A generalized risk-based approach to segmentation based on hidden Markov models

Jüri Lember*
Institute of Mathematical Statistics
Tartu University
J. Liivi 2-507, Tartu, 50409, Estonia

Alexey A. Koloydenko
Department of Mathematics
Royal Holloway, University of London
TW20 0EX, UK

Abstract
Motivated by the continuing interest in discrete time hidden Markov models (HMMs), this paper reexamines these models using a risk-based approach. Simple modifications of the classical optimization criteria for hidden path inference lead to a new class of hidden path estimators. The estimators are efficiently computed in the usual forward-backward manner and a corresponding dynamic programming algorithm is also presented. A particularly interesting subclass of such alignments are sandwiched between the most common maximum a posteriori (MAP), or Viterbi, path estimator and the minimum error, or point-wise maximum a posteriori (PMAP), estimator. Similar to previous work, the new class is parameterized by a small number of tunable parameters. Unlike their previously proposed relatives, the new parameters and class are more explicit and have clear interpretations, and bypass the issue of numerical scaling, which can be particularly valuable for applications.

Keywords: risk, HMM, hybrid, interpolation, MAP sequence, Viterbi algorithm, symbol-by-symbol, posterior decoding

1. Introduction

Besides their classical and traditional applications in signal processing and communications (Bahl et al., 1974; Brushe et al., 1998; Hayes et al., 1982; Viterbi, 1967) (cf. also further references in (Cappé et al., 2005)) and speech recognition (Huang et al., 1990; Jelinek, 2001, 1976; McDermott and Hazen, 2004; Ney et al., 1994; Padmanabhan and Picheny, 2002; Rabiner and Juang, 1993; Rabiner et al., 1986; Shu et al., 2003; Steinbiss et al., 1995; Ström et al., 1999), hidden Markov models have recently become indispensable in computational biology and bioinformatics (Brejová et al., 2008; Burge and Karlin, 1997; Durbin et al., 1998; Eddy, 2004; Krogh, 1998; Majoros and Ohler, 2007) as well as in natural language modelling (Ji and Bilmes, 2006; Och and Ney, 2000) and information security (Mason et al., 2006).

* Jüri Lember is supported by the Estonian Science Foundation Grant number 7553
At the same time, their spatial extensions, known as hidden Markov random field models (HMRFM), have also been immensely influential in spatial statistics (Besag and Green, 1993; Green and Richardson, 2002; Künsch et al., 1995; McGrory et al., 2009), and particularly in image analysis, restoration, and segmentation (Besag, 1986; Geman and Geman, 1984; Li et al., 2000; Marroquin et al., 2003; Winkler, 2003). Indeed, hidden Markov models are called ‘one of the most successful statistical modelling ideas that have [emerged] in the last forty years’ (Cappé et al., 2005).

HM(RF)Ms owe much of their success on the one hand to the persistence of the Markov property of the unobserved, or hidden, layer in the presence of observed data, and on the other, to the richness of the observed system (Künsch et al., 1995). Namely, in Bayesian terms, in addition to the prior, the posterior distribution of the hidden layer also possesses a Markov property (albeit generally inhomogeneous even with homogeneous priors), whereas the marginal law of the observed layer can still include global, i.e. non-Markovian, dependence.

The Markov property of the posterior distribution and the conditional independence of the observed variables given the hidden ones, have naturally led to a number of computationally feasible methods for inference about the hidden realizations as well as model parameters (if any). HMMs are naturally a special case of graphical models (Lauritzen, 1996), (Bishop, 2006, ch. 8).

HMMs, or one dimensional HMRFMs, have been particularly popular not least due to the fact that linear order of the indexing set (usually associated with time) makes exploration of hidden realizations relatively straightforward from the computational viewpoint. In contrast, higher dimensional HMRFMs generally require approximate, possibly stochastic, techniques in order to compute optimal configurations of the hidden field (Cocozza-Thivent and Bekkhoucha, 1993; Joshi et al., 2006; McGrory et al., 2009; Winkler, 2003). In particular, maximum a posteriori (MAP) estimator of the hidden layer of an HMM is efficiently and exactly computed by a dynamic programming algorithm bearing the name of Viterbi, whereas a general higher dimensional HMRFM would commonly employ a simulated annealing type method (Geman and Geman, 1984; Winkler, 2003) to produce approximate solutions to the same task.

1.1 Notation and main ingredients

We adopt the machine and statistical learning convention and therefore refer to the hidden and observed processes as $Y$ and $X$, respectively, in effect reversing the convention that is more commonly used in the HMM context. Thus, let $Y = \{Y_t\}_{t \geq 1}$ be a Markov chain with state space $S = \{1, \ldots, K\}$, $K > 1$. Even though we include inhomogeneous chains in most of what follows, for brevity we will still be suppressing the time index wherever this does not cause ambiguity. Hence, we write $P = (p_{ij})_{i,j \in S}$ for all transition matrices. Let $X = \{X_t\}_{t \geq 1}$ be a process with the following properties. First, given $\{Y_t\}_{t \geq 1}$, the random variables $\{X_t\}_{t \geq 1}$ are conditionally independent. Second, for each $t = 1, 2, \ldots$, the distribution of $X_t$ depends on $\{Y_t\}_{t \geq 1}$ (and $t$) only through $Y_t$. The process $X$ is sometimes called the hidden Markov process (HMP) and the pair $(Y, X)$ is referred to as a hidden Markov model (HMM). The name is motivated by the assumption that the process $Y$ (sometimes called a regime) is generally non-observable. The conditional distribution of
Given a set $A$, integers $m$ and $n$, and a sequence $a_1, a_2, \ldots \in A^\infty$, we write $a^n_m$ for the subsequence $(a_m, \ldots, a_n)$. When $m = 1$, it will be often suppressed.

Thus, $x^T := (x_1, \ldots, x_T)$ and $y^T := (y_1, \ldots, y_T)$ stand for the fixed observed and unobserved realizations, respectively, of the HMM $(X_t, Y_t)_{t \geq 1}$ up to time $T \geq 1$. Any sequence $s^T \in S^T$ is called a path. We shall denote by $p(x^T, y^T)$ the joint probability density of $(x^T, y^T)$, i.e.

$$p(x^T, y^T) := P(Y^T = y^T) \prod_{t=1}^{T} f_{y_t}(x_t).$$

Overloading the notation, for every $s^T \in S^T$ and for every sequence of observations $x^T$, let $p(s^T)$ and $p(x^T)$ stand for the marginal probability mass function $P(Y^T = s^T)$ of path $s^T$ and probability density function $\sum_{s^T} p(x^T, s^T)$ of the data $x^T$, respectively. It is a standard (see, e.g. (Cappé et al., 2005; Ephraim and Merhav, 2002), (Bishop, 2006, ch. 13)) in this context to define the so-called forward and backward variables

$$\alpha_t(s) := p(x^t|Y_t = s)P(Y_t = s), \quad \beta_t(s) := \begin{cases} 1, & \text{if } t = T \\ p(x^T_{t+1}|Y_t = s), & \text{if } t < T \end{cases},$$

where $p(x^t|Y_t = s)$ and $p(x^T_{t+1}|Y_t = s)$ are the conditional densities of the data segments $x^t$ and $x^T_{t+1}$, respectively, given $Y_t = s$.

### 1.2 Segmentation

*Segmentation* here refers to estimation of the hidden path $y^T$. Treating $y^T$ as missing data (Rabiner, 1989), or parameters, a classical and by far the most popular solution to the segmentation problem is to maximize $p(x^T, s^T)$ in $s^T \in S^T$. Often, especially in the digital communication literature (Brushe et al., 1998; Lin and Costello Jr., 1983), $p(x^T, s^T)$ is called the likelihood function which might become potentially problematic in the presence of any genuine model parameters. Such “maximum likelihood” paths are also called *Viterbi paths* or *alignments* after the Viterbi algorithm (Rabiner, 1989; Viterbi, 1967) commonly used for their computation. If $p(s^T)_{s^T \in S^T}$ is thought of as the prior distribution of $Y^T$, then Viterbi path also maximizes $p(s^T|x^T) := P(Y^T = s^T|X^T = x^T)$, the probability mass function of the posterior distribution of $Y^T$, hence the term ‘maximum a posteriori (MAP) path’.

In spite of its computational attractiveness, Viterbi inference may be unsatisfactory for a number of reasons, including its suboptimality with regard to the number of correctly estimated states $y_t$. Also, using the language of information theory, there is no reason to expect a Viterbi path to be typical (Lember and Koloydenko, 2010). Indeed, “there might be many similar paths through the model with probabilities that add up to a higher probability than the single most probable path” (Käll et al., 2005). The fact that a MAP estimate need not be representative of the posterior distribution has also been recently discussed in
a more general context in (Carvalho and Lawrence, 2008). Atypicality of Viterbi paths particularly concerns situations when estimation of $y^T$ is combined with inference about model parameters, e.g. transition probabilities $p_{ij}$ (Lember and Koloydenko, 2010). Even when estimating, say, the probability of heads from independent tosses of a biased coin, we naturally hope to observe a typical realization and not the constant one of maximum probability.

An alternative and very natural way to estimate $y^T$ is by maximizing the posterior probability $p_t(s|x^T) := \mathbb{P}(Y_t = s|X^T = x^T)$ of each individual hidden state $Y_t$, $1 \leq t \leq T$. We refer to the corresponding estimator as pointwise maximum a posteriori (PMAP). PMAP is well-known to maximize the expected number of correctly estimated states (Section 2), hence the characterization ‘optimal accuracy’ (Holmes and Durbin, 1998). In statistics, especially spatial statistics and image analysis, this type of estimation is known as Marginal Posterior Mode (Winkler, 2003) or Maximum Posterior Marginals (Rue, 1995) (MPM) estimation. In computational biology, this is also known as the posterior decoding (PD) (Břejová et al., 2008) and has been reported to be particularly successful in pairwise sequence alignment (Holmes and Durbin, 1998). In the wider context of biological applications of discrete high-dimensional probability models, this has also been called consensus estimation, and in the absence of constraints, centroid estimation (Carvalho and Lawrence, 2008). In communications applications of HMMs, largely influenced by (Bahl et al., 1974), the terms ‘optimal symbol-by-symbol decoding’ (Hayes et al., 1982), ‘symbol-by-symbol MAP estimation’ (Robertson et al., 1995), and ‘MAP state estimation’ (Brushe et al., 1998) have been used for this.

Although optimal in the sense of maximizing the expected number of correctly estimated states, a PMAP path might at the same time have low, in principle zero, probability (Rabiner, 1989). It is actually not difficult to constrain the PMAP decoder to admissible paths, i.e. of positive posterior probabilities as described in (Käll et al., 2005) (albeit in a slightly more general form allowing for state aggregation) and also in Subsection 2.2, (7), below. A variation on this idea has been applied in (Fariselli et al., 2005) for prediction of membrane proteins, giving rise to the term ‘posterior Viterbi decoding (PVD)’ (Fariselli et al., 2005). PVD, however, maximizes the product $\prod_{t=1}^{T} p_t(s|x^T)$ (Fariselli et al., 2005) (and also (10) below) and not the sum $\sum_{t=1}^{T} p_t(s|x^T)$, whereas the two criteria are no longer equivalent in the presence of path constraints (Subsection 2.2). In (Holmes and Durbin, 1998), a PMAP decoder is proposed to obtain optimal pairwise sequence alignments. The authors of (Holmes and Durbin, 1998) use the term “a legitimate alignment” which suggests admissibility, but the description of the actual algorithm (Holmes and Durbin, 1998, Section 3.8) appears to be insufficiently detailed to verify if the algorithm indeed enforces admissibility, or, if inadmissible solutions are altogether an issue in that context.

In many applications, e.g. gene identification, the pointwise (e.g. nucleotide level) error rate is not necessarily the main measure of accuracy, hence the constrained PMAP need not be an ultimate answer. Together with the above problem of atypicality of MAP paths, this has been addressed by moving from single path inference towards envelopes (Holmes and Durbin, 1998). Thus, for example, in computational biology the most common approach would be to aggregate individual states into a smaller number of semantic labels (e.g. codon, intron, intergenic). In effect, this would realise the notion of path similarity by
mapping many “similar” state paths to a single label path, or annotation (Brejová et al., 2008; Fariselli et al., 2005; Käll et al., 2005; Krogh, 1997). However, this leads to the problem of multiple paths, which in many practically important HMMs renders the dynamic programming approach of the Viterbi algorithm NP-hard (Brejová et al., 2007). Unlike the Viterbi/MAP decoder, the PMAP decoder handles annotations as easily as it does state paths, including the enforcement of admissibility (Käll et al., 2005). A number of alternative heuristic approaches are also known in computational biology, but none appears to be fully satisfactory (Brejová et al., 2008). Evidently, mapping optimal state paths to the corresponding annotations need not lead to optimal annotation and can actually give poor results (Brejová et al., 2007). Overall, although the original Viterbi decoder has still been the most popular paradigm in many applications, and in computational biology in particular, alternative approaches have demonstrated significantly higher performance, e.g., in predicting various biological features. For example, (Krogh, 1997) suggested the 1-best algorithm for optimal labelling. More recently, (Fariselli et al., 2005) have demonstrated PVD to be superior to the 1-best algorithm, and not surprisingly, to the Viterbi and PMAP decoders, on tasks of predicting membrane proteins.

A starting point of this paper is that restricting the PMAP decoder to paths of positive probability is but one of numerous ways to combine the useful features of the MAP and PMAP path estimators. Indeed, as a sensible remedy against vanishing probabilities, in his popular tutorial (Rabiner, 1989) Rabiner briefly mentions maximization of the expected number of correctly decoded (overlapping) blocks of length two or three, rather than single states. With \( k \geq 1 \) and \( \hat{y}^T(k) \) being the block length and corresponding path estimate, respectively, this approach yields Viterbi inference as \( k \) increases to \( T \) (with \( \hat{y}^T(1) \) corresponding to PMAP). Therefore, this approach could be interpreted as interpolating between the PMAP and Viterbi inferences. Intuitively, one might also expect \( p(x^T, \hat{y}^T(k)) \) to be strictly increasing with \( k \). This is not exactly so as can be seen from Example A where \( p(x^T, \hat{y}^T(2)) = p(x^T, \hat{y}^T(1)) = 0 \). However, we find the idea of interpolation between the PMAP and Viterbi inferences worth a further investigation. To the best of our knowledge, the only published work which explicitly proposes a solution to such interpolation is (Brushe et al., 1998). The approach of (Brushe et al., 1998) is algorithmic, directly based on continuous mappings, and also deserves an analysis which we present in Subsection 4.2.

1.3 Organization of the rest of the paper

In this paper, we consider the segmentation problem in the more general framework of statistical learning. Namely, we consider sequence classifier mappings

\[
g : X^T \rightarrow S^T
\]

and optimality criteria for their selection. In Section 2, criteria for optimality of \( g \) are naturally formulated in terms of risk minimization whereby \( R(s^T|x^T) \), the risk of \( s^T \), derives from a suitable loss function. In Section 3, we consider families of risk functions which naturally generalize those corresponding to the Viterbi and PMAP solutions (Subsection 2.1). Furthermore, as shown in Section 4, these risk functions define a family of path decoders parameterized by an integer \( k \) with \( k = 1 \) and \( k \rightarrow \infty \) corresponding to the PMAP and Viterbi cases, respectively (Theorem 4). We also show the close connection between
the aforementioned family of decoders and the Rabiner k-block approach. If needed, the new family of decoders can easily be embedded into a yet wider class with a principled criterion of optimality.

All these decoders (classifiers) would only be of theoretical interest if they could not be easily calculated. In Section 3, we show that all of the newly defined decoders can be implemented efficiently as a dynamic programming algorithm in the usual forward-backward manner with essentially the same (computational as well as memory) complexity as the PMAP or Viterbi decoders (Theorem 3).

2. Risk-based segmentation

Given a sequence of observations \( x^T \), we define the (posterior) risk to be a function

\[ R(\cdot | x^T) : S^T \mapsto [0, \infty]. \]

Naturally, we seek a state sequence with minimum risk:

\[ g^*(x^T) := \arg \min_{s^T \in S^T} R(s^T | x^T). \]

Following the statistical decision and pattern recognition theories, the classifier \( g^* \) will be referred to as the Bayes classifier (relative to risk \( R \)). Within the same framework, the risk is often specified via a loss-function

\[ L : S^T \times S^T \rightarrow [0, \infty], \]

interpreting \( L(a^T, b^T) \) as the loss incurred by the decision to predict \( b^T \) when the actual state sequence was \( a^T \). Therefore, for any state sequence \( s^T \in S^T \), the risk is given by

\[ R(s^T | x^T) := E[L(Y^T, s^T) | X^T = x^T] = \sum_{a^T \in S^T} L(a^T, s^T) p(a^T | x^T). \]

### 2.1 Standard path inferences

The most popular loss function is the so-called symmetrical or zero-one loss \( L_{\infty} \) defined as follows:

\[ L_{\infty}(a^T, b^T) = \begin{cases} 1, & \text{if } a^T \neq b^T; \\ 0, & \text{if } a^T = b^T. \end{cases} \]

We shall denote the corresponding risk by \( R_{\infty} \). With this loss, clearly

\[ R_{\infty}(s^T | x^T) = P(Y^T \neq s^T | X^T = x^T) = 1 - p(s^T | x^T), \]

thus \( R_{\infty}(\cdot | x^T) \) is minimized by a Viterbi path, i.e. a sequence of maximum posterior probability. Let \( v(\cdot ; \infty) \) stand for the corresponding classifier, i.e.

\[ v(x^T ; \infty) := \arg \max_{s^T \in S^T} p(s^T | x^T), \]

with a suitable tie-breaking rule.
Evidently, Viterbi paths also minimize the following risk
\[ \bar{R}_\infty(s^T|x^T) := -\frac{1}{T} \log p(s^T|x^T). \] (3)

It can actually be advantageous to use the log-likelihood based risk (3) since, as we shall see later, it leads to various natural generalizations (Sections 3 and 4).

When sequences are compared pointwise, it is common to use additive loss-functions of the form
\[ L_1(a^T, b^T) = \frac{1}{T} \sum_{t=1}^{T} l(a_t, b_t), \] (4)
where \( l(a_t, b_t) \geq 0 \) is the loss associated with classifying the \( t \)-th element \( a_t \) as \( b_t \). Typically, for every state \( s \), \( l(s, s) = 0 \). It is not hard to see that, with \( L_1 \) as in (4), the corresponding risk can be represented as follows
\[ R_1(s^T|x^T) = \frac{1}{T} \sum_{t=1}^{T} R_t(s_t|x^T), \]
where \( R_t(s|x^T) = \sum_{a \in S} l(a, s)p_t(a|x^T) \). Most commonly, \( l \) is again symmetrical, or zero-one, i.e. \( l(s, s') = \mathbb{I}_{\{s \neq s'\}} \), where \( \mathbb{I}_A \) stands for the indicator function of set \( A \). In this case, \( L_1 \) is naturally related to the Hamming distance (Carvalho and Lawrence, 2008). Then also \( R_t(s_t|x^T) = 1 - p_t(s_t|x^T) \) so that the corresponding risk is
\[ R_1(s^T|x^T) := 1 - \frac{1}{T} \sum_{t=1}^{T} p_t(s_t|x^T). \] (5)

Let \( v(\cdot; 1) \) stand for the Bayes classifier relative to the \( R_1 \)-risk. It is easy to see from the above definition of \( R_1 \), that \( v(\cdot; 1) \) delivers PMAP paths, which clearly minimize the expected number of misclassification errors. In addition to maximizing \( \sum_{t=1}^{T} p_t(s_t|x^T) \), \( w \) also maximizes the pseudolikelihood \( \prod_{t=1}^{T} p_t(s_t|x^T) \), and therefore minimizes the following log-pseudolikelihood risk
\[ \bar{R}_1(s^T|x^T) := -\frac{1}{T} \sum_{t=1}^{T} \log p_t(s_t|x^T). \] (6)

2.2 Generalizations

Recall (Subsection 1.2) that PMAP paths can be of zero probability (i.e. not admissible). To ensure admissibility, \( R_1 \)-risk can simply be minimized over the admissible paths:
\[ \min_{s^T : p(s^T|x^T) > 0} R_1(s^T|x^T) \iff \max_{s^T : p(s^T|x^T) > 0} \sum_{t=1}^{T} p_t(s_t|x^T). \] (7)

Assuming that \( p_t(j|x^T), 1 \leq t \leq T, j \in S \), have been precomputed (by the classical forward-backward recursion (Rabiner, 1989)), the solution of (7) can be easily found by a
Viterbi-like recursion (8)

\[
\begin{align*}
\delta_1(j) &:= p_1(j|x^T)r_j, \quad \forall \ j \in S, \\
\delta_{t+1}(j) &:= \max_i \delta_t(i) r_{ij} + p_{t+1}(j|x^T), \text{ for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S,
\end{align*}
\]

where \( r_j := \mathbb{I}_{\{\pi_j > 0\}} \), \( r_{ij} := \mathbb{I}_{\{p_{ij} > 0\}} \). The recursion (8) is also equivalent to

\[
\begin{align*}
\delta_1(j) &:= p_1(j|x^T) + \log r_j, \quad \forall \ j \in S, \\
\delta_{t+1}(j) &:= \max_i (\delta_t(i) + \log r_{ij}) + p_{t+1}(j|x^T) \text{ for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S.
\end{align*}
\]

However, in the presence of path constraints, minimization of the \( R_1 \)-risk is no longer equivalent to minimization of the \( \bar{R}_1 \)-risk. In particular, the problem (7) is not equivalent to the following problem (posterior-Viterbi decoding)

\[
\min_{s^T, p(s^T|x^T) > 0} \bar{R}_1(s^T|x^T) \iff \max_{s^T, p(s^T|x^T) > 0} \sum_{t=1}^{T} \log p_t(s_t|x^T). \tag{10}
\]

A solution to (10) can be computed by a related recursion given in (11) below

\[
\begin{align*}
\delta_1(j) &:= p_1(j|x^T)r_j, \quad \forall \ j \in S, \\
\delta_{t+1}(j) &:= \max_i \delta_t(i) r_{ij} \times p_{t+1}(j|x^T), \text{ for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S.
\end{align*}
\]

Recursion (11) is clearly equivalent to

\[
\begin{align*}
\delta_1(j) &:= \log p_1(j|x^T) + \log r_j, \quad \forall j \in S, \\
\delta_{t+1}(j) &:= \max_i (\delta_t(i) + \log r_{ij}) + \log p_{t+1}(j|x^T), \text{ for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S.
\end{align*}
\]

Although admissible minimizers of \( R_1 \) and \( \bar{R}_1 \) risk are by definition of positive probability, this probability might still be very small. Indeed, in the above recursions, the weight \( r_{ij} \) is 1 even when \( p_{ij} \) is very small. We next replace \( r_{ij} \) (or \( r_j \)) by the true transition (initial) probability \( p_{ij} (\pi_j) \) in minimizing the \( \bar{R}_1 \)-risk (i.e., maximization of \( \prod_{t=1}^{T} p_t(s_t|x^T) \)). Then the solutions remain admissible and now also tend to maximize the prior path probability. With the above replacements, recursions (11) and (12) now solve the following seemingly unconstrained optimization problem (see Theorem 3)

\[
\max_{s^T} \left[ \sum_{t=1}^{T} \log p_t(s_t) + \log p(s^T) \right] \iff \min_{s^T} \left[ \bar{R}_1(s^T|x^T) + h(s^T) \right], \tag{13}
\]

where the penalty term

\[
h(s^T) = -\frac{1}{T} \log p(s^T) =: \bar{R}_\infty(s^T) \tag{14}
\]

is the prior log-likelihood risk which does not depend on the data. The thereby modified recursions immediately generalize as follows:

\[
\begin{align*}
\delta_1(j) &:= \log p_1(j|x^T) + C \log \pi_j, \quad \forall j \in S, \\
\delta_{t+1}(j) &:= \max_i (\delta_t(i) + C \log p_{ij}) + \log p_{t+1}(j|x^T) \text{ for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S,
\end{align*}
\]
solving

\[
\min_{s^T} \left[ \bar{R}_1(s^T|x^T) + Ch(s^T) \right],
\tag{16}
\]

where \( C > 0 \) is a regularization constant and \( h(s^T) = \bar{R}_\infty(s^T) \) (see Section 3 and Theorem 3). Then, PVD, i.e. the problem solved by the original recursions (11) and (12), can be recovered by taking \( C \) sufficiently small. (Alternatively, the PVD problem can also be formally written in the form (16) with \( C = \infty \) and \( h(s^T) \) given, for example, by \( \mathbb{I}_{\{p(s^T)=0\}} \).

What if the actual probabilities \( p_{ij} (\pi_j) \) were also used in the optimal accuracy/PMAP decoding, i.e. optimization (8)-(9)? It appears more sensible to replace the indicators \( r_{ij} (\pi_j) \) with \( p_{ij} (\pi_j) \) in (9) (and not in (8)). This solves the following problem:

\[
\max_{s^T} \left[ \sum p_t(s_t) + \log p(s^T) \right] \Leftrightarrow \min_{s^T} \left[ R_1(s^T|x^T) + \bar{R}_\infty(s^T) \right].
\tag{17}
\]

A more general problem can be written in the form

\[
\min_{s^T} \left[ R_1(s^T|x^T) + Ch(s^T) \right],
\tag{18}
\]

where \( h \) is some penalty function (independent of the data \( x^T \)). Thus, the problem (7) of optimal accuracy/PMAP decoding over the admissible paths is obtained by taking \( C \) sufficiently small and \( h(s^T) = \bar{R}_\infty(s^T) \). (Setting \( C \times h(s^T) = \infty \times \mathbb{I}_{\{p(s^T)=0\}} \) also reduces the problem (18) back to (7).)

3. Combined risks

Motivated by the previous section, we consider the following general problem

\[
\min_{s^T} \left[ C_1 \bar{R}_1(s^T|x^T) + C_2 \bar{R}_\infty(s^T|x^T) + C_3 \bar{R}_1(s^T) + C_4 \bar{R}_\infty(s^T) \right],
\tag{19}
\]

where \( C_i \geq 0, i = 1, 2, 3, 4, \sum_{i=1}^{4} C_i > 0 \). This is also equivalent to

\[
\min_{s^T} \left[ C_1 \bar{R}_1(s^T|x^T) + C_2 \bar{R}_\infty(s^T, x^T) + C_3 \bar{R}_1(s^T) + C_4 \bar{R}_\infty(s^T) \right],
\tag{20}
\]
where

\[
\tilde{R}_1(s^T|x^{T}) = -\frac{1}{T} \sum_{t=1}^{T} \log p_t(s_t|x^{T}), \quad \text{recalling (6)},
\]

\[
\tilde{R}_\infty(s^T, x^{T}) := -\frac{1}{T} \log p(x^{T}, s^T)
\]

\[
= -\frac{1}{T} \left[ \log p(s^T) + \sum_{t=1}^{T} \log f_{s_t}(x_t) \right],
\]

\[
= -\frac{1}{T} \left[ \log \pi_s + \sum_{t=1}^{T-1} \log p_{s_{t}s_{t+1}} + \sum_{t=1}^{T} \log f_{s_t}(x_t) \right],
\]

\[
\tilde{R}_\infty(s^T|x^{T}) = \frac{1}{T} \log p(s^T|x^{T}), \quad \text{recalling (3)},
\]

\[
= \tilde{R}_\infty(s^T, x^{T}) - \frac{1}{T} \log p(x^{T}),
\]

\[
\tilde{R}_1(s^T) := -\frac{1}{T} \sum_{t=1}^{T} \log p_t(s_t),
\]

\[
\tilde{R}_\infty(s^T) = -\frac{1}{T} \log p(s^T), \quad \text{recalling (14)},
\]

\[
= -\frac{1}{T} \left[ \log \pi_s + \sum_{t=1}^{T-1} \log p_{s_{t}s_{t+1}} \right].
\]  

(21)

The newly introduced risk \(\tilde{R}_1(s^T)\) is the prior log-pseudo-likelihood. Evidently, the combination \(C_1 = C_3 = C_4 = 0\) corresponds to the MAP/Viterbi decoding; the combination \(C_2 = C_3 = C_4 = 0\) yields the PMAP case, whereas the combinations \(C_1 = C_2 = C_3 = 0\) and \(C_1 = C_2 = C_4 = 0\) give the \textit{maximum a priori} decoding and \textit{marginal prior mode} decoding, respectively. The case \(C_2 = C_3 = 0\) subsumes (16) and the case \(C_1 = C_3 = 0\) is the problem

\[
\min_{s^T} \left[ \tilde{R}_\infty(s^T|x^{T}) + C \tilde{R}_\infty(s^T) \right].
\]  

(22)

Thus, a solution to (23) is a generalization of the Viterbi decoding that allows one to suppress \((C > 0)\) contribution of the data. \textit{It is important to note that with }\(C_2 > 0\) \textit{every solution of (19) is admissible. No less important, and perhaps a bit less obvious, is that }\(C_1, C_4 > 0\) \textit{also guarantees admissibility of the solutions, as stated in Proposition 1 below.}

**Proposition 1** Let \(C_1, C_4 > 0\). Then, for almost every realization \((x^T, y^T)\) of the HMM process \((X^T, Y^T)\), the minimized risk (19) is finite and any minimizer \(s^T\) is admissible, i.e. satisfies \(p(s^T|x^{T}) > 0\).

**Proof** Without loss of generality, assume \(C_2 = C_3 = 0\). Suppose the problem has no finite solution. Then for any \(s^T\) with \(p(s^T) > 0\), we would have some \(t, 1 \leq t \leq T\), such \(p_t(s_t|x^{T}) = 0\). This would imply that \(p(x^T, s^T) = 0\) for all \(s^T\) with \(p(s^T) > 0\), contradicting the hypothesis that \(p(x^T, y^T) > 0\). Now, suppose that \(s^T\) is a minimizer of (19) but \(p(s^T|x^{T}) = 0\). Since \(p(s^T) > 0\), we would have some \(t, 1 \leq t \leq T\), such \(f_{s_t}(x_t) = 0\). This
would imply \( \alpha_t(s_t) = 0 \) (1), and subsequently that \( p_t(s_t|x^T) = 0 \) and \( \tilde{R}_1(s^T|x^T) = \infty \), contradicting optimality of \( s^T \).

**Remark 2** Thus, note that the Posterior-Viterbi decoding (Fariselli et al., 2005) can be obtained by either setting \( C_3 = C_4 = 0 \) and taking \( 0 < C_2 \ll C_1 \), or setting \( C_2 = C_3 = 0 \) and taking \( 0 < C_4 \ll C_1 \).

If the smoothing probabilities \( p_t(j|x^T) \), \( t = 1, \ldots, T \) and \( j \in S \), have been already computed (say, by the usual forward-backward algorithm), a solution to (19) can be found by a standard dynamic programming algorithm. Let us first introduce more notation. For every \( t \in 1, \ldots, T \) and \( j \in S \), let

\[
g_t(j) := C_1 \log p_t(j|x^T) + C_2 \log f_j(x_t) + C_3 \log p_t(j).
\]

Note that the function \( g_t \) depends on the entire data \( x^T \). Next, let us also define the following scores

\[
\begin{align*}
\delta_1(j) & := C_1 \log p_1(j|x^T) + (C_2 + C_3 + C_4) \log \pi_j + C_2 \log f_j(x_1), \ \forall j \in S, \\
\delta_{t+1}(j) & := \max_i (\delta_t(i) + (C_2 + C_4) \log p_{ij}) + g_{t+1}(j) \\
& \quad \text{for } t = 1, 2, \ldots, T - 1, \text{ and } \forall j \in S.
\end{align*}
\]

Using the above scores \( \delta_t(j) \) and a suitable tie-breaking rule, below we define the back-pointers \( i_t(j) \), terminal state \( i_T \), and the optimal path \( \hat{s}^T(i_T) \).

\[
\begin{align*}
& i_t(j) := \arg \max_{i \in S} [\delta_t(i) + (C_2 + C_4) \log p_{ij}], \quad \text{when } t = 1, \ldots, T - 1; \\
& i_T := \arg \max_{i \in S} \delta_T(i). \\
& \hat{s}^t(j) := \begin{cases} i_1(j), & \text{when } t = 1; \\
& (\hat{s}^{t-1}(i_{t-1}(j)), j) & \text{when } t = 2, \ldots, T.
\end{cases}
\end{align*}
\]

The following theorem formalizes the dynamic programming argument; its proof is standard and we state it below for completeness only.

**Theorem 3** Any solution to (19) can be represented in the form \( \hat{s}^T(i_T) \) provided the ties in (25) are broken accordingly.

**Proof** With a slight abuse of notation, for every \( s^t \in S^t \), let

\[
U(s^t) = \sum_{u=1}^{t} [g_u(s_u) + (C_2 + C_4) \log p_{s_u-1s_u}],
\]

where \( s_0 := 0 \) and \( p_{0s} := \pi_s \). Hence,

\[
-T[C_1 \tilde{R}_1(s^T|x^T) + C_2 \tilde{R}_\infty(s^T, x^T) + C_3 \tilde{R}_1(s^T) + C_4 \tilde{R}_\infty(s^T)] = U(s^T)
\]

and any maximizer of \( U(s^T) \) is clearly a solution to (19) and (20).
Next, note that \( \delta_1(j) = U(j) \) for all \( j \in S \), and that

\[
U(s^{t+1}) = U(s^t) + (C_2 + C_4) \log p_{st,s_{t+1}} + g_{t+1}(s_{t+1}),
\]

for \( t = 1, 2, \ldots, T - 1 \) and also \( s^t \in S^t \). By induction on \( t \), these yield

\[
\delta_t(j) = \max_{s^t : s_t = j} U(s^t)
\]

for every \( t = 1, 2, \ldots, T \) and for all \( j \in S \). Clearly, every maximizer \( \hat{s}_T \) of \( U(s^T) \) over the set \( S^T \) must satisfy \( \hat{s}_T = i_T \), or, more precisely \( \hat{s}_T \in \arg \max_{j \in S} \delta_T(j) \), allowing for non-uniqueness. Continuing to interpret \( \arg \max \) as a set, recursion (25) implies recursions (26) and (27), hence any maximizer \( \hat{s}_T \) can indeed be computed in the form \( \hat{s}_T(s_T) \) via the forward (recursion (25))-backward (recursion (26)) procedure.

Similarly to the generalized risk minimization of (19), the generalized problem of accuracy optimization (18) can also be further generalized as follows:

\[
\min_{s^T} \left[ C_1 R_1(s^T|x^T) + C_2 R_\infty(s^T|x^T) + C_3 R_1(s^T) + C_4 R_\infty(s^T) \right], \tag{27}
\]

where risk

\[
R_1(s^T) := \frac{1}{T} \sum_{t=1}^{T} P(Y_t \neq s_t) = 1 - \frac{1}{T} \sum_{t=1}^{T} p_t(s_t) \tag{28}
\]

is the error rate relative to the prior distribution. This problem apparently can be solved by the following recursion

\[
\delta_1(j) := C_1 p_1(j|x^T) + (C_2 + C_4) \log \pi_j + C_2 \log f_j(x_1) + C_3 \pi_j, \ \forall j \in S,
\]

\[
\delta_{t+1}(j) := \max_i \left( \delta_t(i) + (C_2 + C_4) \log p_{ij} + g_{t+1}(j) \right), \tag{29}
\]

where now

\[
g_t(s) = C_1 p_t(s|x^T) + C_2 \log f_s(x_t) + C_3 p_t(j).
\]

As in the generalized posterior-Viterbi decoding (19), here \( C_2 > 0 \) also implies admissibility of the optimal paths. However, unlike in (19), \( C_1, C_4 > 0 \) is not sufficient to guarantee admissibility of the solutions.

4. Other approaches to hybridization of PMAP and Viterbi

We have been discussing a set of related ideas which allow us to balance path accuracy against path probabilities. Next, we extend this discussion by presenting a couple of notably different approaches.
4.1 The \( k \)-block Posterior-Viterbi decoding

Recall (Subsection 1.2) that Rabiner’s compromise between MAP and PMAP is to maximize the expected number of correctly decoded pairs or triples of (adjacent) states. With \( k \) being the length of the overlapping block (\( k = 2, 3, \ldots \)) this means to minimize the conditional risk

\[
R_k(s^T|x^T) := \sum_{t=1}^{T-k+1} p(s_{t+k-1}^{t+1} \mid x^T)
\]

which derives from the following loss function:

\[
L_k(y^T, s^T) := \sum_{t=1}^{T-k+1} I\{s_{t+k-1}^{t+1} \neq y_{t+k-1}^{t+1}\}.
\]

Obviously, for \( k = 1 \) this gives the usual \( R_1 \) maximization – the PMAP decoding – which is known to fault by allowing inadmissible paths. It is natural to think that minimizers of \( R_k(s^T|x^T) \) “move” towards Viterbi paths “monotonically” as \( k \) increases to \( T \). Indeed, when \( k = T \), minimization of \( R_k(s^T|x^T) \) (30) is equivalent to minimization of \( \bar{R}_{\infty}(s^T|x^T) \) achieved by the Viterbi decoding. However, as Example A shows below, minimizers of (30) are not guaranteed to be admissible for \( k > 1 \), which is a drawback of using the loss \( L_k \).

We now show that this drawback is easily overcome when the sum in (30) is replaced by the product. Certainly, these problems are not equivalent, and in particular with the product in place of the sum the \( k \)-block idea works well. Namely, the longer the block, the larger the resulting path probability, which is also now guaranteed to be positive already for \( k = 2 \). Moreover, this gives another interpretation of the risks \( \bar{R}_1(s^T|x^T) + C\bar{R}_{\infty}(s^T|x^T) \) (see also Remark 2 above) and, though perhaps less interestingly, the prior risks \( \bar{R}_1(s^T) + C\bar{R}_{\infty}(s^T) \).

Let \( k \) be a positive integer. For the time being, let \( p \) represent any first order Markov chain on \( S^T \), and let us define

\[
\bar{U}_k(s^T) := \prod_{j=1-k}^{T-1} p(s_{(j+1)\wedge T}^{(j+k)\wedge T} \mid s_{(j+1)\vee 1}^{(j+k)\vee 1}), \quad \bar{R}_k(s^T) := -\frac{1}{T} \ln \bar{U}_k(s^T).
\]

Thus

\[
\bar{U}_k(s^T) = U_1^k \cdot U_2^k \cdot U_3^k,
\]

where

\[
U_1^k := p(s_1) \cdot \ldots \cdot p(s_1^{k-2})p(s_1^{k-1}) \quad U_2^k := p(s_1^k)p(s_2^{k+1}) \cdot \ldots \cdot p(s_{T-k}^{T-1})p(s_{T-k+1}^{T}) \quad U_3^k := p(s_{T-k+2}^{T-3})p(s_{T-k+3}^{T}) \cdot \ldots \cdot p(s_T).
\]

Thus, \( \bar{R}_k \) is a natural generalization of \( \bar{R}_1 \) (introduced first for the posterior distribution in (6)) since when \( k = 1 \), \( \bar{R}_k = \bar{R}_1 \).
Theorem 4 Let $k$ be such that $T \geq k > 1$. Then the following recursion holds
\[ \bar{R}_k(s^T) = \bar{R}_\infty(s^T) + \bar{R}_{k-1}(s^T), \quad \forall s^T \in S^T. \]

Proof Note that
\[ U_1^k = U_1^k p(s_1^{k-1}), \quad U_3^k = p(s_{T-k+2}^T) U_3^{k-1}. \]
Next, for all $j$ such that $j + k \leq T$, the Markov property gives
\[ p(s_{j+k}^T) = p(s_{j+k}|s_{j+k-1}) p(s_{j+k-1}^T) \]
and
\[ U_2^k p(s_{T-k+2}^T) = p(s_1^k) p(s_2^{k-1}) \cdots p(s_{T-k+1}^T) p(s_{T-k+2}) = p(s_k|s_{k-1}) p(s_{k+1}|s_k) p(s_{k+2}|s_{k+1}) \cdots p(s_{T-1}|s_{T-2}) p(s_T|s_{T-1}) \]
\[ p(s_k|s_{k-1}) p(s_{k+1}|s_k) \cdots p(s_{T-1}|s_{T-2}) p(s_T|s_{T-1}) p(s_{T-k+2}) = p(s_k|s_{k-1}) \cdots p(s_T|s_{T-1}) U_2^{k-1}. \]
Hence,
\[ U_k(s^T) = U_1^{k-1} p(s_1^{k-1}) p(s_k|s_{k-1}) \cdots p(s_{T-1}|s_{T-2}) p(s_T|s_{T-1}) U_2^{k-1} U_3^{k-1} = p(s_T^T) \bar{U}_{k-1}(s^T). \]
The second equality above also follows from the Markov property. Taking logarithms on both sides and dividing by $-T$ completes the proof.

Now, we specialize this result to our HMM context, and, thus, $p(s^T)$ and $p(s^T|x^T)$ are again the prior and posterior hidden path distributions.

Corollary 5 Let $k$ be such that $T \geq k > 1$. For all paths $s^T \in S^T$ the prior risks $\bar{R}_k$ and $\bar{R}_\infty$ satisfy (32). For every $x^T \in X^T$ and for all paths $s^T \in S^T$, the posterior risks $\bar{R}_k$ and $\bar{R}_\infty$ satisfy (33).
\[ \bar{R}_k(s^T) = \bar{R}_\infty(s^T) + \bar{R}_{k-1}(s^T), \quad (32) \]
\[ \bar{R}_k(s^T|x^T) = \bar{R}_\infty(s^T|x^T) + \bar{R}_{k-1}(s^T|x^T). \quad (33) \]

Proof Clearly, conditioned on the data $x^T$, $Y^T$ remains a first order Markov chain (generally inhomogeneous even if it was homogeneous a priori). Hence, Theorem 4 applies.

Below, we focus on the posterior distribution and risks, even though the following would readily extend to any first order Markov chain.

Let $v(x^T; k)$ be a classifier that minimizes $\bar{R}_k(s^T|x^T)$. Thus,
\[ v(x^T; k) = \arg \max_{s^T \in S^T} \bar{U}_k(s^T|x^T) = \arg \min_{s^T \in S^T} \bar{R}_k(s^T|x^T). \]
We refer to such classifiers as $k$-block PVD or $k$-block PMAP as they naturally extend $v(\cdot; 1)$, the PMAP/optimal accuracy/posterior decoder (Section 2.1). To be consistent with applications (Fariselli et al., 2005), the term ‘$k$-block posterior-Viterbi decoding’, however, is perhaps more accurate given the use of the product-based risk $\bar{R}_k(s^T|x^T)$ as opposed to $R_1(s^T|x^T)$.

Now, we present some properties of the new risks and decoders.
Corollary 6 For every \( x^T \in X^T \), and for every \( s^T \in S^T \), we have

\[
R_k(s^T|x^T) = (k-1)\bar{R}_\infty(s^T|x^T) + \tilde{R}_1(s^T|x^T)
\]

for every \( k \geq 1, 2, \ldots, T \). (34)

\( v(x^T; k) \) is admissible

\[
\bar{R}_\infty(v(x^T; k)|x^T) \leq \bar{R}_\infty(v(x^T; k-1)|x^T)
\]

for every \( k > 1, 2, \ldots, T \). (35)

\[
\tilde{R}_1(v(x^T; k)|x^T) \geq \tilde{R}_1(v(x^T; k-1)|x^T)
\]

for every \( k > 1, 2, \ldots, T \). (36)

Proof Equation (34) follows immediately from equation (33) of Corollary 5. Inequality (35) follows from inequalities (37) below

\[
\bar{R}_\infty(v(x^T; k-1)|x^T) - \bar{R}_\infty(v(x^T; k)|x^T) \geq 0\]

which in turn follow from equation (33) of Corollary 5. Also, equation (34) implies that for every \( k \geq 2 \),

\[
(k-2)\bar{R}_\infty(v(x^T; k-1)|x^T) + \tilde{R}_1(v(x^T; k-1)|x^T) \leq (k-2)\bar{R}_\infty(v(x^T; k)|x^T) + \tilde{R}_1(v(x^T; k)|x^T),
\]

which, together with inequality (35), implies (36).

Inequality (35) means that the posterior path probability \( p(v(x^T; k)|x^T) \) increases with \( k \). Equation (34) is also of practical significance showing that \( v(x^T; k) \) is a solution to (19) with \( C_1 = 1, C_2 = k - 1, C_3 = C_4 = 0 \), and as such can be computed in the same fashion for all \( k \) (see Theorem 3 above).

Thus, increasing \( k \) increases \( \bar{R}_1 \)-risk, i.e. decreases the product of the (conditional) marginal probabilities of states along the path \( v(x^T; k) \). Inequalities (35) and (36) clearly show that as \( k \) increases, \( v(\cdot; k) \) monotonically moves from \( v(\cdot; 1) \) (PMAP) towards the Viterbi decoder, i.e. \( v(\cdot; \infty) \). However, the maximum block length is \( k = T \). A natural way to complete this bridging of PMAP with MAP is by embedding the collection of risks \( \bar{R}_k \) into the family \( \bar{R}_\alpha \) via \( \alpha = 1/k \in [0, 1] \). Thus, (34) extends to

\[
\bar{R}_\alpha(s^T|x^T) := (1 - \alpha)\bar{R}_\infty(s^T|x^T) + \alpha\tilde{R}_1(s^T|x^T)
\]

with \( \alpha = 0 \) and \( \alpha = 1 \) corresponding to the Viterbi and PMAP cases, respectively. Given \( x^T \) and a sufficiently small \( \alpha \) (equivalently, large \( k \)), \( v(x^T; k) \), the minimizer of \( \bar{R}_\alpha(s^T|x^T) \) (38) (or, the right hand side of (34)) would produce a Viterbi path \( v(x^T; \infty) \) (since \( S^T \) is finite). However, such \( \alpha \) (and \( k \)) would generally depend on \( x^T \), and in particular \( k \) may need to be larger than \( T \), i.e. \( v(x^T; T) \) may be different from \( v(x^T; \infty) \). At the same time, we clearly have

\[
\bar{R}_\infty(v(x^T; \infty)|x^T) \leq \bar{R}_\infty(v(x^T; k)|x^T) \leq \bar{R}_\infty(v(x^T; \infty)|x^T) + \frac{\tilde{R}_1(v(x^T; \infty)|x^T)}{k-1},
\]

on which we comment more in Section (5) below.
4.2 Algorithmic approaches

An alternative that does not involve the risk functions is to simply transform the forward and backward variables $\alpha_t(i)$ and $\beta_t(j)$ defined in (1). Consider, for one example, the recursively applied power transformations given in (40) below

$$\alpha_1(j; q) := \alpha_1(j)$$

$$\alpha_{t+1}(j; q) := \left[ \sum_{i \in S} (\alpha_t(i; q)p_{ij})^q \right]^{\frac{1}{q}} f_j(x_{t+1}), \quad 1 \leq t < T$$

$$\beta_t(j; q) := \left[ \sum_{i \in S} (p_{ji}f_i(x_{t+1})\beta_{t+1}(i; q))^q \right]^{\frac{1}{q}}, \quad 1 \leq t < T$$

$$\beta_T(j; q) := \beta_T(j) = 1.$$  

Clearly, $\alpha_t(j; 1) = \alpha_t(j)$ and $\beta_t(j; 1) = \beta_t(j)$, for all $j \in S$ and all $t = 1, 2, \ldots, T$. Thus, $q = 1$ leads to the PMAP decoding. Using induction on $t$ and continuity of the power transform, it can also be seen that the following limits exist and are finite for all $j \in S$ and all $t = 1, 2, \ldots, T$: $\lim_{q \to \infty} \alpha_t(j; q) =: \alpha_t(j; \infty)$ and $\lim_{q \to \infty} \beta_t(j; q) =: \beta_t(j; \infty)$, where for $1 < t < T,$

$$\alpha_t(j; \infty) = \max_{s_t:s_{t+1} = j} p(x_t, s_t),$$

$$\beta_t(j; \infty) = \max_{s_t:s_{t+1} = j} p(x_t^T, s_t^T | Y_t = j),$$

and therefore, any Viterbi path $v(x_T; \infty) = (v_1, \ldots, v_T)$ has the following property:

$$v_t = \arg \max_{j \in S} \{ \alpha_t(j; \infty)\beta_t(j; \infty) \}. \quad (42)$$

This has been already been pointed out by (Brushe et al., 1998), who, to the best of our knowledge, (Brushe et al., 1998) have been the only group to publish on the idea of hybridization of the PMAP and Viterbi decoders via a continuous transformation. Ignoring potential non-uniqueness of Viterbi paths, (Brushe et al., 1998) state, based on (42), that the Viterbi path can be found symbol-by-symbol. Certainly, when Viterbi paths are non-unique, symbol-by-symbol decoding based on (42) can produce suboptimal, and in principle inadmissible, paths. In contrast to Viterbi, non-unique PMAP paths (in the absence of constraints) can certainly be found symbol-by-symbol.

Note also that for their hybrid of PMAP with Viterbi, (Brushe et al., 1998) use the following transformations:

$$\frac{1 + (N - 1)\exp(-\mu)}{\mu} \ln \left( \frac{1}{N} \sum_{j=1}^{N} \exp(\mu d_j(\mu)) \right), \quad (43)$$

where $N = K$ (in our notation) and $d_j(\mu)$ are required to be continuous on $[0, \infty)$ with finite limits as $\mu \to \infty$. These $d_j$ are then substituted for by appropriate expressions in terms of the recursively transformed forward and backward variables.

The following points with regard to this hybridization idea have also motivated our present work:
1. Transformations (43) appear to be somewhat more sophisticated than the power transforms (40). It appears that the only reason explicitly stated in (Brushe et al., 1998) for making their choice of transformation is to deliver the correct limits (in that case PMAP with $\mu = 0$ and Viterbi with $\mu = \infty$). Besides (43) and (40), there are other (single parameter) transformations meeting this condition.

2. Implicitly, the authors of (Brushe et al., 1998) do recognize the usual problem of numerical underflow, but somehow appear to suggest that the rescaling trick

$$\log(e^a + e^b) = \max\{a, b\} + \log\left(1 + e^{-|a - b|}\right)$$

would be sufficient to resolve this problem when computing their transformed variables. We could not experimentally confirm this optimism even with simple models. In fact, without addressing the scaling issue in full, the expressions (43) and (40) are short of defining practically meaningful path decoders. True, it is not difficult to renormalize these, or similar, expressions while preserving their limiting behavior. However, unlike in the original, i.e. untransformed, forward-backward algorithm, renormalization of the transformed forward and backward variables will alter the original decoders for intermediate values of the tuning parameter. This can already be suspected by examining equations (44) below

$$\tilde{\alpha}_1(j; q) := \frac{\alpha_1(j)}{\sum_{l \in S} \alpha_1(l)}$$

$$\tilde{\alpha}_{t+1}(j; q) := \frac{\left[\sum_{i \in S} (\tilde{\alpha}_t(i; q)p_{ij})]q^{\frac{1}{q}} f_j(x_{t+1})\right]}{\left[\sum_{i \in S} p_{ij} f_i(x_{t+1})\right] f_j(x_{t+1})}, \quad 1 \leq t < T$$

$$\tilde{\beta}_t(j; q) := \frac{\left[\sum_{i \in S} p_{ji} f_i(x_{t+1})\tilde{\beta}_{t+1}(i; q)\right]^{\frac{1}{q}} f_i(x_{t+1})}{\left[\sum_{i \in S} (\tilde{\alpha}_t(i; q)p_{ji})]q^{\frac{1}{q}} f_j(x_{t+1})\right]}, \quad 1 \leq t < T$$

$$\tilde{\beta}_T(j; q) := \beta_T(j) = 1.$$ 

which implement the usual rescaling (Rabiner, 1989).

3. Moreover, algorithmically defined estimators are generally hard to analyze rigorously (Winkler, 2003, pp. 25, 129-131). In our context, optimization criteria for intermediate members of the above interpolating families are indeed not clear, making it difficult to interpret the corresponding decoders. This might be discouraging should such decoders be included in more complex inference cycles (i.e. when any genuine model parameters are estimated as well, e.g. Viterbi Training (Koski, 2001; Lember and Koloydenko, 2008, 2010)).

4. Other recursion schemes (for example, cf. (Koski, 2001, pp. 272-273) for Derin's formula) can surely be experimented with in a similar manner. However, now more than ten years after appearance of (Brushe et al., 1998), we find that the value of any such interpolation is yet to be demonstrated.
5. As already mentioned above, the symbol-by-symbol implementation of the transform-based hybrids is problematic when the solution is non-unique and full path probability is a factor.

5. Asymptotic risks

Given a classifier $g$ and a risk function $R$, the quantity $R(g(x^T)|x^T)$ evaluates the risk when $g$ is applied to a given sequence $x^T$. When $g$ is optimal in the sense of risk minimization, then $R(g(x^T), x^T) = \min_{s^T} R(s^T|x^T) =: R(x^T)$. We are also interested in the random variables $R(g(X^T), X^T)$. Thus, in (Kuljus and Lember, 2010), convergence of several risks of the Viterbi decoding has been considered. Based on the asymptotic theory of Viterbi processes $v(X^\infty; \infty)$ (Lember and Koloydenko, 2008, 2010), it has been shown that under fairly general assumptions on the HMM, the random variables $R_1(v(X^T; \infty)|X^T)$, $R_1(v(X^T; \infty)|X^T)$, $R^\infty(X^T) = \bar{R}_\infty(v(X^T; \infty)|X^T)$ as well as $\bar{R}_\infty(v(X^T; \infty))$, $R_1(v(X^T; \infty)$ and $\bar{R}_1(v(X^T; \infty))$ all converge to constant limits, a.s.. Convergence of these risks obviously imply convergence of

$$C_1\bar{R}_1(v(X^T; \infty)|X^T) + C_2\bar{R}_\infty(v(X^T; \infty)|X^T) + C_3\bar{R}_1(v(X^T; \infty)) + C_4\bar{R}_\infty(v(X^T; \infty)),$$

and

$$C_1R_1(v(X^T; \infty)|X^T) + C_2\bar{R}_\infty(v(X^T; \infty)|X^T) + C_3R_1(v(X^T; \infty)) + C_4\bar{R}_\infty(v(X^T; \infty)),$$

the risks appearing in the generalized problems (19) and (27), respectively. Actually, convergence of $\bar{R}_\infty(v(X^T; \infty)|X^T)$ is also proved (and used in the proof of convergence of $\bar{R}_\infty(v(X^T; \infty)|X^T)$). Hence, the risk in (20), evaluated at Viterbi paths, converges as well.

The limits – asymptotic risks – are (deterministic) constants that depend only on the model and evaluate the Viterbi inference. For example, let $R_1(v(\infty))$ be the limit of $R_1(v(X^T; \infty)|X^T)$, which is the asymptotic misclassification rate of the Viterbi decoding. Thus, for big $T$, the Viterbi decoding makes about $TR_1(v(\infty))$ misclassification errors. The asymptotic risks might be, in principle, found theoretically, but as the limit theorems show, the limiting risks can also be estimated by simulations.

In (Lember, 2009), it has been also shown that under the same assumptions $R_1(X^T) = R_1(v(X^T; 1)|X^T)$ converges to a constant limit, say $R_1(v(1))$. In (Kuljus and Lember, 2010), $\bar{R}_1(X^T) = \bar{R}_1(v(X^T; 1)|X^T)$ has been also shown to converge. Clearly $R_1(v(\infty)) \geq R_1(v(1))$, and even if their difference is small, the number of errors made by the Viterbi decoder in excess of PMAP in the long run can still be found significant.

Presently, we are not aware of a universal method for proving the limit theorems for these risks. Convergence of the risks of the Viterbi decoding is possible due to the existence of the so-called Viterbi process (see (Koloydenko and Lember, 2008; Lember and Koloydenko, 2008, 2010)) that has nice ergodic properties. The question whether infinite PMAP processes have similar properties, is still open. Therefore, convergence of $R_1(X^T) = R_1(v(X^T; 1)|X^T)$ was proven with a completely different method based on the smoothing probabilities. In fact, all of the limit theorems obtained thus far have been proven with different methods. We conjecture that these different methods can be combined so that convergence of the minimized combined risk (19) or (27) could be proven as well. In
Certainly, the logarithmic risks (39) gives
\[
0 \leq \bar{R}_\infty(v(x^T; C)|x^T) - \bar{R}_\infty(v(x^T; \infty)|x^T) \leq \frac{\bar{R}_1(v(x^T; \infty)|x^T)}{C}.
\]
This, together with the a.s. convergence of \( \bar{R}_1(v(X^T; \infty)|X^T) \), implies that in the long run, for most sequences \( x^T \), \( \bar{R}_\infty(v(x^T; C)|x^T) \) will not exceed \( \bar{R}_\infty(v(x^T; \infty)|x^T) \) by more than \( \lim_{T \to \infty} \bar{R}_1(v(X^T; \infty)|X^T)/C \). Since this limit is finite, letting \( C \) increase with \( T \), \( R_\infty(v(X^T; C(T))) \) obviously approaches \( \lim_{T \to \infty} R_\infty(v(X^T; \infty)) \) a.s., i.e. as the intuition predicts, the likelihood of \( v(X^T; C(T)) \) approaches to that of \( v(X^T; \infty) \).

6. Discussion

Certainly, the logarithmic risks (3), (6), (14), (21) on the one hand, and the ordinary risks (2), (5), \( R_\infty(s^T) = 1 - p(s^T) \), (28), on the other, can be respectively combined into a single parameter family by, for example, the power transformation as shown below. Let \( p \) for the moment be any probability distribution on \( S^T \).

\[
R_1(s^T; \beta) = \begin{cases} 
-\frac{1}{T} \sum_{t=1}^{T} \frac{p_t(s_t)^{\beta-1}}{\beta}, & \text{if } \beta \neq 0 \\
-\frac{1}{T} \sum_{t=1}^{T} \log p_t(s_t), & \text{if } \beta = 0
\end{cases}
\]

\[
R_\infty(s^T; \beta) = \begin{cases} 
-\frac{1}{T} \frac{p(s^T)^{\beta-1}}{\beta}, & \text{if } \beta \neq 0 \\
-\frac{1}{T} \log p(s^T), & \text{if } \beta = 0
\end{cases}
\]

Thus, our two generalized problems (19) and (27) are naturally members of the same family of problems:

\[
\min_{s^T} \left[ C_1 R_1(s^T|x^T; \beta_1) + C_2 R_\infty(s^T|x^T; \beta_2) + C_3 R_1(s^T; \beta_3) + C_4 R_\infty(s^T; \beta_4) \right],
\]

where \( C_i \geq 0 \) and \( \beta_i \geq 0 \), \( i = 1, 2, 3, 4 \), and \( \sum_{i=1}^{4} C_i > 0 \). Clearly, the dynamic programming approach of Theorem 3 and (29) immediately applies to any member of the above family (46) with \( \beta_2 = \beta_4 = 0 \).

Theorem 4 and Corollaries 5 and 6 obviously generalize to higher order Markov chains as can be seen below.

**Proposition 7** Let \( p \) represent a Markov chain of order \( m \), \( 1 \leq m \leq T \), on \( S^T \). Then for any \( s^T \in S^T \) and for any \( k \in \{m, m+1, \ldots\} \), we have

\[
\bar{R}_k(s^T) = \bar{R}_m(s^T) + (k - m) \bar{R}_\infty(s^T).
\]
The present risk-based discussion of HMM path inference also naturally extends to the problem of optimal labelling or annotation (Section 1.2). Namely, the state space $S$ can be partitioned into subsets $S_1, S_2, \ldots, S_{\Lambda}$, for some $\Lambda \leq K$, in which case $\lambda(s)$ assigns label $\lambda$ to every state $s \in S_{\lambda}$. The fact that the PMAP problem is as easily solved over the label space $\Lambda_T$ as it is over $S_T$ has already been used in practice. Indeed, adding the admissibility constraint, (Käll et al., 2005) in effect average $p_t(s_t|x^T)$'s within the label classes and then use recursions (8) to obtain the PMAP labelling, say, $\lambda^T(x^T; 1^+)$, of admissible state paths. This approach clearly corresponds to using the point loss $l(s, s') = \mathbb{I}_{\{\lambda(s) \neq \lambda(s')\}}$ in (4) when solving $\min_{s_T: p_t(s_t|x^T) > 0} R_1(s^T|x^T)$ (7). Importantly, our generalized problem (46) also immediately incorporates the above pointwise label-loss in either the prior $R_1(\cdot; \beta_3)$ or posterior risk $R_1(\cdot; \beta_1)$, or both. Since computationally these problems are essentially as light as (29) and since (Käll et al., 2005) report their special case to be useful in practice, we believe that the above generalizations offer yet more useful possibilities to practitioners. Note the different kinds of averaging corresponding to different values of $\beta$ to be used with the $R_1$ risks:

$$
\bar{p}_t(s; \beta) \propto \begin{cases} 
\left( \frac{\sum_{s' \in S_{\lambda(s)}} p_t(s')}{|S_{\lambda(s)}|} \right)^{\beta} & \text{if } \beta \neq 0, \\
\left( \prod_{s' \in S_{\lambda(s)}} p_t(s') \right)^{\frac{1}{|S_{\lambda(s)}|}} & \text{if } \beta = 0.
\end{cases}
$$

Certainly, the choice of the basic loss functions, inflection parameters $\beta_i$ and weights $C_i$ of the respective risks, is application dependent, and can be tuned with the help of labelled data, using cross-validation. Finally, these generalizations are presented for the standard HMM setting, and work on extending them to more complex and practically more useful HMM-based settings (e.g. semi-Markov, autoregressive, etc.) is underway.

**Appendix A. Example:** Optimal recognition of pairs can still produce inadmissible solutions.

Consider the following four-state MC transition matrix

$$
\begin{pmatrix}
1 & 0 & 4 & 2 \\
0 & \frac{1}{8} & 4 & 1 & 1 & 2 \\
0 & 0 & 2 & 1 & 1 & 4 \\
0 & 0 & 0 & 2 & 2 & 4 & 0
\end{pmatrix}
$$

Suppose observations $x_1, x_2, x_3, x_4$ and the emission densities $f_s s = 1, 2, 3, 4$ are such that

$$
f_s(x_1) = f_s(x_4) = \begin{cases} 
1, & \text{if } s = 2; \\
0, & \text{if } s \neq 2.
\end{cases}, \quad f_s(x_3) = f_s(x_2) = \begin{cases} 
A > 1, & \text{if } s = 1; \\
1, & \text{if } s \neq 1.
\end{cases}
$$

Hence every admissible path begins and ends with 2. Thus, to simplify the notation, we assume without loss of generality that $P(Y_1 = 2) = 1$. Amongst the paths that begin
and end with 2, the paths whose probabilities are listed below, are the only ones of positive (prior) probability (the probabilities below are calculated up to the normalization constant):

\[
\begin{align*}
p(2, 1, 2, 2) &= p(2, 2, 1, 2) \propto 16, \\
p(2, 1, 3, 2) &= p(2, 3, 1, 2) \propto 8, \\
p(2, 1, 4, 2) &= p(2, 4, 1, 2) \propto 16, \\
p(2, 2, 1, 2) &= p(2, 1, 2, 2) \propto 16, \\
p(2, 2, 2, 2) &= \propto 1, \\
p(2, 2, 3, 2) &= p(2, 3, 2, 2) \propto 1, \\
p(2, 2, 4, 2) &= p(2, 4, 2, 2) \propto 4, \\
p(2, 3, 1, 2) &= p(2, 1, 3, 2) \propto 8, \\
p(2, 3, 2, 2) &= p(2, 2, 3, 2) \propto 1, \\
p(2, 3, 3, 2) &= p(2, 4, 3, 2) \propto 8, \\
p(2, 4, 1, 2) &= p(2, 1, 4, 2) \propto 16, \\
p(2, 4, 2, 2) &= p(2, 2, 4, 2) \propto 4, \\
p(2, 4, 3, 2) &= p(2, 3, 4, 2) \propto 8. \\
\end{align*}
\]

For every pair \(s_1, s_2\) of states and for every \(i = 1, 2, 3\), let \(p_{ii+1}(s_1, s_2) := P(Y_i = s_1, Y_{i+1} = s_2|x^4)\). Then, we have

\[
\begin{align*}
p_{1,2}(1, 1) &= p_{3,4}(1, 2) \propto 40A, \\
p_{1,2}(2, 2) &= p_{3,4}(2, 2) \propto 16A + 6, \\
p_{1,2}(3, 1) &= p_{3,4}(3, 2) \propto 8A + 10, \\
p_{1,2}(4, 2) &= p_{3,4}(4, 2) \propto 16A + 12, \\
p_{2,3}(1, 2) &= p_{2,3}(2, 1) \propto 16A, \\
p_{2,3}(1, 3) &= p_{2,3}(3, 1) \propto 8A, \\
p_{2,3}(2, 1) &= p_{2,3}(4, 1) \propto 16A, \\
p_{2,3}(2, 2) &= \propto 1, \\
p_{2,3}(2, 3) &= p_{2,3}(3, 2) \propto 1, \\
p_{2,3}(2, 4) &= p_{2,3}(4, 2) \propto 4, \\
p_{2,3}(3, 2) &= p_{2,3}(3, 3) \propto 1, \\
p_{2,3}(3, 4) &= p_{2,3}(4, 3) \propto 8. \\
\end{align*}
\]

Let

\[
W_2(s^4) := p_{1,2}(s_1, s_2) + p_{2,3}(s_2, s_3) + p_{3,4}(s_3, s_4).
\]
Then we also have

\[ W_2(2, 1, 1, 2) \propto 40A + 40A = 80A \]
\[ W_2(2, 1, 2, 2) \propto 40A + 16A + 16A + 6 = 72A + 6 \]
\[ W_2(2, 1, 3, 2) \propto 40A + 8A + 8A + 10 = 56A + 10 \]
\[ W_2(2, 1, 4, 2) \propto 40A + 16A + 16A + 12 = 72A + 12 \]
\[ W_2(2, 1, 1, 2) = W_2(2, 1, 2, 2) \propto 72A + 6 \]
\[ W_2(2, 2, 2, 2) \propto 16A + 6 + 1 + 16A + 6 = 32A + 13 \]
\[ W_2(2, 2, 3, 2) \propto 16A + 6 + 1 + 8A + 10 = 24A + 17 \]
\[ W_2(2, 2, 4, 2) \propto 16A + 6 + 4 + 16A + 12 = 32A + 22 \]
\[ W_2(2, 3, 1, 2) = W_2(2, 1, 3, 2) \propto 56A + 10 \]
\[ W_2(2, 3, 2, 2) = W_2(2, 2, 3, 2) \propto 24A + 17 \]
\[ W_2(2, 3, 3, 2) \propto 8A + 10 + 1 + 8A + 10 = 16A + 21 \]
\[ W_2(2, 3, 4, 2) \propto 8A + 10 + 8 + 16A + 12 = 24A + 30 \]
\[ W_2(2, 4, 1, 2) = W_2(2, 1, 4, 2) \propto 72A + 12 \]
\[ W_2(2, 4, 2, 2) = W_2(2, 2, 4, 2) \propto 32A + 22 \]
\[ W_2(2, 4, 3, 2) = W_2(2, 3, 4, 2) \propto 24A + 30 \]
\[ W_2(2, 4, 4, 2) = 32A + 24. \]

Hence, when \( A \) is sufficiently big, then

\[
\arg \max_{s^4} W_2(s^4) = (2, 1, 1, 2),
\]

but \( p(2, 1, 1, 2) = 0 \), i.e. the path is not admissible.

The Viterbi paths here are (2, 1, 2, 2), (2, 2, 1, 2), (2, 1, 4, 2), (2, 4, 1, 2). Also note that, for every \( i = 1, 2, 3, 4 \), \( p_2(i) = p_{1,2}(2, i) \) and \( p_3(i) = p_{3,4}(i, 2) \). Thus (2, 1, 1, 2) is also the unique PMAP path. Minimization of \( \bar{R}_2(s^4|x^4) \) over all \( s^4 \) here is equivalent to maximization of \( p_2(s_2)p_{2,3}(s_2, s_3)p_3(s_3) \), and the optimal paths are (2, 1, 4, 2) and (2, 4, 1, 2).

References

L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate (corresp.). *Information Theory, IEEE Transactions on*, 20(2):284–287, 1974. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1055186.

Julian Besag. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B*, 48(3):259–302, 1986. ISSN 0035-9246.

Julian Besag and Peter J. Green. Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B*, 55(1):25–37, 1993. ISSN 0035-9246.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0387-31073-2; 0-387-31073-8. doi: 10.1007/978-0-387-45528-0. URL http://dx.doi.org/10.1007/978-0-387-45528-0.
Broňa Brejová, Daniel G. Brown, and Tomáš Vinař. The most probable annotation problem in hmms and its application to bioinformatics. *J Comput Syst Sci*, 73(7):1060 – 1077, 2007. ISSN 0022-0000. doi: DOI:10.1016/j.jcss.2007.03.011. URL http://www.sciencedirect.com/science/article/B6WJ0-4N85BD6-1/2/ffd66145030eb2db405304f3de9bf6c4. Bioinformatics III.

Broňa Brejová, Daniel G. Brown, and Tomáš Vinař. Advances in hidden Markov models for sequence annotation. In I. I. Mandoiu and A. Zelikovski, editors, *Bioinformatics Algorithms: Techniques and Applications*, chapter 4, pages 55–92. Wiley, 2008.

Gary D. Brushe, Robert E. Mahony, and John B. Moore. A soft output hybrid algorithm for ML/MAP sequence estimation. *IEEE Trans. Inform. Theory*, 44(7):3129–3140, 1998. ISSN 0018-9448.

Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology*, 268(1):78 – 94, 1997. ISSN 0022-2836. doi: DOI:10.1006/jmbi.1997.0951. URL http://www.sciencedirect.com/science/article/B6WK7-45VG77-9/2/0db813275493b6dd0d07e85a6276d801.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2; 0-387-40264-0. With Randal Douc’s contributions to Chapter 9 and Christian P. Robert’s to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

Luis E. Carvalho and Charles E. Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *P Natl Acad Sci USA*, 105(9):3209–3214, 2008. doi: 10.1073/pnas.0712329105. URL http://www.pnas.org/content/105/9/3209. abstract.

Christiane Cocozza-Thivent and Abdelkrim Bekkhoucha. Estimation in Pickard random fields and application to image processing. *Pattern Recognition*, 26(5):747–761, 1993.

R. Durbin, S. Eddy, Krogh A., and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

S. Eddy. What is a hidden Markov model? *Nature Biotechnology*, 22(10):1315 – 1316, 2004.

Yariv Ephraim and Neri Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6):1518–1569, June 2002. ISSN 0018-9448. Special issue on Shannon theory: perspective, trends, and applications.

Piero Fariselli, Pier Martelli, and Rita Casadio. A new decoding algorithm for hidden markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, 6(Suppl 4):S12, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-S4-S12.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
Peter J. Green and Sylvia Richardson. Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.*, 97(460):1055–1070, 2002. ISSN 0162-1459.

J. Hayes, T. Cover, and J. Riera. Optimal sequence detection and optimal symbol-by-symbol detection: Similar algorithms. *Communications, IEEE Transactions on*, 30(1):152–157, January 1982. ISSN 0090-6778. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1095391.

Ian Holmes and Richard Durbin. Dynamic programming alignment accuracy. *J Comput Biol*, 5(3):493–504, 1998.

X. Huang, Y. Ariki, and M. Jack. *Hidden Markov models for speech recognition*. Edinburgh University Press, Edinburgh, UK, 1990.

F. Jelinek. *Statistical methods for speech recognition*. The MIT Press, Cambridge, MA, USA, 2001.

F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64:532–556, April 1976.

G. Ji and J. Bilmes. Backoff model training using partially observed data: Application to dialog act tagging. In *Proc. Human Language Techn. Conf. NAACL, Main Conference*, pages 280–287, New York City, USA, 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N/N06/N06-1036.

Dhiraj Joshi, Jia Li, and James Z. Wang. A computationally efficient approach to the estimation of two- and three-dimensional hidden Markov models. *IEEE Trans. Image Process.*, 15(7):1871–1886, 2006. ISSN 1057-7149.

Lukas Käll, Anders Krogh, and Erik L. L. Sonnhammer. An hmm posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21(suppl_1):i251–257, 2005. doi: 10.1093/bioinformatics/bti1014. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/suppl%_1/i251.

Alexey Koloydenko and Jüri Lember. Infinite Viterbi alignments in the two state hidden Markov models. *Acta Comment. Univ. Tartu. Math.*, (12):109–124, 2008. ISSN 1406-2283. Proc. 8th Tartu Conf. Multivariate Statist. June 2007.

Timo Koski. *Hidden Markov models for bioinformatics*, volume 2 of *Computational Biology Series*. Kluwer Academic Publishers, Dordrecht, 2001. ISBN 1-4020-0135-5; 1-4020-0136-3.

A. Krogh. *Computational Methods in Molecular Biology*, chapter An Introduction to Hidden Markov Models for Biological Sequences. Elsevier Science, 1998.

Anders Krogh. Two methods for improving performance of an hmm and their application for gene finding. In *ISMB-97 Proceedings*, pages 179–186. AAAI, 1997.

Kristi Kuljus and Jüri Lember. Asymptotic Risks of Viterbi Segmentation. preprint, February 2010. URL http://arxiv.org/abs/1002.3509.
Generalized risk-based path inference in HMMs

Hans Künsch, Stuart Geman, and Athanasios Kehagias. Hidden Markov random fields. *Ann. Appl. Probab.*, 5(3):577–602, 1995. ISSN 1050-5164.

Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

Jüri Lember. On approximation of smoothing probabilities for hidden Markov models. preprint, October 2009. URL http://arxiv.org/abs/0910.4636.

Jüri Lember and Alexey Koloydenko. The Adjusted Viterbi training for hidden Markov models. *Bernoulli*, 14(1):180–206, February 2008. ISSN 1350-7265. doi: 10.3150/07-BEJ105.

Jüri Lember and Alexey Koloydenko. A constructive proof of the existence of Viterbi processes. *IEEE Trans. Inform. Theory*, 56(4):2017–2033, 2010.

Jia Li, Robert M. Gray, and Richard A. Olshen. Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models. *IEEE Trans. Inform. Theory*, 46(5):1826–1841, 2000. ISSN 0018-9448. Information-theoretic imaging.

Shu Lin and Daniel J. Costello Jr. *Error Control Coding: Fundamental and Applications*. Computer Applications in Electrical Engineering. Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1983. ISBN 0-13-283796-X.

William H. Majoros and Uwe Ohler. *Knowledge Discovery and Emergent Complexity in Bioinformatics*, volume 4366/2007 of *Lecture Notes in Computer Science*, chapter Advancing the State of the Art in Computational Gene Prediction, pages 81–106. Springer Berlin / Heidelberg, 2007. doi: 10.1007/978-3-540-71037-0_6. URL http://www.springerlink.com/content/X7H8126560375R25.

Jose L. Marroquin, Edgar Arce Santana, and Salvador Botello. Hidden markov measure field models for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(11):1380–1387, 2003. ISSN 0162-8828. doi: http://dx.doi.org/10.1109/TPAMI.2003.1240112.

Joshua Mason, Kathryn Watkins, Jason Eisner, and Adam Stubblefield. A natural language approach to automated cryptanalysis of two-time pads. In *CCS ’06: Proceedings of the 13th ACM conference on Computer and communications security*, pages 235–244, New York, NY, USA, 2006. ACM. ISBN 1-59593-518-5. doi: http://doi.acm.org/10.1145/1180405.1180435.

E. McDermott and T. Hazen. Minimum classification error training of landmark models for real-time continuous speech recognition. In *Proc. ICASSP*, 2004.

C. A. Megroty, D. M. Titterington, R. Reeves, and A. N. Pettitt. Variational bayes for estimating the parameters of a hidden potts model. *Statistics and Computing*, 19(3):329–340, 2009. ISSN 0960-3174. doi: http://dx.doi.org/10.1007/s11222-008-9095-6.

H. Ney, V. Steinbiss, R. Haeb-Umbach, and *et. al*. An overview of the Philips research system for large vocabulary continuous speech recognition. *Int. J. Pattern Recognit. Artif. Intell.*, 8(1):33–70, 1994.
F. Och and H. Ney. Improved statistical alignment models. In Proc. 38th Ann. Meet. Assoc. Comput. Linguist., pages 440 – 447. Assoc. Comput. Linguist., 2000.

M. Padmanabhan and M. Picheny. Large-vocabulary speech recognition algorithms. Computer, 35(4):42 – 50, 2002.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77(2):257–286, February 1989.

L. Rabiner and B. Juang. Fundamentals of speech recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

L. Rabiner, J. Wilpon, and B. Juang. A segmental K-means training procedure for connected word recognition. AT&T Tech. J., 64(3):21–40, 1986.

Patrick Robertson, Emmanuelle Villebrun, and Peter Hoeher. A comparison of optimal and sub-optimal map decoding algorithms operating in the log domain. In Communications, 1995. ICC ’95 Seattle, ’Gateway to Globalization’, 1995 IEEE International Conference on, volume 2, pages 1009–1013, 1995. doi: 10.1109/ICC.1995.524253.

Havard Rue. New loss functions in bayesian imaging. J Am Stat Assoc, 90(431):900–908, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291324.

I. Shu, L. Hetherington, and J. Glass. Baum-Welch training for segment-based speech recognition. In Proc. IEEE ASRU Workshop, pages 43–48, St. Thomas, U. S. Virgin Islands, http://groups.csail.mit.edu/sls/publications/2003/ASRU_Shu.pdf, 2003.

V. Steinbiss, H. Ney, X. Aubert, and et. al. The Philips research system for continuous-speech recognition. Philips J. Res., 49:317–352, 1995.

N. Ström, L. Hetherington, T. Hazen, E. Sandness, and J. Glass. Acoustic modeling improvements in a segment-based speech recognizer. In Proc. IEEE ASRU Workshop, pages 139–142, 1999.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inform. Theory, 13(2):260–269, April 1967.

Gerhard Winkler. Image analysis, random fields and Markov chain Monte Carlo methods, volume 27 of Applications of Mathematics (New York). Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-44213-8. A mathematical introduction, With 1 CD-ROM (Windows), Stochastic Modelling and Applied Probability.