Equivalence of Linear Boltzmann Chains and Hidden Markov Models

David J.C. MacKay
Cavendish Laboratory
Madingley Road
Cambridge CB3 0HE
United Kingdom
mackay@mrao.cam.ac.uk

March 31, 1995. To appear in *Neural Computation*.

Abstract

Several authors have studied the relationship between hidden Markov models and 'Boltzmann chains' with a linear or 'time-sliced' architecture. Boltzmann chains model sequences of states by defining state-state transition energies instead of probabilities. In this note I demonstrate that, under the simple condition that the state sequence has a mandatory end state, the probability distribution assigned by a strictly linear Boltzmann chain is identical to that assigned by a hidden Markov model.

Several authors have made a link between hidden Markov models for time series and energy-based models (Luttrell 1989, Williams 1990, Saul and Jordan 1995). Saul and Jordan (1995) discuss a linear Boltzmann chain model with state-state transition energies $A_{ii'}$ (going from state i to state i') and symbol emission energies B_{ij}, under which the probability of an entire state $\{i_l, j_l\}_1^L$ given the length of the sequence L, is:

$$P(\{i_l, j_l\}_1^L | \Pi, \mathbf{A}, \mathbf{B}, L, \mathcal{H}_{\mathrm{BC}}) = \frac{1}{Z(\Pi, \mathbf{A}, \mathbf{B}, L)} \exp\left(-\Pi_{i_1} - \sum_{l=1}^{L-1} A_{i_l i_{l+1}} - \sum_{l=1}^{L} B_{i_l j_l} \right) \tag{1}$$

where $Z(\Pi, \mathbf{A}, \mathbf{B}, L)$ is the obvious normalizing constant. Here the symbol i runs over n discrete 'hidden' states, and j runs over m visible states. In contrast, a hidden Markov model (HMM) assigns a probability distribution of the form:

$$P(\{i_l, j_l\}_1^L | \pi, \mathbf{a}, \mathbf{b}, L, \mathcal{H}_{\mathrm{HMM}}) = \pi_{i_1} \prod_{l=1}^{L-1} a_{i_l i_{l+1}} \prod_{l=1}^{L} b_{i_l j_l} \tag{2}$$

where π_i is a prior probability vector for the initial state, $a_{ii'}$ is a transition probability matrix, and b_{ij} is a matrix of emission probabilities satisfying respectively:

$$\sum_i \pi_i = 1, \quad \sum_{i'} a_{ii'} = 1 \ \ \forall i, \ \text{ and } \ \sum_j b_{ij} = 1 \ \ \forall i. \tag{3}$$

Here again, the symbol i runs over an alphabet of n hidden states, and j runs over m visible states.

Whilst any HMM can be written as a linear Boltzmann chain by setting $\exp(-A_{ii'}) = a_{ii'}$, $\exp(-B_{ij}) = b_{ij}$ and $\exp(-\Pi_i) = \pi_i$, not all linear Boltzmann chains can be represented as HMMs (Saul and Jordan 1995). However, the difference between the two models is minimal. To be precise, *if the final hidden state i_L of a linear Boltzmann chain is constrained to be a particular end state, then the distribution over sequences is identical to that of a hidden Markov model.*

PROOF

Start from the distribution (1) and consider the quantity in the exponent. The probability distribution over states $\{i_l, j_l\}_1^L$ is unchanged if we subtract arbitrary constants μ, ν from this exponent. The distribution will also be unaffected if we add arbitrary terms β_{i_l} to every appearance of $B_{i_l j_l}$, provided we also subtract β_{i_l} from every term $A_{i_l i_{l+1}}$. And we may similarly add $\alpha_{i_{l+1}}$ to every term $A_{i_l i_{l+1}}$ if we also subtract $\alpha_{i_{l+1}}$ from the following term $A_{i_{l+1} i_{l+2}}$. The probability distribution may therefore be rewritten unchanged (except for the normalizing constant) as:

$$P(\{i_l, j_l\}_1^L | \Pi, \mathbf{A}, \mathbf{B}, L, \mathcal{H}_{\mathrm{BC}}) \propto$$
$$\exp\left(-(\Pi_{i_1} + \alpha_{i_1} + \mu) - \sum_{l=1}^{L-1} (-\alpha_{i_l} - \beta_{i_l} + A_{i_l i_{l+1}} + \alpha_{i_{l+1}} + \nu) - \sum_{l=1}^{L} (B_{i_l j_l} + \beta_{i_l}) + \alpha_{i_L} + \beta_{i_L} \right), \quad (4)$$

where $\mu, \nu, \{\alpha_i, \beta_i\}$ are arbitrary quantities. This probability distribution has the form of an HMM (equation 2) if

1. the quantities $\pi_i \equiv \exp(-(\Pi_i + \alpha_i + \mu))$, $a_{ii'} \equiv \exp(\alpha_i + \beta_i - A_{ii'} - \alpha_{i'} - \nu)$ and $b_{ij} \equiv \exp(-(B_{ij} + \beta_i))$ satisfy the normalization conditions (3).

2. the trailing term $\alpha_{i_L} + \beta_{i_L}$ can be treated as a constant, which holds if we assume that i_L is fixed to a particular end state $i_L = n$, say (a commonly applied constraint in the HMM literature).

Does a solution over $\mu, \nu, \{\alpha_i, \beta_i\}$ of the normalization conditions (3) exist? Trivially, we find for β_i:

$$\sum_j b_{ij} = 1 \;\Rightarrow\; \beta_i = \log\left[\sum_j \exp(-B_{ij}) \right].$$

The normalization condition that $\{\alpha_i\}$ and ν must satisfy is:

$$\sum_{i'} \exp(\alpha_i + \beta_i - A_{ii'} - \alpha_{i'} - \nu) = 1 \;\; \forall i$$

Rearranging, we obtain:

$$\sum_{i'} [\exp(\beta_i - A_{ii'}))] [\exp(-\alpha_{i'})] = \exp(\nu) [\exp(-\alpha_i)] \quad \forall i$$

which can be recognised as an eigenvector/eigenvalue equation for the matrix $M_{ii'} \equiv [\exp(\beta_i - A_{ii'})]$, with $\exp(\nu)$ being the eigenvalue and $[\exp(-\alpha_i)]$ being the eigenvector. This eigenproblem has a solution, by the Perron-Frobenius theorem (Seneta 1973, p.1), which states that a positive matrix (*i.e.*, one in which every element $M_{ii'}$ is positive) has a positive eigenvector with positive eigenvalue. A solution for $\{\alpha_i\}$ and ν therefore exists. Finally μ is given by:

$$\mu = \log\left[\sum_i \exp(-(\Pi_i + \alpha_i)) \right].$$

This completes the proof.

The linear Boltzmann chain therefore can only differ from an HMM in having a pseudo-prior over its final state as well as a pseudo-prior over its initial state. However the equivalence of linear Boltzmann chains to HMMs may prove fruitful in stimulating the development of new optimization methods for these models. And it may be found that Saul and Jordan's generalizations to Boltzmann chains with more complex architectures provide useful new modelling capabilities.

The Boltzmann chain, and its relationship to HMMs, have also been studied by Luttrell (1989) who calls it the 'Gibbs machine', and by Williams (1990), who calls it a 'Boltzmann machine with a time-sliced architecture and Potts units'. Luttrell discusses an alternative optimization algorithm to the decimation method suggested by Saul and Jordan, and notes that the Gibbs machine is only an improvement on the HMM when generalized to architectures with loops and other non-tree structures. Williams also shows how to translate an HMM into a Boltzmann machine and notes that a generalized Boltzmann machine with a 'componential' structure (similar to the 'coupled parallel Boltzmann chains' of Saul and Jordan) has greater representational power than a single HMM of the same size.

Acknowledgements

I thank Radford Neal and Chris Williams for comments on the manuscript. This work is supported by a Royal Society research fellowship.

References

Luttrell, S. P. (1989). The Gibbs machine applied to hidden Markov model problems. part 1: Basic theory, *Technical Report 99*, SP4 division, RSRE, Malvern, U.K.

Saul, L. and Jordan, M. (1995). Boltzmann chains and hidden Markov models, *Advances in Neural Information Processing Systems 7*, M.I.T. Press (in press).

Seneta (1973). *Non-negative Matrices*, Wiley, New York.

Williams, C. K. I. (1990). *Using deterministic Boltzmann machines to discriminate temporally distorted strings*, Master's thesis, Dept. of Computer Science, Univ. of Toronto; see also Williams, C. K. I. and Hinton, G. E. (1990): "Mean field networks that learn to discriminate temporally distorted strings" Proc. of the 1990 Connectionist Models Summer School eds. Touretzky, D. S., Elman, J. L., Sejnowski, T. S. and Hinton, G. E. San Mateo, CA, Morgan Kaufmann.