Ath. Kehagias.
"Bayesian Classification of Hidden Markov Models".

This paper has appeared in the journal:
Mathematical and Computer Modelling, Vol.23, pp.25-43, 1996.

Abstract

We develop a recursive Maximum A Posteriori Classification algorithm for discrete valued Stochastic Processes modelled by Hidden Markov Models. The classification algorithm solves recursively the following problem: given a collection of HMM's (P^θ, Q^θ), $\theta \in \Theta$, and a sequence of observations y_1, ... , y_t from a stochastic process $\{Y_t\}_{t=1}^{\infty}$, find the HMM that has Maximum Posterior Probability of producing y_1 ,..., y_t. This algorithm is a modification (for discrete valued stochastic processes) of the Lainiotis Partition algorithm [10], [15]. We prove that, subject to ergodicity and positivity assumptions on $\{Y_t\}_{t=1}^{\infty}$, our algorithm will converge to the "right" (in the cross entropy sense) HMM as $t \rightarrow \infty$, for almost all sequences y_1, y_2, ... Finally, we give an example of the application of our algorithm to the classification of speech signals.

Bayesian Classification of Hidden Markov Models

Athanasios Kehagias
Division of Electronics and Computer Engineering
Department of Electrical Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
e-mail: kehagias@egnatia.ee.auth.gr

June 17, 1997

This paper appeared in *Mathematical and Computer Modelling*, Vol.10, pp.25-43, 1996

1 Introduction

Consider the following classification problem. We observe a discrete valued, stationary ergodic stochastic process for which there is a countable (finite or infinite) number of Hidden Markov Models. At time t the observations y_1, ... , y_t are available, and we use these to classify the process to the model that maximizes the Bayesian posterior probability of y_1, ... , y_t. In this paper we do the following.

1. We develop a recursive algorithm to compute the posterior probabilities. This algorithm combines elements of two algorithms previously reported in the literature: the Backward - Forward algorithm of Baum [5, 6] and the Partition Algorithm of Lainiotis [13, 17].

2. We prove the convergence of our classification algorithm: the posterior probability of the "best" (in a precisely defined sense) candidate model tends to one almost surely.

3. We present examples of classification using speech data and phoneme Hidden Markov Models.

Hidden Markov Models (HMM) are used widely for speech recognition [1, 25, 24, 27, 28] and they have lately been introduced in a number of other applications, e.g. shape recognition [26], arterial modelling [10], biological applications [2], etc. Here we concentrate on HMM's with discrete valued hidden (or state) and observable process. The importance of these models is underscored by recent results about their universal representation and consistent estimation properties (see [14]).

Our classification algorithm is modelled after the Partition Algorithm of Lainiotis, which has been used for parameter estimation of stochastic control systems. An early version of this algorithm can be found in [13, 17]. This is a very general algorithm, which applies to continuous- as well as discrete-valued stochastic processes. However, details for the discrete valued case have not been worked out in the literature. Control theoretic applications can be found in [11, 17, 18, 19] and computational issues are treated in [29, 30]. In all of these papers the algorithm is applied to continuous valued stochastic processes; this is also the case for more recent theoretical developments in [20, 21], as well as applications to seismic signal processing [22, 23]. The convergence of the Partition Algorithm for continuous valued processes is discussed in [15, 31]. To the best of our knowledge, convergence for discrete valued processes has not been studied so far.

The classification problem can be formulated as a problem of parameter estimation: we introduce a parameter θ such that to every value of θ corresponds a HMM and it is required to estimate (see next section) the optimal value of θ. In this sense, we present here a parameter estimation algorithm; perhaps it should be briefly compared to the Backward-Forward (BF) parameter estimation algorithm by Petrie and Baum (see [5, 6, 27, 28]). While our algorithm borrows from BF the computation of forward probabilities, in most respects the two algorithms are quite different. Our algorithm assumes a *countable* number of possible parameter values and proceeds to find the optimal value by an *online* process; the BF algorithm can operate with an *uncountably infinite* (i.e. continuous valued) set of parameters, but is an *offline* algorithm. While our algorithm uses only the forward probabilities, the BF algorithm (as the name indicates) uses forward *and* backward probabilities. In terms of implementation, this means that our algorithm performs only a forward integration in time and finds an *exact* maximizing parameter value; while the BF algorithm performs a sequence of forward and backward integrations to *approximate* a maximizing parameter value.

This paper is organized as follows. In Section 2 we present some definitions and notation that relate to Hidden Markov Models. In Section 3 we present our classification algorithm. In Section 4 we present our convergence results: the algorithm converges almost surely to the best HMM, if $\{Y_t\}_{t=1}^{\infty}$ satisfies some ergodicity and positivity conditions. The proofs of these results are deferred to the Appendix. In Section 5 we present some numerical experiments of classification, using speech data. In Section 6 we present the conclusions this paper. Finally, the proof of the algorithm convergence is presented in the Appendix.

2 Preliminaries

In this section we present some definitions and notation that relate to stationary stochastic processes and Hidden Markov Models. The material is standard (see e.g. Billingsley [7, 8]).

We study discrete valued, stationary stochastic processes, e.g. $\{Y_t\}_{t=1}^{\infty}$, taking values in $\Omega_Y = \{1, 2, ..., L\}$. There is no loss of generality in assuming Y_t to be integer valued. We define

$$\Omega_Y^N = \{y_1...y_N, y_n \in \Omega_Y \ \ 1 \leq n \leq N\} \ \ N = 1, 2, ... , \tag{1}$$

$$\Omega_Y^{\infty} = \{y = y_1 y_2 ..., y_n \in \Omega_Y \ \ 1 \leq n\}. \tag{2}$$

$\{Y_t\}$ has probability function p defined for all $N \geq 0$, $y_1 y_2 ... y_N \in \Omega_Y^N$ by

$$p(y_1...y_N) \doteq Prob(Y_1 = y_1, ..., Y_N = y_N). \tag{3}$$

By stationarity we have $p(y_1...y_N) = Prob(Y_{t+1} = y_1, ..., Y_{t+N} = y_N)$. Note that p is a complete description of $\{Y_t\}$.

Define $\sigma(\Omega_Y)$ to be the smallest sigma-algebra that contains all *cylinder sets* (see Billingsley [8]) of Ω_y^∞. The probability function p can be used to define a measure on the cylinder sets and this measure can be uniquely extended on $\sigma(\Omega_Y)$. Define this measure by π. By the previous discussion π is determined by p; conversely we can recover p from the value of π on cylinder sets. Hence $\{Y_t\}$, p and π are equivalent descriptions of a discrete valued stationary stochastic process.

We use a somewhat restricted definition of Hidden Markov Models. A Hidden Markov Model (HMM) is a pair of stationary stochastic processes $(\{X_t\}, \{Y_t\})$. $\{X_t\}$ takes values on $\Omega_X = \{1, 2, ..., K\}$ and is Markov with transition matrix P, defined by $P_{xz} \doteq Prob(X_t = z \mid X_{t-1} = x)$. $\{Y_t\}$ takes values on $\Omega_Y = \{1, 2, ..., L\}$ and depends *instantaneously* on X_t via the emission matrix Q: $Q_{xy} \doteq Prob(Y_t = y \mid X_t = x, X_s, s \neq t)$, independent of X_s, $s \neq t$.

In what follows we use a countable parameter set Θ such that for every $\theta \in \Theta$ we have a HMM $(\{X_t^\theta\}, \{Y_t^\theta\})$, with transition and emission matrices (P^θ, Q^θ). We will always assume that for all $\theta \in \Theta$ $P^\theta > 0$; in this case $\{X_t^\theta\}$ is ergodic and has a unique stationary probability distribution (determined by P^θ, see [9]), call it p^θ. For all $x \in \Omega_X$ we have $p^\theta(x) \doteq Prob(X_t^\theta = x)$. Using $p^\theta(x)$, P_{xz}^θ, Q_{xy}^θ, $x, z \in \Omega_X$ $y \in \Omega_Y$ we can obtain the probability functions of $\{X_t^\theta\}$ and $\{Y_t^\theta\}$. Hence (P^θ, Q^θ) is yet another complete description of $(\{X_t^\theta\}, \{Y_t^\theta\})$.

We use the following rather abusive (but hopefully not confusing) notation:

$$p^\theta(x_1...x_t) \doteq Prob(X_1^\theta = x_1...X_t^\theta = x_t), \tag{4}$$

$$p^\theta(y_1...y_s) \doteq Prob(Y_1^\theta = y_1...Y_s^\theta = y_s), \tag{5}$$

$$p^\theta(x_1...x_t, y_1...y_s) \doteq Prob(X_1^\theta = x_1...X_t^\theta = y_t, Y_1^\theta = y_1...Y_s^\theta = y_s), \qquad (6)$$

$$p^\theta(y_1 \mid x_0, y_0) \doteq Prob(Y_1^\theta = y_1 \mid X_0^\theta = x_0, Y_0^\theta = y_0) \quad \text{etc.} \qquad (7)$$

ATTENTION: The symbol $p_t^\theta(y_1...y_t)$ (i.e. with an additional t subscript) is used to denote a completely different quantity, the *model posterior probability*, which will be defined in the next section.

Also, we commit another gross but harmless offense: sometimes we write

$$p^\theta(y_0 \mid y_{-1}...y_{-t}) = Prob(Y_0^\theta = y_0 \mid Y_{-1}^\theta = y_{-1}, ..., Y_{-t}^\theta = y_{-t}). \qquad (8)$$

Now, strictly speaking, our processes are not defined for $t < 0$. However it is a simple matter to extend a one-sided ($0 \leq t < \infty$) stationary stochastic process to a two-sided one ($-\infty < t < \infty$) (see [7]), so the expression in eq.(8) is meaningful.

Finally we define the entropy $H(p)$ of a process p and the cross entropy $H(q;p)$ of a process q with respect to a process p.

$$H(p) \doteq -\int \log p(y_0 \mid y_{-1}y_{-2}...)d\pi(y_0y_{-1}...). \qquad (9)$$

$$H(q;p) \doteq -\int \log q(y_0 \mid y_{-1}y_{-2}...)d\pi(y_0y_{-1}...) - H(p). \qquad (10)$$

These are *formally* defined in eqs.(9,10); but for any specific processes it must be proven that the integrals above are well defined, otherwise the definition is vacuous.

3 The Classification Algorithm

We consider the following problem of stochastic process classification. Start with a collection of Hidden Markov Models: $\{(P^\theta, Q^\theta)\}_{\theta\in\Theta}$, where Θ is a countable (finite or infinite) parameter set. Next, introduce a random variable Z which takes values in Θ. At time t the value of Z is chosen randomly, according to the probability distribution $p_0^\theta \doteq Prob(Z = \theta)$,

$\theta \in \Theta$. Z remains fixed for $t \geq 0$, but its value is unknown to us. [1] For time $t = 1, 2, ...$ the HMM (P^Z, Q^Z) produces an observable stochastic process $\{Y_t\}_{t=1}^{\infty}$. We observe a realization of $\{Y_t\}$, call it y_1, y_2, ... and we want to infer the value of Z from this sequence. A problem of this type is, for example, phoneme classification, with Θ being the finite set of the English language phonemes.

To solve this problem, we adopt a Bayesian point of view. Before any observations y_1, y_2, ... are available, we can express our *prior belief* of the value of Z in terms of the prior probability distribution $\{p_0^\theta\}_{\theta \in \Theta}$. At time t the observations y_1, ... , y_t are available and our knowledge of Z has improved. This knowledge is now expressed in terms of the posterior probability (at time t):

$$p_t^\theta(y_1...y_t) \doteq Prob(Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t). \tag{11}$$

Given $p_t^\theta(y_1...y_t)$, our estimate of Z at time t is

$$\hat{\theta}_t(y_1, ..., y_t) \doteq \arg\max_{\theta \in \Theta} p_t^\theta(y_1...y_t). \tag{12}$$

Thus, at time t we claim that the data $y_1, ..., y_t$ was produced by $\hat{\theta}_t(y_1, ..., y_t)$, which maximizes the posterior probability. This is a reasonable choice, usually referred to as *Maximum A Posteriori* (MAP) estimate. [2]

Remark: Notice the difference between Z, which is fixed for all time $t \geq 0$ (reflecting the fact that $y_1, y_2, ...$ are indeed produced by a fixed HMM) and $\hat{\theta}_t$ which changes as more data is collected (reflecting our *belief* about what is the true model). Our belief may change over time, so for $t \neq \tau$ we may have $\hat{\theta}_\tau \neq \hat{\theta}_t$. However, it is desirable that as $t \to \infty$ we have $\hat{\theta}_t \to Z$, in some appropriate sense of stochastic convergence.

Remark: The assumption that the observations y_1, y_2, ... are actually produced by a Hidden Markov Model, is not really necessary. We adopt it here for simplicity of exposition; soon we will remove it. However, it is not

[1] In what follows we assume that $p_0^\theta > 0$ for all $\theta \in \Theta$. It will soon become obvious that this can be done without loss of generality.

[2] For brevity of notation, sometimes we will write simply p_t^θ, $\hat{\theta}_t$, dropping $y_1...y_t$.

excessively restrictive, in light of the following result (proven in [14]): subject to certain ergodicity and positivity assumptions, every discrete-valued stochastic process can be approximated arbitrarily well by a sequence of Hidden Markov Models; these models can also be consistently estimated.

The classification problem has now been reduced to finding an efficient way to compute p_t^θ, $\theta \in \Theta$, $t = 1, 2, \ldots$. To do this, we develop a recursive algorithm. First, note that

$$p_{t+1}^\theta = Prob(Z = \theta \mid Y_1 = y_1, ..., Y_{t+1} = y_{t+1}) =$$

$$\frac{Prob(Y_{t+1} = y_{t+1}, Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t)}{Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t)} =$$

$$\frac{Prob(Y_{t+1} = y_{t+1}, Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t)}{\sum_{\zeta\in\Theta} Prob(Y_{t+1} = y_{t+1}, Z = \zeta \mid Y_1 = y_1, ..., Y_t = y_t)}. \tag{13}$$

Also note that

$$Prob(Y_{t+1} = y_{t+1}, Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t) =$$

$$Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = \theta) \cdot Prob(Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t) =$$

$$Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = \theta) \cdot p_t^\theta. \tag{14}$$

Now (13,14) imply the recursion:

$$p_{t+1}^\theta = \frac{Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = \theta) \cdot p_t^\theta}{\sum_{\zeta\in\Theta} Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = m) \cdot p_t^\zeta}. \tag{15}$$

Remark: The last equation shows why we can assume $p_0^\theta > 0$ for all $\theta \in \Theta$ without loss of generality. It is obvious that if there is a θ such that $p_0^\theta = 0$, then $p_t^\theta = 0$ for all $t \geq 0$; hence models with zero prior probability need not be discussed at all.

What remains to be done is finding a recursive way to compute the quantity $Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = \theta)$ for $\theta \in \Theta$, $t =$

$1, 2, ...$. Note that

$$Prob(Y_{t+1} = y_{t+1} \mid Y_1 = y_1, ..., Y_t = y_t, Z = \theta) =$$

$$\frac{p^\theta(y_1 ... y_{t+1})}{p^\theta(y_1 ... y_t)}. \tag{16}$$

However, if $Z = \theta$ then the probabilities in eq.(16) can be computed in terms of the matrices P^θ, Q^θ. We present a method for recursive computation of $p^\theta(y_1, ..., y_t)$; this is developed in [25] as part of the Forward - Backward algorithm. For every $\theta \in \Theta$, define the *forward probabilities* for given observations $y_1, y_2, ...$, for $x = 1, 2, ..., K$ and $t = 1, 2, ...$

$$\alpha_t^\theta(x) \doteq Prob(Y_1 = y_1, ..., Y_t = y_t, X_t = x \mid Z = \theta), \tag{17}$$

It is easily checked that the evolution equation for the forward probabilities is:

$$\alpha_t^\theta(x) = \sum_{z=1}^{K} \alpha_{t-1}^\theta(z) P_{zx}^\theta Q_{xy_t}^\theta \qquad x = 1, 2, ..., K. \tag{18}$$

Assume all initial states to be equally likely – then the initial condition for α^θ is $\alpha_0^\theta(x) \doteq Prob(X_0 = x \mid Z = \theta) = \frac{1}{K}$ for $x = 1, 2, ..., K$.

Now use the α^θ's to compute the p^θ's:

$$p^\theta(y_1 ... y_t) = \sum_{x=1}^{K} \alpha_t^\theta(x). \tag{19}$$

This completes the description of the recursive classification algorithm. Putting all the pieces together we get:

Maximum A Posteriori Classification Algorithm

Given an observation sequence $y_1, y_2, ...$ and a set of HMM's (P^θ, Q^θ), $\theta \in \Theta$, assume that the sequence has been produced by the HMM (P^Z, Q^Z), where

Z is a random variable with probability distribution $p_0^\theta = Prob(Z = \theta)$, $\theta \in \Theta$. The MAP estimate of Z at time t is $\hat{\theta}_t$, defined as

$$\hat{\theta}_t \doteq \arg\max_{\theta \in \Theta} p_t^\theta(y_1 ... y_t) \tag{20}$$

where p_t^θ is defined for all $\theta \in \Theta$, $t = 1, 2, ...$ by

$$p_t^\theta(y_1 ... y_t) \doteq Prob(Z = \theta \mid Y_1 = y_1, ..., Y_t = y_t). \tag{21}$$

To obtain the MAP estimate, compute p_t^θ for $\theta \in \Theta$, $t = 1, 2, ..$ as follows. For $t = 0$

$$p_0^\theta = Prob(Z = \theta) \qquad \theta \in \Theta. \tag{22}$$

$$\alpha_0^\theta(x) = \frac{1}{K} \qquad x = 1, 2, ..., K, \qquad \theta \in \Theta. \tag{23}$$

Then for $t = 1, 2, ...$:

1. Compute for all $\theta \in \Theta$, $x = 1, 2, ..., K$

$$\alpha_t^\theta(x) = \sum_{z=1}^{K} \alpha_{t-1}^\theta(z) P_{zx}^\theta Q_{xy_t}^\theta. \tag{24}$$

2. Compute for all $\theta \in \Theta$

$$Prob(Y_t = y_t \mid Y_1 = y_1, ..., Y_{t-1} = y_{t-1}, Z = \theta) = \frac{\sum_{x=1}^{K} \alpha_t^\theta(x)}{\sum_{x=1}^{K} \alpha_{t-1}^\theta(x)} \tag{25}$$

   and

$$p_t^\theta = \frac{Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^\theta}{\sum_{\zeta \in \Theta} Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^\zeta} \tag{26}$$

3. Finally, set

$$\hat{\theta}_t \doteq \arg\max_{\theta \in \Theta} p_t^\theta(y_1, ..., y_t). \tag{27}$$

Remark: The process that produces the observation sequence y_1, y_2, ... need not be Hidden Markov. The classification algorithm applies to any discrete valued process and locates the HMM (P^θ, Q^θ) that has maximum posterior likelihood with respect to y_1, y_2, ... This is the subject of Theorems 1, 2 of Section 4. Experimental verification of this fact is given in Section 5.

Remark: The MAP classification algorithm is the adaptation of the Lainiotis Partition algorithm [13, 17] to discrete valued stochastic processes.

4 Convergence of the Classification Algorithm

We now state our convergence results. Proofs will be deferred to the Appendix. The results hold true under the following assumptions.

A Take a stochastic process $\{Y_t\}_{t=1}^{\infty}$ (with probability p) that satisfies the following assumptions.

A1 $\{Y_t\}_{t=1}^{\infty}$ is stationary ergodic.

A2 $\forall t \quad Y_t \in \Omega_Y = \{1, 2, ..., L\}$.

A3 $\exists \alpha > 0$ such that $\forall y = y_1 y_2 ... \in \Omega_Y^{\infty}$, $p(y_t \mid y_1 ... y_{t-1}) \geq \alpha$.

B Also take a collection of HMM's $\{(P^\theta, Q^\theta)\}_{\theta \in \Theta}$, Θ countable, that satisfy the following assumptions.

B1 $\forall \theta \in \Theta$ P^θ is K-by-K.

B2 $\forall \theta \in \Theta$ Q^θ is K-by-L.

B3 $\exists \beta > 0$ such that $\forall \theta \in \Theta$, $\forall x, z \in \Omega_X = \{1, 2, ..., K\} \quad P_{xz}^\theta \geq \beta$.

B4 $\exists \gamma > 0$ such that $\forall \theta \in \Theta$, $\forall x \in \Omega_X = \{1, 2, ..., K\}$ and $\forall y \in \Omega_Y = \{1, 2, ..., L\} \quad Q_{xy}^\theta \geq \gamma$.

Then we have the following

Theorem 1 *If conditions [A1-A3] and [B1-B4] are satisfied, then for any* $\theta, \zeta \in \Theta$ *such that* $H(p^{\zeta}; p) < H(p^{\theta}; p)$ *we have*

$$\lim_{t \to \infty} \frac{p_t^{\theta}(y_1 ... y_t)}{p_t^{\zeta}(y_1 ... y_t)} = 0 \tag{28}$$

for π*-almost all* $y = y_1 y_2 ...$.

Remark: Theorem 1 is similar to martingale convergence theorems for likelihood ratios [16], but there is an important difference: neither p^{θ} nor p^{ζ} is assumed to be the true process from which $y_1, y_2, ...$ come. In fact, as already mentioned, we do not even need assume that $\{Y_t\}$ is HMM.

From Theorem 1 we can easily get a "consistent classification" theorem. First we need the following notation. For every $\delta \geq 0$ define $\Theta_{\delta} \doteq \{\theta \in \Theta : H(p^{\theta}; p) \leq \delta\}$ and $\Theta^{\delta} \doteq \{\theta \in \Theta : H(p^{\theta}; p) > \delta\}$. Obviously, $\forall \delta \geq 0$ we have $\Theta_{\delta} \cup \Theta^{\delta} = \Theta$. Now we can state the following theorem.

Theorem 2 *If conditions [A1-A3] and [B1-B4] are satisfied, then*

Case 1: $\mid \Theta \mid < \infty$**.** *Define* $h^0 \doteq \min_{\theta \in \Theta} H(p^{\theta}; p)$*. Then*

$$\lim_{t \to \infty} \sum_{\theta \in \Theta_{h^0}} p_t^{\theta}(y_1 ... y_t) = 1$$

$$\lim_{t \to \infty} \sum_{\theta \in \Theta^{h^0}} p_t^{\theta}(y_1 ... y_t) = 0$$

for π*-almost all* $y = y_1 y_2 ... \in \Omega_Y^{\infty}$.

Case 2: $\mid \Theta \mid = \aleph_0$**.** *If there are* δ, ϵ *such that* $0 \leq \delta < \epsilon$ *and* $\Theta = \Theta_{\delta} \cup \Theta^{\epsilon}$, *then*

$$\lim_{t \to \infty} \sum_{\theta \in \Theta_{\delta}} p_t^{\theta}(y_1 ... y_t) = 1 \tag{29}$$

$$\lim_{t \to \infty} \sum_{\theta \in \Theta^{\epsilon}} p_t^{\theta}(y_1 .. y_t) = 0 \tag{30}$$

for π*-almost all* $y = y_1 y_2 ... \in \Omega_Y^{\infty}$..

Remark: In Case 1 (finite Θ) there is at least one value of θ which achieves the minimum cross entropy $H(p^\theta;p)$. There may be more than one such θ; Θ_{h^0} is the set of all such minimizing θ's. If Θ_{h^0} is not a singleton, it is clear that we cannot , in general, expect $\hat{\theta}_t \rightarrow Z$. For instance, if model (P^θ, Q^θ) produces the observations, and there is a model (P^ζ, Q^ζ) such that P^ζ is a permutation of P^θ, Q^ζ is the same permutation of Q^θ, then (P^θ, Q^θ) and (P^ζ, Q^ζ) have identical output behavior. At any rate, Theorem 2 states that the posterior probability will almost surely concentrate all its mass on Θ_{h^0} (the "good" models) as t goes to infinity.

Remark: In Case 2 (countably infinite Θ), it is not guaranteed that the minimum (the infimum, really) of $H(p^\theta;p)$ will be achieved. A further complication is that the "bad" values of θ may give a slow increase of cross entropy. Then (28) alone does not guarantee that all the probability mass is concentrated on the "good" models. Therefore, we need to impose the additional δ - ϵ condition, which ensures a sharp separation of the good and bad θ's. To clarify the nature of this condition, consider the case where the observations y_1, y_2, ... are actually produced by $(P^{\theta^*}, Q^{\theta^*})$, $\theta^* \in \Theta$. Then $p^{\theta^*} = p$, $H(p^{\theta^*};p) = 0$, and we can take $\delta = 0$. If there is also an $\epsilon > 0$ such that $\Theta_0 \cup \Theta^\epsilon = \Theta$, then for all θ in Θ^ϵ, (P^θ, Q^θ) is at least ϵ distant from p^{θ^*} (in the cross entropy sense) and, by Theorem 2 convergence is guaranteed.

Remark: Extensions to continuous valued processes are possible. These are the subject of current research and will be reported elsewhere. Let us briefly mention a simple case. Consider HMM with discrete state and continuous observable process. At every time step t a y_t is emitted, according to a probability *density* $q_x(y)$, where x is the current state. If (a) $q_x(y)$ is bounded below by $\gamma > 0$ on an interval A, with $A \subset (-\infty, \infty)$, (b) $q_x(y) = 0$ outside of A, and (c) γ and A are *independent* of x, then we can prove convergence results in exactly the same manner as Theorems 1 and 2. A look at the Appendix will convince the reader that the proof of the discrete-valued case carries over to this special continuous-valued case. Extension to more general cases, e.g. unbounded densities, is harder.

In Case 2 of Theorem 2 we prove convergence for the case of countably infinite parameter set. In a real world application we cannot actually apply the algorithm to an infinite set, because we cannot implement the computations of eqs.(20) to (27) for an infinite number of terms. Instead, we can truncate Θ to a finite subset (say by removing of consideration models with low prior probability or models that consistently perform poorly and hence receive low p_t^θ). The question then arises: what is the relationship of the new posterior probabilities (computed by operating on the truncated set) to the correct ones (obtained from operating on the original, infinite set)?

To make the question precise, consider the following case. First, for simplicity, assume that Θ is the set of positive integers: $\Theta = \{1, 2, 3, ...\}$. There is no loss of generality in this assumption. Also assume there is only one model of minimum cross entropy, and (again without loss of generality) that it is the first model ($\theta = 1$). In short:

C1 $H(p^1; p) < H(p^\theta; p)$ for $\theta = 2, 3, ...$.

Then the update of p_t^θ is given as usual, by eq.(26). Since we have taken Θ to be the positive integers, we can rewrite eq.(26) as

$$p_t^\theta = \frac{Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^\theta}{\sum_{\zeta=1}^\infty Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^\zeta}. \qquad (31)$$

We will also use the *truncated posteriors* $p_t^{\theta,N}$, for θ= 1, 2, ... , N, which are updated by

$$p_t^{\theta,N} = \frac{Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^{\theta,N}}{\sum_{\zeta=1}^N Prob(Y_t = y_t \mid Y_1 = y_1, .., Y_{t-1} = y_{t-1}, Z = \theta) \cdot p_{t-1}^{\zeta,N}}. \qquad (32)$$

In other words, to compute $p_t^{\theta,N}$ we start from the same initial values as for p_t^θ, but perform the summations only up to the N-th term. Since we also compute only a finite number of truncated posteriors, the total number of computations required is finite. This of course turns the problem to a finite one, and by Theorem 2 we will have $p_t^{1,N} \to 1$, $p_t^{\theta,N} \to 0$, and this for all N and π-a.a. So in this sense, truncating the parameter set Θ causes no damage. However, we now present a stronger result.

Theorem 3 *Suppose conditions [A1-A3], [B1-B4] and C1 are satisfied, and take* $\Theta = \{1, 2, 3, ...\}$, $\Theta_N = \{1, 2, ..., N\}$. *Then for all* θ *in* Θ_N *we have some* t_0 *such that* $\forall t \geq t_0$

$$\lim_{N\to\infty} \frac{p_t^{\theta,N}}{p_t^{\theta}} = 1 \quad \pi - a.a. \text{ and } \lim_{N\to\infty} \frac{p_t^{\theta} - p_t^{\theta,N}}{p_t^{\theta}} = 0 \quad \pi - a.a. \tag{33}$$

Remark: Theorem 3 shows that for every θ there is a value N such that the "relative error" (between the true and the N-truncated posterior) goes to zero. Also, by Theorem 2 we see that convergence to the "best" model is preserved for the truncated posteriors. Hence truncation gives quite good approximation and is an efficient method for dealing with infinite parameter sets.

5 Numerical Experiments

We now test our algorithm on some simple classification experiments. The observations y_1, y_2, ... are speech data. It must be emphasized that the point of these experiments is not to compare our algorithm to large speech recognizers currently in use, but simply to test our algorithm on real world data. The algorithm might be incorporated and evaluated as a component of a speech recognizer but this is not pursued here.

We start with an utterance of the word "one". This is sampled at 10 KHz and gives a continuous valued signal. To keep things simple, we use only two models: $\Theta = \{\theta_1, \theta_2\}$. θ_1 corresponds to the phoneme [ah]. We pick the relevant portion of the signal, subsample this (take every 5-th sample) and quantize at L=16 levels. We obtain two such sequences, each 120 steps long. We use one to train a HMM with $K = 12$ states and $L = 16$ observables. Note that this HMM produces output in the range $\{1, 2, ..., L\}$. In other words it reproduces the quantized speech signal, and not derivative quantities, such as LPC or FFT coefficients. The second sequence, plotted in Fig.??, is our observation sequence $y_1, \dots\ , y_{120}$. Similarly, θ_2 corresponds to the phoneme [n]. After subsampling and quantization, we get two sequences, each 120 steps long. We use one to train HMM with $K = 17$ and $L = 16$. The other sequence, appearing in Fig.??. is our observation sequence. We proceed to apply our classification algorithm.

In the first experiment we use the [ah] signal and want the probability $p_t^{\theta_1}$, as computed by our algorithm, to converge to 1. This indeed happens, as displayed in Fig.??.

Exactly similar results obtain in the second experiment, which is identical to the first one, except that the sequence $y_1, ..., y_{120}$ is now the [n] signal and we want the probability $p_t^{\theta_2}$ to go to one. This result is achieved, as displayed in Fig.??.

Finally we apply the classification algorithm to a signal which has an [ah] to [n] transition. We use the previously trained HMM's and a sequence form the transition region between the [ah] and [n] phonems. This sequence, after subsampling and quantization, is plotted in Fig.??.

There is a crucial difference between this and the previous experiments. The derivation of our algorithm was based on the assumption that the observations were produced by a single HMM. This assumption is obviously violated in this case, and there is no theoretical guarantee that the algorithm will work (Theorems 1 and 2 do not apply here). Still, for fairly long time intervals, the observation sequence *does* come from a *fixed* HMM. The desired behavior of the of the p_t^θ probabilities is that $p_t^{\theta_1}$ goes to 1 and stays there for a while, then it decays to 0, while $p_t^{\theta_2}$ rises to 1.

The classification algorithm can, in theory, reproduce this behavior, as is obvious from eq.(26). However, looking at this equation we discover a practical difficulty: if $p_t^\theta = 0$, for some θ and t, then $p_s^\theta = 0$ for all $s > t$. In practice, if p_t^θ goes below machine precision, it is set to 0. Even when no underflow occurs, if p_t^θ, becomes too small, the classification algorithm can be very slow to respond to a change of signal source (such as the [ah] to [n] transition).

To avoid these problems, we introduce the ad hoc precaution of keeping

p_t^θ above a threshold ϵ; ϵ is chosen small, but well above machine precision (say $\epsilon = 10^{-3}$). So, values of p_t^θ are not significantly changed, except when they become so small that they do not matter anyway. In this way, underflow and slow classification are avoided. Now we run our modified algorithm and obtain the results of Fig.**??** which are exactly what we wanted.

6 Conclusions

We have developed a MAP classification algorithm for discrete valued HMM's. This algorithm is recursive (and hence suitable for online implementation) and classifies quickly and accurately, as we have demonstrated using real speech data examples. We have also proven that it converges with probability 1, under certain mild assumptions.

A Appendix: Proof of Theorems

Here we prove Theorems 1, 2 and 3 of Section 4. The following lemmas will be needed; they all assume [A1-A3] and [B1-B4].

Lemma 4 *For all $\theta \in \Theta$*

$$\lim_{t\to\infty} -\frac{\log p^\theta(y_1 ... y_t)}{t} = H(p^\theta; p) + H(p) \qquad \pi - a.a. \ \ y_1 y_2 ... . \tag{34}$$

The proof of Lemma 4 will be given last, since it requires the following lemmas.

Lemma 5 *For all $\theta \in \Theta$, $l, m > 0$ x_{-m}, x_{-1}, x_0, $y_{-l}, ..., y_{-m}$ we have*

$$p^\theta(y_0 ... y_m, x_0 \mid x_{-1}, y_{-1} ... y_{-l}) = p^\theta(y_0 ... y_m, x_0 \mid x_{-1}) \tag{35}$$

$$p^\theta(y_0 \mid y_{-1} ... y_{-m} ... y_{-l}, x_{-m}) = p^\theta(y_0 \mid y_{-1} ... y_{-m+1}, x_{-m}). \tag{36}$$

Proof: This is a simple consequence of the Markov conditioning and can be verified by writing out the left and right sides of eq.(35) and cancelling equal terms. The proof of eq.(36) is similar. •

Lemma 6 *For all $\theta \in \Theta$, $k, l, m, n > 0$ and $x, x_k, ..., x_l$, $y_m, ..., y_n$ we have*

$$p^\theta(x) \geq \beta. \tag{37}$$

$$p^\theta(x_k...x_l, y_m...y_n) \geq 0, \tag{38}$$

$$p^\theta(y_0 \mid y_{-1}...y_{-m}) \geq \gamma, \tag{39}$$

Proof: Fix some $\theta \in \Theta$; since it is fixed, drop it from the notation for the rest of the proof, for simplicity. Now, since $\{p(x)\}_{x\in\Omega_X}$ is the stationary equilibrium probability of P, we have:

$$p(x) = \sum_{z\in\Omega_X} p(z)P_{zx} \geq \{\min_{z\in\Omega_X} P_{zx}\} \cdot \sum_{z\in\Omega_Y} p(z) \geq \beta \cdot 1.$$

This proves eq.(37). Next, take any $m < k < l < n$, x_k, ... , x_l and y_m, ... , y_n. We have:

$$p(x_k....x_l, y_m....y_n) =$$

$$\sum_{x_j\in\Omega_X, m\leq j<k \text{ or } l<j\leq n} p(x_m)Q_{x_m y_m}P_{x_m x_{m+1}}Q_{x_{m+1}y_{m+1}}...P_{x_{k-1}x_k}Q_{x_k y_k}...P_{x_l x_{l+1}}Q_{x_{l+1}y_{l+1}}...P_{x_{n-1}x_n}Q_{x_n y}$$

This proves eq.(38) for the particular ordering of k, l, m, n that we chose. The proof for any other ordering is exactly the same. Finally, to prove eq.(39), note that

$$p(y_0 \mid y_{-1}...y_{-m}) = \sum_{x_0,x_{-1}} p(y_0, x_0, x_{-1} \mid y_{-1}...y_{-m}) =$$

(using eq.(35))

$$\sum_{x_0,x_{-1}} p(y_0, x_0 \mid x_{-1}) \cdot p(x_{-1} \mid y_{-1}...y_{-m}) =$$

$$\sum_{x_0,x_{-1}} P_{x_{-1}x_0}Q_{x_0y_0}p(x_{-1} \mid y_{-1}...y_{-m}) \geq$$

$$\gamma \cdot \sum_{x_0,x_{-1}} p(x_0, x_{-1} \mid y_{-1}...y_{-m}) = \gamma > 0$$

which completes the proof of the Lemma. •

The next lemma appears in Petrie and Baum [4]; but the basic idea is standard in the treatment of Markov chains (see, e.g. Doob [9]).

Lemma 7 *For all* $\theta \in \Theta$, $y_0 y_{-1} y_{-2} ... \in \Omega_Y^\infty$, $m > 0$ *define*

$$D_m^\theta(y) \doteq \max_{x_{-m} \in \Omega_X} p^\theta(y_0 \mid y_{-1} ... y_{-m}, x_{-m}),$$

$$d_m^\theta(y) \doteq \min_{x_{-m} \in \Omega_X} p^\theta(y_0 \mid y_{-1} ... y_{-m}, x_{-m}).$$

Then

$$0 \leq D_m^\theta(y) - d_m^\theta(y) \leq (1-2\beta)^m. \tag{40}$$

Proof: Fix some $\theta \in \Theta$; since it is fixed, drop it from the notation for the rest of the proof, for simplicity. Also, choose any n, with $0 \leq n \leq m$ and fix it. Now define

$$a_{n,x_{-n}}(y) \doteq p(y_0 \mid y_{-1} ... y_{-n}, x_{-n}),$$

$$b_{n,x_{-n},x_{-n-1}}(y) \doteq p(x_{-n} \mid y_{-1} ... y_{-n-1}, x_{-n-1}).$$

Note that, because of eq.(36) we have

$$a_{n,x_{-n}}(y) \doteq p(y_0 \mid y_{-1} ... y_{-n-1}, x_{-n-1} x_{-n}).$$

Also define

$$c_n(y) \doteq \min_{x_{-n-1}, x_{-n}} b_{n,x_{-n-1},x_{-n}}(y)$$

$$x_*(y) \doteq \arg\min_{x_{-n}} a_{n,x_{-n}}(y)$$

$$x^*(y) \doteq \arg\max_{x_{-n}} a_{n,x_{-n}}(y).$$

Now

$$p(y_0 \mid y_{-1} ... y_{-n-1}, x_{-n-1}) = \sum_{x_{-n}} b_{n,x_{-n-1},x_{-n}}(y) \cdot a_{n,x_{-n}}(y) =$$

$$\left\{\sum_{x_{-n}\neq x_*(y)} b_{n,x_{-n-1},x_{-n}}(y)\cdot a_{n,x_{-n}}(y)\right\}+\{b_{n,x_*(y),x_{-n-1}}(y)-c_n(y)\}\cdot a_{n,x_*}(y)+c_n(y)\cdot a_{n,x_*}(y)\leq$$

$$D_n(y)\cdot\left\{\sum_{x_{-n}\neq x_*(y)} b_{n,x_{-n-1},x_{-n}}(y)\right\}+\{b_{n,x_*(y),x_{-n-1}}(y)-c_n(y)\}\cdot D_n(y)+c_n(y)\cdot d_n(y)=$$

$$D_n(y)\cdot\left\{\sum_{x_{-n}} b_{n,x_{-n-1},x_{-n}}(y)\right\}-c_n(y)\cdot D_n(y)+c_n(y)\cdot d_n(y)=$$

$$(1-c_n(y))\cdot D_n(y)+c_n(y)\cdot d_n(y) \qquad \Rightarrow$$

$$p(y_0 \mid y_{-1}...y_{-n-1}, x_{-n-1}) \leq (1-c_n(y))\cdot D_n(y)+c_n(y)\cdot d_n(y). \tag{41}$$

In exactly the same way we obtain

$$p(y_0 \mid y_{-1}...y_{-n-1}, x_{-n-1}) \geq (1-c_n(y))\cdot d_n(y)+c_n(y)\cdot D_n(y). \tag{42}$$

Taking the max / min in eq.(41) / eq.(42) we get

$$D_{n+1}(y) \leq (1-c_n(y))\cdot D_n(y)+c_n(y)\cdot d_n(y), \tag{43}$$

$$d_{n+1}(y) \geq (1-c_n(y))\cdot d_n(y)+c_n(y)\cdot D_n(y). \tag{44}$$

Eqs.(43,44) in turn imply

$$0 \leq D_{n+1}(y)-d_{n+1}(y) \leq (1-2c_n(y))\cdot(D_n(y)-d_n(y)) \leq (1-2\beta)\cdot(D_n(y)-d_n(y)) \qquad \Rightarrow$$

(by repeated application for $n=0,1,2,...,m$)

$$0 \leq D_m(y)-d_m(y) \leq (1-2\beta)^m$$

which completes the proof of the Lemma. •

Lemma 8 *For all* $\theta\in\Theta$, $y_0y_{-1}y_{-2}...\in\Omega_Y^\infty$, $0<t<s$

$$\mid p^\theta(y_0 \mid y_{-1}...y_{-t}) - p^\theta(y_0 \mid y_{-1}...y_{-s}) \mid < (1-2\beta)^t \tag{45}$$

$$\mid p^{\theta}(y_0 \mid y_{-1}...y_{-t}) - p^{\theta}(y_0 \mid y_{-1}y_{-2}...) \mid < (1-2\beta)^t \tag{46}$$

Proof: Fix some $\theta \in \Theta$; since it is fixed, drop it from the notation for the rest of the proof, for simplicity. First we prove eq.(45). We have

$$p(y_0 \mid y_{-1}...y_{-t}) = \sum_{x_{-t}} p(y_0 \mid y_{-1}...y_{-t}, x_{-t}) p(x_{-t} \mid y_{-1}...y_{-t}) \leq$$

$$D_t(y) \cdot \sum_{x_{-t}} p(x_{-t} \mid y_{-1}...y_{-t}) = D_t(y).$$

Similarly we get

$$p(y_0 \mid y_{-1}...y_{-t}) \geq d_t(y).$$

In short we have

$$D_t(y) \geq p(y_0 \mid y_{-1}...y_{-t}) \geq d_t(y). \tag{47}$$

In the same way we get

$$D_t(y) \geq p(y_0 \mid y_{-1}...y_{-s}) \geq d_t(y). \tag{48}$$

Notice that in both eqs.(47,48) we have $D_t(y)$, $d_t(y)$ (with a t subscript), despite the different conditioning. Combining eqs.(47,48) we have

$$-(D_t(y) - d_t(y)) \leq \ \mid p^{\theta}(y_0 \mid y_{-1}...y_{-t}) - p^{\theta}(y_0 \mid y_{-1}...y_{-s}) \mid \ \leq (D_t(y) - d_t(y)) \Rightarrow$$

(using Lemma 7)

$$-(1-2\beta)^t \leq \ \mid p^{\theta}(y_0 \mid y_{-1}...y_{-t}) - p^{\theta}(y_0 \mid y_{-1}...y_{-s}) \mid \ \leq (1-2\beta)^t \tag{49}$$

which completes the proof of (45). This also implies that for all y_0, y_{-1}, ... $\{p(y_0 \mid y_{-1}...y_{-t})\}_{t=1}^{\infty}$ is a Cauchy sequence and has a limit. Define

$$p(y_0 \mid y_{-1}y_{-2}...) \doteq \lim_{t \to \infty} p(y_0 \mid y_{-1}...y_{-t}).$$

From this and eq.(49), eq.(46) follows immediately and the proof of the lemma is completed. •

Now we will prove Lemma 4; this in turn will be used to prove Theorems 1 and 2.

Proof of Lemma 4 We have to prove three things: (a) that $H(p)$ exists, (b) that, for all $\theta \in \Theta$, $H(p^\theta; p)$ exists and (c) that for all $\theta \in \Theta$ $\lim_{t\to\infty} -\log p_t^\theta(y_1...y_t)/t$ exists and is equal to $H(p^\theta; p) + H(p)$.

First look at $H(p)$. This is defined to be

$$H(p) \doteq -\int \log p(y_0 \mid y_{-1}y_{-2}...)d\pi(y_0y_{-1}...). \tag{50}$$

This integral is well defined, because, by **A3** and the convergence of conditional probabilities (see Billingsley [7])

$$0 < \alpha \leq p(y_0 \mid y_{-1}...y_{-t}) \to p(y_0 \mid y_{-1}y_{-2}...)$$

for π-a.a. $y_0y_{-1}y_{-2}...$. Now we can use Dominated Convergence Theorem to ensure the existence of (50). So $H(p)$ is well-defined. Next pick any $\theta \in \Theta$ and look at $H(p^\theta; p)$. This is defined by

$$H(p^\theta; p) \doteq -\int \log p^\theta(y_0 \mid y_{-1}y_{-2}...)d\pi(y_0y_{-1}...) - H(p). \tag{51}$$

Now, from Lemmas 6 and 8,

$$p^\theta(y_0 \mid y_{-1}y_{-2}...) = \lim_{t\to\infty} p^\theta(y_0 \mid y_{-1}...y_{-t}) \geq \beta > 0.$$

Once again, we can use the Dominated Convergence Theorem to ensure that (51) is well defined. So we have proven the existence of $H(p^\theta; p)$.

Finally consider

$$\frac{\log p^\theta(y_1...y_t)}{t} =$$

(for any N)

$$\frac{\log p^\theta(y_1...y_N)}{t} + \frac{\sum_{s=N+1}^t \log p^\theta(y_s \mid y_{s-1}...y_1)}{t} \leq$$

(using Lemma 6 and Lemma 8)

$$\frac{\log \gamma^N}{t} + \frac{\sum_{s=N+1}^{t} \log p^\theta(y_s \mid y_{s-1}...y_{s-N})}{t} + \frac{\log(1-2\beta)^N}{t}.$$

Letting $t \to \infty$ and using the Ergodic Theorem we get that for π-a.a. $y_1y_2...$

$$\lim_{t\to\infty} \frac{\log p^\theta(y_1...y_t)}{t} \leq \int \log p^\theta(y_0 \mid y_{-1}...y_{-N}) d\pi(y_0y_{-1}...) \ .$$

Now letting $N \to \infty$, using Lemma 8, the convergence of $p^\theta(y_0 \mid y_{-1}...y_{-N})$ and the Dominated Convergence Theorem we get

$$\lim_{t\to\infty} \frac{\log p^\theta(y_1...y_t)}{t} \leq \int \log p^\theta(y_0 \mid y_{-1}y_{-2}...) d\pi(y_0y_{-1}....).$$

In exactly symmetric manner we can bound $\lim_{t\to\infty} \frac{\log p^\theta(y_1...y_t)}{t}$ from below, to obtain

$$\lim_{t\to\infty} \frac{\log p^\theta(y_1...y_t)}{t} \geq \int \log p^\theta(y_0 \mid y_{-1}y_{-2}...) d\pi(y_0y_{-1}....).$$

But then we must have

$$\lim_{t\to\infty} \frac{\log p^\theta(y_1...y_t)}{t} = \int \log p^\theta(y_0 \mid y_{-1}y_{-2}...) d\pi(y_0y_{-1}....).$$

This, together with (50) and (51) shows that

$$\lim_{t\to\infty} -\frac{\log p_t^\theta(y_1...y_t)}{t} = H(p^\theta;p) + H(p) \qquad \pi - a.a. \ y_1y_2... \qquad (52)$$

which completes the proof of the Lemma. •

Proof of Theorem 1 We are interested in the ratio

$$\frac{Prob(Z=\theta \mid Y_1 = y_1...Y_t = y_t)}{Prob(Z=\zeta \mid Y_1 = y_1...Y_t = y_t)} = \frac{Prob(Y_1 = y_1...Y_t = y_t \mid Z=\theta)}{Prob(Y_1 = y_1...Y_t = y_t \mid Z=\zeta)} \cdot \frac{p_0^\theta}{p_0^\zeta} =$$

$$\frac{p^\theta(y_1....y_t)}{p^\zeta(y_1...y_t)} \cdot \frac{p_0^\theta}{p_0^\zeta}.$$

Obviously the ratio p_0^θ / p_0^ζ will not affect convergence, so we can concentrate on the ratio $p^\theta(y_1....y_t)/p^\zeta(y_1...y_t)$.

Now, from Lemma 4 we have

$$\lim_{t\to\infty} -\frac{\log p^\theta(y_1...y_t)}{t} = H(p^\theta; p) + H(p) \tag{53}$$

$$\lim_{t\to\infty} -\frac{\log p^\zeta(y_1...y_t)}{t} = H(p^\zeta; p) + H(p) \tag{54}$$

Therefore

$$\frac{1}{t}\log\frac{p^\theta(y_1....y_t)}{p^\zeta(y_1...y_t)} \to -H(p^\theta; p) + H(p^\zeta; p) \qquad \Rightarrow$$

(assuming $H(p^\theta; p) > H(p^\zeta; p)$)

$$0 \le \sqrt[t]{\frac{p^\theta(y_1....y_t)}{p^\zeta(y_1...y_t)}} \to e^{-H(p^\theta;p)+H(p^\zeta;p)} \qquad \Rightarrow$$

(for all $\varepsilon > 0$ such that $e^{H(p^\theta;p)-H(p^\zeta;p)} + \varepsilon < 1$, for all t greater than some appropriate t_ε)

$$0 \le \sqrt[t]{\frac{p^\theta(y_1....y_t)}{p^\zeta(y_1...y_t)}} \le e^{-H(p^\theta;p)+H(p^\zeta;p)} + \varepsilon \qquad \Rightarrow$$

$$0 \le \frac{p^\theta(y_1....y_t)}{p^\zeta(y_1...y_t)} \le (e^{-H(p^\theta;p)+H(p^\zeta;p)} + \varepsilon)^t \to 0$$

for π-a.a. $y_1 y_2 ...$ as $t \to \infty$. This completes the proof of the theorem. •

Proof of Theorem 2

Case 1: $\mid \Theta \mid < \infty$

$$\frac{\sum_{\theta\in\Theta^{h^0}} p_t^\theta(y_1...y_t)}{\sum_{\theta\in\Theta_{h^0}} p_t^\theta(y_1...y_t)} =$$

$$\frac{\sum_{\theta\in\Theta^{h^0}} p^\theta(y_1...y_t)\cdot p_0^\theta}{\sum_{\theta\in\Theta_{h^0}} p^\theta(y_1...y_t)\cdot p_0^\theta} \le$$

$$\frac{(\max_{\theta\in\Theta^{h^0}} p^\theta(y_1...y_t))\cdot\sum_{\theta\in\Theta^{h^0}} p_{\mathbf{0}}^\theta}{(\min_{\theta\in\Theta_{h^0}} p^\theta(y_1...y_t))\cdot\sum_{\theta\in\Theta_{h^0}} p_{\mathbf{0}}^\theta}\leq$$

$$\left\{\max_{\theta\in\Theta^{h^0},\zeta\in\Theta_{h_0}}\frac{p^\theta(y_1...y_t)}{p^\zeta(y_1...y_t)}\right\}\cdot\frac{\sum_{\theta\in\Theta^{h^0}} p_{\mathbf{0}}^\theta}{\sum_{\zeta\in\Theta_{h^0}} p_{\mathbf{0}}^\zeta}\leq$$

(from Theorem 1, for small enough $\varepsilon > 0$, $t \geq t_\varepsilon$)

$$\left\{\max_{\theta\in\Theta^{h^0},\zeta\in\Theta_{h_0}}(e^{-H(p^\theta;p)+H(p^\zeta;p)})+\varepsilon\right\}^t\cdot\frac{\sum_{\theta\in\Theta^{h^0}} p_{\mathbf{0}}^\theta}{\sum_{\zeta\in\Theta_{h^0}} p_{\mathbf{0}}^\zeta}\qquad\Rightarrow$$

(choosing ε small enough, the term in braces is smaller than one)

$$\lim_{t\to\infty}\frac{\sum_{\theta\in\Theta^{h^0}} p_t^\theta(y_1...y_t)}{\sum_{\zeta\in\Theta_{h^0}} p_t^\zeta(y_1...y_t)}=0$$

for π-a.a. $y_1y_2....$ Since the numerator and denominator must add up to 1, the desired result is proven.

Case 2: $\mid\Theta\mid=\aleph_0$

$$\frac{\sum_{\theta\in\Theta^{h^0}} p_t^\theta(y_1...y_t)}{\sum_{\theta\in\Theta_{h^0}} p_t^\theta(y_1...y_t)}=$$

$$\frac{\sum_{\theta\in\Theta^\epsilon} p^\theta(y_1...y_t)\cdot p_{\mathbf{0}}^\theta}{\sum_{\theta\in\Theta_\delta} p^\theta(y_1...y_t)\cdot p_{\mathbf{0}}^\theta}\leq$$

$$\frac{(\sup_{\theta\in\Theta^\epsilon} p^\theta(y_1...y_t))\cdot\sum_{\theta\in\Theta^\epsilon} p_{\mathbf{0}}^\theta}{(\inf_{\theta\in\Theta_\delta} p^\theta(y_1...y_t))\cdot\sum_{\theta\in\Theta_\delta} p_{\mathbf{0}}^\theta}\leq$$

$$\left\{\sup_{\theta\in\Theta^\epsilon,\zeta\in\Theta_\delta}\frac{p^\theta(y_1...y_t)}{p^\zeta(y_1...y_t)}\right\}\cdot\frac{\sum_{\theta\in\Theta^\epsilon} p_{\mathbf{0}}^\theta}{\sum_{\zeta\in\Theta_\delta} p_{\mathbf{0}}^\zeta}\leq$$

(for all $\varepsilon > 0$, $t \geq t_\varepsilon$)

$$\left\{\sup_{\theta\in\Theta^\epsilon,\zeta\in\Theta_\delta}(e^{-H(p^\theta;p)+H(p^\zeta;p)})+\varepsilon\right\}^t\cdot\frac{\sum_{\theta\in\Theta^\epsilon} p_{\mathbf{0}}^\theta}{\sum_{\zeta\in\Theta_\delta} p_{\mathbf{0}}^\zeta}\leq$$

$$(e^{\delta-\epsilon}+\varepsilon)^t \cdot \frac{\sum_{\theta\in\Theta^\epsilon} p_0^\theta}{\sum_{\zeta\in\Theta_\delta} p_0^\zeta} \qquad \Rightarrow$$

(choosing ε small enough, the term in parentheses is smaller than one)

$$\lim_{t\to\infty} \frac{\sum_{\theta\in\Theta^\epsilon} p_t^\theta(y_1...y_t)}{\sum_{\zeta\in\Theta_\delta} p_t^\zeta(y_1...y_t)} = 0$$

for π-a.a. $y_1y_2....$ Since the numerator and denominator must add up to 1, the desired result is proven and we are done. •

Finally let us prove Theorem 3; for this we first need the following Lemma.

Lemma 9 *Under the conditions of Theorem 3 we have for all $\theta\neq 1$ and π-a.a.*

$$\lim_{t\to\infty} \frac{p^\theta(y_1...y_t)}{p^1(y_1...y_t)} = 0. \tag{55}$$

Proof: We know that for all θ and t

$$A_t = \frac{p(Z=\theta \mid Y_1=y_1,...,Y_t=y_t)}{p(Z=1 \mid Y_1=y_1,...,Y_t=y_t)} = \frac{p(Z=\theta \mid Y_1=y_1,...,Y_t=y_t)p(Y_1=y_1,...,Y_t=y_t)}{p(Z=1 \mid Y_1=y_1,...,Y_t=y_t)p(Y_1=y_1,...,Y_t=y_t)}$$

$$\frac{p(Z=\theta, Y_1=y_1,...,Y_t=y_t)}{p(Z=1, Y_1=y_1,...,Y_t=y_t)} \cdot \frac{Pr(Z=1)}{Pr(Z=\theta)} \cdot \frac{Pr(Z=\theta)}{Pr(Z=1)} =$$

(using the definition of conditional probability, with Z conditioning)

$$\frac{p^\theta(Y_1=y_1,...,Y_t=y_t)}{p^1(Y_1=y_1,...,Y_t=y_t)} \cdot \frac{Pr(Z=\theta)}{Pr(Z=1)}.$$

Since A_t tends to 0 (by Theorem 1) and $Pr(Z=\theta)/Pr(Z=1)$ is a constant, it follows that $p^\theta(Y_1=y_1,...,Y_t=y_t)/\ p^1(Y_1=y_1,...,Y_t=y_t)$ must also tend to zero and we are done. •

Proof of Theorem 3 by dividing the update equations for p_t^θ and p_t^1 we get

$$\frac{p_t^\theta}{p_t^1} = \frac{Pr(Y_t=y_t \mid Y_{t-1}=y_{t-1},...,Y_1=y_1,Z=\theta)}{Pr(Y_t=y_t \mid Y_{t-1}=y_{t-1},...,Y_1=y_1,Z=1)} \cdot \frac{p_{t-1}^\theta}{p_{t-1}^1} =$$

$$\frac{Pr(Y_t = y_t \mid Y_{t-1} = y_{t-1}, ..., Y_1 = y_1, Z = \theta)}{Pr(Y_t = y_t \mid Y_{t-1} = y_{t-1}, ..., Y_1 = y_1, Z = 1)} \cdot \frac{Pr(Y_{t-1} = y_{t-1} \mid Y_{t-1} = y_{t-1}, ..., Y_1 = y_1, Z = \theta)}{Pr(Y_{t-1} = y_{t-1} \mid Y_{t-1} = y_{t-1}, ..., Y_1 = y_1, Z = 1)} \cdot \frac{p^\theta_{t-2}}{p^1_{t-2}} :$$

$$\frac{Pr(Y_t = y_t, Y_{t-1} = y_{t-1} \mid Y_{t-2} = y_{t-2}, ..., Y_1 = y_1, Z = \theta)}{Pr(Y_t = y_t, Y_{t-1} = y_{t-1} \mid Y_{t-2} = y_{t-2}, ..., Y_1 = y_1, Z = 1)} \cdot \frac{p^\theta_{t-2}}{p^1_{t-2}}.$$

By repeated application of the above reasoning we finally get

$$\frac{p^\theta_t}{p^1_t} = \frac{Pr(Y_t = y_t, Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, ..., Y_1 = y_1 \mid Z = \theta)}{Pr(Y_t = y_t, Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, ..., Y_1 = y_1 \mid Z = 1)} \cdot \frac{p^\theta_0}{p^1_0} = w^\theta_t, \tag{56}$$

where $\lim_{t\to\infty} w^\theta_t = 0$ for all $\theta \neq 1$, by Lemma 9. From eq.(56) follows immediately that $p^\theta_t = w^\theta_t \cdot p^1_t$. Using the fact that the sum of the p^θ_t's must equal one for every t, with a little algebra we finally get the expressions

$$p^1_t = \frac{1}{\sum_{\zeta=1}^\infty w^\zeta_t} \quad \text{and} \quad p^\theta_t = \frac{w^\theta_t}{\sum_{\zeta=1}^\infty w^\zeta_t}. \tag{57}$$

From Theorem 2 we know that $p^1_t \to 1$, hence there is some t_0 such that for all $t \geq t_0$ we have $B_t = \sum_{\zeta=1}^\infty w^\zeta_t < \infty$.

Using exactly the same reasoning as in the previous paragraph we also obtain expressions for $p^{\theta,N}_t$. these are as follows

$$p^{1,N}_t = \frac{1}{\sum_{\zeta=1}^N w^\zeta_t} \quad \text{and} \quad p^{\theta,N}_t = \frac{w^\theta_t}{\sum_{\zeta=1}^N w^\zeta_t}. \tag{58}$$

It is important to note that the w^θ_t in (57) and (58) are the same quantities; the only thing that changes is the limit of summation. Now let us divide (57) by (58); we get

$$\frac{p^{\theta,N}_t}{p^\theta_t} = \frac{\sum_{\zeta=1}^N w^\zeta_t}{\sum_{\zeta=1}^\infty w^\zeta_t}. \tag{59}$$

But for every $t \geq t_0$ we have $\lim_{N\to\infty} \sum_{\zeta=1}^N w^\zeta_t = \sum_{\zeta=1}^\infty w^\zeta_t < \infty$; from this and (59) it follows immediately that

$$\lim_{N\to\infty} \frac{p^{\theta,N}_t}{p^\theta_t} = \lim_{N\to\infty} \frac{\sum_{\zeta=1}^N w^\zeta_t}{\sum_{\zeta=1}^\infty w^\zeta_t} = 1 \tag{60}$$

and we are done; the second part of (33) follows immediately. In fact, from (60) we also notice that the rate of convergence is the same for all θ. $\bullet$

References

[1] L.R. Bahl et al, “A Maximum Likelihood approach to continuous speech recognition”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, March 1983, pp. 179–190.

[2] P. Baldi et al, “Hidden Markov Models in Molecular Biology: new algorithms and applications”, preprint, 1993.

[3] P. Baldi and Y. Chauvin, “Smooth on-line learning algorithms for Hidden Markov Models”, preprint, 1993.

[4] L.E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov Chains”, *Ann. of Math. Stat.*, Vol. 37, 1966, pp.1554–1663.

[5] L.E. Baum and J.A. Eagon, “An inequality with applications to statistical estimation for probabilistic functions of Markov Processes”, *Ann. of Math. Stat.*, Vol. 38, 1967, pp. 356–363.

[6] L.E. Baum et al, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chain”, *Ann. of Math. Stat.*, Vol. 41, No. 1, 1970, pp. 164–171.

[7] P. Billingsley, *Ergodic Theory and Information*, Wiley, New York, 1965.

[8] P. Billingsley, *Probability and Measure*, Wiley, New York, 1965.

[9] J.L. Doob, *Stochastic Processes*, Wiley, New York, 1953.

[10] S. Geman and K. Manbeck, “Machine recognition of human coronary arteries by deformable templates”, preprint, Division of Applied Mathematics, Brown University, 1991.

[11] C.G. Hilborn and D.G. Lainiotis, "Optimal adaptive filter realizations for stochastic processes with an unknown parameter", *IEEE Trans. on Automatic Control*, Vol. 14, December 1969, pp. 767-770.

[12] C.G. Hilborn and D.G. Lainiotis, "Optimal estimation in the Presence of unknown parameters", *IEEE Trans. on Systems Science and Cybernetics*, Vol. 5, No. 1, January 1969, pp. 38-43.

[13] C.G. Hilborn and D.G. Lainiotis, "Unsupervised learning minimum risk pattern classification for dependent hypotheses and dependent measurements, *IEEE Trans. on Systems Science and Cybernetics*, Vol.5, No.2, April 1969, pp. 109-115.

[14] A. Kehagias, *Approximation and estimation of stochastic processes by Hidden Markov Models*, PhD thesis, Division of Applied Mathematics, Brown Un., Providence, Rhode Island, May 1992.

[15] A. Kehagias, "Convergence properties of the Lainiotis partition algorithm", *Control and Computers*, Vol. 19, No. 1, 1991, pp. 1-6.

[16] P.R. Kumar and P. Varaiya, *Stochastic Systems*, Prentice Hall, Englewood Cliffs, 1986.

[17] D.G. Lainiotis, "Optimal adaptive estimation: structure and pattern adaptation", *IEEE Transactions on Automatic Control*, Vol. 16, No. 2, 1971, pp. 160–170.

[18] D.G. Lainiotis, "Partitioning: a unifying framework for adaptive systems I: estimation", *Proc. IEEE*, Vol. 64, No. 8, August 1976, pp. 1126-1143.

[19] D.G. Lainiotis, "Partitioning: a unifying framework for adaptive systems II: control", *Proc. IEEE*, Vol. 64, No. 8, August 1976, pp. 1182-1198.

[20] D.G. Lainiotis, "A unifying framework for linear estimation: generalized partitioned algorithms", *J. Inform. Sciences*, Vol. 10, No. 3, April 1976, pp. 243-278.

[21] D.G. Lainiotis, "Partitioning filters", *J. Inform. Sciences*, Vol. 17, No. 2, 1979, pp. 177-193.

[22] D.G. Lainiotis et al., "Adaptive deconvolution of seismic signals – performance, computational analysis, paaralellism", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 36, No.11, November 1988, pp. 1715-1734.

[23] D.G. Lainiotis et al., "Optimal seismic deconvolution", *Signal Processing*, Vol. 15, No. 4, December 1988, pp. 375-404.

[24] K.F. Lee et al, "An overview of the SPHINX speech recognition system", *IEEE Transactions on Acoustics, Signals and Speech Processing*, Vol. 38, No. 1, January 1990, pp. 35–45.

[25] S.E. Levinson et al, "An introduction to the application of the theory of probabilistic functions of a Markov Chain.", *The Bell Sys. Tech. J.*, Vol. 62, No. 4, April 1983, pp. 1034-1074.

[26] W.D. Mao and S.Y. Kung, "Shape recognition by ring HMMs", *Proc. of the Int. Joint Conf. on Neural Networks*, Vol. 2, January 1990, pp.409–412.

[27] L.R. Rabiner et al, "On the application of vector quantization and HMM to speaker independent isolated word recognition", *The Bell Sys. Tech. J.*, Vol. 62, No. 4, April 1983, pp. 1075–1105.

[28] L.R. Rabiner, "A tutorial on HMM and selected applications in speech recognition", *Proc. IEEE* , Vol. 77, No. 2, February 1989, pp. 257-286.

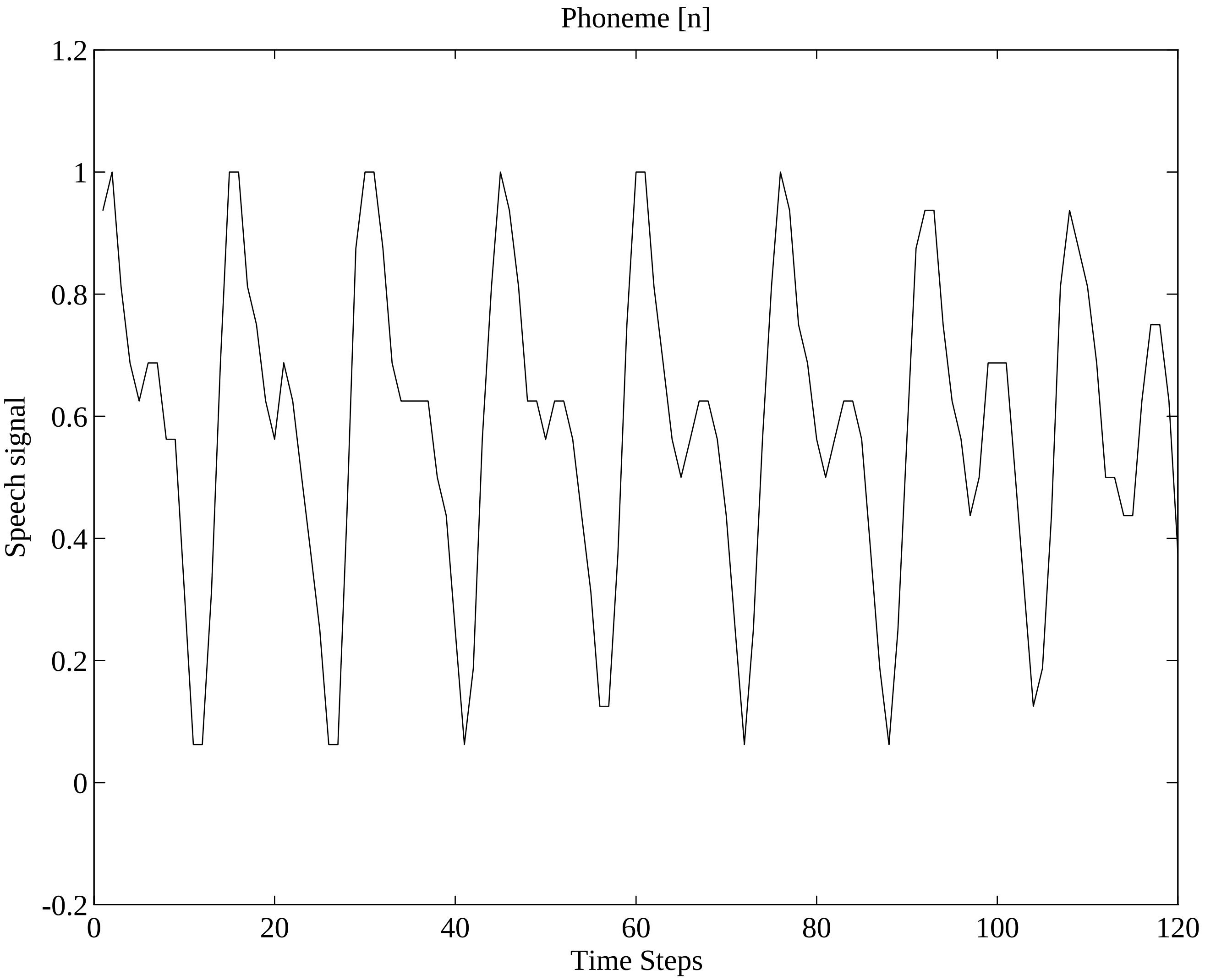
[29] RF.L. Sengbush and D.G. Lainiotis, "Simplified parameter quantization procedure for adaptive estimation", *IEEE Trans. on Automatic Control*, Vol. 14, June 1969, pp. 424-425.

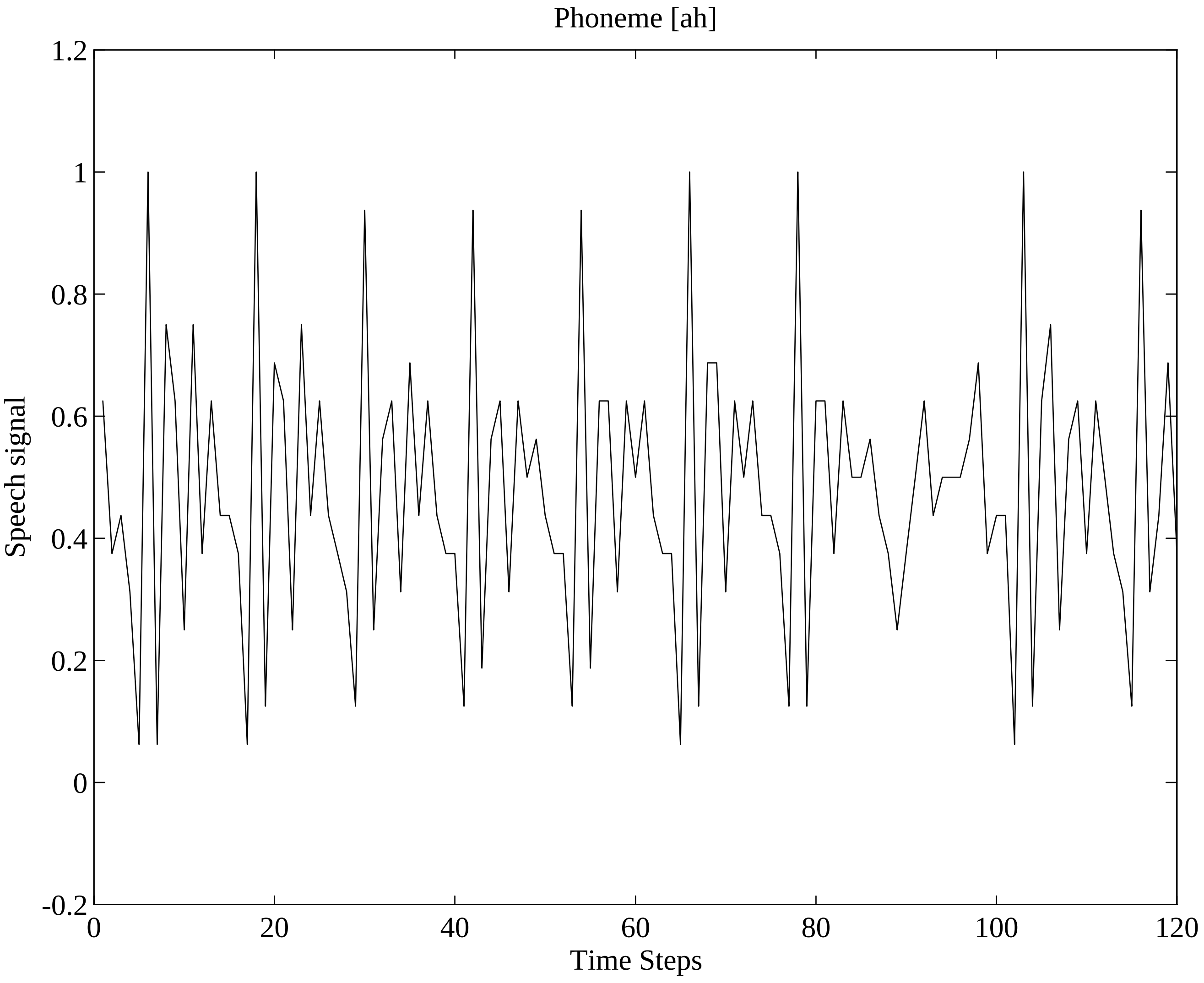
[30] F.L. Sims, D.G. Lainiotis and D.T. Magill, “Recursive algorithm for the calculation of the adaptive Kalman filter coefficients”. *IEEE Trans. on Automatic Control*, Vol. 14, April 1969, pp. 215-218.

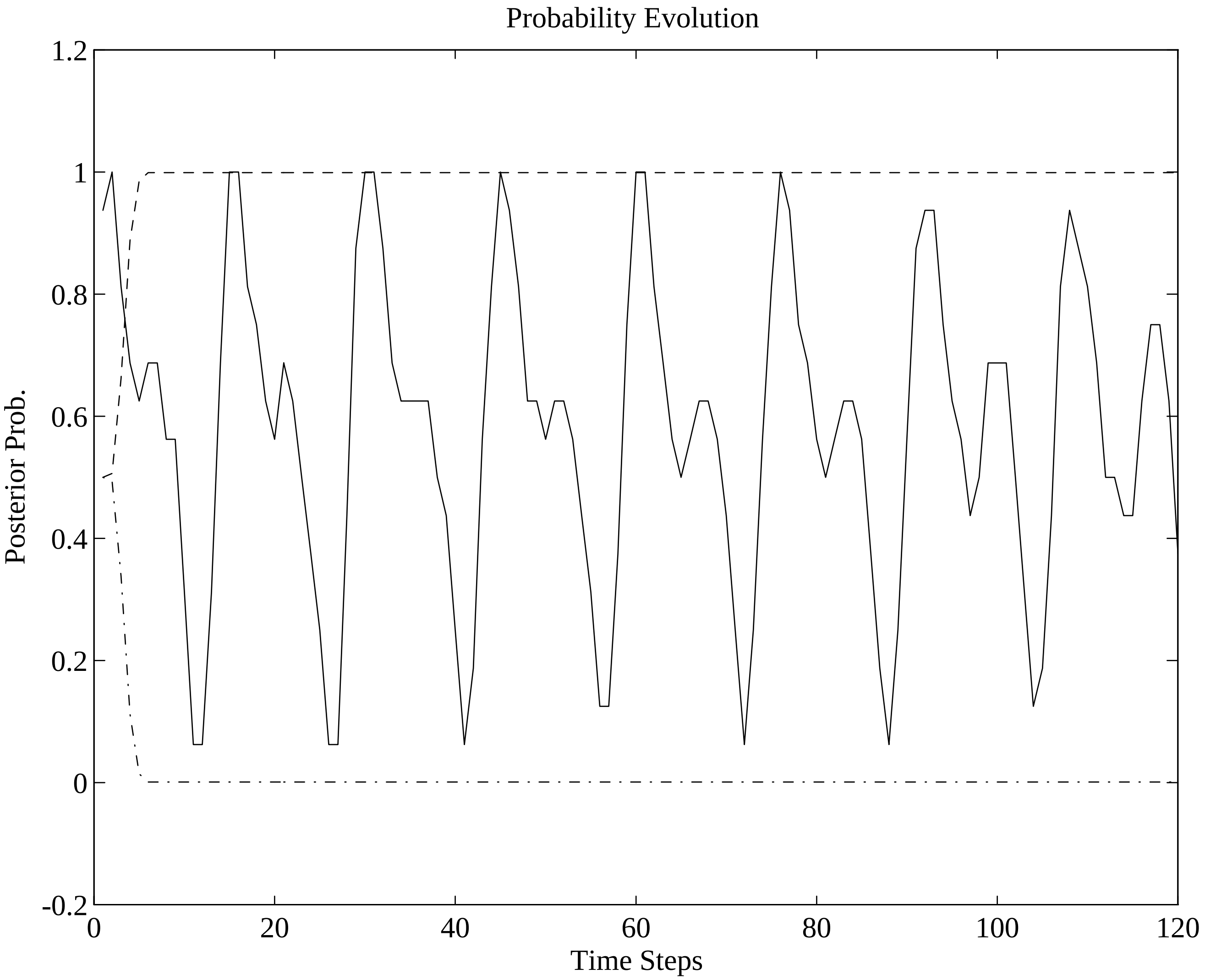
[31] J.K. Tugnait, “ Convergence analysis of partitioned adaptive estimators under continuous parameter uncertainty”, *IEEE Trans. on Automatic Control*, Vol. 25, June 1980, pp. 569-573.

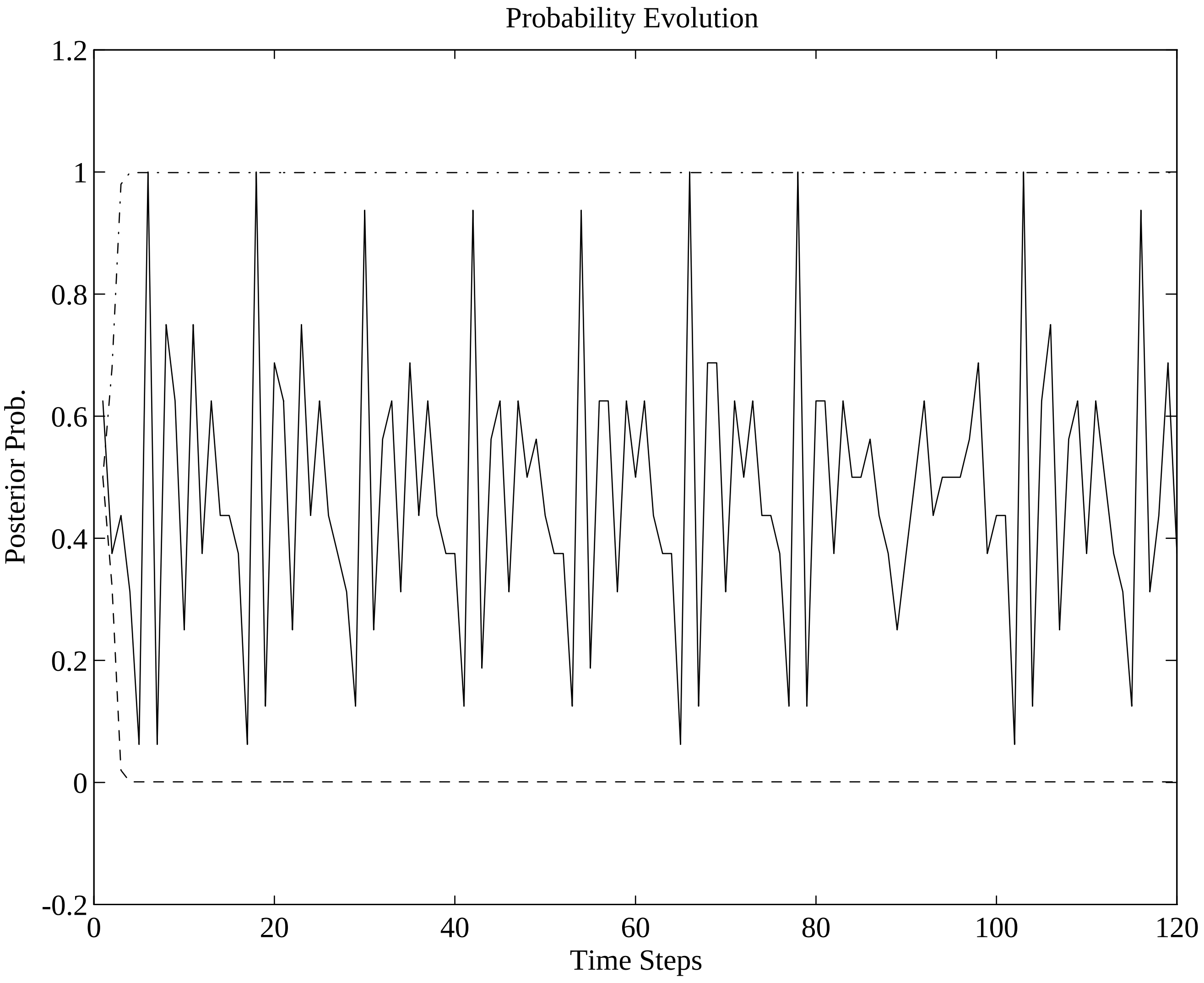
List of Captions

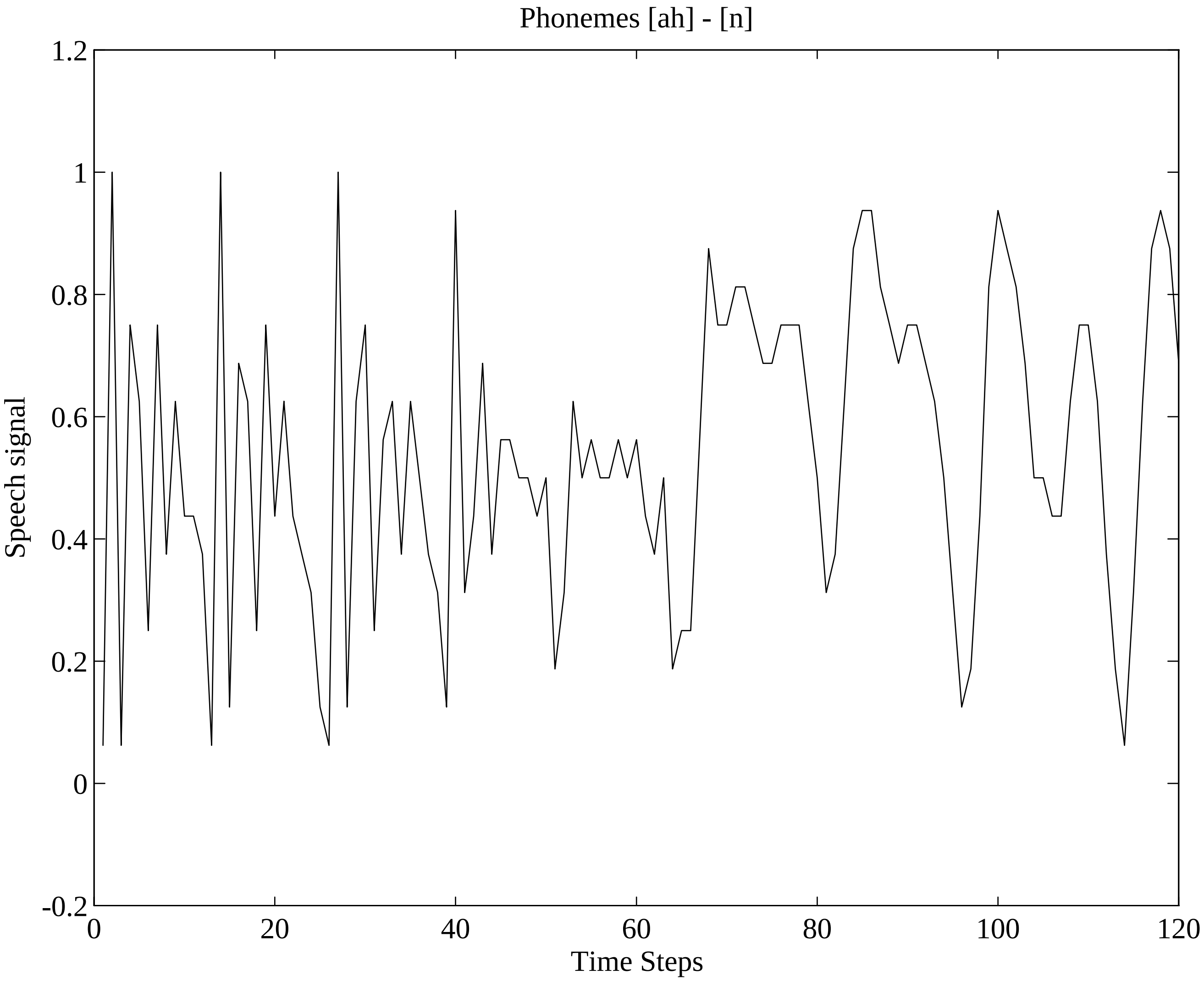
- Fig.??: Quantized version of [ah] signal.
- Fig.??: Quantized version of [n] signal.
- Fig.??: Evolution of $p_t^{\theta_1}, p_t^{\theta_2}$. Solid line is signal; dash-dotted line is $p_t^{\theta_1}$, dashed line is $p_t^{\theta_2}$.
- Fig.??: Evolution of $p_t^{\theta_1}, p_t^{\theta_2}$. Solid line is signal; dash-dotted line is $p_t^{\theta_1}$, dashed line is $p_t^{\theta_2}$.
- Fig.??: Quantized version of [ah] - [n] signal.
- Fig.??: Evolution of $p_t^{\theta_1}, p_t^{\theta_2}$. Solid line is signal; dash-dotted line is $p_t^{\theta_1}$, dashed line is $p_t^{\theta_2}$.

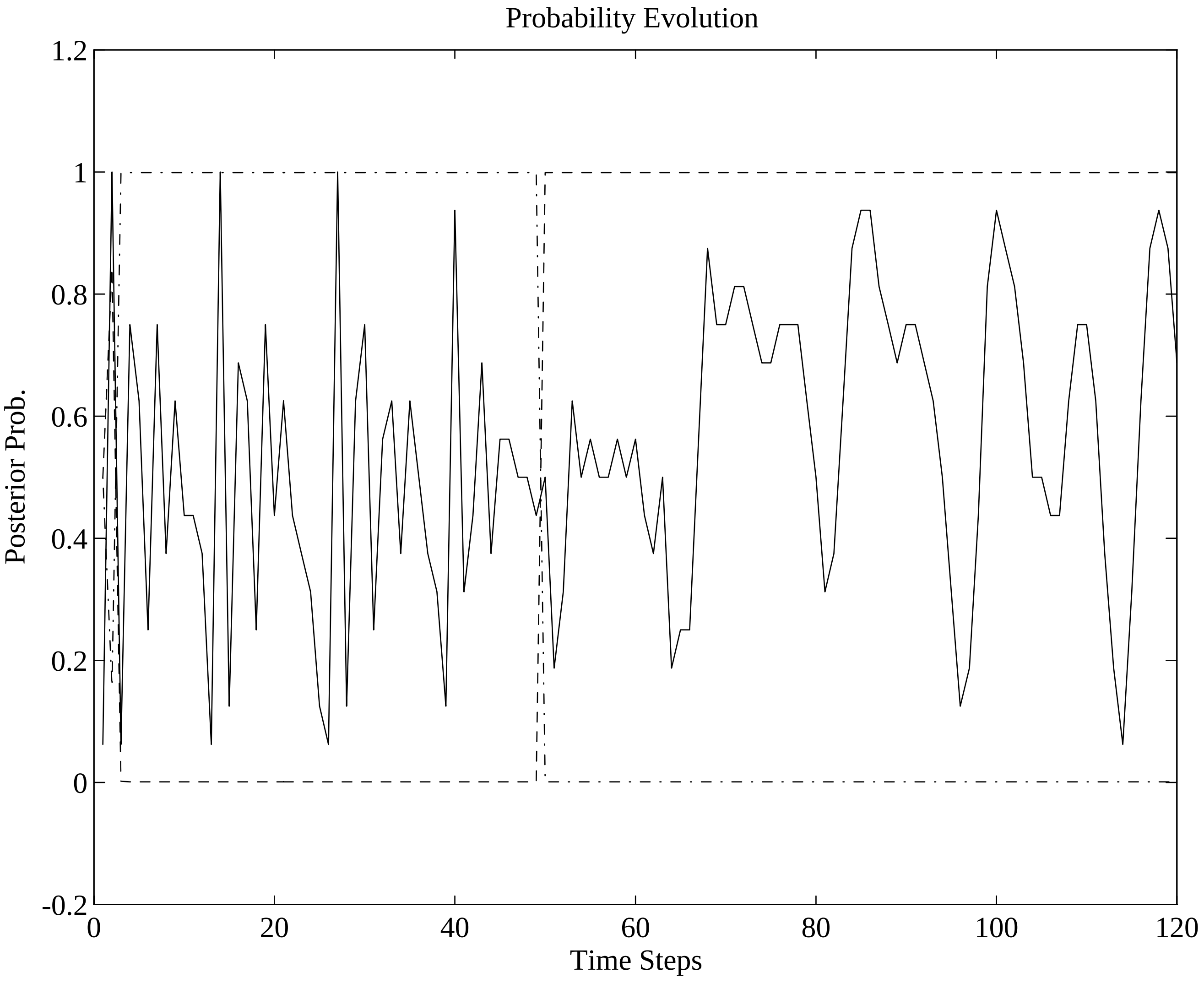

Phoneme [n]
Speech signal
Time Steps
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120


Phoneme [ah]
Speech signal
Time Steps
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120


Probability Evolution
Posterior Prob.
Time Steps
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120


Probability Evolution
Posterior Prob.
Time Steps
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120


Phonemes [ah] - [n]
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
Speech signal
0
20
40
60
80
100
120
Time Steps


Probability Evolution
Posterior Prob.
Time Steps
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120