MINIMUM-ENTROPY DECISION ANALYSIS

Alfonso Gonzalez-Serrano

Abstract

The concept of minimum-entropy decision analysis is reviewed. Maximum likelihood is defined as a process which optimizes error probability, thus minimizing entropy. Some bounds on error probability are derived based on the Tchebyscheff inequality and the union bound. The concepts are illustrated with the additive white Gaussian noise channel. An example for a picking algorithm is discussed, where performance is optimized by introducing memory in the system.

Introduction

Entropy, as is discussed in a paper by Claerbout in this report (p.) is a measure of the expected information of a system. Any data processing dealing with minimum-entropy tries to optimize the probability distributions associated with the system.

This paper reviews some ideas on minimum-entropy decision analysis. Starting from definitions we find bounds and the associated probability distributions for maximum entropy situations. An important result shows that for a given system, memory reduces entropy, while data processing increases it.

Maximum-likelihood decision is defined as a minimum-entropy process which optimizes error probability when nothing is known about the distribution of the inputs. The structure of the decision process makes ***dynamic programming***

techniques, namely the Viterbi algorithm, feasible to solve for the most probable state of a system from noisy data. To evaluate the performance of this decision process some bounds on error probability are presented: they are either generalizations of the Tchebyscheff inequality or extensions of the union bound for probability distributions. The particular case of an additive white Gaussian noise channel is used as an example.

Finally, an example of practical application is shown. We take the problem of picking times to align a given event through a suite of traces. We show that a procedure based on crosscorrelation followed by picking maximum amplitudes assumes a maximum-entropy situation. We show that this process is severely limited in performance by the signal-to-noise ratio. Introducing memory in the system helps to improve performance.

The theoretical part follows in particular the treatments by Gallager (1968) and Viterbi (1969).

Theory

i) Entropy

Consider a random variable u taking values from a finite alphabet $U = (a_1, a_2, \cdots, a_A)$ with probabilities $P(a_k)$ $\forall$ $k = 1,2,\cdots,A$. The ***self-information*** of the event α_k when the random variable takes the value $u(\alpha_k) = a_k$, is defined as

$$I(\alpha_k) \equiv -\log_\beta P(a_k) \tag{1}$$

The base β of the logarithm is arbitrary and determines the numerical scale used to measure information. For base $\beta = 2$ the numerical value of I is given in ***bits*** and intuitively will equal the number of ***yes-no*** questions we would need to ask to completely identify the event α_k.

A more important quantity than self-information is its expectation, or ***entropy***

$$H(U) \equiv E\{I(\alpha_k)\} = \sum_{k=1}^{A} P(a_k) I(\alpha_k)$$

$$= \sum_{u} P(u) \log_\beta \frac{1}{P(u)} \qquad (2)$$

For a continuous distribution the alternative definition is

$$H(U) = \int_{-\infty}^{+\infty} P(u) \log_\beta \frac{1}{P(u)} \, du \qquad (3)$$

From these definitions it is easy to find bounds for the entropy of an information source.

Consider the inequality between the functions ln x and x-1 sketched in figure 1

$$\ln x \leq x - 1 \qquad (4)$$

In the discrete case we have that, for any two arbitrary distributions P(u) and Q(u) over the alphabet U,

$$\sum_{u} P(u) \log_\beta \frac{Q(u)}{P(u)} \leq (\ln\beta)^{-1} \sum_{u} P(u) \left[\frac{Q(u)}{P(u)} - 1\right] = 0 \qquad (5)$$

Therefore,

$$0 \leq \sum_{u} P(u) \log_\beta \frac{1}{P(u)} \leq \sum_{u} P(u) \log_\beta \frac{1}{Q(u)} \qquad (6)$$

In particular let Q(u) be uniformly distributed:

$$Q(u) = \frac{1}{A} \qquad \forall\, u \in U = (a_1, a_2, \cdots, a_A)$$

Using this distribution, in inequality (6) we get

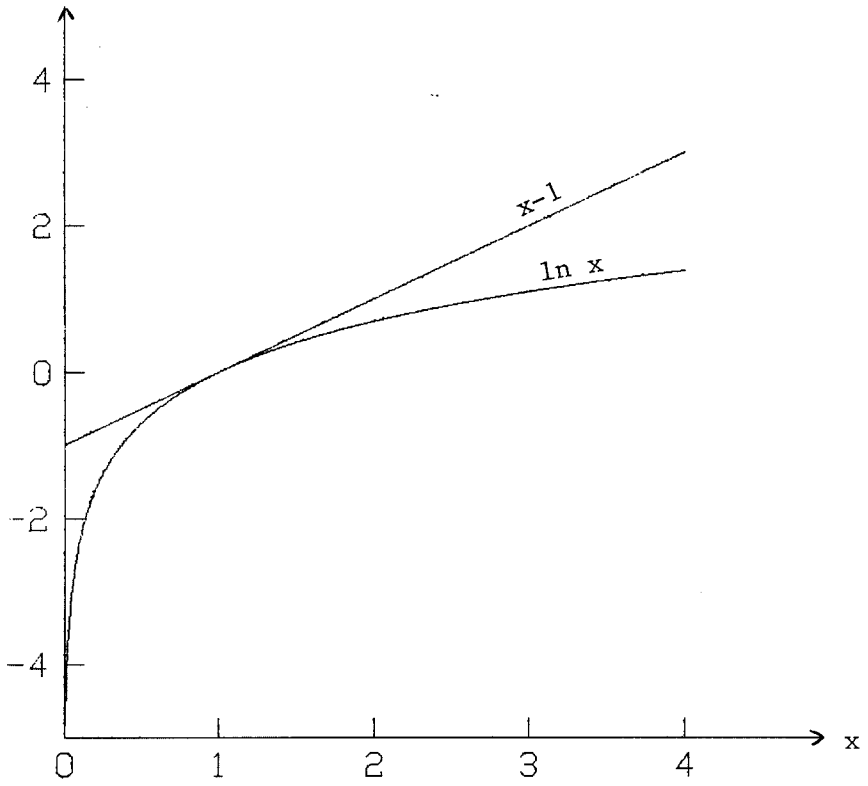


FIG. 1. Sketch of the functions ln x and x-1.

$$0 \leq \sum_{u} P(u) \log_{\beta} \frac{1}{P(u)} = H(U) \leq \log_{\beta} A \qquad (7)$$

therefore, for discrete distributions the uniform will attain the upper bound, i.e. there will be total uncertainty about outcomes.

In the continuous case we can use again the inequality (4). For a distribution P(u) with moments

$$\int_{-\infty}^{+\infty} u\, P(u)\, du = 0$$

$$\int_{-\infty}^{+\infty} u^2\, P(u)\, du = \sigma^2$$

let Q(u) be ***Normally*** distributed as $N(0,\sigma^2)$

$$Q(u) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{u^2}{2\sigma^2}}$$

using these distributions we get

$$\int_{-\infty}^{+\infty} P(u) \log_\beta \frac{1}{Q(u)} du = \int_{-\infty}^{+\infty} P(u) \log_\beta (2\pi\sigma^2)^{\frac{1}{2}} \exp\left[\frac{u^2}{2\sigma^2}\right] du$$

$$= \int_{-\infty}^{+\infty} P(u) \left[\log_\beta (2\pi\sigma^2)^{\frac{1}{2}} + \frac{u^2}{2\sigma^2} \log_\beta e\right] du$$

$$= \tfrac{1}{2} \log_\beta 2\pi e\sigma^2$$

Therefore,

$$H(U) \leq \tfrac{1}{2} \log_\beta 2\pi e\sigma^2 \qquad (8)$$

In the continuous case a normal distribution has maximum-entropy.

A useful inequality for non-independent probability distributions is also obtained from inequality (6). Whenever there is dependence among events, we have memory on the system. Let

$$Q_N(u) = \prod_{n=1}^{N} P(u_n)$$

where

$$P(u_n) = \sum_{u_1} \cdots \sum_{u_{n-1}} \sum_{u_{n+1}} \cdots \sum_{u_N} P_N(u)$$

Substituting this distribution in inequality (6),

$$
\begin{aligned}
H(U_N) &= \sum_u P_N(u) \log_\beta \frac{1}{P_N(u)} \\
&\leq \sum_u P_N(u) \log_\beta \frac{1}{Q_N(u)} \\
&= \sum_u P_N(u) \log_\beta \frac{1}{\prod_{n=1}^{N} P(u_n)} \\
&= \sum_{n=1}^{N} P(u_n) \log_\beta \frac{1}{P(u_n)} \\
&= N\, H(U) \qquad (9)
\end{aligned}
$$

we see that memory decreases entropy.

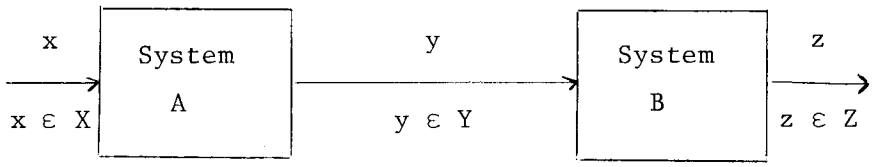


FIG. 2. Two systems in cascade. Discrete memoryless systems.

To prove that data processing can only increase the entropy of a system consider the definition of ***mutual-information*** between events x and y

$$
I(x;y) \equiv \log_\beta \frac{P(x|y)}{P(x)} \qquad (10)
$$

This is the information provided about event x by the occurrence of event y. Its expectation

$$I(X;Y) \equiv E\{I(x;y)\} \tag{11}$$

$$= \sum_{x}\sum_{y} p(y|x)\, q(x)\, \log_{\beta} \frac{p(y|x)}{\sum_{x'} p(y|x')\, q(x')}$$

where

$$p(x,y) = q(x)\, p(y|x)$$

is the average mutual information. Consider the systems in cascade shown in figure 2. The random variables x,y,z have a joint probability distribution $p(x,y,z)$ $\forall\, x\epsilon X,\ y\epsilon Y,\ z\epsilon Z$. We need a relationship between the mutual-information of inputs and outputs to the systems. Taking the difference,

$$I(X;Z) - I(X;Y) = \sum_{z}\sum_{y}\sum_{x} p(x,y,z)\, \log_{\beta} \frac{p(z|x)p(y)}{p(z)p(y|x)}$$

$$= (\ln\beta)^{-1} \sum_{z}\sum_{y}\sum_{x} p(x,y,z)\, \ln \frac{p(z|x)p(y)}{p(z)p(y|x)}$$

$$\leq (\ln\beta)^{-1} \sum_{z}\sum_{y}\sum_{x} p(x,y,z) \left[\frac{p(z|x)p(y)}{p(z)p(y|x)} - 1\right]$$

and using Bayes' rule

$$\frac{p(x,y,z)p(z|x)p(y)}{p(z)p(y|x)} = p(z|y)p(x|z)p(y)$$

where we used the fact implied by the system

$$p(z|x,y) = p(z|y)$$

it can be seen that

$$I(X;Z) - I(X;Y) \leq (\ln\beta)^{-1} \left[\sum_z \sum_y \sum_x p(z|y)p(x|z)p(y) - 1 \right]$$

$$= (\ln\beta)^{-1} \left[\sum_z \sum_y p(z|y)p(y) - 1 \right] = 0 \qquad (12)$$

From this result it follows that

$$I(X;Z) \leq I(X;Y) \qquad (13)$$

$$I(X;Z) \leq I(Y;Z) \qquad (14)$$

Similarly, for the three systems in cascade shown in figure 3,

$$I(W;Z) \leq I(X;Y) \qquad (15)$$

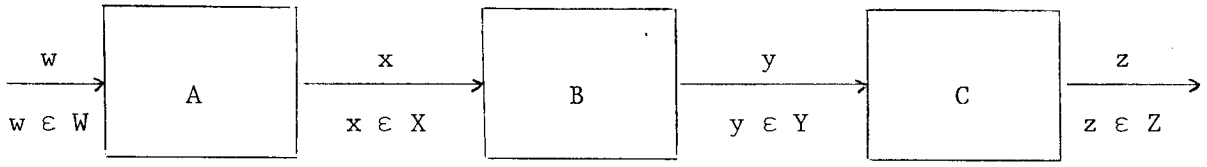


FIG. 3. Data-processing systems in cascade. Discrete memoryless systems.

This result is known as the ***data-processing theorem*** and states that in a cascade of systems, the introduction of a new system (or process) will weaken the dependence between input and output, therefore reducing the average mutual-information.

*ii) **Maximum-likelihood decision***

Consider ths situation arising in communication systems theory. A message m from a finite set of messages H_m $\forall$ $m\in\{1,2,\cdots,M\}$ is chosen and encoded into a digital vector $x_m = \{x_{m1}, x_{m2}, \cdots, x_{mN}\}$, where the x_{mn} are elements from a finite discrete alphabet of length A

$$x_m \in X^N = \{a_1, a_2, \cdots, a_A\}^N$$

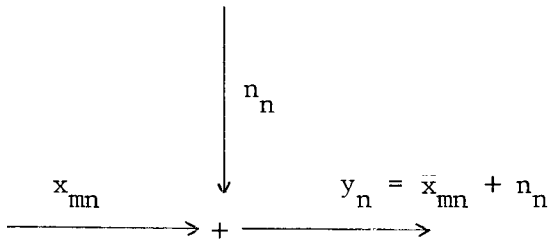


FIG. 4. Additive white Gaussian noise channel.

The message vector is transmitted over a noisy channel and the sequence y is observed (figure 4). The channel can be characterized by a transition probability function from all possible inputs x_m to all possible observations y

$$p_N(y|x_m)$$

where the subscript N denotes the dimension of the vectors in question.

It is clear that only when the channel is noiseless will the vector of observations y uniquely determine the input message, but in the general situation there is some uncertainty about the input.

Suppose that when the vector y takes on some particular value we make the decision $H_{\hat{m}} = H_m$. The problem is to decide which message was transmitted in order to optimize the performance of the channel. In other words we want to minimize the error of the decision

$$y \rightarrow x_{\hat{m}} \rightarrow H_{\hat{m}} = H_m$$

The probability of error in this decision, denoted as $P_E(H_m;y)$, is

$$P_E(H_m;y) = Pr\{H_m \text{ not sent}|y\}$$

$$= 1 - Pr\{H_m \text{ sent}|y\} \qquad (16)$$

An optimum decision rule would be

$$H_{\hat{m}} = H_m \quad \text{if} \quad Pr\{H_m \text{ sent}|y\} \geq Pr\{H_{m'} \text{ sent}|y\} \quad \forall\, m \neq m' \qquad (17)$$

Using Bayes' rule on the ***a priori*** probability term,

$$P\{H_m \text{ sent}|y\} = \frac{P\{H_m \text{ sent}\}\, P\{y|H_m \text{ sent}\}}{P\{y\}} \qquad (18)$$

we can rewrite inequality (17) as a function of ***a posteriori*** probabilities

$$H_{\hat{m}} = H_m \quad \text{if} \quad \frac{P\{H_m \text{sent}\}\ P\{y|H_m \text{sent}\}}{P\{y\}} > \frac{P\{H_{m'} \text{sent}\}\ P\{y|H_{m'} \text{sent}\}}{P\{y\}} \quad \forall\ m \neq m' \qquad (19)$$

where we have considered ties as errors. It is clear the decision does not depend on $P\{y\}$.

Next, since we know the mapping from message to code vector $H_m \rightarrow x_m$ is one to one, we can rewrite this last equation as

$$H_{\hat{m}} = H_m \quad \text{if} \quad p_N(x_m)\ p_N(y|x_m) > p_N(x_{m'})\ p_N(y|x_{m'}) \quad \forall\ m \neq m' \qquad (20)$$

The ***a priori*** probabilities $p(x_m)$, if known, reduce the entropy in the decision because they are information from the source we know in advance. In the general case, however, we do not know any ***a priori*** information about the source; to avoid introducing spurious information we have to assume a maximum-entropy source. For this kind of source we talk about ***maximum-likelihood decision*** and the optimal rule is simply

$$H_{\hat{m}} = H_m \quad \text{if} \quad p_N(y|x_m) > p_N(y|x_{m'}) \quad \forall\ m \neq m' \qquad (21)$$

For the particular case of a memoryless channel we can rewrite inequality (21) as

$$H_{\hat{m}} = H_m \quad \text{if} \quad \prod_{n=1}^{N} p(y_n|x_{mn}) > \prod_{n=1}^{N} p(y_n|x_{m'n}) \quad \forall\ m \neq m' \qquad (22)$$

Finally a useful form is obtained taking logarithms; this is called the ***metric***:

$$H_{\hat{m}} = H_m \quad \text{if} \quad \sum_{n=1}^{N} \ln p(y_n|x_{mn}) > \sum_{n=1}^{N} \ln p(y_n|x_{m'n}) \quad \forall\ m \neq m' \qquad (23)$$

The maximum-likelihood decoder therefore looks at all the metrics for each possible signal, compares them, and decides in favor of the maximum.

iii) Error Probability

Gallager (1968) computes some bounds on error probability for maximum-likelihood decoding. These bounds are generalizations of the Tchebyscheff inequality, or follow from the union bound on probabilities. They are useful because it is not always practical or even possible to compute the P_E exactly, and because in some situations they give a tighter bound than the Tchebyscheff. We follow Gallager (1968) and Viterbi (1969) in the derivation.

For a random variable u, taking only nonnegative values, the Tchebyscheff inequality states

$$\Pr(u \geq \delta) \leq \frac{E\{u\}}{\delta} \quad \forall\, \delta > 0 \tag{24}$$

Another form of the inequality can be obtained defining new variables $u = (w - E\{w\})^2$ and $\epsilon = \delta^{\frac{1}{2}}$:

$$\Pr[|w - E\{w\}| \geq \epsilon] \leq \frac{E\{(w - E\{w\})^2\}}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2} \tag{25}$$

Another inequality can be obtained by letting $u = e^{sw}$ $\forall\, s \in R$, denoting by $g_w(s)$ the moment generating function of w

$$g_w(s) \equiv E\{e^{sw}\}$$

and defining $\delta = e^{sA}$ $\forall\, A \in R$ we get the so called Chernoff bound

$$\Pr[e^{sw} \geq e^{sA}] \leq e^{-sA} g_w(s) \quad \forall\, s, A \in R$$

Equivalently,

$$\Pr[w \geq A] \leq e^{-sA} g_w(s) \quad \forall\, A \in R, \quad s > 0 \tag{26}$$

$$\Pr[w \leq A] \leq e^{-sA} g_w(s) \quad \forall\, A \in R, \quad s < 0 \tag{27}$$

These bounds are useful only for tails of distributions. Therefore inequality (26) is useful only if $A > E\{w\}$ and inequality (27) when $A < E\{w\}$. Gallager (1968) gives some examples on the behavior of these bounds for different normal distributions for the important case when w is a sum of random variables. His conclusion is that the Chernoff bound is a poor approximation when A is close to $E\{w\}$ (for small s), but for large A the bound is good.

Next consider the union bound on probabilities, for a set of events with probabilities $P(\mathbf{A}_1), \ldots, P(\mathbf{A}_M)$, the probability of their union is

$$P\left[\bigcup_{m=1}^{M} \mathbf{A}_m\right] \tag{28}$$

It follows that

$$P\left[\bigcup_{m=1}^{M} \mathbf{A}_m\right] \leq \left[\sum_{m=1}^{M} P(\mathbf{A}_m)\right]^{\rho} \qquad \forall\ 0<\rho\leq 1 \tag{29}$$

For maximum-likelihood decoding we can define decision regions using inequality (23) as

$$\mathbf{A}_m \equiv \{y:\ \ln p_N(y|x_m) > \ln p_N(y|x_{m'}) \qquad \forall\ m'\neq m\} \tag{30}$$

From the definition the sets are disjoint:

$$\mathbf{A}_k \cap \mathbf{A}_j = \varnothing \qquad \forall\ k\neq j$$

We will make an error in the decision if $y\in\mathbf{A}_{m'}$ given that x_m was sent. Using the union bound we have

$$P_E(y|x_m) \leq P\left[\bigcup_{m'\neq m} \mathbf{A}_{m'}\right]$$

$$\leq \left[\sum_{m'\neq m} P(\mathbf{A}_{m'})\right]^{\rho} \qquad \forall\ 0<\rho\leq 1 \tag{31}$$

The inequality follows from the fact that the maximum-likelihood decision does not necessarily make an error in case of ties.

Since for an erroneous decision we have

$$\frac{p_N(y|x_{m'})}{p_N(y|x_m)} \geq 1 \qquad \text{for some } m' \neq m \tag{32}$$

This implies

$$\sum_{m' \neq m} \left[\frac{p_N(y|x_{m'})}{p_N(y|x_m)}\right]^{\lambda} \geq 1 \qquad \forall\, y \in \bar{A}_m,\ \lambda > 0 \tag{33}$$

where we have defined

$$A_m \equiv \{y:\ \ln p_N(y|x_m) > \ln p_N(y|x_{m'}) \qquad \forall\, m' \neq m\} \tag{34}$$

Using this result we can write for $P(A_{m'})$

$$P(A_{m'}) \leq \sum_{m' \neq m} \left[\frac{p_N(y|x_{m'})}{p_N(y|x_m)}\right]^{\lambda} \qquad \forall\, \lambda > 0 \tag{35}$$

Substituting this into equation (31) we get

$$P_E(y|x_m) \leq \left\{\left[(M-1) \sum_{m' \neq m} \frac{p_N(y|x_{m'})}{p_N(y|x_m)}\right]^{\lambda}\right\}^{\rho} \qquad \forall\, \lambda > 0,\ \rho > 0 \tag{36}$$

Using $\lambda = 1/(1+\rho)$ we get the Gallager bound

$$P_E(y|x_m) \leq \sum_y [p_N(y|x_m)]^{1/(1+\rho)} \left\{\sum_{m' \neq m} [p_N(y|x_{m'})]^{1/(1+\rho)}\right\}^{\rho} \qquad \forall\, \rho > 0 \tag{37}$$

For the particular case $\rho = 1$ this bound is known as the ***Bhattacharyya bound***

$$P_E(y|x_m) \leq \sum_y [p_N(y|x_m)]^{\frac{1}{2}} \sum_{m' \neq m} [p_N(y|x_{m'})]^{\frac{1}{2}} \tag{38}$$

iv) Additive white Gaussian noise channel

To illustrate an application of the decision theory the additive white Gaussian noise channel is used, figure 4. This process with white spectral density is defined to have the covariance

$$R(\tau) = \frac{N_0}{2} \delta(\tau) \tag{39}$$

where $\delta(\tau)$ is the Dirac delta function, and N_0 the one-sided noise power spectral density.

The conditional probability function for this process is therefore

$$p_N(y|x_m) = \prod_{n=1}^{N} P(y_n|x_{mn})$$

$$= \frac{1}{(\pi N_0)^{\frac{1}{2}}} e^{-[y_n - x_{mn}]^2/N_0} \tag{40}$$

The decoding regions for this channel are given by

$$H_{\hat{m}} = H_m \qquad \text{if} \qquad y \in A_m$$

where

$$A_m \equiv \{y: \ln p_N(y|x_m) > \ln p_N(y|x_{m'}) \quad \forall\, m' \neq m\}$$

$$= \left\{ y: \sum_{n=1}^{N} \ln \frac{p(y_n|x_{mn})}{p(y_n|x_{m'n})} \geq 0 \quad \forall\, m' \neq m \right\}$$

$$= \left\{ y: \frac{1}{N_0} ||y - x_{m'}||^2 - \frac{1}{N_0} ||y - x_m||^2 > 0 \quad \forall\, m' \neq m \right\} \tag{41}$$

and

$$||z||^2 = \sum_{n=1}^{N} z_n^2 \tag{42}$$

Define the energy of the signal m as

$$\mathcal{E}_m \equiv ||x_m||^2 \tag{43}$$

and denoting the dot product between vectors x and y as $\langle x|y \rangle$, we can rewrite equation (41) as

$$\Lambda_m = \left\{ y: \frac{2}{N_0} \langle x_m - x_{m'}|y \rangle - \frac{(\mathcal{E}_m - \mathcal{E}_{m'})}{N_0} > 0 \quad \forall\, m' \neq m \right\} \tag{44}$$

For this channel the expected probability of error can be calculated exactly. From equation (44) the probability $P_E(m \to m')$ of an error when the message m is sent and m' is the only alternative, is

$$P_E(m \to m') = \Pr\left\{ \frac{2}{N_0} \langle x_m - x_{m'}|y \rangle \leq \frac{(\mathcal{E}_m - \mathcal{E}_{m'})}{N_0} \,\middle|\, x_m \right\} \tag{45}$$

Defining a new random variable

$$Z_{mm'} \equiv \frac{2}{N_0} \langle x_m - x_{m'}|y \rangle \tag{46}$$

its distribution can be found remembering that y_n is a Gaussian random variable by assumption, with mean x_{mn} and variance $N_0/2$. $Z_{mm'}$ is a linear combination of independent Gaussian random variables; therefore it is Gaussian with

mean

$$E\{Z_{mm'} | x_m\} = \frac{2}{N_0}\left[\Xi_m - \langle x_{m'} | x_m \rangle\right] \equiv \mu_z \tag{47}$$

and variance

$$E\{(Z_{mm'} - \mu_z)^2\} = \frac{2}{N_0}\,||x_m - x_{m'}||^2 \equiv \sigma_z^2 \tag{48}$$

Using these values in equation (45) we get

$$P_E(m \to m') = \int_{-\infty}^{-\frac{\Xi_m - \Xi_{m'}}{N_0}} \frac{\exp\left[-(Z_{mm'} - \mu_z)^2/2\sigma_z^2\right]}{(2\pi\sigma_z^2)^{\frac{1}{2}}}\, dZ_{mm'}$$

$$= Q\left(\frac{||x_m - x_{m'}||}{(2N_0)^{\frac{1}{2}}}\right) \tag{49}$$

where Q(.) is the Gaussian integral

$$Q(\beta) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{-\beta} e^{-x^2/2}\, dx \tag{50}$$

Finally, consider the particular case when all the events have the same energy given by

$$x_{mn} = \Xi^{\frac{1}{2}}\,\delta_{mn} \tag{51}$$

For this signal set we can use equation (40) to find its probability distribution: it is given by

$$p_N(y|x_m) = \frac{1}{(\pi N_0)^{\frac{1}{2}}} \exp\left[-(y_m - \Xi^{\frac{1}{2}})^2/N_0\right] \frac{\exp\left[-\sum_{n=1,N}^{n\neq m} y_n^2/N_0\right]}{(\pi N_0)^{(M-1)/2}} \tag{52}$$

Therefore

$$p(y_m|x_m) = \int_{-\infty}^{+\infty} dy_1 \cdots \int_{-\infty}^{+\infty} dy_{m-1} \int_{-\infty}^{+\infty} dy_{m+1} \cdots \int_{-\infty}^{+\infty} dy_N \, p_N(y|x_m)$$

$$= \int_{-\infty}^{+\infty} \frac{e^{-y_1^2/N_0}}{(\pi N_0)^{\frac{1}{2}}} dy_1 \cdots \int_{-\infty}^{+\infty} \frac{e^{-y_N^2/N_0}}{(\pi N_0)^{\frac{1}{2}}} dy_N \frac{1}{(\pi N_0)^{\frac{1}{2}}} e^{-[y_m - \Xi^{\frac{1}{2}}]^2/N_0}$$

$$= \frac{1}{(\pi N_0)^{\frac{1}{2}}} e^{-[y_m - \Xi^{\frac{1}{2}}]^2/N_0} \tag{53}$$

The probability of error is given as

$$P_E(m \to m') = \Pr\{y_m \leq y_{m'}\}$$

$$= \int_{-\infty}^{-y_{m'}} \frac{1}{(\pi N_0)^{\frac{1}{2}}} e^{-[y_m - \Xi^{\frac{1}{2}}]^2/N_0} dy_m \tag{54}$$

Changing variables

$$x \equiv \frac{y_m - \Xi^{\frac{1}{2}}}{(N_0/2)^{\frac{1}{2}}} \tag{55}$$

we get

$$P_E(m \to m') = \int_{-\infty}^{-\frac{y_{m'} + \Xi^{\frac{1}{2}}}{(N_0/2)^{\frac{1}{2}}}} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-x^2/2} dx$$

$$= Q\left[\left(\frac{2}{N_0}\right)^{\frac{1}{2}} y_{m'} + \left(\frac{2\Xi}{N_0}\right)^{\frac{1}{2}}\right] \tag{56}$$

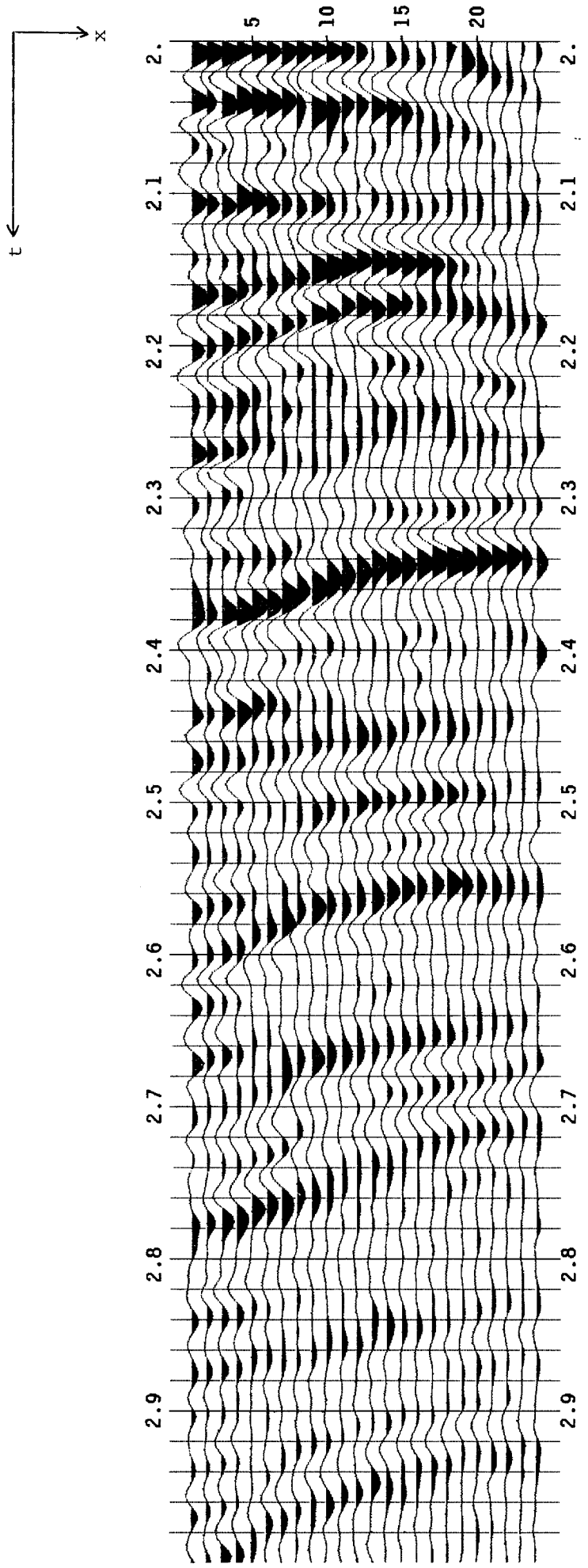


FIG. 5. Window of data. The event at 2.34 sec in the first trace was used as a reference for the croscorrelation with the remaining traces; the operator length was 0.5 sec.

Next, the overall probability of error for message m can be found as follows:

$$P_{E_m} = Pr\{y_m < y_{m'}, \quad \forall\, m \neq m' \mid x_1\}$$

$$= 1 - Pr\{y_{m'} \leq y_m \quad \forall\, m \neq m' \mid x_m\}$$

$$= 1 - \prod_{m \neq m'}^{M} Pr\{y_{m'} \leq y_m | x_m\}$$

$$= 1 - \prod_{m' \neq m}^{M} \left[1 - P_E(m \rightarrow m')\right] \tag{57}$$

and using equation (56)

$$P_{E_m} = 1 - \left\{1 - Q\left[\left(\frac{2}{N_0}\right)^{\frac{1}{2}} y_{m'} + \left(\frac{2\mathcal{E}}{N_0}\right)^{\frac{1}{2}}\right]\right\}^{M-1} \tag{58}$$

The final form is obtained by taking the expectation, since

$$E\{y_{m'} | x_m\} = 0$$

$$E\{(y_{m'} - E\{y_{m'}\})^2 | x_m\} = \frac{N_0}{2}$$

Using

$$x = \frac{y_{m'}}{(N_0/2)^{\frac{1}{2}}}$$

we get

$$E\{P_E\} = 1 - \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{+\infty} e^{-x^2/2} \left\{1 - Q\left[x + (2\mathcal{E}/N_0)^{\frac{1}{2}}\right]\right\}^{M-1} dx \tag{59}$$

Using this result Viterbi (1979) finds the optimum expected performance of the system, taking the limit

$$\lim_{M\to\infty} E\{P_E\} = \begin{cases} 1 & \text{if } \Xi_b/N_0 < \ln 2 \\ 0 & \text{if } \Xi_b/N_0 > \ln 2 \end{cases} \tag{60}$$

where

$$\Xi = \Xi_b \log_2 M \tag{61}$$

is the energy per bit. This is a remarkable bound since it states that for a maximum-entropy situation in a system under the assumed constraints, ***if the signal-to-noise ratio per bit of information is below*** ln 2, ***then, regardless of both our method for finding a decision and the sampling rate, the probability of error goes asymptotically to one.*** If the signal-to-noise ratio is above this threshold, the probability of error goes asymptotically to zero.

Example

Consider the following problem. Take a window of data with M time samples and a single event. In the absence of noise we can get the event in any of the M positions as a spike of amplitude $\Xi^{\frac{1}{2}}$ and zeroes everywhere else. For an ensemble of N traces let us consider the maximum-entropy case and denote the time of the event in trace n as t_{nm} $\forall\ 0 \le n \le N\text{-}1 \quad 0 \le m \le M$. This times are uniformly distributed as

$$P(t_{nm}) = \frac{1}{M} \qquad \forall \quad 0 \le n \le N \quad 0 \le m \le M$$

The problem is to align all events along a given t_0. Therefore we need to find the time shift to perform to each trace

$$\Delta t_{nm} = t_0 - t_{nm}$$

The problem is trivial in the absence of noise. However if there is noise in the channel, by equation (60) we know that the probability of error (picking an erroneous arrival as an event), will go to one for low signal-to-noise ratios.

The conventional procedure for picking time shifts to be used in solving the statics equation, is equivalent to this maximum-entropy situation, and is therefore very sensitive to the signal-to-noise ratio.

However, we can improve the situation. From equation (9) we know that any conditioning or memory in the system will reduce entropy. This immediately suggests a Markoff situation for the time-picking process. It is clear we want to keep coherency for the events through the traces for the selected event; thus a decision considering both the times and amplitudes of events will be more robust than the memoryless maximum-entropy situation.

For the implementation, we start crosscorrelating with a reference trace, and instead of looking just at the maximum amplitude, we select a set of events which qualify as allowable choices, both in terms of amplitude and time position. Call each picked event a state. Then define a transition probability function which will be a measure of the cost in going from a current state to the next one. We can use the maximum-likelihood decision procedure of equation (21) and the Viterbi algorithm (Forney, 1973) to find the optimum decision.

Results

As an example we used the data window shown in figure 5. The first trace was used for the crosscorrelation and the goal was to align all events along t_0 = 2.34 sec from the event in the first trace.

We picked as states all events with amplitudes above a threshold $A \geq \epsilon A_{max}$. In the examples $\epsilon = 0.6$ was used. The probability function was defined as

$$\Pr\{\text{transition from state k of trace i to state f of trace i+1}\} =$$

$$P(\Delta t_{i,k \rightarrow i+1,f}) = \alpha P(A) + (1-\alpha)P(T) \qquad (62)$$

where

$$P(A) = \Pr\{\text{Amplitude}\} = \frac{A^{\eta}_{i+1,f}}{\sum_j A^{\eta}_{j,f}} \tag{63}$$

$$P(T) = \Pr\{\text{Time difference}\} = \begin{cases} \dfrac{|\Delta t_{i,k \to i+1,f}|^{\xi}}{\sum_j |\Delta t_{i,k \to i+1,j}|^{\xi}} & \forall\ \Delta t \neq 0 \\ 1 & \Delta T = 0 \end{cases} \tag{64}$$

Figures 6 and 7 sketch these probability distributions for different values of η and ξ.

The probability function as defined is very flexible in its parameters. In general, however, the probability as a function of time difference is strongly weighted toward small jumps, so the choice of the weight α in equation (62) must compromise this bias with amplitudes.

The results of applying the algorithm with different values of α and $\eta = 2$, $\xi = -2$, are shown in figures 8 and 9. Figure 8 plots the curve defined as the minimum cost route for the given data and probability functions; the range goes from the conventional procedure of looking only at amplitudes ($\alpha = 1$), to the case of neglecting them ($\alpha = 0$). The corresponding traces with the time shifts defined from these curves are shown in figure 9. Note in particular the curve when no amplitude information was considered at all: the results are better than when we only consider amplitudes.

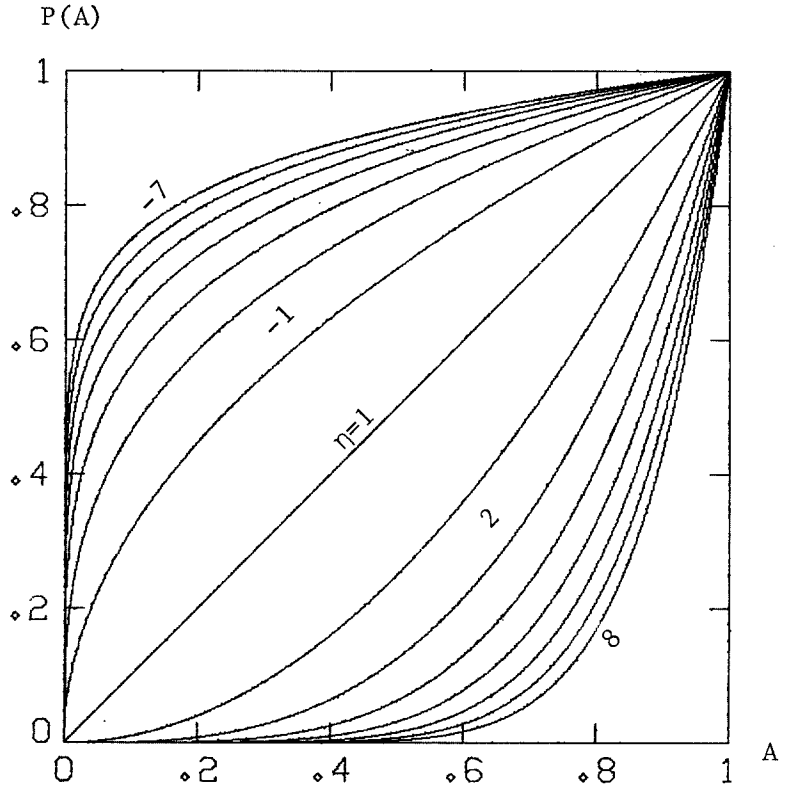


FIG. 6. Probability distributions for the amplitude of an event plotted for different values of exponent.

REFERENCES

Gallager, R.G., 1968, Information theory and reliable communication: New York, John Wiley & Sons.

Forney, G.D., Jr., 1973, The Viterbi algorithm: Proc. IEEE, v. 61, p. 268-276.

Viterbi, A.J., and Omura, J.K., 1979, Principles of digital communication and coding: New York, McGraw-Hill.

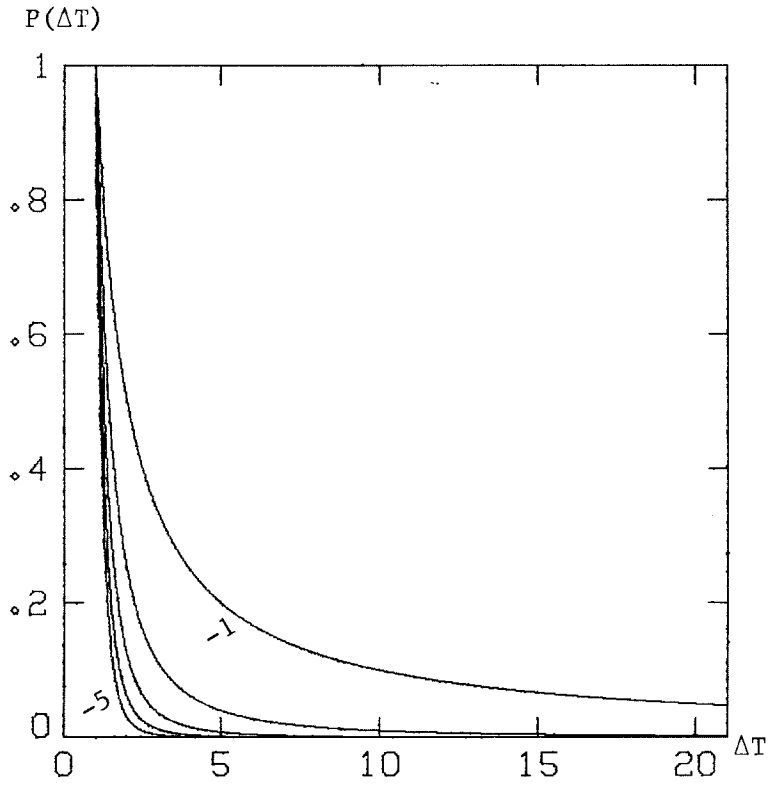
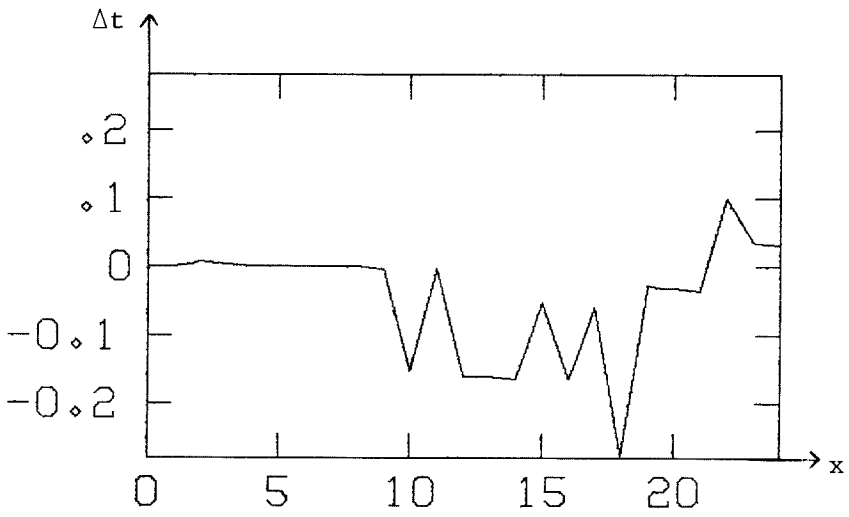
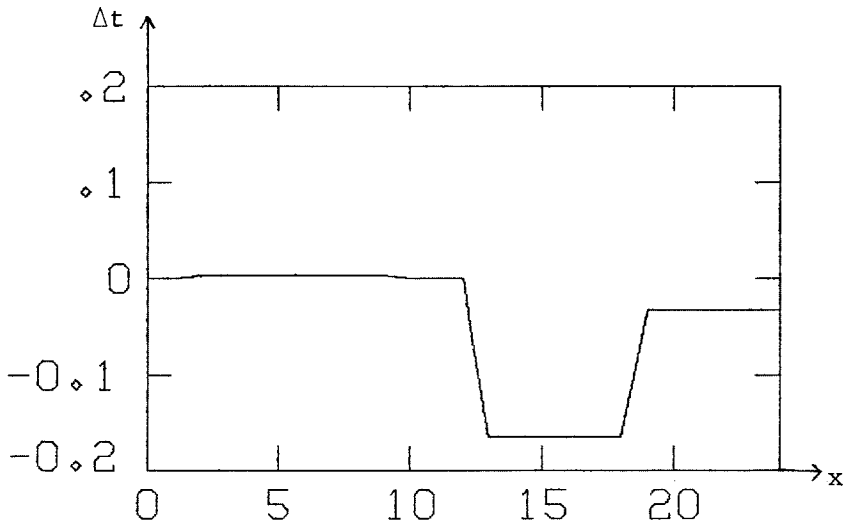


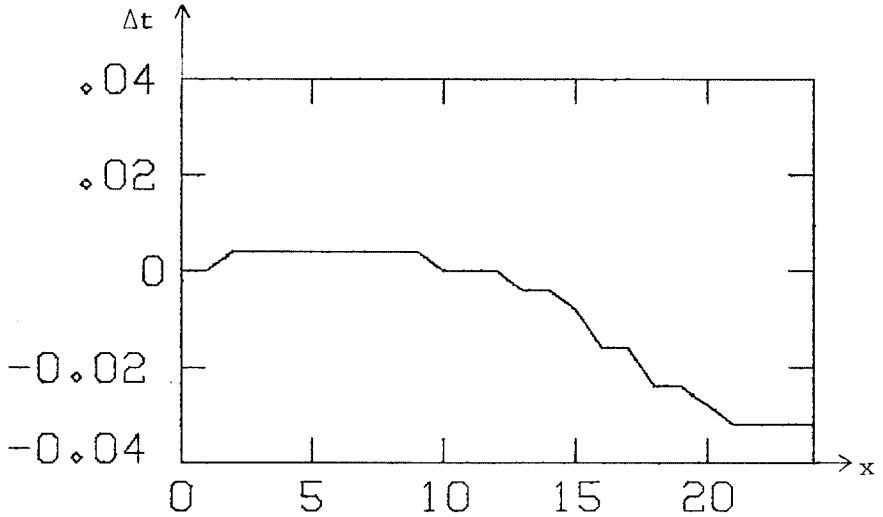
FIG. 7. Probability distributions for time difference between pairs of events for different values of exponent.



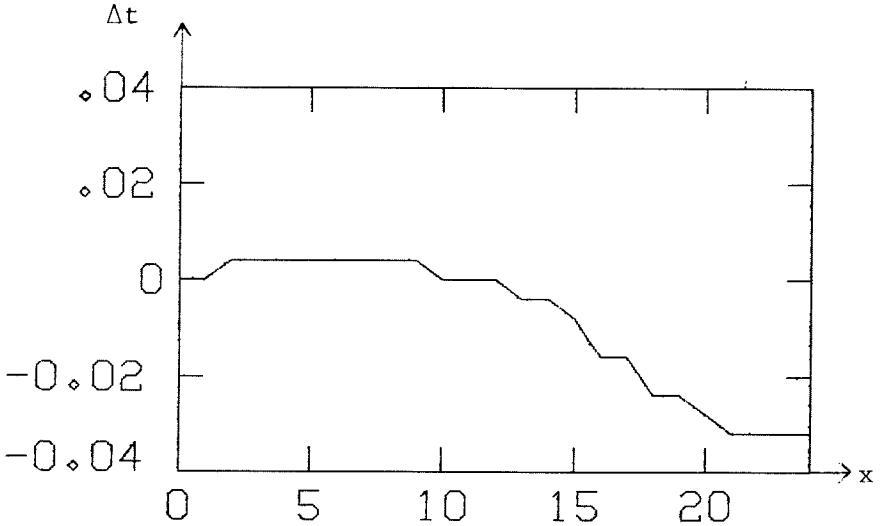
(a) $\alpha = 1.$



(b) $\alpha = .75$



(c) $\alpha = .50$



(d) $\alpha = 0.$

FIG. 8. Optimal curves for different values of α and $\eta = 2$, $\xi = -2$. (a) corresponds to the case of looking only at amplitude information, (b) and (c) have some memory considering both time difference and amplitude in the decision. (d) is the case when no amplitude information is included in the decision.

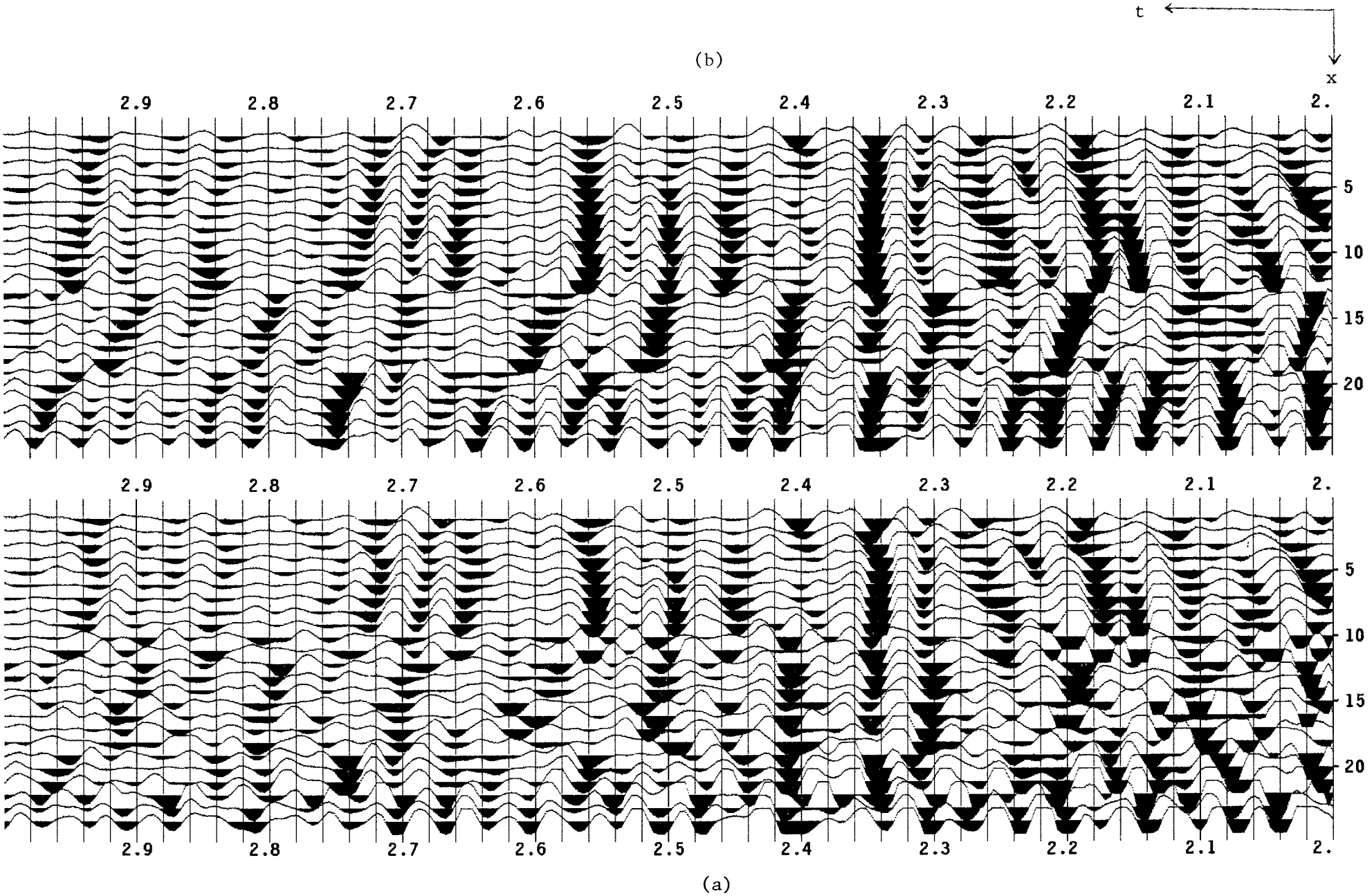
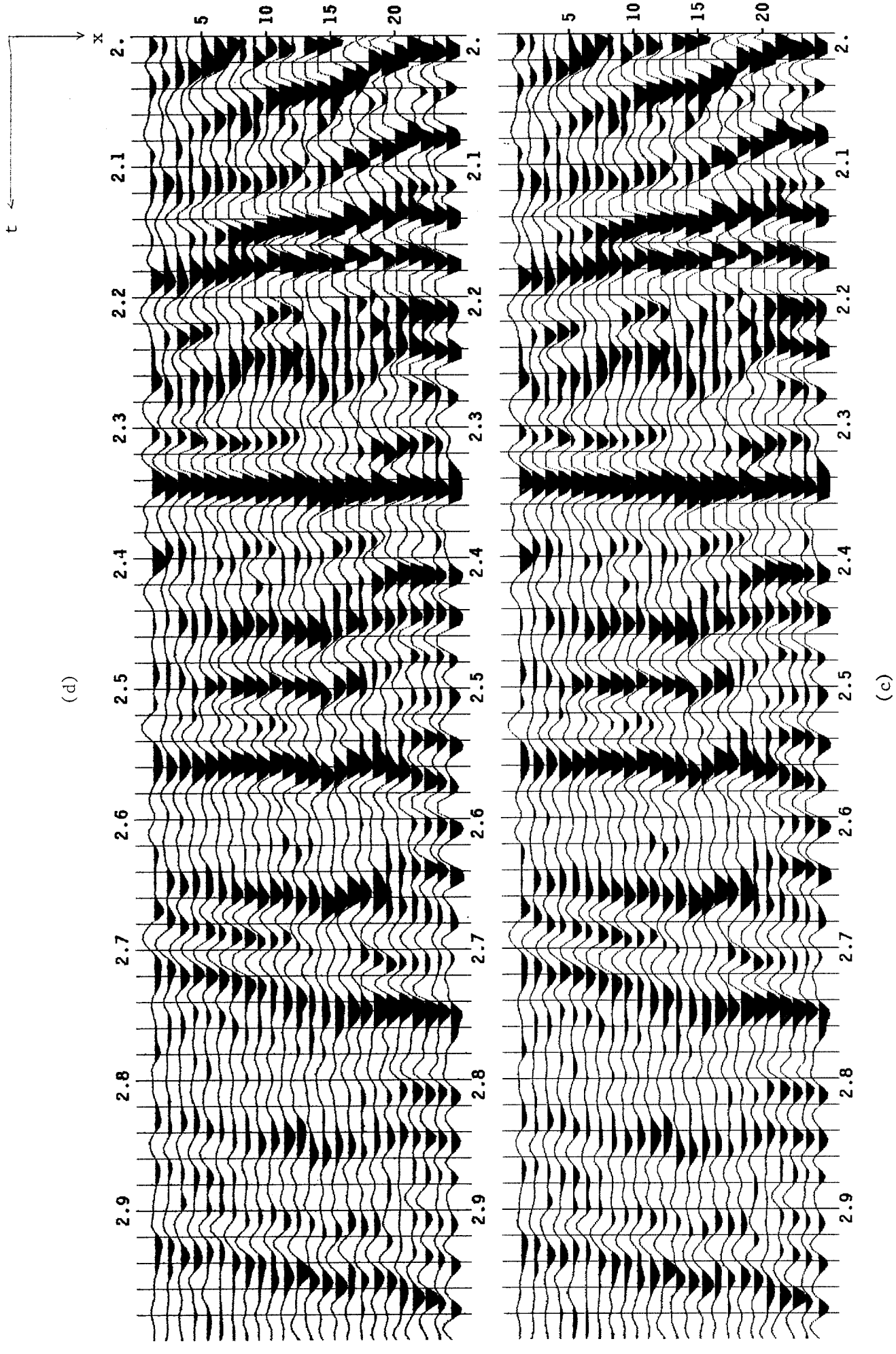


FIG. 9. Data with the time shifts defined in figure 8. Compare in particular (a) and (d) which are associated with the cases when only amplitude and only time differences are considered in the decision process.


x
t
5
10
15
20
2.
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
(d)
(c)