Infinite Viterbi alignment

J. Lember

University of Tartu, Estonia

with

A. Koloydenko

University of Nottingham, UK

M. Käärik

University of Tartu, Estonia

Hidden Markov Model (HMM)

Y – irreducible, aperiodic Markov Chain with finite state space S, $|S| = k$, transition matrix (p_{ij}).

Y is sometimes called as the regime.

To each state $l \in S$ corresponds an emission distribution P_l with densities f_l w.r.t. some reference measure λ on $\mathcal{B}(\mathbb{R}^d)$.

HMM:

To any realization $y_1, y_2, \ldots$ of Y corresponds a sequence of independent random variables $X_1, X_2, \ldots$, where $X_n \sim P_{y_n}$.

HMM's are used (among others):

Speech recognition:

Acoustic-phonetic modelling (complex)

Computational molecular biology:

1. DNA-sequence alignment

Y has 3 states: mach, deletion, insertion

2. Modelling DNA regions

3. ....

Assume that the first n elements $x_1^n := x_1, \ldots, x_n$ of a realization of X are observed.

The corresponding outcomes of Y, $y_1, \ldots, y_n$ are not observed (Y is hidden).

One possible way to estimate hidden $y_1, \ldots, y_n$ is to use the state sequence $q_1^n := q_1, \ldots, q_n \in S^n$ with maximum likelihood. This sequence is called (Viterbi) alignment.

To every observation-sequence corresponds a Viterbi alignment (ignore ties), so we consider a mapping or coding

$$v : \mathbb{R}^n \mapsto S^n, \quad v(x_1^n) = \arg\max_{q_1^n \in S^n} p(q_1^n | x_1^n).$$

In general, it is conceptionally wrong to make the statistical inferences using the max. likelihood sequence as the substitution of the truth. However, when you know the differences between the max. likelihood sequence and truth, and when you take those differences into account, you can still ripe the benefit from it.

The underlying MC $Y_1, Y_2, \ldots$ is very well-studied process.
The properties of $X_1, X_2, \ldots$ can be studied as well.
What are the (long run) properties of Viterbi alignment?

Note: adding one more observation, x_{n+1} can, in principle, change the whole alignment. Formally, if $v(x_1, \ldots, x_n) = (v_1, \ldots, v_n)$ and $v(x_1, \ldots, x_{n+1}) = (w_1, \ldots, w_n, w_{n+1})$, then it can be so that $w_i \neq v_i$ for every $i = 1, \ldots, n$. What about the asymptotics in this case?

1. Is there anything like infinite alignment $v(X_1, X_2, \ldots)$?
2. If yes, what are the properties of the process $v(X_1, X_2, \ldots)$?

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations:

$$\underline{\quad x_1 \quad x_2 \quad x_3 \quad {\color{red}a} \quad x_5 \quad x_6 \quad x_7 \quad x_8 \quad {\color{red}a} \quad x_{10} \quad x_{11} \quad x_{12} \quad x_{13} \quad x_{14} \quad {\color{red}a} \quad x_{17} \quad x_{18}}$$

What can we say about the Viterbi (max likelihood) alignment?

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations:

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
?	?	?	1	?	?	?	?	1	?	?	?	?	?	1	?	?

The a's correspond to the state 1. Then use the optimality principle.

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations:

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
q_1	q_2	q_3	1	?	?	?	?	1	?	?	?	?	?	1	?	?

The observations to first a can be used to determine the first piece.

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations:

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
q_1	q_2	q_3	1	q_5	q_6	q_7	q_8	1	?	?	?	?	?	1	?	?

The observations from first to second a can be used to determine the second piece.

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

Suppose we have observations:

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
q_1	q_2	q_3	1	q_4	q_5	q_6	q_7	1	q_{10}	q_{11}	q_{12}	q_{13}	q_{14}	1	?	?

The observations from second to third a can be used to determine the second piece.

An easy but yet insightful special case.

Suppose there $\exists$ set $A : P_1(A) > 0$ but $P_2(A) = \cdots = P_K(A) = 0$. To emit an observation from A, Y has to be in the state 1, a.s.

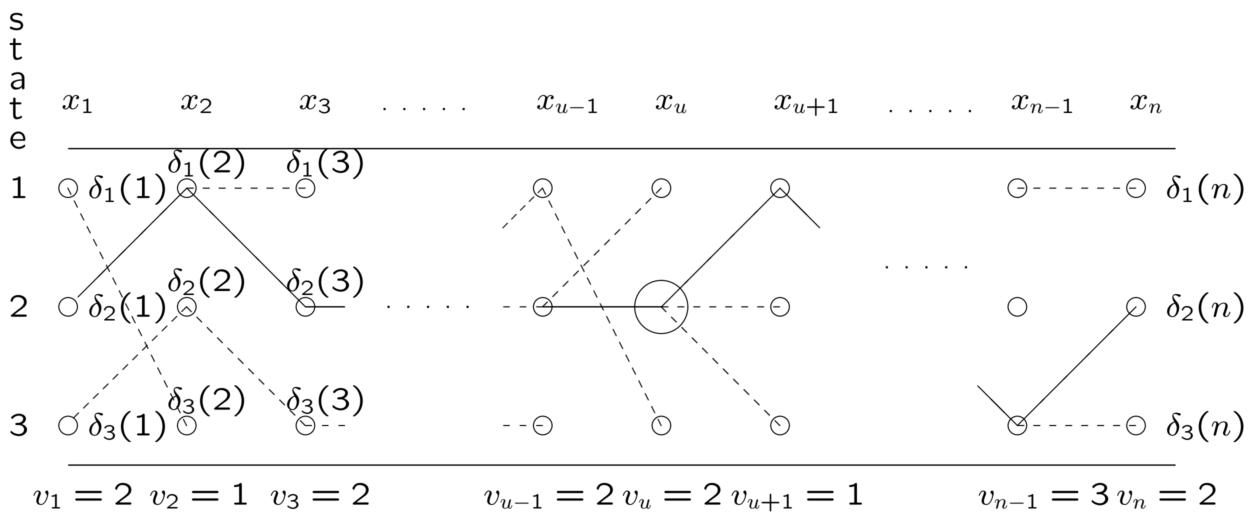
Suppose we have observations:

x_1	x_2	x_3	a	x_5	x_6	x_7	x_8	a	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	a	x_{16}	x_{17}
q_1	q_2	q_3	1	q_4	q_5	q_6	q_7	1	q_{10}	q_{11}	q_{12}	q_{13}	q_{14}	1	q_{16}	q_{17}

Finally the last piece. So, the whole alignment can be constructed piecewise. The process X is ergodic: every realization of the process has infinitely many a's. Hence, the piecewise alignment can be extended to infinity – we have an infinite (piecewise) alignment!

How to generalize the concept of a? The answer lies in the Viterbi aligorithm – the dynamic programming algorithm to find the (max-likelihood) alignment.

$$\delta_l(u) := \max_{q_1,\ldots,q_{u-1}} p(q_1,\ldots,q_{u-1},q_n = l; x_1,\ldots,x_u).$$

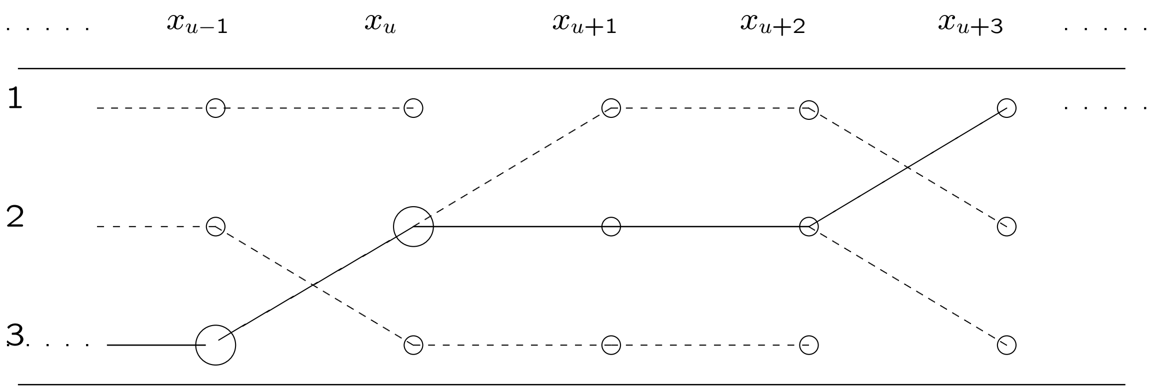


Let $x_1,\ldots,x_u$ be the first u observations. We call x_u an l-node if

$$\delta_l(u)p_{lj} \geq \delta_i(u)p_{ij}, \quad \forall i,j \in S. \tag{1}$$

Restriction of the concept of node: in order an l-node to exists, it is necessary $p_{lj} > 0 \ \forall j$. But HMM can have a 0 in every row.

Generalization of the node – r-order node:



2-nd order 2-node

A node – 0-order node.

In general, to understand that x_u is an r-order node, one has to look at the observations

$$x_1, x_2, \ldots, x_u, x_{u+1}, \ldots, x_{u+r}.$$

On the other hand, a was a node independently of the previous observations. Could we have something like that as well?

A barrier is a block of observations that contains a (r-order) node independently of the observations before (and after) it.

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, \ldots$$

The block $x_6, x_7, x_8, x_9, x_{10}, x_{11}$ is a barrier of length 6.

a – barrier of length 1.

Thm (Koloydenko, L.; simplified version)

Assume:

1) for each state $l \in S$

$$P_l\Big(x : f_l(x) \max_j\{p_{jl}\} > \max_{i,i\neq l}\{f_i(x) \max_j\{p_{ji}\}\}\Big) > 0.$$

2) the supports of f_l have non-empty intersection;

Then there exists:

1) a set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ such that every vector $(x_1, \ldots, x_M) \in \mathcal{X}$ is a barrier with x_{M-r} being the corresponding r-order l-node;

2) a M-tuple of states $(y_1, \ldots, y_M) \in S^M$ such that $y_{M-r} = l$ and

$$\mathbf{P}\Big((X_1, \ldots, X_M) \in \mathcal{X} \Big| Y_1 = y_1, \ldots, Y_M = y_M\Big) > 0$$
$$\mathbf{P}(Y_1 = y_1, \ldots, Y_M = y_M) > 0.$$

Thm. applies for a large class of HMM's. Essentially generalizes the earlier results by Caliebe and Rösler (2002).

1) there is a positive probability that the process X generates a barrier from $\mathcal{X}$ in observations (observable);
2) there is a positive probability that the process X generates a barrier from $\mathcal{X}$ in observations <u>and</u> the underlying MC Y generates $(y_1, \ldots, y_M)$ at the same time .

By <u>ergodic argument</u> that means:
1) almost every realization of X has infinitely many barriers (and, hence, nodes);
2) almost every realization of X has infinitely many barriers generated by the block $(y_1, \ldots, y_M)$.

X : – – – – – – – – * – – – – – – – – – * – – – – – – – – * – – – – –

Y : – – – * – – – – – – – – – – * – – – – – – – – – – – * – – – – –

τ

By 1), for almost every realization we can define piecewise infinite alignment. Formally, we have a map (coding)

$$v : \mathbb{R}^{\infty} \mapsto S^{\infty}.$$

The process $V := v(X)$ is called the alignment process. So

$$V_1, V_2, \ldots = v(X_1, X_2, \ldots)$$

From 2), it follows:

a) the process X is regenerative with respect to τ;

b) the process V is regenerative with respect to τ;

c)the process (X, V) is regenerative with respect to τ;

$V :$ − − − ∗ − − − − ∗ − − − − − − − − ∗ − − − − − − − − − ∗ − − − − −

$X :$ − − − ∗ − − − − ∗ − − − − − − − − − ∗ − − − − − − − − − ∗ − − − − −

$Y :$ − − − ∗ − − − − − − − − − ∗ − − − − − − − − − − − − − ∗ − − − − −

τ_i $\qquad\qquad\qquad\qquad$ τ_{i+1}

V is not stationary, but can be easily stationarized by embedding into double-sided process. Then V as well as (X, V) ergodic.

Regenerativity (ergodicity) immediately gives SLLN type of theorems.

Example: States of Y: 1 2. Observations $x_1, \ldots, x_n$. Subsamples based on Viterbi alignment $P_l^n, l \in S$.

X :	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
v :	1	2	2	1	2	1	2	2	2	1

The subsamples (empirical measures) are

$x_1 \quad x_4 \quad x_6 \quad x_{10}$ $\qquad P_1^{10}$

$x_2 \quad x_3 \quad x_5 \quad x_7 \quad x_8 \quad x_9$ $\qquad P_2^{10}$

Using the regenerativity (or ergodicity) of (X, V), it easily follows that there exists probability measures Q_l such that a.s.

$$P_l^n \Rightarrow Q_l, \quad \forall l$$

Important: $\boxed{Q_l \text{ might be very different from } P_l}$

This difference is not taken into account in Viterbi training.

This difference is taken into account in adjusted Viterbi training.