Supplement to "An Infinite Hidden Markov Model with Similarity-Biased Transitions", ICML 2017

Colin Dawson, Chaofan Huang, Clayton Morrison

June 13, 2017

This supplement contains additional derivations to accompany the model definition and Gibbs updates for the paper "An Infinite Hidden Markov Model with Similariy-Biased Transitions", published in ICML 2017. Section 1 concerns the derivation of the augmented data representation referred to as the "Markov Jump Process with Failed Transitions" (MJP-FT). Section 2 fills in details for the Gibbs sampling steps to sample the rescaled HDP used by the HDP-HMM-LT. Section 3 gives a derivation for the updates to the binary state vectors, $\boldsymbol{\theta}$, in the version of the HDP-HMM-LT used in the cocktail party experiment. Finally, section 4 gives the details for the Hamiltonian Monte Carlo update for $\boldsymbol{\ell}$ in the version of the model used in the Bach chorale experiment.

1. Details of the Markov Jump Process with Failed Transitions Representation

We can gain stronger intuition, as well as simplify posterior inference, by re-casting the HDP-HMM-LT as a continuous time Markov Jump Process where some of the attempts to jump from one state to another fail, and where the failure probability increases as a function of the "distance" between the states.

Let ϕ be defined as in the last section, and let $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ be defined as in the Normalized Gamma Process representation of the ordinary HDP-HMM. That is,

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma) \tag{1}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{2}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathsf{Gamma}(\alpha\beta_{j'}, 1) \tag{3}$$

Now suppose that when the process is in state j, jumps to state j' are made at rate $\pi_{jj'}$. This defines a continuous-time Markov Process where the off-diagonal elements of the transition rate matrix are the off diagonal elements of $\boldsymbol{\pi}$. In addition, self-jumps are allowed, and occur with rate π_{jj}. If we only observe the jumps and not the durations between jumps, this is an ordinary Markov chain, whose transition matrix is obtained by appropriately normalizing $\boldsymbol{\pi}$. If we do not observe the jumps themselves, but instead an observation is generated once per jump from a distribution that depends on the state being jumped to, then we have an ordinary HMM.

We modify this process as follows. Suppose that each jump attempt from state j to state j' has a chance of failing, which is an increasing function of the "distance" between the states. In particular, let the success probability be $\phi_{jj'}$ (recall that we assumed above that $0 \leq \phi_{jj'} \leq 1$ for all j, j'). Then, the rate of successful jumps from j to j' is $\pi_{jj'}\phi_{jj'}$, and the corresponding rate of unsuccessful jump attempts is $\pi_{jj'}(1-\phi_{jj'})$. To see this, denote by $N_{jj'}$ the total number of jump attempts to j' in a unit interval of time spent in state j. Since we are assuming the process is Markovian, the total number of attempts is $\mathsf{Poisson}(\pi_{jj'})$ distributed. Conditioned on $N_{jj'}$, $n_{jj'}$ will be successful, where

$$n_{jj'} \,|\, N_{jj'} \sim \mathsf{Binom}(N_{jj'}, \phi_{jj'}) \tag{4}$$

It is easy to show (and well known) that the marginal distribution of $n_{jj'}$ is $\mathsf{Poisson}(\pi_{jj'}\phi_{jj'})$, and the marginal distribution of $\tilde{q}_{jj'} := N_{jj'} - n_{jj'}$ is $\mathsf{Poisson}(\pi_{jj'}(1-\phi_{jj'}))$. The rate of successful jumps from state j overall is then $T_j := \sum_{j'} \pi_{jj'}\phi_{jj'}$.

Let t index jumps, so that z_t indicates the tth state visited by the process (couting self-jumps as a new time step). Given that the process is in state j at discretized time $t-1$ (that is, $z_{t-1} = j$), it is a standard property of

Markov Processes that the probability that the first successful jump is to state j' (that is, $z_t = j'$) is proportional to the rate of successful attempts to j', which is $\pi_{jj'}\phi_{jj'}$.

Let $\tilde{u}_t$ indicate the time elapsed between the tth and and $t-1$th successful jump (where we assume that the first observation occurs when the first successful jump from a distinguished initial state is made). We have

$$\tilde{u}_t \mid z_{t-1} \sim \mathsf{Exp}(T_{z_{t-1}}) \tag{5}$$

where $\tilde{u}_t$ is independent of z_t.

During this period, there will be $\tilde{q}_{j't}$ unsuccessful attempts to jump to state j', where

$$\tilde{q}_{j't} \mid z_{t-1} \sim \mathsf{Poisson}(\tilde{u}_t \pi_{z_{t-1}j'}(1-\phi_{z_{t-1}j'})) \tag{6}$$

Define the following additional variables

$$\mathcal{T}_j = \{t \mid z_{t-1} = j\} \tag{7}$$

$$q_{jj'} = \sum_{t \in \mathcal{T}_j} \tilde{q}_{j't} \tag{8}$$

$$u_j = \sum_{t \in \mathcal{T}_j} \tilde{u}_t \tag{9}$$

and let $\mathbf{Q} = (q_{jj'})_{j,j' \geq 1}$ be the matrix of unsuccessful jump attempt counts, and $\mathbf{u} = (u_j)_{j \geq 1}$ be the vector of the total times spent in each state.

Since each of the $\tilde{u}_t$ with $t \in \mathcal{T}_j$ are i.i.d. $\mathsf{Exp}(T_j)$, we get the marginal distribution

$$u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\phi} \overset{ind}{\sim} \mathsf{Gamma}(n_{j\cdot}, T_j) \tag{10}$$

by the standard property that sums of i.i.d. Exponential distributions has a Gamma distribution with shape equal to the number of variates in the sum, and rate equal to the rate of the individual exponentials. Moreover, since the $\tilde{q}_{j't}$ with $t \in \mathcal{T}_j$ are Poisson distributed, the total number of failed attempts in the total duration u_j is

$$q_{jj'} \overset{ind}{\sim} \mathsf{Poisson}(u_j \pi_{jj'}(1-\phi_{jj'})). \tag{11}$$

Thus if we marginalize out the individual $\tilde{u}_t$ and $\tilde{q}_{j't}$, we have a joint distribution over $\mathbf{z}$, $\mathbf{u}$, and $\mathbf{Q}$, conditioned on the transition rate matrix $\boldsymbol{\pi}$ and the success probability matrix $\boldsymbol{\phi}$, which is

$$\begin{aligned}
p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) &= \left(\prod_{t=1}^{T} p(z_t \mid z_{t-1})\right) \prod_j p(u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\phi}) \prod_{j'} p(q_{jj'} \mid u_j \pi_{jj'}, \phi_{jj'}) && (12)\\
&= \left(\prod_t \frac{\pi_{z_{t-1}z_t}\phi_{z_{t-1}z_t}}{T_{z_{t-1}}}\right) \prod_j \frac{T_j^{n_{j\cdot}}}{\Gamma(n_{j\cdot})} u_j^{n_{j\cdot}-1} e^{-T_j u_j} && (13)\\
&\qquad \times \prod_{j'} e^{-u_j \pi_{jj'}(1-\phi_{jj'})} u_j^{q_{jj'}} \pi_{jj'}^{q_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} && (14)\\
&= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} && (15)\\
&\qquad \times \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}\phi_{jj'}u_j} e^{-\pi_{jj'}(1-\phi_{jj'})u_j} (q_{jj'}!)^{-1} && (16)\\
&= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}u_j} (q_{jj'}!)^{-1} && (17)
\end{aligned}$$

Setting aside terms that do not depend on $\boldsymbol{\pi}$, we get the conditional likelihood function used in sampling $\boldsymbol{\pi}$:

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) \propto \prod_j \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} e^{-\pi_{jj'}u_j} \tag{18}$$

which, combined with the independent Gamma priors on π yields conditionally independent Gamma posteriors:

$$\pi_{jj'} \mid \mathbf{z}, \mathbf{u}, \mathbf{Q}, \boldsymbol{\beta}, \alpha \overset{\text{ind.}}{\sim} \mathsf{Gamma}(\alpha\boldsymbol{\beta}_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{19}$$

2. Inference details for hyperparameters of the rescaled HDP

2.1. Sampling $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, α and γ

The joint conditional over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given the augmented data $\mathcal{D} = (\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \mathbf{r}, w)$ factors as

$$p(\gamma, \alpha, \beta, \pi \mid \mathcal{D}) = p(\gamma \mid \mathcal{D}) p(\alpha \mid \mathcal{D}) p(\beta \mid \gamma, \mathcal{D}) p(\pi \mid \alpha, \beta, \mathcal{D}) \tag{20}$$

We will derive these four factors in reverse order.

Sampling $\boldsymbol{\pi}$ The entries in $\boldsymbol{\pi}$ are conditionally independent given α and β, so we have the prior

$$p(\boldsymbol{\pi} \mid \boldsymbol{\beta}, \alpha) = \prod_j \prod_{j'} \Gamma(\alpha\boldsymbol{\beta}_{j'})^{-1} \pi_{jj'}^{\alpha\boldsymbol{\beta}_{j'}-1} \exp(-\pi_{jj'}), \tag{21}$$

and the likelihood given $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}\}$ given by (17). Combining these, we have

$$p(\boldsymbol{\pi}, \mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \beta, \alpha, \boldsymbol{\phi}) = \prod_j u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \Gamma(\alpha\beta_{j'})^{-1} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} \tag{22}$$

$$\times e^{-(1+u_j)\pi_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{23}$$

Conditioning on everything except $\boldsymbol{\pi}$, we get

$$p(\boldsymbol{\pi} \mid \mathbf{Q}, \mathbf{u}, \mathbf{z}, \boldsymbol{\beta}, \alpha) \propto \prod_j \prod_{j'} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} \exp(-(1+u_j)\pi_{jj'}) \tag{24}$$

and thus we see that the $\pi_{jj'}$ are conditionally independent given u, z and Q, and distributed according to

$$\pi_{jj'} \mid n_{jj'}, q_{jj'}, \beta_{j'}, \alpha \overset{ind}{\sim} \mathsf{Gamma}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{25}$$

Sampling $\boldsymbol{\beta}$ Consider the conditional distribution of β having integrated out $\boldsymbol{\pi}$. The prior density of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta} \mid \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \prod_j \beta_j^{\frac{\gamma}{J}-1} \tag{26}$$

After integrating out π in (22), we have

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \mid \boldsymbol{\beta}, \alpha, \gamma, \boldsymbol{\phi}) = \prod_{j=1}^{J} u_j^{-1} \prod_{j'=1}^{J} u^{n_{jj'}+q_{jj'}-1} (1+u_j)^{-(\alpha\beta_{j'}+n_{jj'}+q_{jj'})} \tag{27}$$

$$\times \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{28}$$

$$= \prod_{j=1}^{J} \Gamma(n_{j\cdot})^{-1} u_j^{-1} (1+u_j)^{-\alpha} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{29}$$

$$\times \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{30}$$

where we have used the fact that the β_j sum to 1. Therefore

$$p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})}. \tag{31}$$

Following (Teh et al., 2006), we can write the ratios of Gamma functions as polynomials in β_j, as

$$p(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \sum_{m_{jj'}=1}^{n_{jj'}} s(n_{jj'} + q_{jj'}, m_{jj'})(\alpha\beta_{j'})^{m_{jj'}} \tag{32}$$

where $s(m, n)$ is an unsigned Stirling number of the first kind, which is used to represent the number of permutations of n elements such that there are m distinct cycles.

This admits an augmented data representation, where we introduce a random matrix $\mathbf{M} = (m_{jj'})_{1 \le j, j' \le J}$, whose entries are conditionally independent given $\boldsymbol{\beta}$, $\mathbf{Q}$ and $\mathbf{z}$, with

$$p(m_{jj'} = m \,|\, \beta_{j'}, \alpha, n_{jj'}, q_{jj'}) = \frac{s(n_{jj'} + q_{jj'}, m)\alpha^m \beta_{j'}^m}{\sum_{m'=0}^{n_{jj'}+q_{jj'}} s(n_{jj'} + q_{jj'}, m')\alpha^{m'} \beta_{j'}^{m'}} \tag{33}$$

for integer m ranging between 0 and $n_{jj'} + q_{jj'}$. Note that $s(n, 0) = 0$ if $n > 0$, $s(0, 0) = 1$, $s(0, m) = 0$ if $m > 0$, and we have the recurrence relation $s(n + 1, m) = ns(n, m) + s(n, m - 1)$, and so we could compute each of these coefficients explicitly; however, it is typically simpler and more computationally efficient to sample from this distribution by simulating the number of occupied tables in a Chinese Restaurant Process with n customers, than it is to enumerate its probabilities.

For each $m_{jj'}$ we simply draw $n_{jj'}$ assignments of customers to tables according to the Chinese Restaurant Process and set $m_{jj'}$ to be the number of distinct tables realized; that is, assign the first customer to a table, setting $m_{jj'}$ to 1, and then, after n customers are assigned, assign the $n + 1$th customer to a new table with probability $\alpha\beta_{j'}/(n+\alpha\beta_{j'})$, in which case we increment $m_{jj'}$, and to an existing table with probability $n/(n+\alpha)$, in which case we do not increment $m_{jj'}$.

Then, we have joint distribution

$$p(\boldsymbol{\beta}, \mathbf{M} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'})\alpha^{m_{jj'}} \beta_{j'}^{m_{jj'}} \tag{34}$$

which yields (32) when marginalized over $\mathbf{M}$. Again discarding constants in $\boldsymbol{\beta}$ and regrouping yields

$$p(\beta \,|\, M, z, u, \theta, \alpha, \gamma) \propto \prod_{j'=1}^{J} \beta_{j'}^{\frac{\gamma}{J} + m_{\cdot j'} - 1} \tag{35}$$

which is Dirichlet:

$$\beta \,|\, M, \gamma \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \ldots, \frac{\gamma}{J} + m_{\cdot J}) \tag{36}$$

Sampling α and γ Assume that α and γ have Gamma priors, parameterized by shape, a and rate, b:

$$p(\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) \tag{37}$$

$$p(\gamma) = \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) \tag{38}$$

Having integrated out $\boldsymbol{\pi}$, we have

$$p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\phi}) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \alpha^{m_{\cdot\cdot}} \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J} + m_{\cdot j} - 1} \Gamma(n_{j\cdot})^{-1} u_j^{-1} (1 + u_j)^{-\alpha} \left(\frac{u_j}{1 + u_j}\right)^{n_{j\cdot} + q_{j\cdot}} \tag{39}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{40}$$

We can also integrate out $\boldsymbol{\beta}$, to yield

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\phi}) = \alpha^{m_{\cdot\cdot}} e^{-\sum_{j''} \log(1+u_{j''})\alpha} \frac{\Gamma(\gamma)}{\Gamma(\gamma + m_{\cdot\cdot})} \tag{41}$$

$$\times \prod_j \frac{\Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})\Gamma(n_{j\cdot})} u_j^{-1} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{42}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{43}$$

demonstrating that α and γ are independent given $\boldsymbol{\phi}$ and the augmented data, with

$$p(\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}) \propto \alpha^{a_\alpha + m_{\cdot\cdot}} \exp(-(b_\alpha + \sum_j \log(1+u_j))\alpha) \tag{44}$$

and

$$p(\gamma \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) \frac{\Gamma(\gamma) \prod_{j=1}^{J} \Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})^J \Gamma(\gamma + m_{\cdot\cdot})} \tag{45}$$

So we see that

$$\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \sim \mathsf{Gamma}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1+u_j)) \tag{46}$$

To sample γ, we introduce a new set of auxiliary variables, $\mathbf{r} = (r_1, \ldots, r_J)$ and w with the following distributions:

$$p(r_{j'} = r \,|\, m_{\cdot j'}, \gamma) = \frac{\Gamma(\frac{\gamma}{J})}{\Gamma(\frac{\gamma}{J} + m_{\cdot j'})} s(m_{\cdot j'}, r) \left(\frac{\gamma}{J}\right)^r \qquad r = 1, \ldots, m_{\cdot j} \tag{47}$$

$$p(w \,|\, m_{\cdot\cdot}\gamma) = \frac{\Gamma(\gamma + m_{\cdot\cdot})}{\Gamma(\gamma)\Gamma(m_{\cdot\cdot})} w^{\gamma-1}(1-w)^{m_{\cdot\cdot}-1} \qquad w \in (0,1) \tag{48}$$

so that

$$p(\gamma, r, w \,|\, \mathbf{M}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) w^{\gamma-1} (1-w)^{m_{\cdot\cdot}-1} \prod_{j'=1}^{J} s(m_{\cdot j'}, r_{j'}) \left(\frac{\gamma}{J}\right)^{r_{j'}} \tag{49}$$

and

$$p(\gamma \,|\, r, w) \propto \gamma^{a_\gamma + r_{\cdot} - 1} \exp(-(b_\gamma - \log(w))\gamma), \tag{50}$$

which is to say

$$\gamma \,|\, \mathbf{r}, w, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \sim \mathsf{Gamma}(a_\gamma + r_{\cdot}, b_\gamma - \log(w)) \tag{51}$$

3. Derivation of $\boldsymbol{\eta}$ update in the Cocktail Party and Synthetic Data Experiments

In principle, $\boldsymbol{\eta}$ can have any distribution over binary vectors, but we will suppose for simplicity that it can be factored into D independent coordinate-wise Bernoulli variates. Let μ_d be the Bernoulli parameter for the dth coordinate.

The similarity function $\phi_{jj'}$ is the Laplacian kernel:

$$\phi_{jj'} = \Phi(\boldsymbol{\eta}_j, \boldsymbol{\eta}_{j'}) = \exp(-\lambda d_{jj'}) \tag{52}$$

where $d_{jj'd} = |\boldsymbol{\eta}_{jd} - \boldsymbol{\eta}_{j'd}|$ is Hamming distance in the dth coordinate, $d_{jj'} := \sum_{d=1}^{D} d_{jj'}$ is the total Hamming distance between $\boldsymbol{\eta}_j$ and $\boldsymbol{\eta}_{j'}$, and $\lambda \geq 0$ (if $\lambda = 0$, the $\phi_{jj'}$ are identically 1, and so do not have any influence, reducing the model to an ordinary HDP-HMM).

Let

$$\phi_{jj'-d} = \exp(-\lambda(d_{jj'} - d_{jj'd})) \tag{53}$$

so that $\phi_{jj'} = \phi_{jj'-d} e^{-\lambda d_{jj'd}}$.

Since the matrix $\boldsymbol{\phi}$ is assumed to be symmetric, we have

$$\frac{p(\mathbf{z}, \mathbf{Q} \,|\, \eta_{jd} = 1, \boldsymbol{\eta} \setminus \eta_{jd})}{p(\mathbf{z}, \mathbf{Q} \,|\, \eta_{jd} = 0, \boldsymbol{\eta} \setminus \eta_{jd})} \propto \prod_{j' \neq j} \frac{e^{-\lambda(n_{jj'} + n_{j'j})\left|1-\theta_{j'd}\right|}(1 - \phi_{jj'-d} e^{-\lambda\left|1-\theta_{j'd}\right|})^{q_{jj'}+q_{j'j}}}{e^{-\lambda(n_{jj'} + n_{j'j})\left|\theta_{j'd}\right|}(1 - \phi_{jj'-d} e^{-\lambda\left|\theta_{j'd}\right|})^{q_{jj'}+q_{j'j}}} \tag{54}$$

$$= e^{-\lambda(c_{jd0} - c_{jd1})} \prod_{j' \neq j} \left(\frac{1 - \phi_{jj'-d} e^{-\lambda}}{1 - \phi_{jj'-d}} \right)^{(-1)^{\theta_{j'd}}(q_{jj'} + q_{j'j})} \tag{55}$$

where c_{jd0} and c_{jd1} are the number of successful jumps to or from state j, to or from states with a 0 or 1, respectively, in position d. That is,

$$c_{jd0} = \sum_{\{j' \,|\, \theta_{j'd} = 0\}} n_{jj'} + n_{j'j} \qquad c_{jd1} = \sum_{\{j' \,|\, \theta_{j'd} = 1\}} n_{jj'} + n_{j'j} \tag{56}$$

Therefore, we can Gibbs sample η_{jd} from its conditional posterior Bernoulli distribution given the rest of $\boldsymbol{\eta}$, where we compute the Bernoulli parameter via the log-odds

$$\log\left(\frac{p(\eta_{jd} = 1 \,|\, \mathbf{Y}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\eta} \setminus \eta_{jd})}{p(\eta_{jd} = 0 \,|\, \mathbf{Y}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\eta} \setminus \eta_{jd})} \right) = \log\left(\frac{p(\eta_{jd} = 1) p(\mathbf{z}, \mathbf{Q} \,|\, \eta_{jd} = 1, \boldsymbol{\eta} \setminus \eta_{jd}) p(\mathbf{Y} \,|\, \mathbf{z}, \eta_{jd} = 1, \boldsymbol{\eta} \setminus \eta_{jd})}{p(\eta_{jd} = 0) p(\mathbf{z}, \mathbf{Q} \,|\, \eta_{jd} = 0, \boldsymbol{\eta} \setminus \eta_{jd}) p(\mathbf{Y} \,|\, \mathbf{z}, \eta_{jd} = 0, \boldsymbol{\eta} \setminus \eta_{jd})} \right) \tag{57}$$

$$= \log\left(\frac{\mu_d}{1 - \mu_d} \right) + (c_{jd1} - c_{jd0})\lambda + \sum_{j' \neq j} (-1)^{\theta_{j'd}} (q_{jj'} + q_{j'j}) \log\left(\frac{1 - \phi_{jj'}^{(-d)} e^{-\lambda}}{1 - \phi_{jj'}^{(-d)}} \right) \tag{58}$$

$$+ \sum_{\{t \,|\, z_t = j\}} \log\left(\frac{f(\mathbf{y}_t; \eta_{jd} = 1, \boldsymbol{\eta}_j \setminus \eta_{jd})}{f(\mathbf{y}_t; \eta_{jd} = 0, \boldsymbol{\eta}_j \setminus \eta_{jd})} \right) \tag{59}$$

Suppose also that the observed data $\mathbf{Y}$ consists of a $T \times K$ matrix, where the tth row $\mathbf{y}_t = (y_{t1}, \ldots, y_{tK})^{\mathsf{T}}$ is a K-dimensional feature vector associated with time t, and let $\mathbf{W}$ be a $D \times K$ weight matrix with kth column $\mathbf{w}_k$, such that

$$f(\mathbf{y}_t; \boldsymbol{\eta}_j) = g(\mathbf{y}_t; \mathbf{W}^{\mathsf{T}} \boldsymbol{\eta}_j) \tag{60}$$

for a suitable parametric function g. We assume for simplicity that g factors as

$$g(\mathbf{y}_t; \mathbf{W}^{\mathsf{T}} \boldsymbol{\eta}_j) = \prod_{k=1}^{K} g_k(y_{tk}; \mathbf{w}_k \cdot \boldsymbol{\eta}_j) \tag{61}$$

Define $x_{tk} = \mathbf{w}_k \cdot \theta_{z_t}$, and $x_{tk}^{(-d)} = \mathbf{w}_k^{-d} \cdot \theta_{z_t}^{-d}$, where θ_j^{-d} and $\mathbf{w}_k^{-d}$ are θ_j and $\mathbf{w}_k$, respectively, with the dth coordinate removed. Then

$$\log\left(\frac{f(\mathbf{y}_t; \eta_{jd} = 1, \boldsymbol{\eta}_j \setminus \eta_{jd})}{f(\mathbf{y}_t; \eta_{jd} = 0, \boldsymbol{\eta}_j \setminus \eta_{jd})} \right) = \sum_{k=1}^{K} \log\left(\frac{g_k(y_{tk}; x_{tk}^{(-d)} + w_{dk})}{g_k(y_{tk}; x_{tk}^{(-d)})} \right). \tag{62}$$

If $g_k(y; x)$ is a Normal density with mean x and unit variance, then

$$\log\left(\frac{g_k(y_{tk}; x_{tk}^{(-d)} + w_{dk})}{g_k(y_{tk}; x_{tk}^{(-d)})} \right) = -w_{dk}(y_{tk} - x_{tk}^{(-d)} + \frac{1}{2} w_{dk}) \tag{63}$$

4. Derivation of HMC update for $\boldsymbol{\ell}$ in the Bach Chorale Experiment

We have a set of states with parameters $\boldsymbol{\ell}_j$, $j = 1, \dots, J$. In the previous version of the model, $\boldsymbol{\ell}_j$ was a binary state vector on which both the similarities $\phi_{jj'}$ and the emission distribution F_j depended. Here, we define the latent locations $\boldsymbol{\ell}_j = (\ell_{j1}, \ell_{jD})$ to be locations in $\mathbb{R}^D$, independent of the emission distributions, so that during inference they are informed solely by the transitions.

We set

$$\phi_{jj'}(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'}) = \exp\left(-\frac{\lambda}{2} d_{jj'}^2\right)$$

where $d_{jj'}$ is the Euclidean distance between $\boldsymbol{\ell}_j$ and $\boldsymbol{\ell}_{j'}$; that is,

$$d_{jj'}^2 = \sum_d (\ell_{jd} - \ell_{j'd})^2$$

Since now $\boldsymbol{\ell}_j$ are continuous locations, we use Hamlitonian Monte Carlo (Duane et al., 1987; Neal et al., 2011) to sample them jointly. HMC is a variation on Metropolis-Hastings algorithm which is designed to more efficiently explore a high-dimensional continuous distribution by adopting a proposal distribution which incorporates an auxiliary "momentum" variable to make it more likely that proposals will go in useful directions and improve mixing compared to naive movement.

To do Hamiltonian Monte Carlo to sample from the conditional posterior of $\boldsymbol{\ell}$ given $\mathbf{z}$ and $\mathbf{Q}$, we need to compute the gradient of the log posterior, which is just the sum of the gradient of the log prior and the gradient of the log likelihood.

Assume independent and isotropic Gaussian priors on each $\boldsymbol{\ell}_j$, so we have

$$p(\boldsymbol{\ell}_j) \propto \exp\left(-\frac{h_\ell}{2} \sum_d \ell_{jd}^2\right),$$

where h_ℓ is the prior precision which does not depend on d.

Then the log prior density, up to an additive constant c, is

$$\log p(\boldsymbol{\ell}_j) = c - \frac{h_\ell}{2} \sum_d \ell_{jd}^2$$

The relevant log likelihood is the log of the probability of the $\mathbf{z}$ and $\mathbf{Q}$ variables given the $\phi_{jj'}$. In particular, we have

$$L := p(\mathbf{z}, \mathbf{Q} \mid \boldsymbol{\phi}) \propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}}$$

so that

$$\log L = \sum_j \sum_{j'} \left(n_{jj'} \log(\phi_{jj'}) + q_{jj'} \log(1 - \phi_{jj'})\right)$$

The j, d coordinate of the gradient of the log prior is simply $-h_\ell \ell_{jd}$.

To get the j, d coordinate of the gradient of the log likelihood, we can apply the chain rule to terms as is convenient. In particular,

$$\frac{\partial L}{\partial \ell_{jd}} = \sum_j \sum_{j'} n_{jj'} \frac{\partial \log(\phi_{jj'})}{\partial d_{jj'}^2} \frac{\partial d_{jj'}^2}{\partial \ell_{jd}} + \sum_j \sum_{j'} q_{jj'} \frac{\partial \log(1 - \phi_{jj'})}{\partial (1 - \phi_{jj'})} \frac{\partial (1 - \phi_{jj'})}{\partial d_{jj'}^2} \frac{\partial d_{jj'}^2}{\partial \ell_{jd}}$$

We have the following components:

$$\begin{aligned}
\frac{\partial \log(\phi_{jj'})}{\partial d_{jj'}^2} &= -\frac{\lambda}{2} \\
\frac{\partial d_{jj'}^2}{\partial \ell_{jd}} &= 2 d_{jj'd} I(j \neq j') \\
\frac{\partial \log(1-\phi_{jj'})}{\partial (1-\phi_{jj'})} &= \frac{1}{1-\phi_{jj'}} \\
\frac{\partial (1-\phi_{jj'})}{\partial d_{jj'}^2} &= \frac{\lambda}{2}\phi_{jj'}
\end{aligned}$$

which yields

$$\begin{aligned}
\frac{\partial L}{\partial \ell_{jd}} &= -\lambda \sum_{j}\sum_{j'} n_{jj'} d_{jj'd} \mathbb{I}(j \neq j') + \lambda \sum_{j}\sum_{j'} q_{jj'} d_{jj'd} \frac{\phi_{jj'}}{1-\phi_{jj'}} \mathbb{I}(j \neq j) \\
&= -\lambda \sum_{(j,j'):j\neq j'} d_{jj'd} \left(n_{jj'} - q_{jj'} \frac{\phi_{jj'}}{1-\phi_{jj'}} \right)
\end{aligned}$$

References

Duane, Simon, Kennedy, Anthony D, Pendleton, Brian J, and Roweth, Duncan. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Neal, Radford M et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2: 113–162, 2011.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.