IDIAP RESEARCH REPORT

STATISTICAL MODELS FOR HMM/ANN HYBRIDS

Philip N. Garner David Imseng

Idiap-RR-11-2013

APRIL 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Statistical models for HMM/ANN hybrids

Philip N. Garner, David Imseng

March 24, 2013

Abstract

We present a theoretical investigation into the use of normalised artificial neural network (ANN) outputs in the context of hidden Markov models (HMMs). The work is motivated by the pursuit of a more theoretically rigorous understanding of the Kullback-Liebler (KL)-HMM. Two possible models are considered based respectively on the HMM states storing categorical distributions and Dirichlet distributions. Training and recognition algorithms are derived, and possible relationships with KL-HMM are briefly discussed.

Contents

1 MLPs in ASR **2**
- 1.1 The softmax . . . 2
- 1.2 Tandem . . . 2
- 1.3 KL-HMM . . . 2

2 Categorical model **3**
- 2.1 Model formulation . . . 3
- 2.2 Priors . . . 4
- 2.3 Parameter estimation . . . 5
- 2.4 Model evaluation . . . 6
- 2.5 Corollaries . . . 6
  - 2.5.1 Relationship with conventional hybrids . . . 6
  - 2.5.2 Relationship with scalar product . . . 7
  - 2.5.3 Relationship with KL-HMM . . . 7

3 Dirichlet model **8**
- 3.1 Model formulation . . . 8
- 3.2 Priors . . . 8
- 3.3 Parameter estimation . . . 8
- 3.4 Model evaluation . . . 9
- 3.5 Corollary . . . 9

4 Conclusions **10**

5 Acknowledgements **10**

A Kullback Leibler **12**

1 MLPs in ASR

1.1 The softmax

A common approach in multi-layer perceptron (MLP) based pattern recognition is to use a "softmax" (sometimes called multiple logistic) activation function at the output layer (Bridle, 1990). The softmax activation function results in an observation, $\mathbf{o}_t$, at time t:

$$\mathbf{o}_t = (o_{t,1}, o_{t,2}, \ldots, o_{t,P})^\mathsf{T}, \qquad \sum_{i=1}^{P} o_{t,i} = 1, \qquad o_{t,i} \geqslant 0. \tag{1}$$

Notice that the output of softmax has the form of a categorical probability distribution, so it can be used in certain operations that are only well-defined if that is so. The outputs may be more-or-less useful estimates of posterior probabilities, depending on the way it is trained. Generally, however, the softmax is taken as an estimate of posterior probabilities. In particular, when incorporated into an HMM as an HMM/ANN hybrid, each state of the HMM is associated with a specific one of the labels represented by the output of the softmax (Bourlard and Morgan, 1994; Morgan and Bourlard, 1995a,b). It is conventional to compute a number to use as the HMM output likelihood for a specific state by dividing the appropriate output from the softmax by the prior probability for the class.

1.2 Tandem

In the tandem model of Hermansky et al. (2000), the softmax outputs are passed through a logarithm that allows them to be treated as Gaussian random variates. This form is then modelled using a normal HMM/GMM system.

Tandem is not a rigorous statistical model. However, it is known to perform very well in speech recognition, demonstrating that the softmax representation carries useful information.

1.3 KL-HMM

More recently, the Kullback-Leibler (KL) HMM was introduced by Aradilla et al. (2007); Aradilla Zapata (2008). In the KL-HMM, the output distribution of each state of the HMM is replaced by a multinomial distribution (constrained to be a categorical distribution) that can, in principle, emit any of the labels represented by the softmax. Rather than evaluating an output likelihood, a score is calculated being the KL divergence between the multinomial stored in the state, and the one represented by the softmax output.

In this report, we take the stance that the KL-HMM derives its performance from its ability to separate the state labels and the softmax labels, rather than the KL formulation. We derive two *generative* models that use broadly the same parameterisation, but have a different statistical interpretation.

One goal is to understand what generative statistical assumption could lead to the KL formulation, or something similar to it.

2 Categorical model

2.1 Model formulation

Assume that each HMM state, $q^d, d \in \{1, 2, \ldots, D\}$, stores a categorical distribution; that is, a multinomial distribution where only one sample is drawn.

$$p(\rho \mid \theta_d) = \theta_{d,\rho}, \qquad \sum_{\rho=1}^{P} \theta_{d,\rho} = 1. \tag{2}$$

The softmax outputs can also interpreted as categorical distributions,

$$p(\rho \mid \mathbf{o}_t) = o_{t,\rho}, \qquad \sum_{\rho=1}^{P} o_{t,\rho} = 1, \tag{3}$$

but note that the softmax do not actually generate phones. Rather, they encode the belief that a phone was emitted by the state. Given the deterministic relationship between $\mathbf{o}_t$ and $\mathbf{x}_t$, the values $o_{t,\rho}$ can be referred to as either $p(\rho \mid \mathbf{x}_t)$ or $p(\rho \mid \mathbf{o}_t)$. In either case, they have the form of posterior probabilities.

q^d	Identifier for the state
q_t	State at time t
$\mathbf{q}$	The state sequence $q_1, q_2, \ldots, q_T$
θ_d	The categorical distribution on state d
$\boldsymbol{\alpha}_d$	The hyperparameter of the Dirichlet prior for θ_d
ρ_t	The class generated by the state at time t
$\boldsymbol{\rho}$	The sequence of classes generated by the state sequence
$\mathbf{o}_t$	The softmax output at time t
$\mathbf{O}$	The set (matrix) of softmax outputs
$\mathbf{x}_t$	The observed audio representation at time t
$\mathbf{X}$	The set (matrix) of observed audio representations
$\boldsymbol{\omega}$	The prior on the softmax outputs

Table 1: Notation.

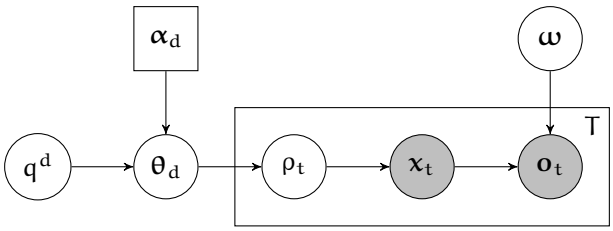

Figure 1: Generative diagram in plate notation for categorical model before training. Arrows mean "influences", rectangles are assumed fixed, circles are state variables. Shading means observable. The large box means "repeat T times".

Figure 1 shows the model in plate notation[1]. A Dirichlet prior can be assumed for θ_d:

$$p(\theta_d \mid \boldsymbol{\alpha}_d) = \frac{1}{B(\boldsymbol{\alpha}_d)} \theta_{d,1}^{\alpha_{d,1}-1} \theta_{d,2}^{\alpha_{d,2}-1} \ldots \theta_{d,P}^{\alpha_{d,P}-1}, \tag{4}$$

where $\alpha_{d,\rho} = 1$ represents a flat prior.

[1] http://en.wikipedia.org/wiki/Plate_notation

The full joint distribution can be read off the inference diagram as:

$$p(\theta_d, \rho, \mathbf{O}) = p(\theta_d \mid \alpha_d) \prod_{t=1}^{T} p(\rho_t \mid \theta_d)\, p(\mathbf{o}_t \mid \rho_t, \omega) \tag{5}$$

$$= p(\theta_d \mid \alpha_d) \prod_{t=1}^{T} p(\rho_t \mid \theta_d)\, \frac{p(\mathbf{o}_t \mid \omega)\, p(\rho_t \mid \mathbf{o}_t)}{p(\rho_t \mid \omega)} \tag{6}$$

$$= p(\theta_d \mid \alpha_d) \prod_{t=1}^{T} p(\mathbf{o}_t \mid \omega)\, \frac{1}{p(\rho_t \mid \omega)} p(\rho_t \mid \theta_d)\, p(\rho_t \mid \mathbf{o}_t), \tag{7}$$

where

$$p(\rho_t \mid \omega) = \int d\mathbf{o}_t\, p(\mathbf{o}_t \mid \omega)\, p(\rho_t \mid \mathbf{o}_t), \tag{8}$$

and the integral is over the unit hypercube.

It is tempting to drop the term $p(\mathbf{o}_t \mid \omega)$ because it is observed. However, doing so would reduce the integral to simply $p(\rho_t \mid \mathbf{o}_t)$ which would then cancel. In fact, this is a reminder that $\mathbf{o}_t$ is simply an estimate of $p(\rho_t \mid \theta_d)$, and has a distribution.

2.2 Priors

The prior term, $p(\rho_t)$, is normally calculated directly from the data. However, in this framework it is more involved. Assume a Dirichlet prior:

$$p(\mathbf{o}_t \mid \omega_t) = \frac{1}{B(\omega_t)} o_{t,1}^{\omega_{t,1}-1} o_{t,2}^{\omega_{t,2}-1} \dots o_{t,P}^{\omega_{t,P}-1}. \tag{9}$$

Equation 8 then evaluates to a beta function with a single hyperparameter incremented by 1 that in turn cancels with the one in the Dirichlet:

$$\int d\mathbf{o}_t\, p(\mathbf{o}_t \mid \omega_t)\, p(\rho_t \mid \mathbf{o}_t) = \int d\mathbf{o}_t\, \frac{1}{B(\omega_t)} o_{t,1}^{\omega_{t,1}-1} o_{t,2}^{\omega_{t,2}-1} \dots o_{t,P}^{\omega_{t,P}-1} o_{t,\rho_t} \tag{10}$$

$$= \frac{\Gamma(\omega_{t,1} + \omega_{t,2} + \dots + \omega_{t,P})}{\Gamma(1 + \omega_{t,1} + \omega_{t,2} + \dots + \omega_{t,P})} \frac{\Gamma(\omega_{t,\rho_t} + 1)}{\Gamma(\omega_{t,\rho_t})} \tag{11}$$

$$p(\rho_t \mid \omega) = \frac{\omega_{t,\rho_t}}{\omega_{t,1} + \omega_{t,2} + \dots + \omega_{t,P}}. \tag{12}$$

As for the value of ω, we can attempt to come up with a maximum likelihood value:

$$p(\omega \mid \mathbf{O}) \propto \prod_{t=1}^{T} p(\mathbf{o}_t \mid \omega) \tag{13}$$

$$= \prod_{t=1}^{T} \frac{1}{B(\omega)} o_{t,1}^{\omega_1} o_{t,2}^{\omega_2} \dots o_{t,P}^{\omega_P}. \tag{14}$$

Differentiating w.r.t. ω and equating to zero has no closed form solution, even after logarithm. However, a solution is given in section 3.3.

It is possible, however, to make an informed approximation. Consider $\mathbf{o}_t$ being a generative (categorical) distribution generating labelled training data where the labels are the same as the classes produced by the softmax.

$$p(\mathbf{o}_t \mid l) = \frac{1}{p(l)} p(\mathbf{o}_t) \prod_{t=1}^{T} p(l_t \mid \mathbf{o}_t) \tag{15}$$

$$= \frac{1}{p(l)} o_{t,1}^{n_1} o_{t,2}^{n_2} \dots o_{t,P}^{n_P} \tag{16}$$

where there are n_ρ frames of class ρ_t in the training data. So, we have a posterior of sorts on $\mathbf{o}_t$, and by inspection $\omega_t \approx n_t$.

In practice, the Dirichlet of equation 9 with $\omega_t = n_t$ represents very strong prior information (N is large). This will cause it to dominate any posterior calculation. It is possible (and usual) to reduce the effect of such a prior by scaling the counts; that is, write

$$\omega_\rho = \frac{n_\rho}{N}\nu, \tag{17}$$

where ν is an arbitrary scale factor being a hypothetical number of samples that the prior is taken to represent. Notice that $p(\rho_t \mid \boldsymbol{\omega})$ is unchanged by the re-parameterisation.

2.3 Parameter estimation

To train the model, a-priori we expect the solution to be similar to that for Gaussian mixture weights. we assume Viterbi training here, i.e., we don't consider all state sequences. However, we need EM for the summation inside a product. The auxiliary function is

$$Q(\theta'_d, \theta_d) = \sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{O}) \log\left(p\left(\theta'_d, \boldsymbol{\rho}, \mathbf{O}\right)\right), \tag{18}$$

where the summation is over all possible combinations of sequence $\boldsymbol{\rho}$. Combining the auxiliary function with a Lagrange multiplier, λ, to enforce a sum to one constraint and substituting in the full joint of equation 7,

$$\frac{\partial}{\partial \theta'_{d,\rho}}\left[\sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{O}) \log\left(p\left(\theta'_d, \boldsymbol{\rho}, \mathbf{O}\right)\right) - \lambda\left(1 - \sum_{\rho=1}^{P} \theta'_{d,\rho}\right)\right] = 0 \tag{19}$$

$$\frac{\partial}{\partial \theta'_{d,\rho}}\left[\sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{O})\left(\log p\left(\theta'_d \mid \boldsymbol{\alpha}_d\right) + \sum_{t=1}^{T} \log p\left(\rho_t \mid \theta'_d\right) + C\right) - \lambda\left(1 - \sum_{\rho=1}^{P} \theta'_{d,\rho}\right)\right] = 0 \tag{20}$$

$$\frac{\partial}{\partial \theta'_{d,\rho}}\left[\sum_{\rho=1}^{P} \log \theta'^{\alpha_{d,\rho}-1}_{d,\rho} + \sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{O})\left(\sum_{t=1}^{T} \log \theta'_{d,\rho} + C\right) - \lambda\left(1 - \sum_{\rho=1}^{P} \theta'_{d,\rho}\right)\right] = 0 \tag{21}$$

where the constant C represents the terms independent of θ'_d. Differentiating,

$$\frac{\alpha_{d,\rho} - 1}{\theta'_{d,\rho}} + \sum_{\boldsymbol{\rho}} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{O}) \sum_{t:\rho_t=\rho} \frac{1}{\theta'_{d,\rho}} + \lambda = 0 \tag{22}$$

$$\frac{\alpha_{d,\rho} - 1}{\theta'_{d,\rho}} + \frac{1}{\theta'_{d,\rho}} \sum_{t=1}^{T} \sum_{\boldsymbol{\rho}:\rho_t=\rho} p(\boldsymbol{\rho} \mid \theta_d, \mathbf{o}_t) + \lambda = 0 \tag{23}$$

$$1 - \alpha_{d,\rho} - \sum_{t=1}^{T} p(\rho_t \mid \theta_d, \mathbf{o}_t) = \lambda \theta'_{d,\rho}. \tag{24}$$

$$\tag{25}$$

Summing over ρ gets rid of $\theta'_{d,\rho}$. So,

$$\lambda = \sum_{\rho=1}^{P}\left(1 - \alpha_{d,\rho} - \sum_{t=1}^{T} p(\rho_t \mid \theta_d, \mathbf{o}_t)\right), \tag{26}$$

and

$$\theta'_{d,\rho} = \frac{1}{\lambda}\left(1 - \alpha_{d,\rho} - \sum_{t=1}^{T} p(\rho_t \mid \theta_d, \mathbf{o}_t)\right). \tag{27}$$

The expression for the latent class follows directly from the full joint:

$$p(\rho_t \mid \theta_d, \mathbf{o}_t) = \frac{1}{p(\mathbf{o}_t, \theta_d)} p(\rho_t, \theta_d, \mathbf{o}_t) \tag{28}$$

$$= \frac{1}{p(\rho_t \mid \boldsymbol{\omega})} p(\rho_t \mid \theta_d)\, p(\rho_t \mid \mathbf{o}_t). \tag{29}$$

Substituting,

$$\theta'_{d,\rho} = \frac{1}{\sum_{\rho=1}^{P}\left(1-\alpha_{d,\rho}-\sum_{t=1}^{T}\frac{1}{p(\rho_t\mid\omega)}\theta_{d,\rho}o_{t,\rho}\right)}\left(1-\alpha_{d,\rho}-\sum_{t=1}^{T}\frac{1}{p(\rho_t\mid\omega)}\theta_{d,\rho}o_{t,\rho}\right) \tag{30}$$

$$= \frac{\alpha_{d,\rho}-1+\frac{1}{p(\rho\mid\omega)}\theta_{d,\rho}\sum_{t=1}^{T}o_{t,\rho}}{\sum_{\rho=1}^{P}\left(\alpha_{d,\rho}-1+\frac{1}{p(\rho\mid\omega)}\theta_{d,\rho}\sum_{t=1}^{T}o_{t,\rho}\right)}. \tag{31}$$

In principle, it is possible to set all $\alpha_{d,\rho}$ equal. However, it does make sense to set it more like

$$\alpha_{d,\rho} = \begin{cases} 0.2 & \text{if } d=\rho, \\ 0.1 & \text{otherwise.} \end{cases} \tag{32}$$

2.4 Model evaluation

In recognition (decoding), we are typically interested in maximising the probability of a state sequence, $\mathbf{q}$, given the parameters and observation. This expands via Bayes as the quantity

$$\mathbf{q} = \underset{\mathbf{q}}{\operatorname{argmax}}\, p(\mathbf{q}\mid\mathbf{X}); \quad p(\mathbf{q}\mid\mathbf{X}) \propto p(\mathbf{q})\,p(\mathbf{X}\mid\mathbf{q}). \tag{33}$$

If $q_t = d$, this can be written in terms of θ_d and $\mathbf{o}_t$ given the deterministic relationships above. The actual class, ρ_t, that is emitted is not known and must be marginalised. Writing in terms of the full joint and substituting equation 7:

$$p(\mathbf{X}\mid\mathbf{q}) = p(\mathbf{O}\mid\theta_d) = \frac{p(\theta_d,\mathbf{O})}{p(\theta_d)} \tag{34}$$

$$= \prod_{t=1}^{T} p(\mathbf{o}_t\mid\omega)\sum_{\rho_t=1}^{P}\frac{1}{p(\rho_t\mid\omega)}p(\rho_t\mid\theta_d)\,p(\rho_t\mid\mathbf{o}_t). \tag{35}$$

Or, written as a logarithm over T frames, we have:

$$\log p(\{\mathbf{o}\}_T\mid\theta_d) = \sum_{t=1}^{T}\log\left(p(\mathbf{o}_t\mid\omega)\sum_{\rho_t=1}^{P}\frac{1}{p(\rho_t\mid\omega)}p(\rho_t\mid\theta_d)\,p(\rho_t\mid\mathbf{o}_t)\right) \tag{36}$$

$$= \sum_{t=1}^{T}\log p(\mathbf{o}_t\mid\omega) + \sum_{t=1}^{T}\log\left(\sum_{\rho_t=1}^{P}\frac{1}{p(\rho_t\mid\omega)}p(\rho_t\mid\theta_d)\,p(\rho_t\mid\mathbf{o}_t)\right). \tag{37}$$

Notice that there is still a prior, $p(\mathbf{o}_t\mid\omega)$, on the softmax output (equivalently on the data, $\mathbf{x}_t$). There are two interesting things:

1. It is independent of the state sequence, so can be ignored in a maximisation.
2. It is actually the same term in the denominator of equation 33. In this sense it just cancels, and the proportionality symbol of equation 33 becomes equality:

$$p(\mathbf{q}\mid\mathbf{X}) = \prod_{t=1}^{T}\sum_{\rho_t=1}^{P}\frac{1}{p(\rho_t\mid\omega)}p(\rho_t\mid\theta_d)\,p(\rho_t\mid\mathbf{o}_t). \tag{38}$$

2.5 Corollaries

2.5.1 Relationship with conventional hybrids

Notice that when the state does not store a distribution, only a label, we have $\rho_t = q_t$ and $p(\rho_t\mid\theta_{q_t}) = 1$. Hence,

$$\prod_{t=1}^{T}p(\mathbf{x}_t\mid q_t) = \prod_{t=1}^{T}\frac{p(q_t\mid\mathbf{o}_t)}{p(q_t\mid\omega)}, \tag{39}$$

which is the usual hybrid HMM/ANN expression.

2.5.2 Relationship with scalar product

Writing the core part of equation 5 as

$$p(\mathbf{o}_t \mid \theta_d) = \sum_{\rho_t=1}^{P} p(\rho_t \mid \theta_d)\, p(\mathbf{o}_t \mid \rho_t), \tag{40}$$

the operative term is clearly a scalar product. This is nearly identical to the *Posterior Scalar Product* of Picart (2009).

It is also identical to the formulation of Rottland and Rigoll (2000), who simply treat the softmax outputs divided by the prior $p(\rho_t)$ in the same manner as Gaussian outputs in a mixture.

2.5.3 Relationship with KL-HMM

Writing

$$L = \log \sum_{\rho_t=1}^{P} \frac{1}{p(\rho_t \mid \boldsymbol{\omega})} p(\rho_t \mid \theta_d)\, p(\rho_t \mid \mathbf{o}_t) \tag{41}$$

$$= \log \sum_{\rho_t=1}^{P} Q(\rho_t) \frac{p(\rho_t \mid \theta_d)}{Q(\rho_t)} \frac{p(\rho_t \mid \mathbf{o}_t)}{p(\rho_t \mid \boldsymbol{\omega})} \tag{42}$$

This holds for any arbitrary distribution $Q(\rho_t)$. Now use Jensen's inequality to give a lower bound

$$L \geqslant \sum_{\rho_t=1}^{P} Q(\rho_t) \log \left(\frac{p(\rho_t \mid \theta_d)}{Q(\rho_t)} \frac{p(\rho_t \mid \mathbf{o}_t)}{p(\rho_t \mid \boldsymbol{\omega})} \right) \tag{43}$$

$$= \sum_{\rho_t=1}^{P} Q(\rho_t) \log \frac{p(\rho_t \mid \theta_d)}{Q(\rho_t)} + \sum_{\rho_t=1}^{P} Q(\rho_t) \log \frac{p(\rho_t \mid \mathbf{o}_t)}{p(\rho_t \mid \boldsymbol{\omega})}. \tag{44}$$

If we set $Q(\rho_t) = p(\rho_t \mid \mathbf{o}_t)$, this is very close to the (symmetric) KL divergence. Certainly the form is the same. The differences are that the terms before the logarithms are the same, and the final denominator is a prior. Nevertheless, the relationship is evident.

3 Dirichlet model

3.1 Model formulation

We seek a probability distribution that outputs observations of the same form as the softmax. Such a distribution is the Dirichlet distribution:

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\Gamma(\theta_1 + \theta_2 + \cdots + \theta_P)}{\Gamma(\theta_1)\Gamma(\theta_2)\ldots\Gamma(\theta_P)} x_1^{\theta_1 - 1} x_2^{\theta_2 - 1} \ldots x_P^{\theta_P - 1}, \tag{45}$$

$$= \frac{1}{B(\boldsymbol{\theta})} \prod_{\rho=1}^{P} x_\rho^{\theta_\rho - 1}, \qquad \sum_{\rho=1}^{P} x_\rho = 1. \tag{46}$$

In this case, the softmax outputs are interpreted as normalised numbers. There is no probabilistic interpretation required. This is analogous to the Tandem interpretation. A slightly more involved model is the Dirichlet mixture of Chen et al. (2007), apparently duplicated by V. et al. (2011).

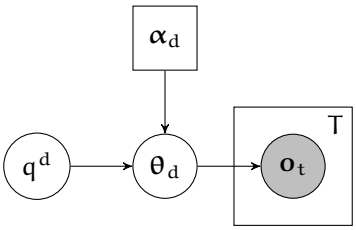

Figure 2: Generative diagram in plate notation for Dirichlet model before training. Arrows mean "influences", rectangles are assumed fixed, circles are state variables. Shading means observable. The large box means "repeat T times".

The full joint distribution follows directly from figure 2:

$$p(\boldsymbol{\theta}_d, \mathbf{O}) = p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}_d) \prod_{t=1}^{T} p(\mathbf{o}_t \mid \boldsymbol{\theta}_d). \tag{47}$$

3.2 Priors

The prior in this case is on the parameter of a Dirichlet distribution. There is not an obvious distribution to use in this case, so we proceed with

$$p(\boldsymbol{\theta}_d) \propto 1. \tag{48}$$

3.3 Parameter estimation

Say we have T observations;

$$\frac{\partial}{\partial \theta_{d,\rho}} \log \prod_{t=1}^{T} p(\mathbf{o}_t \mid \boldsymbol{\theta}) = T\Psi\left(\sum_\rho \theta_{d,\rho}\right) - T\Psi(\theta_{d,\rho}) + \sum_{t=1}^{T} \log o_{t,\rho} = 0, \tag{49}$$

where $\Psi(\cdot)$ is the digamma function. There is no closed form solution. The "normal" approach is to use Newton-Raphson iterations to solve this equation; this is detailed by Chen et al. (2007). The situation is also discussed by Wicker et al. (2008).

Another possibility is the EM-like solution given by Minka[2], evaluated by Huang[3] where he points out a lower bound on the gamma function

$$\Gamma(x) \geqslant \Gamma(y) \exp\left[(x - y)\Psi(y)\right] \tag{50}$$

[2] http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf
[3] http://www.stanford.edu/~jhuang11/research/dirichlet/dirichlet.pdf

so

$$\log \Gamma(x) \geqslant x\Psi(y) + \log \Gamma(y) - y\Psi(y) \tag{51}$$

where equality holds for $x = y$. This allows us to write

$$\log \prod_{t=1}^{T} p\left(\mathbf{o}_t \mid \boldsymbol{\theta}'\right) \geqslant T\left(\sum_{\rho=1}^{P} \theta'_{d,\rho}\right) \Psi\left(\sum_{\rho=1}^{P} \theta_{d,\rho}\right) - T\sum_{\rho=1}^{P} \log \Gamma(\theta'_{d,\rho}) + \sum_{\rho=1}^{P} (\theta'_{d,\rho} - 1) \sum_{t=1}^{T} \log o_{t,\rho} + C \tag{52}$$

so

$$\frac{\partial}{\partial \theta'_{d,\rho}} \log \prod_{t=1}^{T} p\left(\mathbf{o}_t \mid \boldsymbol{\theta}'\right) = T\Psi\left(\sum_{\rho=1}^{P} \theta_{d,\rho}\right) - T\Psi(\theta'_{d,\rho}) + \sum_{t=1}^{T} \log o_{t,\rho} = 0 \tag{53}$$

$$\Psi(\theta'_{d,\rho}) = \Psi\left(\sum_{\rho=1}^{P} \theta_{d,\rho}\right) + \frac{1}{T}\sum_{t=1}^{T} \log o_{t,\rho}. \tag{54}$$

One point here is that it is necessary to invert the Ψ function, but Minka shows that it can be done using Newton-Raphson.

3.4 Model evaluation

Given an observation set, $\mathbf{O}$, we can write the log-likelihood as

$$\log \prod_{t=1}^{T} p\left(\mathbf{o}_t \mid \boldsymbol{\theta}_d\right) = T \log \Gamma\left(\sum_{\rho=1}^{P} \theta_{d,\rho}\right) - T\sum_{\rho=1}^{P} \log \Gamma(\theta_{d,\rho}) + \sum_{\rho=1}^{P} (\theta_{d,\rho} - 1) \sum_{t=1}^{T} \log o_{t,\rho}. \tag{55}$$

3.5 Corollary

Writing the final term of equation 55 as

$$\sum_{t=1}^{T}\sum_{\rho=1}^{P} (\theta_{d,\rho} - 1) \log o_{t,\rho} \tag{56}$$

it is clear that the accumulated value is a scalar product involving the parameter and the logarithm of the observation. So, it has information-theoretic overtones.

4 Conclusions

We have detailed two possible generative models with the same structure as the KL-HMM of Aradilla Zapata (2008):

1. A categorial model, in which HMM states are assumed to store a categorical distribution. Such a distribution emits phones.
2. A Dirichlet model, in which HMM states are assumed to store a Dirichlet distribution. Such a distribution emits vectors of the form in equation 1.

5 Acknowledgements

The authors are grateful to John S. Bridle for comments on an earlier version of this manuscript.

References

Guillermo Aradilla, Jithendra Vepa, and Hervé Bourlard. An acoustic model based on Kullback-Leibler divergence for posterior features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, April 2007.

Guillermo Aradilla Zapata. *Acoustic Models for Posterior Features in Speech Recognition*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2008. Thèse no 4164 (2008), presentée a là Faculté des sciences et techniques de l'ingénieur.

Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994. ISBN 0-7923-9396-1.

John S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, number 2, pages 211–217. Morgan Kaufmann, San Mateo, 1990.

Li Chen, David Barber, and Jean-Marc Odobez. Dynamical Dirichlet mixture model. IDIAP-RR 2007-02, Idiap Research Institute, Centre du Parc, Rue Marconi 19, P.O. Box 592, CH-1920 Martigny, April 2007.

Hynek Hermansky, Daniel P. W. Ellis, and Sangita Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1635–1638, Istanbul, Turkey, June 2000.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1):79–86, 1951. URL http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729694.

Solomon Kullback. The Kullback-Leibler distance. *The American Statistician*, 41(4):340–341, November 1987. in Letters to the Editor.

Nelson Morgan and Hervé Bourlard. Neural networks for statistical recognition of continuous speech,. *Proceedings of the IEEE*, 83(5):741–770, May 1995a. Invited paper.

Nelson Morgan and Hervé Bourlard. Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach. *IEEE Signal Processing Magazine*, 12(3):25–42, May 1995b. IEEE Award paper.

Benjamin Picart. Improved phone posterior estimation through K-NN and MLP-based similarity. IDIAP-RR 18-2009, Idiap Research Institute, Centre du Parc, Rue Marconi 19, P.O. Box 592, CH-1920 Martigny, August 2009.

Jörg Rottland and Gerhard Rigoll. Tied posteriors: An approach for efficient introduction of context dependency in hybrid NN/HMM LVCSR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1241–1244, Istanbul, Turkey, June 2000.

Balakrishnan V., G.S.V.S. Sivaram, and Sanjeev Khudanpur. Dirichlet mixture models of neural net posteriors for HMM-based speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5028–5031, Prague, Czech Republic, May 2011.

Nicolas Wicker, Jean Muller, Ravi Kiran Reddy Kalathur, and Olivier Poch. A maximum likelihood approximation method for Dirichlet's parameter estimation. *Computational Statistics & Data Analysis*, 52(3):1315–1322, January 2008.

A Kullback Leibler

There is some uncertainty over how to handle the asymmetry of the KL divergence. In fact, Kullback and Leibler (1951) defined the divergence to be symmetrical; they referred to the asymmetric versions as "mean information for discrimination" measures.

From Kullback and Leibler (1951), we have "the mean information for discrimination between H_1 and H_2 per observation from distribution μ_1" is

$$I(1:2) = I_{1:2}(X) = \int dx\, p\,(x \mid H_1) \log \frac{p\,(x \mid H_1)}{p\,(x \mid H_2)}, \tag{57}$$

where H_1 is the hypothesis that the data x originated from distribution μ_1. Kullback (1987) expresses a preference for the term $I(1:2)$ to be called "discrimination information" rather than "distance". Then the "divergence between μ_1 and μ_2" is

$$\begin{aligned} J(1,2) &= J_{12}(X) && (58)\\ &= I_{1:2}(X) + I_{2:1}(X) && (59)\\ &= \int dx\, \left(p\,(x \mid H_1) - p\,(x \mid H_2)\right) \log \frac{p\,(x \mid H_1)}{p\,(x \mid H_2)}. && (60) \end{aligned}$$

So, clearly

$$J(1,2) = \int dx\, p\,(x \mid H_1) \log \frac{p\,(x \mid H_1)}{p\,(x \mid H_2)} + \int dx\, p\,(x \mid H_2) \log \frac{p\,(x \mid H_2)}{p\,(x \mid H_1)}. \tag{61}$$