Codes for Unsupervised Learning of Source and Binary Channel Probabilities

E. A. PATRICK AND G. CARAYANNOPoulos

School of Electrical Engineering, Purdue University, Lafayette, Indiana 47907

Codes are constructed for communicating an alphabet consisting of two letters over an unknown, stationary binary channel. The code words, their a priori probabilities, and channel parameters are unknown at the receiver. Consistent estimators for the code words and their a priori probabilities are constructed at the receiver.

LIST OF SYMBOLS

$I_1$: $I_1$ and $I_2$ cover the $l$-dimensional vector space at the receiver

$S_j$: $j$th $l$-dimensional vector space at receiver

$\Omega$: $l$-dimensional signal space

$X_s$: Received, unclassified vector sample

$Q$: A priori probability of $S_1$ being transmitted

$p_j$: Probability $X_s$ is in the region $I_1$ given $S_j$ is transmitted

$q_j$: $1 - p_j$

$E\{\}$: Expectation

VAR: Variance

$X_{s_1}, \cdots, X_{s_v}$: $v$, $l$-dimensional vector samples at the $s$th transmission with the same signal active for each

$\lambda$: Random variable denoting the number of $l$-dimensional samples $X_{s_1}, \cdots, X_{s_v}$ which are in $I_1$ on the $s$th transmission

$\lambda^\alpha$: $\alpha$th sample moment of $\lambda$

INTRODUCTION

Consider the problem of communicating a message to a distant terminal through a stationary, additive noise channel. The message is assumed in the form of binary data bits. This binary alphabet $(0, 1)$
could be encoded into the channel using two time limited signals, \( S_1(t) \) and \( S_2(t) \), to transmit respectively a "0" or a "1". Denote by \( \Omega \) the signal space that contains \( S_1(t) \) and \( S_2(t) \) and let

\[
S_1 = (s_{11}, s_{12}, \cdots, s_{1l})
\]

\[
S_2 = (s_{21}, s_{22}, \cdots, s_{2l})
\]

be their respective vector representation in \( \Omega \).

Denote by

\[
Q = \text{probability of } S_1 \text{ being transmitted.}
\]

\[
1 - Q = \text{probability of } S_2 \text{ being transmitted.}
\]

Assume that the transmitted signal \( S_i (i = 1, 2) \) at the \( n \)th transmission is corrupted by additive, stationary noise, and it is received as a random, \( l \)-dimensional vector

\[
X_n = S_i + N \quad i = 1 \text{ or } 2
\]

where \( N \) is the projection of the noise into the signal space \( \Omega \).

Consider a partition of the signal space \( \Omega \) into two regions such that

\[
\Omega = I_1 \cup I_2,
\]

\[
I_1 \cap I_2 = 0
\]

Given any \( X_n \), there exist the following probabilities

\[
p_1 = \text{Prob} [X_n \in I_1 | S_1]
\]

\[
p_2 = \text{Prob} [X_n \in I_1 | S_2]
\]

\[
\tilde{p}_1 = \text{Prob} [X_n \in I_2 | S_1] = 1 - p_1
\]

\[
\tilde{p}_2 = \text{Prob} [X_n \in I_2 | S_2] = 1 - p_2
\]

which defines a binary channel model. Using this binary channel model and assuming that \( Q, p_1, \) and \( p_2 \) are known, the probability of error in deciding which signal caused \( X_n \) is minimized by deciding in favor of the signal with the largest a posteriori probability.

In this paper the parameters \( Q, p_1, \) and \( p_2 \) are assumed unknown, and a coding scheme is indicated which is sufficient for estimating these parameters. This is not possible if the signals are transmitted as indicated previously in this section, because consistent estimators for the three parameters do not exist (Patrick and Hancock, 1966; Teicher, 1963; and
Since the receiver never is told which signal gave rise to a receiver vector $X$, estimation takes place without supervision.

I. CODING SCHEME

The coding scheme used in the Introduction for defining the binary channel is inadequate for estimating the channel parameters $p_1$, $p_2$, and $Q$. But if the binary alphabet $(0, 1)$ is encoded as follows:

- 0: transmit $S_1(t)$ $v$ consecutive times
- 1: transmit $S_2(t)$ $v$ consecutive times

then consistent estimators for $p_1$, $p_2$, and $Q$ can be constructed. It will be shown (Appendix B) that a necessary and sufficient condition for estimating $p_1$, $p_2$, and $Q$ is $v \geq 3$; when $Q$ is known, a necessary and sufficient condition for estimating $p_1$ and $p_2$ is $v \geq 2$ (Appendix A).

In Section II, this problem is formulated mathematically, and in Sections III and IV moment estimators for the parameters $Q$, $p_1$, and $p_2$ are constructed and their properties studied. Finally, in Section V, two practical applications of the coding scheme are presented.

II. THE MATHEMATICAL PROBLEM

At the $s$th transmission, either $S_1$ or $S_2$ is transmitted $v$ consecutive times. The received vectors $X_{s1}, X_{s2}, \ldots, X_{sv}$ are in $I_1$ $\lambda$ times where $\lambda = 0, 1, 2, \ldots, v$. Because these received vectors are statistically independent, the probability density function for $\lambda$ is binomial given $S_i$ transmitted; that is

$$p[\lambda = a | S_i] = \frac{v!}{a!(v-a)!} p_i^a (1 - p_i)^{v-a},$$

$$a = 0, 1, \ldots, v \quad i = 1, 2$$

The probability density function for $\lambda$ is

$$p[\lambda = a] = \frac{v!}{a!(v-a)!} [Q(p_1)^a (1 - p_1)^{v-a}$$

$$+ (1 - Q)(p_2)^a (1 - p_2)^{v-a}]$$

$$a = 0, 1, \ldots, v$$

which is a parameter-conditional mixture\(^1\) of two binomial distributions.

\(^1\) The phrase "parameter-conditional mixture" was coined in Patrick and Hancock, (1966) for mixture probability density functions such as (2) which should be written $p[\lambda = a | p_1, p_2, Q]$ in Bayes estimation. To avoid notational complexity, the conditioning will be implied in this paper.
(Patrick and Hancock, 1966; Teicher, 1963; Rider, 1961). It is shown in these references that if \( v \geq 3 \), then \( p_1 \), \( p_2 \), and \( Q \) can be uniquely determined given \( p[\lambda = a] \) for all \( a \). The problem is to exhibit moment estimators for \( p_1 \), \( p_2 \), and \( Q \), which are consistent, (that is, the variances of the estimators approach zero as \( n \) approaches infinity). For the rest of this work, \( p_1 \neq p_2 \) and \( Q \neq 0 \) will be assumed.

III. MOMENT ESTIMATORS FOR PARAMETERS OF A PARAMETER-CONDITIONAL MIXTURE OF TWO BINOMIAL DISTRIBUTIONS

The method of estimating the parameters \( (p_1, p_2, Q) \) of a parameter-conditional mixture of binomial distributions, using moment estimators, has been treated by P. Rider, 1961. He did not however, give necessary and sufficient conditions for the moment estimators to be consistent. In this section and Section IV, we summarize Rider’s work, and also consider the special case where \( Q \) is known. Finally necessary and sufficient conditions for the estimator to be consistent are given. A detailed discussion of these conditions is in Appendices A and B in the form of two propositions.

To obtain moment estimators, we first use the moment generating function for the mixture (2) to find its moments. The moment generating function of \( \lambda \) is given by Hogg and Craig, 1965.

\[
M_{\lambda}(t) = E[e^{\lambda t}] = Q(g_1 + p_1 e^t)^v + (1 - Q)(g_2 + p_2 e^t)^v
\]

where

\[
g_j = 1 - p_j
\]

the first, second, and third moment of \( \lambda \) are given respectively by:

\[
E[\lambda] = \left. \frac{dM_{\lambda}(t)}{dt} \right|_{t=0} = v p_1 Q + v p_2 (1 - Q)
\]

\[
E[\lambda^2] = \left. \frac{d^2M_{\lambda}(t)}{dt^2} \right|_{t=0} = v[p_1 Q + p_2 (1 - Q)]
\]

\[
+ v(v - 1)[p_1^2 Q + (1 - Q)p_2^2]
\]

\[
E[\lambda^3] = \left. \frac{d^3M_{\lambda}(t)}{dt^3} \right|_{t=0} = v[p_1 Q + p_2 (1 - Q)]
\]

\[
+ 3v(v - 1)[p_1^2 Q + (1 - Q)p_2^2]
\]

\[
+ v(v - 1)(v - 2)(p_1^3 Q + (1 - Q)p_2^3)
\]
We will consider separately the case when \( Q \) is known and \( p_1 \) and \( p_2 \) are unknown, and the case when \( Q, p_1, \) and \( p_2 \) are all unknown.

**Q Known**

If \( Q \) is assumed known, then Equations (5) and (6) can be solved for \( p_1 \) and \( p_2 \) obtaining
\[
p_2 = A \pm \left( \frac{Q}{1 - Q} (B - A^2) \right)^{1/2} \tag{7a}
\]
and
\[
p_1 = A \mp \left( \frac{1 - Q}{Q} (B - A^2) \right)^{1/2} \tag{7b}
\]
where
\[
A = \frac{E\{\lambda\}}{v} \quad \text{and} \quad B = \frac{E\{\lambda^2\} - E\{\lambda\}^2}{v(v - 1)} \tag{7c}
\]

In Appendix A it is shown that when \( Q \) is known, the sign ambiguity of Equations (7a) and (7b) is resolved if and only if \( v \geq 2 \), then
\[
p_2 = A + \left( \frac{Q}{1 - Q} (B - A^2) \right)^{1/2} \tag{8a}
\]
\[
p_1 = A - \left( \frac{1 - Q}{Q} (B - A^2) \right)^{1/2} \tag{8b}
\]

Moment estimators for \( p_1 \) and \( p_2 \) are obtained if \( E\{\lambda\} \) and \( E\{\lambda^2\} \) in (7) are replaced by the corresponding sample moments. These estimators are:
\[
\hat{p}_1 = \hat{A} - \left( \frac{1 - Q}{Q} (\hat{B} - \hat{A}^2) \right)^{1/2} \tag{9a}
\]
\[
\hat{p}_2 = \hat{A} + \left( \frac{Q}{1 - Q} (\hat{B} - \hat{A}^2) \right)^{1/2} \tag{9b}
\]
where
\[
\hat{A} = \frac{1}{nv} \sum_{s=1}^{n} \lambda_s \tag{10}
\]
\[
\hat{B} = \frac{1}{nv(v - 1)} \sum_{s=1}^{n} \left( \lambda_s^2 - \lambda_s \right)
\]
and $\lambda_s$ is the value of the random variable $\lambda$ at the $s$th observation. Properties of these moment estimators will be considered in Section IV.

Q Unknown

For the case where $Q$ is also unknown, (this case was treated by P. Rider). Equations (4), (5), and (6) are solved for $p_1$, $p_2$, and $Q$ simultaneously as follows:

Rewrite Equations (4), (5), and (6) to obtain

$$vQ(p_1 - p_2) = E(\lambda) - vp_2$$  \hspace{1cm} (11a)

$$vQ(p_1 - p_2)[1 + (v - 1)(p_1 + p_2)] = E(\lambda^2) - vp_2 - v(v - 1)p_2^2$$  \hspace{1cm} (11b)

$$vQ(p_1 - p_2)[1 + 3(v - 1)(p_1 + p_2) + (v - 1)(v - 2)(p_1 + p_2)^2$$

$$- (v - 1)(v - 2)p_1p_2] = E(\lambda^3) - vp_2 - v(v - 1)(v - 2)p_2^3 - vp_2$$  \hspace{1cm} (11c)

solving (11a) for $vQ(p_1 - p_2)$ and substituting the result in Equations (11b) and (11c) we obtain

$$E[\lambda] + (v - 1)E(\lambda)(p_1 + p_2) - v(v - 1)p_1p_2 = E(\lambda^2)$$  \hspace{1cm} (12a)

$$E(\lambda) + 3(v - 1)E(\lambda)(p_1 + p_2)$$

$$+ E(\lambda)(v - 1)(v - 2)[(p_1 + p_2)^2 - p_1p_2] - 3v(v - 1)p_1p_2$$  \hspace{1cm} (12b)

$$- v(v - 1)(v - 2)p_1p_2(p_1 + p_2) = E(\lambda^3)$$

Introducing a new set of variables

$$S \triangleq p_1 + p_2$$  \hspace{1cm} (13a)

$$T \triangleq p_1p_2$$  \hspace{1cm} (13b)

in Equation (12a) and (12b) we obtain

$$(v - 1)E(\lambda)S - v(v - 1)T = E(\lambda^2) - E(\lambda)$$  \hspace{1cm} (14a)

$$3(v - 1)E(\lambda)S + (v - 1)(v - 2)E(\lambda)S^2$$

$$- (v - 1)(v - 2)E(\lambda)T - 3v(v - 1)T$$  \hspace{1cm} (14b)

$$- v(v - 1)(v - 2)TS = E(\lambda^3) - E(\lambda)$$

Solving for $T$ in (14a) gives

$$T = \frac{(v - 1)(E(\lambda)S + (E(\lambda) - E(\lambda^2)))}{v(v - 1)}$$
and substituting it in (14b) gives an expression for $S$ in terms of the moments:

$$S = \frac{v(E(\lambda^3) - E(\lambda)) + (E(\lambda) - E(\lambda^2))[3v + (v - 2)E(\lambda)]}{-(v - 2)[v(E(\lambda) E(\lambda^2)) + (v - 1)E^2(\lambda)]}$$

(15)

The expressions of $p_1$, $p_2$, and $Q$ in terms of $S$ and $T$ are (solving (13a) and (13b)):

$$p_2 = \frac{1}{2} [S \pm (S^2 - 4T)^{1/2}]$$

(16a)

$$p_1 = \frac{1}{2} S \mp (S^2 - 4T)^{1/2}$$

(16b)

$$Q = \frac{2E(\lambda) - vS \mp v(S^2 - 4T)^{1/2}}{\mp 2v(S^2 - 4T)^{1/2}}$$

(16c)

Again there is a sign ambiguity. In the Appendix it is shown that a unique solution for $p_1$, $p_2$, and $Q$ exists if and only if $v \geq 3$. Assuming $p_2 > p_1$ and letting $p_2$ correspond to the plus sign in (16a), we obtain:

$$p_2 = \frac{1}{2} [S + (S^2 - 4T)^{1/2}]$$

(17a)

$$p_1 = \frac{1}{2} [S - (S^2 - 4T)^{1/2}]$$

(17b)

and because $Q$ is the a priori probability of class 1, we take the minus sign in (16c) to obtain

$$Q = \frac{2E(\lambda) - v(S + (S^2 - 4T)^{1/2})}{-2(S^2 - 4T)^{1/2}}$$

$$= \frac{v(S + (S^2 - 4T)^{1/2} - 2E(\lambda))}{2(S^2 - 4T)^{1/2}}$$

(17c)

Actually, the uniqueness problem implied by (16a) and (16b) amounts to a "change of labels;" that is, we can label class 1 as that class with the smallest solution (16a) or the largest solution (16a). Equivalently, we can label class 1 as that class with the smallest a priori probability or the largest. Estimators for $p_1$, $p_2$, and $Q$ result if the sample moments are substituted in the corresponding expressions of $S$ and $T$. The asymptotic properties of the estimators outlined in this section will be considered in the next section.
IV. STATISTICAL PROPERTIES OF MOMENT ESTIMATORS

For the case where the mixing parameter \( Q \) is known the estimators for \( \hat{p}_1 \) and \( \hat{p}_2 \) are given respectively by:

\[
\hat{p}_1 = \hat{\lambda} - \frac{1 - Q}{Q} (\hat{B} - \hat{\lambda}^2) \quad (18a)
\]

\[
\hat{p}_2 = \hat{\lambda} + \frac{Q}{1 - Q} (\hat{B} - \hat{\lambda}^2) \quad (18b)
\]

where

\[
\hat{\lambda} = \frac{\bar{\lambda}}{v} = \frac{1}{n \bar{v}} \sum_{i=1}^{n} \lambda_i
\]

\[
\hat{B} = \frac{\bar{\lambda}^2 - \bar{\lambda}}{v(v - 1)} = \frac{1}{n \bar{v}(v - 1)} \sum_{i=1}^{n} \left( \lambda_i^2 - \lambda_i \right)
\]

are functions of the first and second sample moments.

The estimators \( \hat{p}_1 \) and \( \hat{p}_2 \) may be regarded either as functions of two arguments \( \bar{\lambda}^2 \) and \( \bar{\lambda} \) or replacing \( \bar{\lambda}^2 \) and \( \bar{\lambda} \) by their expressions in terms of the sample values, as function of \( n \) variables \( \lambda_1, \lambda_2, \cdots, \lambda_n \). Furthermore we observe:

(i) In some neighborhood of the point \( \bar{\lambda}^2 = \mathbb{E}[\lambda^2], \bar{\lambda} = \mathbb{E}[\lambda] \) the functions \( \hat{p}_1 \) and \( \hat{p}_2 \) are continuous and have continuous derivatives of the first and second order with respect to the arguments \( \bar{\lambda}^2 \) and \( \bar{\lambda} \).

(ii) for all possible values of \( \lambda_i, i = 1, 2, \cdots, n, |\hat{p}_i| < n^2 \). It can be shown (Cramer, 1946, Rider, 1961) that the expected values of \( \hat{p}_1 \), \( \hat{p}_2 \) and their respective variances are given by:

\[
\mathbb{E}\{\hat{p}_1\} = \mathbb{E}\{\hat{\lambda}\} - \left( \frac{1 - Q}{Q} (\mathbb{E}\{\hat{B}\} - \mathbb{E}^2\{\hat{\lambda}\}) \right)^{1/2} + o(1/n) \quad (20a)
\]

\[
= Qp_1 + (1 - Q)p_2 - (1 - Q)(p_2 - p_1) + o(1/n)
\]

\[
= p_1 + o(1/n)
\]

\[
\mathbb{E}\{\hat{p}_2\} = \mathbb{E}\{\hat{\lambda}\} + \left( \frac{Q}{1 - Q} (\mathbb{E}\{\hat{B}\} - \mathbb{E}^2\{\hat{\lambda}\}) \right)^{1/2} + o(1/n) \quad (20b)
\]

\[
= Qp_1 + (1 - Q)p_2 - Q(p_2 - p_1) + o(1/n)
\]

\[
= p_2 + o(1/n)
\]
\[ \text{VAR}\{\hat{\rho}_1\} = E\{ (\bar{X} - \lambda_0)^2 \} \left( \frac{\partial^2 \hat{\rho}_1}{\partial \lambda_1^2} \right) + E\{ (\bar{X}^2 - \lambda_0^2)^2 \} \left( \frac{\partial^2 \hat{\rho}_1}{\partial \lambda_2^2} \right) \]
\[ + E\{ (\bar{X} - \lambda_0)(\bar{X}^2 - \lambda_0^2) \} \frac{\partial^2 \hat{\rho}_1}{\partial \lambda_1 \partial \lambda_2} + O(1/n^{3/2}) \]
\[ = [4nQ^2v(v - 1)(p_1 - p_2)^2]^{-1}[2(Qp_1^2 + (1 - Q)p_2^2) \]
\[ - (Qp_1^3 + (1 - Q)p_2^3) + Q(v - 1)p_1(p_1 - p_2)^2 \]
\[ + 2(Qp_1^4 + (1 - Q)p_2^4) - 4Q(v - 1)p_1^2(p_1 - p_2)^2 \]
\[ + Q(1 - Q)v(v - 1)(p_1 - p_2)^4] \]  
(21a)

\[ \text{VAR}\{\hat{\rho}_2\} = E\{ (\bar{X} - \lambda_0)^2 \} \left( \frac{\partial^2 \hat{\rho}_2}{\partial \lambda_2^2} \right) + E\{ (\bar{X}^2 - \lambda_0^2)^2 \} \left( \frac{\partial^2 \hat{\rho}_2}{\partial \lambda_1^2} \right) \]
\[ + E\{ (\bar{X} - \lambda_0)(\bar{X}^2 - \lambda_0^2) \} \frac{\partial^2 \hat{\rho}_2}{\partial \lambda_1 \partial \lambda_2} + o(1/n^{3/2}) \]
\[ = [4n(1 - Q)^2v(v - 1)(p_2 - p_1)^2]^{-1}[2(Qp_1^2 + (1 - Q)p_2^2) \]
\[ - (Qp_1^3 + (1 - Q)p_2^3) + (1 - Q)(v - 1)p_2(p_1 - p_2)^2 \]
\[ + 2(Qp_1^4 + (1 - Q)p_2^4) - 4(1 - Q)(v - 1)p_2^2(p_1 - p_2)^2 \]
\[ + Q(1 - Q)v(v - 1)(p_1 - p_2)^4] + o(1/n^{3/2}) \]  
(21b)

In the limit as \( n \to \infty \)
\[ E\{\hat{\rho}_i\} = p_i \]

and
\[ \text{VAR}\{\hat{\rho}_i\} = 0 \quad i = 1, 2 \]

Therefore \( \hat{\rho}_1 \) and \( \hat{\rho}_2 \) are asymptotically consistent estimators.

For the case where the mixing parameter \( Q \) is also unknown, we have
\[ \hat{\rho}_1 = \frac{1}{2}[\hat{S} - (\hat{S}^2 - 4\hat{T})^{1/2}] \]  
(22a)

and
\[ \hat{\rho}_2 = \frac{1}{2}[\hat{S} + (\hat{S}^2 - 4\hat{T})^{1/2}] \]  
(22b)

\[ \hat{Q} = \frac{S + (\hat{S}^2 - 4\hat{T})^{1/2} - 2\bar{X}}{(\hat{S}^2 - 4\hat{T})^{1/2}} \]  
(22c)

where

\[ S = \frac{v[\bar{x}^2 - \bar{x}] + 3v[\bar{x} - \bar{x}^2] + (v - 1)\bar{x}(\bar{x} - \bar{x}^2)}{(v - 2)(v - 1)\bar{x}^2 + v(\bar{x} - \bar{x}^2)} \]

\[ T = \frac{\bar{x} - \bar{x}^2}{v(v - 1)} + \frac{\bar{x}}{v} S \]

the expected value of \( \hat{p}_1, \hat{p}_2 \) and \( Q \), using the extended form of the theorem used in the case of known \( Q \), are given respectively by:

\[ E[\hat{p}_1] = p_1 + o(1/n) \quad (23a) \]
\[ E[\hat{p}_2] = p_2 + o(1/n) \quad (23b) \]
\[ E[\hat{Q}] = Q + o(1/n) \quad (23c) \]

Therefore \( \hat{p}_1, \hat{p}_2 \) and \( Q \) are asymptotically unbiased estimators.

The variances of \( \hat{p}_1, \hat{p}_2 \) and \( \hat{Q} \) can be obtained similarly, but because of the cumbersome algebra involved this was not attempted.

V. APPLICATIONS

In this section two examples of applications of the coding scheme, described in this paper, are presented. In order to achieve simplification, the signals \( S_1 \) and \( S_2 \) in these examples are assumed one-dimensional. Extension to multidimensions is straightforward.

EXAMPLE I

Consider a case where the receiver has knowledge of the channel statistics and the encoded alphabet\(^3\) (i.e., \( S_1 \) and \( S_2 \)), but has no knowledge of the signal probabilities. It is assumed that the channel noise is a gaussian process of zero mean and variance \( \sigma^2 \). It is a well-known result that the decision boundary (threshold) used in deciding which signal is transmitted, with minimum probability of error, is given by

\[ \text{Decision boundary} = \frac{\sigma^2}{S_2 - S_1} \ln \left( \frac{Q}{1 - Q} \right) + \frac{1}{2} (S_2 + S_1) \quad (24) \]

where

\( Q \) and \( (1 - Q) \) are the a priori probabilities of the signals. When the system has no knowledge of \( Q \), it is reasonable to assume \( Q = \frac{1}{2} \); but such an assumption would be disastrous for cases where the signal-to-

\(^3\) In both examples, \( l = 1 \) for convenience.
noise ratio is marginal and the true value of $Q \neq \frac{1}{2}$; because then the first term of (24) is significant. If the new coding scheme is used, then the receiver can start by assuming $Q = \frac{1}{2}$, and update the decision boundary (threshold) with an estimate of $Q$ using equation (22c). A block diagram of such a system is shown in Figure 1.

**Example 2**

In the previous example it is assumed that the channel statistics are known. This restriction is now relaxed. The a priori knowledge available to the receiver is knowledge of the center of inertia of $S_1$ and $S_2$, i.e.

$$S = \frac{1}{2}(S_1 + S_2)$$  \hspace{1cm} (26)

If the new coding scheme described in Section I is used, an optimum decision rule against the available a priori knowledge is

$$\frac{Qf(\lambda \mid S_1)}{(1 - Q)f(\lambda \mid S_2)} > 1: \text{ announce } S_1$$

$$\frac{(1 - Q)f(\lambda \mid S_2)}{(1 - Q)f(\lambda \mid S_2)} \leq 1: \text{ announce } S_2$$  \hspace{1cm} (27)

where $\lambda$ is the number of times the signal exceeds the threshold (see Section II). Substituting the corresponding expressions for $f(\lambda \mid S_1)$ and $f(\lambda \mid S_2)$ in the above decision rule, and then taking the logarithm of both sides, the decision rule can be expressed as follows

announce $S_1$ if

$$\lambda > \frac{\ln \frac{1 - Q}{Q} + \nu \ln \frac{p_2}{1 - p_1}}{\ln \frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}$$  \hspace{1cm} (28)

otherwise announce $S_2$.
If the parameters $p_1$, $p_2$ and $Q$ are unknown they can be estimated using equations (22a), (22b), and (22c) in Section IV. A block diagram of a communication system with the above capabilities is shown in Figure 2.

VI. CONCLUSIONS

Binary codes are constructed for communicating an alphabet consisting of two letters over an unknown, stationary binary channel. The a priori probability ($Q$) of the letter denoted as class 1 is learned without supervision according to Equation (17c), with sample moments substituted in the corresponding expressions for $S$ and $T$. If these codes are used, it is thus possible to learn the a priori class probabilities without supervision. This has application in communication problems where a priori class probabilities are time varying, but may be assumed fixed over a short time period during which unsupervised learning takes place.

APPENDIX A

**Proposition.** Let $\mathcal{F}_i = \{F(\lambda, v, p_i'); i = 1, 2\}, j = 1, 2$ denote two families of binomial distributions, and let $G_1$ and $G_2$ denote parameter conditional mixtures of $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively, with known mixing parameter $Q$. That is

$$G_1 = QF(\lambda, v, p_1') + (1 - Q)F(\lambda, v, p_2') \quad (A-1)$$

and

$$G_2 = QF(\lambda, v, p_1'') + (1 - Q)F(\lambda, v, p_2'') \quad (A-2)$$
then a necessary and sufficient condition for
\[ G_1 = G_2 \]  

(A-3a)

to imply \( p_i' = p_i'' \), \( i = 1, 2 \), given that

\[ p_1' > p_2' \text{ and } p_1'' > p_2'' \]  

(A-3b)
is that

\[ v \geq 2 \]  

(A-3c)

Proof. The moment generating function of the binomial distribution is given by

\[ M\{F(\lambda, v, p)\} = [pe^t + (1 - p)]^v \]  

(A-4)

Let

\[ M_1(t) = Q(p_1'e^t + (1 - p_1')) + (1 - Q)(p_2'e^t + (1 - p_2'))^v \]  

(A-5)

\[ M_2(t) = Q(p_1''e^t + (1 - p_1'')) + (1 - Q)(p_2''e^t + (1 - p_2''))^v \]  

(A-6)

then equation (A-3) is equivalent to

\[ M_1(t) = M_2(t) \]  

(A-7)

Since all derivatives of \( M_1 \) and \( M_2 \) exist, an equivalent way of expressing (A-7) is

\[ \frac{d^n}{dt^n} M_1(t) = \frac{d^n}{dt^n} M_2(t) \]  

(A-8)

for all \( n \) and all \( t \). It should be pointed out that if \( n \geq v \), then only the first \( v \) derivatives are linearly independent. On the other hand, if \( n < v \), all \( n \) derivatives are linearly independent. Without any loss of generality let \( n \geq v \) and consider the first two derivatives of both sides of equation (A-7), evaluated at \( t = 0 \), to obtain

\[ Qp_1' + (1 - Q)p_2' = A \]  

(A-9)

\[ Q(p_1')^2 + (1 - Q)(p_2') = B \]  

(A-10)

where

\[ A = Qp_1'' + (1 - Q)p_2'' \]  

(A-11)

\[ B = Q(p_1'')^2 + (1 - Q)p_2'' \]  

(A-12)
Equations (A-9) and (A-10) are linearly independent and can be solved for \( p_1 \) and \( p_2 \) to obtain:

\[
\begin{align*}
    p_1' &= A - \left( \frac{1 - Q}{Q} (B - A^2) \right)^{1/2} = p_1'' \quad \text{(A-13a)} \\
    p_2' &= A + \left( \frac{Q}{1 - Q} (B - A^2) \right)^{1/2} = p_2'' \quad \text{(A-13b)}
\end{align*}
\]

and

\[
\begin{align*}
    p_1' &= A + \left( \frac{1 - Q}{Q} (B - A^2) \right)^{1/2} \\
        &= p_1'' + 2(1 - Q)(p_2'' - p_1'') \quad \text{(A-14a)} \\
    p_2' &= A - \left( \frac{Q}{1 - Q} (B - A^2) \right)^{1/2} = p_2'' - 2Q(p_2'' - p_1'') \quad \text{(A-14b)}
\end{align*}
\]

Solution (A-14) has to be dismissed because it violates hypothesis (A-3b). The fact that for \( v \geq 2 \) a unique solution for the system of equation (A-9) and (A-10) results, and proves the sufficiency of conditions (A-3c). Now consider the case where \( v = 1 \), then from (A-1), (A-2), and (A-3a) we obtain

\[
\begin{align*}
    Qp_1' + (1 - Q)p_2'' &= Qp_1'' + (1 - Q)p_2'' \quad \text{(A-15a)} \\
    Q(1 - p_1') + (1 - Q)(1 - p_2') \\
        &= Q(1 - p_1'') + (1 - Q)(1 - p_2'') \quad \text{(A-15b)}
\end{align*}
\]

Both equations (A-15a) and (A-15b) simplify to

\[
Q(p_i' - p_1'') = (1 - Q)(p_2'' - p_2'') \quad \text{(A-16)}
\]

or

\[
\frac{p_1' - p_1''}{p_2'' - p_2'} = \frac{1 - Q}{Q} \quad \text{(A-17)}
\]

Equation (A-17) implies that for \( v = 1 \), less than two, there are infinite many solutions, which proves the necessity of condition (A-3c).

APPENDIX B

PROPOSITION. Let \( \mathcal{F}_1 = \{ F(\lambda, v, p'), i = 1, 2 \} \) and \( \mathcal{F}_2 = \{ F(\lambda, v, p_2), i = 1, 2 \} \) denote two families of binomial distributions and let \( G_1 \) and \( G_2 \)
denote parameter conditional mixtures of $\bar{\mathcal{F}}_1$ and $\bar{\mathcal{F}}_2$ respectively, with corresponding mixing parameters $Q'$ and $Q''$, then a necessary and sufficient condition for

$$G_1 \equiv G_2$$

where

$$G = \binom{\nu}{\lambda} [Qp_1 \lambda(1 - p_1)^{\nu - \lambda} + (1 - Q)p_2 \lambda(1 - p_2)^{\nu - \lambda}]$$

and

$$p_1 \neq p_2$$

to imply

$$Q' = Q''$$
$$p_1' = p_1''$$
$$p_2' = p_2''$$

is that $\nu \geq 3$.

Proof. Proceeding as in Appendix A, let

$$M_1(t) = Q'[p_1'e^t + (1 - p_1')]^\nu + (1 - Q')[p_2'e^t + (1 - p_2')]^\nu$$

and

$$M_2(t) = Q''[p_1''e^t + (1 - p_1'')]^\nu + (1 - Q'')[p_2''e^t + (1 - p_2'')]^\nu$$

then equation (B-1) is equivalent to

$$M_1(t) = M_2(t)$$

for all $t$.

Since all derivatives of $M_1$ and $M_2$ exist, then an equivalent way of expressing (B-3) will be

$$\frac{d^n}{dt^n} M_1(t) = \frac{d^n}{dt^n} M_2(t)$$

for all $n$ and all $t$. But as in Appendix A, if $n > \nu$, only the first $\nu$ derivatives of (B-4) are linearly independent. Consider the first three derivatives of both sides of (B-4), evaluated at $t = 0$ to obtain:

$$Q'(p_1' - p_2') = A - p_2'$$

(B-5a)
\[ Q'(p_1' - p_2')(p_1' + p_2') = B - (p_2')^2 \]  
\[ Q(p_1' - p_2')(p_1'^2 + p_1' p_2' + p_2'^2) = C - (p_2')^3 \]

where

\[ A = Q''(p_1'') + (1 - Q'')(p_2'') \]
\[ B = Q''(p_1'') + (1 - Q'')(p_2'') \]
\[ C = Q''(p_1'')^3 + (1 - Q'')(p_2'')^2 \]

Equations (B-5a), (B-5b), and (B-5c) constitute a system of three linearly independent equations with three unknowns \((p_1', p_2', Q')\), which we proceed to solve.

Substituting (B-5a) into (B-5b) and (B-5c) we obtain

\[ A(p_1' + p_2') - p_2' p_1' = B \]  
\[ A(p_1' + p_2')^2 - A p_1' p_2' - p_1' p_2'(p_1' + p_2') = C \]

defining a set of new variables

\[ S = p_1' + p_2' \]
\[ T = p_1' p_2' \]

and substituting into (B-7a) and (B-7b) we obtain

\[ AS - T = B \]  
\[ AS - AT - TS = C \]

Equations (B-9a) and (B-9b) can be solved for \(T\) and \(S\) to obtain

\[ T = \frac{(A - B^2)}{B - A^2} = p_1'' p_2'' \]  
\[ S = \frac{C - BA}{B - A^2} = p_1'' + p_2'' \]

or

\[ p_1' = \frac{1}{2}[p_1'' + p_2'' \mp (p_2'' - p_1'')] \]  
\[ p_2' = \frac{1}{2}[p_1'' + p_2'' \pm (p_2'' - p_1'')] \]

Substituting (B-11a) and (B-11b) into (B-7a) we obtain

\[ Q = \frac{-2Q(p_2'' - p_1'') + (p_2'' - p_1'') \mp ((p_2'' - p_1'')^2)^{1/2}}{(p_2'' - p_1'')^{1/2}} \]
therefore the system of equations (B-7a), (B-7b) and (B-7c), has two solutions:

If we take the minus sign in (B-11a) we have

\[ p_1' = p_1'' \]
\[ p_2' = p_2'' \]
\[ Q' = Q'' \]

if we take the plus sign in (B-11a) we have

\[ p_1' = p_2'' \]
\[ p_2' = p_1'' \]
\[ Q' = 1 - Q'' \]

which is a change of label and not a second solution, thus of the sufficiency of the condition \( v \geq 3 \) is proven. To prove the necessity of \( v \geq 3 \), let \( v = 2 \) in (B-1) to obtain

\[ Q' (1 - p_1')^2 + (1 - Q')(1 - p_2')^2 = Q'' (1 - p_1'')^2 + (1 - Q'')(1 - p_2'')^2 \text{ for } \lambda = 0 \]  
(B-12)

\[ Q' p_1' (1 - p_1') + (1 - Q')p_2' (1 - p_2') = Q'' p_1'' (1 - p_1'') + (1 - Q'')p_2'' (1 - p_2'') \text{ for } \lambda = 1 \]  
(B-13)

and

\[ Q' p_1'^2 + (1 - Q')p_2'^2 = Q'' p_1''^2 + (1 - Q'')p_1''^2 \text{ for } \lambda = 2 \]  
(B-14)

Substituting (B-14) into (B-12) and (B-13) we obtain two identical equations, namely

\[ Q' p_1' + (1 - Q') p_2' = Q'' p_1'' + (1 - Q'') p_2'' \]  
(B-15)

Therefore the original system of equations (B-12), (B-13), (B-14) is undetermined and thus the necessity of \( v \geq 3 \) is proven.

Received: July 29, 1968

References


