall proceed to optimize in sequential fashion:

(i) the code for an arbitrary fixed subinterval (t_j, t_{j+1}),
(ii) the times t_j $(j = 2, \ldots, K)$ at which code changes occur,
(iii) the number K of code changes in $[0, \tau] = [t_1, t_{K+1}]$.

A. OPTIMUM CODE

By n_i^j we shall mean the length of the codeword corresponding to source symbol x_i $(i = 1, \ldots, m)$ in the code used throughout the jth subinterval, (t_j, t_{j+1}). The total dissipation for the time interval $[0, \tau]$ may then be expressed as

$$R = r \int_0^\tau \sum_{i-1}^{m} \left[p_i(t)\, n_i(t) + p_i(t) \log p_i(t) \right] dt \tag{5}$$

where

$$n_i(t) = n_i^j \qquad \text{for} \quad t_j < t < t_{j+1},$$

i.e. the expected minus the optimal expected codeword length at each instant integrated over the coding interval. Note that the term above representing the optimal expected codeword length is independent of the choice of n_i^j and thus has no effect on the choice of n_i^j which will minimize total dissipation R.

Suppose that code changes are to occur at times t_j and t_{j+1} and at no time in between. Then the length n_i^j $(i = 1, \ldots, m)$ of each codeword must be chosen to minimize the expression

$$r \int_{t_j}^{t_{j+1}} \sum_{i=1}^{m} \left[p_i(t)\, n_i^j \right] dt = r \sum_{i=1}^{m} n_i^j \int_{t_j}^{t_{j+1}} p_i(t)\, dt \tag{6}$$

under the restriction that the n_i^j satisfy the condition for the existence of a uniquely decipherable code [1], namely

$$\sum_{i=1}^{m} D^{-n_i^j} \leqslant 1. \tag{7}$$

We shall ignore the fact that the n_i^j represent integers and that equality in (7) is generally not possible. In so doing we establish a minimum dissipation for any real values of n_i^j which is actually a lower bound on that obtainable using integer n_i^j. This lower bound is assured to be quite tight, as already mentioned, if we interpret the "source" symbols to be blocks of symbols of an actual single-character source. Furthermore, the plausibility of this interpretation rests solely on the assumption of separation of time scales already stated, i.e. on the assumption that r is large compared to $|\dot{\mathbf{p}}|$.

With equality in (7) assumed, we form the Lagrangian

$$L(\mathbf{n}^j, \lambda) = r \sum_{i=1}^{m} n_i^j p_i^j + \lambda \left(\sum_{i=1}^{m} D^{-n_i^j} - 1 \right), \tag{8}$$

where we have introduced the notation

$$p_i^j = \int_{t_j}^{t_{j+1}} p_i(t)\, dt = \bar{p}_i^j \, \Delta t_j \tag{9}$$

with $\Delta t_j = t_{j+1} - t_j$. Equation (8) leads to the optimality conditions on n_i^j,

$$n_i^j = -\log \bar{p}_i^j \qquad (i = 1, \ldots, m). \tag{10}$$

The optimality conditions in (10) should hold for each subinterval j ($j = 1, \ldots, K$). In words, the length of the ith codeword for the optimal code on the jth subinterval is minus the logarithm of the average value of $p_i(t)$ on the subinterval.

Substituting the optimal n_i^j into (5) yields

$$R = r \left[\sum_{j=1}^{K} \sum_{i=1}^{m} p_i^j \left(-\log \bar{p}_i^j \right) - \int_0^{\tau} H(\mathbf{p}(t))\, dt \right]. \tag{11}$$

B. OPTIMAL SWITCH TIMES

Equation (11) is now the expression we wish to minimize by choosing the $K-1$ times $t_2,\ldots,t_K$ at which to change the code. We begin by taking the derivatives with respect to one of these times, t_l $l=2,\ldots,K$. Note that the integral in (11) is independent of the switch times:

$$\frac{\partial}{\partial t_l}[R] = \frac{\partial}{\partial t_l}\left[-r\sum_{j=1}^{K}\sum_{i=1}^{m}\int_{t_j}^{t_{j+1}} p_i(t)\log \bar{p}_i^{\,j}\,dt\right]$$

$$= -r\sum_{i=1}^{m} p_i(t_l)\log\left(\frac{\bar{p}_i^{\,l-1}}{\bar{p}_i^{\,l}}\right) \qquad (l=2,\ldots,K). \tag{12}$$

Derivatives of the $\log \bar{p}_i^{\,j}$ factors cancel, since $\sum_i \bar{p}_i^{\,j}=1$. Setting the terms of Equation (12) equal to zero, we find the optimality conditions on the t_l to be

$$\sum_{i=1}^{m} p_i(t_l)\log\left(\frac{p_i(t_l)}{\bar{p}_i^{\,l-1}}\right) = \sum_{i=1}^{m} p_i(t_l)\log\left(\frac{p_i(t_l)}{\bar{p}_i^{\,l}}\right). \tag{13}$$

Equation (13) has the following interpretation in the language of information theory: The time t_l at which to change codes must be such that the mean surprisals of distribution $\mathbf{p}(t_l)$ relative to the average distributions over (t_{l-1}, t_l) and (t_l, t_{l+1}) are equal. [It is easy to see that the $\bar{p}_i^{\,l}$ $(i=1,\ldots,m)$ form a probability distribution.]

Examining (13) further, we expand the logarithms as power series and neglect terms higher than second order in $\Delta p_i^l = \bar{p}_i^{\,l} - p_i(t_l)$, i.e., we assume that $\mathbf{p}(t)$ evolves only slightly in relative terms during each interval:

$$-\sum_{i=1}^{m} p_i(t_l)\left[\frac{\Delta p_i^{l-1}}{p_i(t_l)} - \frac{1}{2}\left(\frac{\Delta p_i^{l-1}}{p_i(t_l)}\right)^2\right]$$

$$= -\sum_{i=1}^{m} p_i(t_l)\left[\frac{\Delta p_i^{l}}{p_i(t_l)} - \frac{1}{2}\left(\frac{\Delta p_i^{l}}{p_i(t_l)}\right)^2\right]. \tag{14}$$

Here the first-order terms sum to zero and we are left with

$$\sum_{i=1}^{m}\left[\frac{(\Delta p_i^{l-1})^2}{p_i(t_l)}\right] = \sum_{i=1}^{m}\left[\frac{(\Delta p_i^{l})^2}{p_i(t_l)}\right]. \tag{15}$$

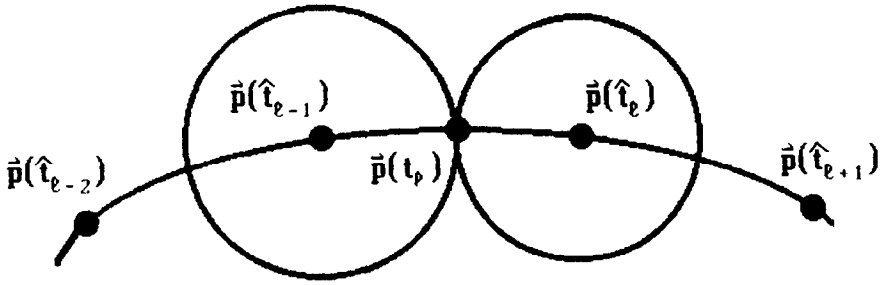

Fig. 2. Optimal switch times should be at points equidistant from the midpoints of previous and upcoming intervals.

Recalling the discussion of Section III, this assumes the form

$$\left(\mathbf{p}(t_l)-\bar{\mathbf{p}}^{l-1}\right)^T\cdot\left(\mathbf{p}(t_l)-\bar{\mathbf{p}}^{l-1}\right)=\left(\mathbf{p}(t_l)-\bar{\mathbf{p}}^{l}\right)^T\cdot\left(\mathbf{p}(t_l)-\bar{\mathbf{p}}^{l}\right) \tag{16}$$

where the dot product is taken relative to the $-D^2H$ metric [Equation (3)] evaluated at $\mathbf{p}(t_l)$. This simply means that for each l, the optimal $\mathbf{p}(t_l)$ will be equidistant, relative to the D^2H metric, from the average $\bar{\mathbf{p}}(t)$ over the $(l-1)$th and lth subintervals (Figure 2).

Our treatment of the optimal-adaptive-coding problem relies heavily on the separability of the time scales. This assumption guarantees that each of the K segments of the path $\mathbf{p}(t)$ is small and justifies our ignoring terms higher than second order in the expansion of (13). To continue with the third step of our sequential optimization, we will need an expression for the minimum dissipation in the lth time interval. Relying once again on the separability of time scales, we approximate the piece of the path $\mathbf{p}(t)$, $t\in(t_l,t_{l+1})$, by a line segment,

$$\begin{aligned}\mathbf{p}(t)&=\mathbf{p}(\hat{t}_l)+\frac{\mathbf{p}(t_{l+1})-\mathbf{p}(t_l)}{\Delta t_l}(t-\hat{t}_l)\\&=\boldsymbol{\alpha}_l+\boldsymbol{\beta}_l(t-\hat{t}_l)\end{aligned} \tag{17}$$

where $\hat{t}_l=(t_l+t_{l+1})/2$ is the midpoint of the lth interval. It is then easy to verify that if $\mathbf{p}(t)$ is linear on a subinterval (t_l,t_{l+1}) then

$$\bar{\mathbf{p}}^{l}=\mathbf{p}(\hat{t}_l)=\boldsymbol{\alpha}_l, \tag{18}$$

so that from (10), optimally the encoder should base the code for a linear

subinterval on the value of $\mathbf{p}(t)$ at the midpoint $\mathbf{p}(\hat{t})$ of the subinterval. With this approximation, the dissipation for the lth subinterval is [cf. equation (11)]

$$R_l = r\int_{t_l}^{t_{l+1}} \sum_{i=1}^{m} p_i(t) \log\left(\frac{p_i(t)}{p_i(\hat{t}_l)}\right) dt$$

$$= r\int_{t_l}^{t_{l+1}} \sum_{i=1}^{m} \left[\alpha_i + \beta_i(t-\hat{t})\right] \log\left[\frac{\alpha_i + \beta_i(t-\hat{t})}{\alpha_i}\right] dt, \tag{19}$$

where all quantities ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\hat{t}$) refer to the lth interval. A second-order expansion of the log terms yields

$$R_l = \frac{r}{\ln D} \sum_{i=1}^{m} \frac{\beta_i^2}{\alpha_i} \frac{\Delta t_l^3}{24}. \tag{20}$$

C. *GEOMETRIC BOUND*

We now compare (18) with the length of the subinterval measured in the D^2H metric:

$$\Delta L_l = \int_{t_l}^{t_{l+1}} \sqrt{\frac{1}{\ln D}} \sqrt{\boldsymbol{\beta}^T \begin{pmatrix} 1/\alpha_i & & 0 \\ & \ddots & \\ 0 & & 1/\alpha_m \end{pmatrix} \boldsymbol{\beta}}\, dt$$

$$= \sqrt{\frac{1}{\ln D}} \sqrt{\sum_{i=1}^{m} \frac{\beta_i^2}{\alpha_i}}\, \Delta t_l. \tag{21}$$

Thus from (20), (21)

$$R_l = \frac{r}{24} (\Delta L_l)^2 \Delta t_l. \tag{22}$$

The total dissipation for a sequence of linear subintervals is

$$R = \sum_{j=1}^{K} R_j = \frac{r}{24} \sum_{j=1}^{K} (\Delta L_j)^2 \Delta t_j$$

$$= \frac{r\overline{\Delta t}}{24} \sum_{j=1}^{K} (\Delta L_j)^2, \tag{23}$$

where

$$\overline{\Delta t} = \frac{\sum_{j=1}^{K+1} (\Delta L_j)^2 \, \Delta t_j}{\sum_{j=1}^{K+1} (\Delta L_j)^2} \tag{24}$$

is the average length of time of the subintervals, each subinterval weighted by its length in the D^2H metric. Using the Cauchy-Schwarz inequality in the form $\sum_j a_j b_j \leqslant (\sum_j a_j^2)^{1/2} (\sum_j b_j^2)^{1/2}$ with $b_j = 1$ for all j and $a_j = \Delta L_j$, we have in (23)

$$R \geqslant \frac{r\overline{\Delta t}}{24K} \left(\sum_{j=1}^{K} \Delta L_j \right)^2 = \frac{r\overline{\Delta t}}{24K} L^2, \tag{25}$$

where L is the total length of $\mathbf{p}(t)$ on $[0, \tau]$.

The inequality in (25) represents a very important intermediate step, and we take a moment here to explore its significance. We have arrived at a general lower bound on the dissipation over an interval of time for the proposed encoder. The terms in this lower bound include the geometric quantity L, the average length of time of the subintervals $\overline{\Delta t}$, and the number K of code changes.

D. OPTIMUM NUMBER OF CODE CHANGES

The last task before us is to optimize the number K of code changes for an optimal encoder. An optimal encoder for us will mean that expression (25) holds with equality. If, for a given optimal encoder, C_1 is the (constant) cost per code change and similarly C_2 is the cost per dissipated channel symbol, then the total cost of operating the encoder over $[0, \tau]$ is

$$C(K) = C_1 K + C_2 \frac{r\overline{\Delta t}}{24K} L^2 + (\text{minimal cost}), \tag{26}$$

where the minimal cost is independent of K. Note that if the ΔL_j are all comparable, i.e. if the drift rate of the source is nearly constant, then to a good approximation

$$\overline{\Delta t} = \frac{\tau}{K}, \tag{27}$$

and (26) takes the simpler form

$$C(K) = C_1 K + C_2 \frac{r\tau L^2}{24K^2} + (\text{minimal cost}), \tag{28}$$

which by setting its first derivative to zero can be shown to have a minimum at

$$K = \sqrt[3]{\frac{C_2}{C_1}\frac{r\tau L^2}{12}} \,. \tag{29}$$

This then is the expression for the number of times the optimal encoder should switch codes (including establishing the code for the first interval).

Using this value of K, we now go back to the expression (22) for the minimum dissipation in a linear step. We make the approximation

$$\Delta L_j = \frac{L}{K}, \qquad j = 1, \ldots, K, \tag{30}$$

which has been justified in (13) and (15). Then (22) becomes

$$\begin{aligned} R_l &= \frac{r}{24}\left(\frac{L}{K}\right)^2 \frac{\tau}{K} \\ &= \frac{rL^2\tau}{24K^3} \\ &= \frac{C_1}{2C_2}, \end{aligned} \tag{31}$$

which states that the optimal encoder should change codes whenever the accumulated cost of transmitting extra bits since the previous code change equals one-half the cost of changing codes.

Finally, the expression for the minimum total cost of encoding $\mathbf{p}(t)$ on the interval $[0, \tau]$ is found using (28), (29), and (31):

$$\begin{aligned} C &= C_1 K + C_2 \sum_{l=1}^{K} R_l + (\text{minimum cost}) \\ &= C_1 K + C_2 \frac{KC_1}{2C_2} + (\text{minimum cost}) \\ &= \tfrac{3}{2} C_1 K + r \int_0^{\tau} H(\mathbf{p}(t))\, dt \end{aligned} \tag{32}$$

for optimal K as in (29).

E. EXAMPLE

The following simple example serves to illustrate the ideas presented here. Let $\mathbf{p}(t)^T = (\sin^2 t, \cos^2 t)$, $t \in [0, \tau]$. Then the length of the path in the D^2H metric is

$$L = \int_0^\tau \left[(2\sin t\cos t, -2\sin t\cos t) \begin{pmatrix} \dfrac{1}{\sin^2 t} & 0 \\ 0 & \dfrac{1}{\cos^2 t} \end{pmatrix} \begin{pmatrix} 2\sin t\cos t \\ -2\sin t\cos t \end{pmatrix} \right]^{1/2} dt$$

$$= \int_0^\tau \sqrt{2\cos^2 t + 2\sin^2 t}\, dt = \sqrt{2}\,\tau. \tag{33}$$

Suppose that $r = 30{,}000$ characters per second to be sent through the channel and that $C_1 = 5$ cents per code change and $C_2 = 1$ cent per dissipated bit. Then the minimum dissipation for the encoder is, from (25),

$$R = \frac{r\tau L^2}{24K^2} = \frac{60{,}000\tau^3}{24K^2}. \tag{34}$$

The optimal choice of K is, from (29)

$$K = \sqrt[3]{\frac{1}{5}\left(\frac{r\tau L^2}{12}\right)} = 10\tau, \tag{35}$$

leading to a minimum dissipation of $R = 25\tau$ bits and a cost of 0.75τ dollars.

The cost of deviating from this optimum can be severe. If a single code is used for the entire duration ($K = 1$), the minimum dissipation is $R = 2500\tau^3$ bits, corresponding to a cost of $25\tau^3$ dollars. On the other hand, if the code is adjusted continuously ($K \to \infty$), the dissipation vanishes but the cost of the code changes is very large. Even if we make the right number of code changes, we can lose significantly by not dividing the interval evenly. Consider for example using $K_1 = 10\tau/3$ from 0 to $L/2$ and $K_2 = 20\tau/3$ from $L/2$ to L. Even if the allocation within these subintervals is otherwise optimal, we end up with a dissipation of $R = 35.16\tau$ bits for a total cost of 0.85τ dollars.

V. IMPLICATION FOR REAL-TIME ALGORITHMS

The results of the preceding section for the optimal retrospective encoder actually have important implications regarding the implementation of minimal cost encoders which are required to work in "real time." We require a real-time

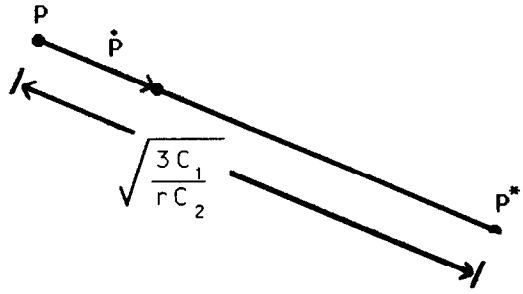

Fig. 3. At a code change, the new code should be a perfect match for an extrapolated state of the source.

encoder to encode characters immediately upon reception from the source, say at time t_1, with knowledge of how $\mathbf{p}(t)$ behaves only for values $t < t_1$. We assume that the time-scale separability is as before and that the real-time encoder has constant fixed unit costs C_1 and C_2 for code changes and characters respectively. Then with (22) and (31) in mind, we can approximate retrospective behavior with real-time behavior in the following fashion.

1. Choose the next code at time t_1 to correspond to the distribution $\mathbf{p}_1^*$ obtained (Figure 3) by moving from $\mathbf{p}(t_1)$ in the direction $\dot{\mathbf{p}}(t_1)$ a distance $\sqrt{3C_1/rC_2}$, i.e.,

$$\mathbf{p}_1^* = \mathbf{p}(t_1) + \frac{\dot{\mathbf{p}}(t_1)}{|\dot{\mathbf{p}}(t_1)|}\sqrt{\frac{3C_1}{rC_2}}. \tag{36}$$

Equivalently, the ith codeword is chosen to have length

$$n_i = -\log\left[p_i(t_1) + \frac{\dot{p}_i(t_1)\sqrt{3C_1/rC_2}}{\sqrt{\sum_{v=1}^{m}\dot{p}_v(t_1)^2/p_v(t_1)}}\right]. \tag{37}$$

2. This code is then retained until such time t_2 as the R calculated using (5) on the interval $[t_1, t_2]$ reaches a value of $C_1/2C_2$. At this time we return to step 1 with $t_1 = t_2$.

This algorithm predicts $\mathbf{p}(t)$ linearly, but any other extrapolation scheme, e.g. quadratic, can be easily accommodated. Step 1 of the algorithm should be changed to move along the extrapolating curve a distance $\sqrt{3C_1/rC_2}$, while step 2 remains unchanged. The algorithm achieves the result of Section IV asymptotically as $r \to \infty$.

We note, finally, that it may be necessary to consider problems with comparable rather than separable time scales. To treat this situation well, the analysis presented here must change drastically. The encoder becomes dependent on statistical estimates of $\mathbf{p}(t)$, so the central concern becomes the tradeoff between sample size [i.e. number of characters observed to estimate $\mathbf{p}(t)$] and the "currentness" of the estimate.

The algorithm presented above can however be slightly modified in an ad hoc fashion to work reasonably well even if the separability of the time scales is far from perfect. In this case we should choose the next $\mathbf{p}^*$ along the extrapolation from $\mathbf{p}(t_1)$ but not "as far." This will reduce losses due to $\mathbf{p}(t)$ following unpredictable paths and not add too much extra dissipation if $\mathbf{p}(t)$ does follow the expected path. This ad hoc algorithm could replace the $\sqrt{3}$ in (36) by a parameter. The parameter should be chosen smaller as the separability gets poorer, and if a historical record of the data is available, it could be optimized empirically.

The criterion for changing codes outlined above is similar to known results for related problems of optimal economic time evolution [3], e.g. optimal portfolio management. In this case there exists a (constant) cost C_1 of changing the allocation of a fixed sum of invested capital. There is also a cost C_2 associated with suboptimality of the current portfolio with respect to the current market conditions and the individual's utility function.

VI. CONCLUSION

We have shown that lengths of paths in a natural geometry on the space of states of the source appear in an expression bounding from below the extra redundancy required of a retrospective encoder due to the fact that the source signal probabilities evolve with time. Prices were introduced for each bit transmitted and for each code change, and the above expression was used to find the minimal-cost coding. Results for the retrospective encoder were then shown to correspond to a natural real-time algorithm for encoding, and it was noted that this algorithm is similar to known results on optimal economic time evolution. The importance of the separability of the time scales (large r) should be recognized.

We thank J. Nulton, T. Feldmann, and J. Robinson for helpful conversations, and the Telluride Summer Research Center for providing a critical audience. Acknowledgement is made to the donors of the Petroleum Research Fund administered by the American Chemical Society for support of this work. B.A. also wishes to acknowledge a grant from the Danish Natural Science Research Council.

REFERENCES

1. R. Ash, *Information Theory*, Interscience, New York, 1965.
2. L. L. Campbell, The relation between information theory and the differential geometry approach to statistics, *Inform. Sci.* 35:199–210 (1985).
3. J. Hirshleifer, *Investment, Interest, and Capital*, Prentice-Hall, Englewood Cliffs, N.J., 1970.
4. S. Kullback, *Information Theory and Statistics*, Dover, New York, 1968.
5. P. Salamon and R. S. Berry, Thermodynamic length and dissipated availability, *Phys. Rev. Lett.* 51:1127 (1983).
6. P. Salamon, J. Komlos, B. Andresen, and J. D. Nulton, A geometric view of consumer surplus with non-instantaneous adjustment, *Math. Social Sci.* 13:153–163 (1987).
7. W. K. Wootters, Statistical distance and Hilbert space, *Phys. Rev. D* 23:357 (1981).

Received 28 January 1986; revised 20 November 1986