Modeling and Unsupervised Learning of Structured Similarity Among Source Contexts in Bayesian Hierarchical Infinite Mixture Models

With Two Applications to Modeling Natural Language Semantics

Colin Reimer Dawson

July 25, 2016

Contents

Abstract

In a classical mixture modeling, each data point is modeled as arising i.i.d. (typically) from a weighted sum of probability distributions, where both the weights and the parameters of the mixture components are targets of inference. When data arises from different sources that may not give rise to the same mixture distribution, a hierarchical model can allow the source contexts to share components while assigning different weights across them (while perhaps coupling the weights to "borrow strength" across contexts). The Dirichlet Process (DP) Mixture Model (e.g., Rasmussen (2000)) is a Bayesian approach to mixture modeling which models the data as arising from a countably infinite number of components: the Dirichlet Process provides a prior on the mixture weights that guards against overfitting. The Hierarchical Dirichlet Process (HDP) Mixture Model (Teh et al., 2006) employs a separate DP Mixture Model for each context, but couples the weights across contexts by using a common base measure which is itself drawn from a top-level DP. This coupling is critical to ensure that mixture components are reused across contexts. For example, in natural language topic modeling, a common application domain for mixture models, the components represent semantic topics, and the contexts are documents, and it is critical that topics be reused across documents.

These models have been widely adopted in Bayesian statistics and machine learning. However, a limitation of DPs is that the atoms are *a priori* exchangeable, and in the case of HDPs, the component weights are independent conditioned on the top-level measure. This is unrealistic in many applications, including topic modeling, where certain components (e.g., topics) are expected to correlate across contexts (e.g., documents). In the case of topic modeling, the Discrete Infinite Logistic Normal model (DILN; Paisley et al. (2011)) addresses this shortcoming by associating with each mixture component a latent location in an abstract

metric, and rescaling each context-specific set of weights, initially drawn from an HDP, by an exponentiated draw from a Gaussian Process (GP), so that components which are nearby in space tend to have their weights be scaled up or down together. However, inference in this model requires the posterior distribution to be approximated by a variational family, as MCMC sampling from the exact posterior was deemed intractable. Thus, one goal of this dissertation is the development of simple MCMC algorithms for HDP models with correlated components.

A second application of HDPs is to time series models, in particular Hidden Markov Models (HMMs), where the HDP can be used as a prior on a doubly infinite transition matrix for the latent Markov chain, giving rise to the HDP-HMM (first developed, as the "Infinite HMM", by Beal et al. (2001), and subsequently shown to be a case of an HDP by Teh et al. (2006)). There, the hierarchy is over rows of the transition matrix, and the distributions across rows are coupled through a top-level Dirichlet Process. The sequential nature of the problem introduces two added wrinkles, namely that: the contexts themselves are random (since the context when generating state t is the state at time $t-1$), and the set of contexts is the same as the set of components. Hence, not only might the components be correlated with each other via locations in some latent space, but we might expect that contexts that correspond to correlated components will overall have similar distributions.

In the first part of the dissertation, I will present a formal overview of Dirichlet Processes and their various representations, as well as associated schemes for tackling the problem of doing approximate inference over an infinitely flexible model with finite computational resources. I will then turn to the Hierarchical Dirichlet Process, and review the literature on modeling correlations between components.

Next, I will present a novel probabilistic model, which I call the Hierarchical Dirichlet Process Hidden Markov Model With Local Transitions, which achieves the goal of simultaneously modeling correlations between contexts and components by assigning each a location

in a metric space and promoting transitions between states that are near each other. I present a Gibbs sampling scheme for inference in this model, employing an augmented data representation to simplify the relevant conditional distributions. I give a intuitive interpretation of the augmented representation by casting the discrete time chain as a continuous time chain in which durations are not observed, and in which some jump attempts fail and are never observed. By tying the success probability of a jump between two states to the distance between them, the first successful (and therefore observed) jump is more likely to be to a nearby state. I refer to this representation as a Markov Process With Failed Jump Attempts. I test this model on both synthetic and real data, including a natural language data set drawn from a corpus of biological research articles, in which the goal is inferences about the semantic scope of assertions about biological processes implicated in cancer (to, e.g., species, organ sites, gene variants, etc.). There, the latent states are sets of entities in the scope, and the data is raw text. It is presumed that succesive assertions in a paper apply in similar scopes.

Chapter 1

A Hierarchical Dirichlet Process Hidden Markov Model With "Local" Transitions (HDP-HMM-LT)

I describe a generalization of the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM; Teh et al. (2006)) which introduces a notion of latent similarity between pairs of hidden states, such that transitions are a priori more likely to occur between states with similar emission distributions. This is achieved by placing a similarity kernel on the space of state parameters, and scaling transition probabilities by the similarity between states. I refer to this model as the Hierarchical Dirichlet Process Hidden Markov Model with Local Transitions (HDP-HMM-LT). Although this achieves the goal of selectively increasing the probability of transitions between similar states, inference is made more complicated since the posterior measure over transition distributions is no longer a Dirichlet Process, due to the heterogenous scale parameters of the Gamma distributed unnormalized weights. I present an alternative representation of this process that facilitates inference by casting the discrete time chain as a continuous time Markov Process in which: (1) some jump

attempts fail, (2) the probability of success is proportional to the similarity between the source and destination states, (3) only successful jumps are observed, and (4) the time elapsed between jumps, as well as the number of unsuccessful jump attempts, are latent variables that are sampled during MCMC inference. By introducing these auxiliary latent variables, all conditional distributions in the model are members of an exponential family, admitting exact Gibbs sampling, with the exception of the parameters of the similarity kernel. The choice of similarity kernel is application-specific, but I present results for an exponential (a.k.a. Laplacian) similarity kernel with a single decay parameter whose conditional posterior density is log-concave, and hence admits Adaptive Rejection Sampling (Gilks and Wild, 1992).

The motivating domain for this model is natural language text, in which sentences in a document are arranged in such a way that the sets of relevant entities in successive sentences have a high degree of overlap, even when they are not identical. The goal is to model the entity set in a sentence using a binary vector, indicating which entities are present in the context, and to constrain the dynamics governing latent state transitions so that transitions between similar entity sets are *a priori* more likely, but where the presence or absence of an entity depends on the state of multiple entities in the previous sentence. The latter property makes an ordinary factorial HMM undesireable.

1.1 Transition Dynamics in the HDP-HMM

The conventional HDP-HMM (Teh et al., 2006) is based on a Hierarchical Dirichlet Process defined as follows:

Each of a countably infinite set of states, indexed by j, receives a location θ_j in emission parameter space, Ω, according to base measure H. A top-level weight distribution, $\boldsymbol{\beta}$, is drawn from a stick-breaking process with parameter $\gamma > 0$, so that state j has overall weight

β_j, and emission distribution parameterized by θ_j.

$$\theta_j \overset{i.i.d.}{\sim} H \tag{1.1}$$

$$\boldsymbol{\beta} \sim GEM(\gamma) \tag{1.2}$$

The actual transition distribution from state j, denoted by $\boldsymbol{\pi}_j$ is then drawn from a DP with concentration α and base measure $\boldsymbol{\beta}$:

$$\boldsymbol{\pi}_j \overset{i.i.d}{\sim} DP(\alpha\boldsymbol{\beta}) \qquad j = 1, 2, \ldots \tag{1.3}$$

The hidden state sequence is then generated according to the π_j. Let z_t be the index of the chain's state at time t. Then we have

$$z_t \,|\, z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \boldsymbol{\pi}_{z_{t-1}} \qquad t = 1, 2, \ldots, T \tag{1.4}$$

where T is the length of the data sequence.

Finally, the emission distribution for state j is a function of θ_j, so that we have

$$y_t \,|\, z_t, \theta_{z_t} \sim F(\theta_{z_t}) \tag{1.5}$$

A shortcoming of this model is that the generative process does not take into account the fact that the set of source states is the same as the set of destination states: that is, the distribution $\boldsymbol{\pi}_j$ has an element which corresponds to state j. Put another way, there is no special treatment of the diagonal of the transition matrix, so that self-transitions are no more likely *a priori* than transitions to any other state. The Sticky HDP-HMM of Fox, et al. (2008) addresses this issue by adding an extra mass of κ at location j to the base measure

of the DP that generates $\boldsymbol{\pi}_j$. That is, they replace (1.3) with

$$\boldsymbol{\pi}_j \sim DP(\alpha\boldsymbol{\beta} + \kappa\delta_j). \tag{1.6}$$

An alternative model is presented by Johnson et al. (2013), wherein state duration distributions are modeled separately, and ordinary self-transitions are ruled out. In both of these models, auxiliary latent variables are introduced to simplify conditional posterior distributions and facilitate Gibbs sampling. However, while both of these models have the useful property that self-transitions are treated as "special", they contain no notion of similarity for pairs of states that are not identical: in both cases, when the transition matrix is integrated out, the prior probability of transitioning to state j' depends only on the top-level stick weight associated with state j', and not on the identity or parameters of the previous state j.

1.2 An HDP-HMM With Local Transitions

The goal is to add to the transition model the concept of a transition to a "nearby" state, where nearness of j and j' is possibly a function of θ_j and $\theta_{j'}$. In order to accomplish this, we first consider an alternative construction of the transition distributions, based on the Normalized Gamma Process representation of the Dirichlet Process (Ferguson, 1973).

1.2.1 A Normalized Gamma Process representation of the HDP-HMM

Define a random measure, $\mu = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$, where

$$\pi_j \overset{ind}{\sim} \mathcal{G}(w_j, 1) \tag{1.7}$$

$$T = \sum_{j=1}^{\infty} \pi_j \tag{1.8}$$

$$\tilde{\pi}_j = \frac{\pi_j}{T} \tag{1.9}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.10}$$

and subject to the constraint that $\sum_{j \geq 1} w_j < \infty$, which ensures that $T < \infty$ almost surely. As shown by Paisley et al. (2011), for fixed $\{w_j\}$ and $\{\theta_j\}$, μ is distributed as a Dirichlet Process with base measure $\mathbf{w} = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$. If we draw $\boldsymbol{\beta}$ from a stick-breaking process and then draw a series $\{\mu_m\}_{m=1}^{M}$ of i.i.d. random measures from the above process, setting $\mathbf{w} = \alpha\boldsymbol{\beta}$ for some $\alpha > 0$, then this defines a Hierarchical Dirichlet Process. If, moreover, there is one μ_m associated with every state j, then we obtain the HDP-HMM.

We can thus write

$$\boldsymbol{\beta} \sim \mathsf{GEM}(\gamma) \tag{1.11}$$

$$\theta_j \overset{i.i.d.}{\sim} H \tag{1.12}$$

$$\pi_{jj'} \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.13}$$

$$T_j = \sum_{j'=1}^{\infty} \pi_{jj'} \tag{1.14}$$

$$\tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{T_j}, \tag{1.15}$$

where γ and α are prior concentration hyperparameters for the two DP levels, where

$$p(z_t \mid z_{t-1}, \boldsymbol{\pi}) = \tilde{\pi}_{z_{t-1} z_t} \tag{1.16}$$

and the observed data $\{y_t\}_{t \geq 1}$ distributed as

$$y_t \mid z_t \overset{ind}{\sim} F(\theta_{z_t}) \tag{1.17}$$

for some family, F of probability measures indexed by values of θ.

1.2.2 Promoting "Local" Transitions

In the preceding formulation, the θ_j and the $\pi_{jj'}$ are independent conditioned on the top-level measure. Our goal is to relax this assumption, in order to allow for prior knowledge that certain "locations", θ_j, are more likely than others to produce large weights. This can be accomplished by letting the rate parameter in the distribution of the $\pi_{jj'}$ be a function of θ_j and $\theta_{j'}$. Let $\Phi : \Omega \times \Omega \to [0, \infty)$ represent a "similarity function", and define a collection of random variables $\{\phi_{jj'}\}_{j,j' \geq 1}$ according to

$$\phi_{jj'} = \phi(\theta_j, \theta'_j) \tag{1.18}$$

We can then generalize (1.11)-(1.15) to

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma) \tag{1.19}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.20}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, \phi_{jj'}^{-1}) \tag{1.21}$$

$$T_j = \sum_{j'=1}^{\infty} \pi_{jj'} \tag{1.22}$$

$$\tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{T_j} \tag{1.23}$$

so that the expected value of $\pi_{jj'}$ is $\alpha\beta_{j'}\phi_{jj'}$. Since a similarity between one object and another should not exceed the similarity between an object and itself, we will assume that $\phi_{jj'} \leq B < \infty$ for all j and j', with equality holding iff $j = j'$. Moreover, there is no loss of generality by taking $B = 1$, since a constant rescaling of $\phi_{jj'}$ gets absorbed in the normalization.

The above model is equivalent to simply drawing the $\pi_{jj'}$ as in (1.11) and scaling each one by $\phi_{jj'}$ prior to normalization.

Unfortunately, this formulation complicates inference significantly, as the introduction of non-constant rate parameters to the prior on $\boldsymbol{\pi}$ destroys the conjugacy between $\boldsymbol{\pi}$ and $\mathbf{z}$, and worse, the conditional likelihood function for $\boldsymbol{\pi}$ contains an infinite sum of the elements in a row, rendering all entries within a row mutually dependent.

1.2.3 The HDP-HMM-LT as a continuous-time Markov Jump Process with "failed" jumps

We can gain stronger intuition, as well as simplify posterior inference, by re-casting the HDP-HMM-LT described in the last section 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,

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma) \tag{1.24}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.25}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.26}$$

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.

I 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 $\mathcal{P}\text{ois}(\pi_{jj'})$ distributed. Conditioned on $N_{jj'}$, $n_{jj'}$ will be successful, where

$$n_{jj'} \,|\, N_{jj'} \sim \mathcal{B}\text{inom}(N_{jj'}, \phi_{jj'}) \tag{1.27}$$

It is easy to show (and well known) that the marginal distribution of $n_{jj'}$ is $\mathcal{P}\text{ois}(\pi_{jj'}\phi_{jj'})$, and the marginal distribution of $\tilde{q}_{jj'} := N_{jj'} - n_{jj'}$ is $\mathcal{P}\text{ois}(\pi_{jj'}(1-\phi_{jj'}))$. 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 τ_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

$$\tau_t \,|\, z_{t-1} \sim \mathcal{E}\text{xp}(T_{z_{t-1}}) \tag{1.28}$$

where τ_t is independent of z_t.

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

$$\tilde{q}_{j't} \,|\, z_{t-1} \sim \mathcal{P}\text{ois}(\tau_t \pi_{z_{t-1}j'}(1-\phi_{z_{t-1}j'})) \tag{1.29}$$

Define the following additional variables

$$\mathcal{T}_j = \{t \,|\, z_{t-1} = j\} \tag{1.30}$$

$$q_{jj'} = \sum_{t\in\mathcal{T}_j} \tilde{q}_{j't} \tag{1.31}$$

$$u_j = \sum_{t\in\mathcal{T}_j} \tau_t \tag{1.32}$$

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 τ_t with $t \in \mathcal{T}_j$ are i.i.d. $\mathcal{E}\text{xp}(T_j)$, we get the marginal distribution

$$u_j \,|\, \mathbf{z}, \boldsymbol{\pi\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, T_j) \tag{1.33}$$

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

$$q_{jj'} \overset{ind}{\sim} \mathcal{P}\text{ois}(u_j \pi_{jj'}(1 - \phi_{jj'})). \tag{1.34}$$

Thus if we marginalize out the individual τ_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

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}) = \left(\prod_{t=1}^{T} p(z_t \,|\, z_{t-1}) \right) \prod_j p(u_j \,|\, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}) \prod_{j'} p(q_{jj'} \,|\, u_j \pi_{jj'}, \phi_{jj'}) \tag{1.35}$$

$$= \left(\prod_t \frac{\pi_{z_{t-1} z_t} \phi_{z_{t-1} z_t}}{T_{z_{t-1}}} \right) \prod_j \frac{T_j^{n_{j\cdot}}}{\Gamma(n_{j\cdot})} u_j^{n_{j\cdot}-1} e^{-T_j u_j} \tag{1.36}$$

$$\times \prod_{j'} e^{-u_j \pi_{jj'}(1-\phi_{jj'})} u_j^{q_{jj'}} \pi_{jj'}^{q_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.37}$$

$$= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} \tag{1.38}$$

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

$$\tag{1.39}$$

$$= \prod_j \Gamma(n_{j\cdot})^{-1} u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}u_j} (q_{jj'}!)^{-1} \tag{1.40}$$

1.2.4 An HDP-HSMM-LT modification

Note that it is trivial to modify the HDP-HMM-LT to allow the number of observations generated each time a state is visited to have a distribution which is not Geometric, by simply fixing the diagonal elements of $\boldsymbol{\pi}$ to be zero, and allowing D_t observations to be emitted *i.i.d.* $F(\theta_{z_t})$ at jump t, where

$$D_t \,|\, \mathbf{z} \overset{ind}{\sim} g(\omega_{z_t}) \qquad \omega_j \overset{i.i.d}{\sim} G \tag{1.41}$$

The likelihood then includes the additional term for the D_t, and the only inference step which is affected is that instead of sampling $\mathbf{z}$ alone, we sample $\mathbf{z}$ and the D_t jointly, by defining

$$z^*_s = z_{\max\{T \,|\, s \leq \sum_{t=1}^{T} D_t\}} \tag{1.42}$$

where s ranges over the number of observations, and associating a $\mathbf{y}_s$ with each z^*_s. Inferences about $\boldsymbol{\phi}$ are not affected, since the diagonal elements are assumed to be 1 anyway.

This is the same construction used in the Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM; Johnson and Willsky (2013)). Unlike in the standard representation of the HDP-HSMM, however, there is no need to introduce additional auxiliary variables as a result of this modification, due to the presence of the (continuous) durations, $\mathbf{u}$, which were already needed to account for the normalization of the $\boldsymbol{\pi}$.

1.2.5 Summary

I have defined the following augmented generative model for the HDP-H(S)MM-LT:

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma) \tag{1.43}$$

$$\theta_j \overset{i.i.d}{\sim} H \tag{1.44}$$

$$\pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathcal{G}(\alpha\beta_{j'}, 1) \tag{1.45}$$

$$z_t \,|\, z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta} \sim \sum_j \left(\frac{\pi_{z_{t-1}j}\phi_{z_{t-1}j}}{\sum_{j'} \pi_{z_{t-1}j'}\phi_{z_{t-1}j'}} \right) \delta_j \tag{1.46}$$

$$u_j \,|\, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, \sum_{j'} \pi_{jj'}\phi_{jj'}) \tag{1.47}$$

$$q_{jj'} \,|\, \mathbf{u}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{P}\text{ois}(u_j(1-\phi_{jj'})\pi_{jj'}) \tag{1.48}$$

$$\mathbf{y}_t \,|\, \mathbf{z}, \boldsymbol{\theta} \sim F(\theta_{z_t}) \tag{1.49}$$

If we are using the HSMM variant, then we simply fix π_{jj} to 0 for each j, draw

$$\omega_j \overset{i.i.d}{\sim} G \tag{1.50}$$

$$D_t \,|\, \mathbf{z} \overset{ind}{\sim} g(\omega_{z_t}), \tag{1.51}$$

for chosen G and g, set

$$z^*_s = z_{\max\{T \,|\, s \leq \sum_{t=1}^{T} D_t\}} \tag{1.52}$$

and replace (1.49) with

$$\mathbf{y}_s \,|\, \mathbf{z}, \boldsymbol{\theta} \sim F(\theta_{z^*_s}) \tag{1.53}$$

1.3 Inference

I develop a Gibbs sampling algorithm based on the Markov Process with Failed Jumps representation, augmenting the data with the duration variables $\mathbf{u}$, the failed jump attempt count matrix, $\mathbf{Q}$, as well as additional auxiliary variables which we will define below. In this representation the transition matrix is not modeled directly, but is a function of the unscaled transition matrix π and the similarity matrix $\boldsymbol{\phi}$. The full set of variables is partitioned into three blocks: $\{\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi}\}$, $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda\}$, and $\{\boldsymbol{\theta}\}$, where Λ represents a set of auxiliary variables that will be introduced below. The variables in each block are sampled jointly conditioned on the other two blocks.

Since we are representing the transition matrix of the Markov chain explicitly, we approximate the stick-breaking process that produces $\boldsymbol{\beta}$ using a finite Dirichlet distribution with a number of components larger than we expect to need, forcing the remaining components to have zero weight. Let J indicate the maximum number of states. Then, we approximate (1.24) with

$$\boldsymbol{\beta} \,|\, \gamma \sim \text{Dirichlet}(\gamma/J, \ldots, \gamma/J) \tag{1.54}$$

This distribution converges weakly to the Stick-Breaking Process as $J \to \infty$. In practice, J is large enough when the vast majority of the probability mass in $\boldsymbol{\beta}$ is allocated to a strict subset of components, or when the latent state sequence $\mathbf{z}$ never uses all J available states, indicating that the data is well described by a number of states less than J.

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

The joint conditional over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given $\mathbf{z}$, $\mathbf{u}$, $\mathbf{Q}$, Λ and $\boldsymbol{\theta}$ will factor as

$$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda, \boldsymbol{\theta}) = p(\gamma \,|\, \Lambda) p(\alpha \,|\, \Lambda) p(\boldsymbol{\beta} \,|\, \gamma, \Lambda) p(\boldsymbol{\pi} \,|\, \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}) \tag{1.55}$$

I will derive these four factors in reverse order.

Sampling $\boldsymbol{\pi}$

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

$$p(\boldsymbol{\pi} \,|\, \boldsymbol{\beta}, \alpha) = \prod_j \prod_{j'} \Gamma(\alpha\beta_{j'})^{-1} \pi_{jj'}^{\alpha\beta_{j'}-1} \exp(-\pi_{jj'}), \tag{1.56}$$

and the likelihood given augmented data $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}\}$ given by (1.40). Combining these, we have

$$p(\boldsymbol{\pi}, \mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\beta}, \alpha, \boldsymbol{\theta}) = \prod_j u_j^{n_{j\cdot}+q_{j\cdot}-1} \prod_{j'} \Gamma(\alpha\beta_{j'})^{-1} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} e^{-(1+u_j)\pi_{jj'}} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.57}$$

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

$$p(\boldsymbol{\pi} \,|\, \mathbf{Q}, \mathbf{u}, \mathbf{Z}, \boldsymbol{\beta}, \alpha, \boldsymbol{\theta}) \propto \prod_j \prod_{j'} \pi_{jj'}^{\alpha\beta_{j'}+n_{jj'}+q_{jj'}-1} \exp(-(1+u_j)\pi_{jj'}) \tag{1.58}$$

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

$$\pi_{jj'} \,|\, n_{jj'}, q_{jj'}, \beta_{j'}, \alpha \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{1.59}$$

Sampling $\boldsymbol{\beta}$

Consider the conditional distribution of $\boldsymbol{\beta}$ having integrated out $\boldsymbol{\pi}$. The prior density of $\boldsymbol{\beta}$ from (1.54) is

$$p(\boldsymbol{\beta} \,|\, \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \prod_j \beta_j^{\frac{\gamma}{J}-1} \tag{1.60}$$

After integrating out $\boldsymbol{\pi}$ in (1.57), we have

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\beta}, \alpha, \gamma, \boldsymbol{\theta}) = \prod_{j=1}^{J} u_j^{-1} \prod_{j'=1}^{J} u^{n_{jj'}+q_{jj'}-1}(1+u_j)^{-(\alpha\beta_{j'}+n_{jj'}+q_{jj'})} \tag{1.61}$$

$$\times \frac{\Gamma(\alpha\beta_{j'}+n_{jj'}+q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}}(1-\phi_{jj'})^{q_{jj'}}(q_{jj'}!)^{-1} \tag{1.62}$$

$$= \prod_{j=1}^{J} \Gamma(n_{j\cdot})^{-1} u_j^{-1}(1+u_j)^{-\alpha} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{1.63}$$

$$\times \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'}+n_{jj'}+q_{jj'})}{\Gamma(\alpha\beta_{j'})} \phi_{jj'}^{n_{jj'}}(1-\phi_{jj'})^{q_{jj'}}(q_{jj'}!)^{-1} \tag{1.64}$$

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

$$p(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \frac{\Gamma(\alpha\beta_{j'}+n_{jj'}+q_{jj'})}{\Gamma(\alpha\beta_{j'})}. \tag{1.65}$$

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

$$p(\boldsymbol{\beta} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} \sum_{m_{jj'}=1}^{n_{jj'}} s(n_{jj'}+q_{jj'}, m_{jj'})(\alpha\beta_{j'})^{m_{jj'}} \tag{1.66}$$

where $s(m,n)$ is an unsigned Stirling number of the first kind. 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

$$p(m_{jj'} = m \,|\, \beta_{j'}, \alpha, n_{jj'}, q_{jj'}) = \frac{s(n_{jj'}+q_{jj'}, m)\alpha^m \beta_{j'}^m}{\sum_{m'=0}^{n_{jj'}+q_{jj'}} s(n_{jj'}+q_{jj'}, m')\alpha^{m'}\beta_{j'}^{m'}} \tag{1.67}$$

for integer m ranging between 0 and $n_{jj'}+q_{jj'}$. Note that $s(n,0) = 0$ if $n > 0$, $s(0,0) = 1$

and $s(0, m) = 0$ if $m > 0$. Then, we have joint distribution

$$p(\boldsymbol{\beta}, \mathbf{M} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \alpha, \gamma, \boldsymbol{\theta}) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}-1} \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'}) \alpha^{m_{jj'}} \beta_{j'}^{m_{jj'}} \tag{1.68}$$

which yields (1.66) when marginalized over $\mathbf{M}$. Again discarding constants in $\boldsymbol{\beta}$ and re-grouping yields

$$p(\boldsymbol{\beta} \,|\, \mathbf{M}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\theta}, \alpha, \gamma) \propto \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}+m_{\cdot j}-1} \tag{1.69}$$

which is Dirichlet:

$$\boldsymbol{\beta} \,|\, \mathbf{M}, \gamma \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \ldots, \frac{\gamma}{J} + m_{\cdot J}) \tag{1.70}$$

Sampling α and γ

Assume that α and γ have Gamma priors, with

$$p(\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) \tag{1.71}$$

$$p(\gamma) = \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \gamma^{a_{\gamma-1}} \exp(-b_\gamma \gamma) \tag{1.72}$$

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

$$p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\theta}) = \frac{\Gamma(\gamma)}{\Gamma(\frac{\gamma}{J})^J} \alpha^{m_{\cdot\cdot}} \prod_{j=1}^{J} \beta_j^{\frac{\gamma}{J}+m_{\cdot j}-1} \Gamma(n_{j\cdot})^{-1} u_j^{-1} (1+u_j)^{-\alpha} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{1.73}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.74}$$

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

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M} \,|\, \alpha, \gamma, \boldsymbol{\theta}) = \alpha^{m_{\cdot\cdot}} e^{-\sum_{j''} \log(1+u_{j''})\alpha} \frac{\Gamma(\gamma)}{\Gamma(\gamma + m_{\cdot\cdot})} \tag{1.75}$$

$$\times \prod_j \frac{\Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})\Gamma(n_{j\cdot})} u_j^{-1} \left(\frac{u_j}{1+u_j}\right)^{n_{j\cdot}+q_{j\cdot}} \tag{1.76}$$

$$\times \prod_{j'=1}^{J} s(n_{jj'} + q_{jj'}, m_{jj'}) \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}} (q_{jj'}!)^{-1} \tag{1.77}$$

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

$$p(\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta}) \propto \alpha^{a_\alpha + m_{\cdot\cdot}} \exp(-(b_\alpha + \sum_j \log(1+u_j))\alpha) \tag{1.78}$$

and

$$p(\gamma \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) \frac{\Gamma(\gamma) \prod_{j=1}^{J} \Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})^J \Gamma(\gamma + m_{\cdot\cdot})} \tag{1.79}$$

So we see that

$$\alpha \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta} \sim \mathcal{G}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1+u_j)) \tag{1.80}$$

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

$$p(r_j = r \,|\, m_{\cdot j}, \gamma) = \frac{\Gamma(\frac{\gamma}{J})}{\Gamma(\frac{\gamma}{J} + m_{\cdot j})} s(m_{\cdot j}, r) \left(\frac{\gamma}{J}\right)^r \qquad r = 1, \ldots, m_{\cdot j} \tag{1.81}$$

$$p(t \,|\, m_{\cdot\cdot}\gamma) = \frac{\Gamma(\gamma + m_{\cdot\cdot})}{\Gamma(\gamma)\Gamma(m_{\cdot\cdot})} t^{\gamma - 1} (1-t)^{m_{\cdot\cdot} - 1} \qquad t \in (0, 1) \tag{1.82}$$

so that

$$p(\gamma, \mathbf{r}, t \,|\, \mathbf{M}) \propto \gamma^{a_\gamma - 1} \exp(-b_\gamma \gamma) t^{\gamma - 1} (1-t)^{m_{\cdot\cdot} + q_\cdot - 1} \prod_{j=1}^{J} s(m_{\cdot j} + q_j, r_j) \left(\frac{\gamma}{J}\right)^{r_j} \tag{1.83}$$

and

$$p(\gamma \,|\, \mathbf{r}, t) \propto \gamma^{a_\gamma + r_\cdot - 1} \exp(-(b_\gamma - \log(t))\gamma), \tag{1.84}$$

which is to say

$$\gamma \,|\, \mathbf{r}, t, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \boldsymbol{\theta} \sim \mathcal{G}(a_\gamma + r_\cdot, b_\gamma - \log(t)) \tag{1.85}$$

Summary

I have made the following additional assumptions about the generative model in this section:

$$\gamma \sim \mathcal{G}(a_\gamma, b_\gamma) \qquad \alpha \sim \mathcal{G}(a_\alpha, b_\alpha) \tag{1.86}$$

The joint conditional over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given $\mathbf{z}$, $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$, t and $\boldsymbol{\theta}$ factors as

$$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{r}, t, \boldsymbol{\theta}) = p(\gamma \,|\, \mathbf{r}, t) p(\alpha \,|\, \mathbf{u}, \mathbf{M}) p(\boldsymbol{\beta} \,|\, \gamma, \mathbf{M}) p(\boldsymbol{\pi} \,|\, \alpha, \boldsymbol{\beta}, \mathbf{z}, \mathbf{u}, \mathbf{Q}) \tag{1.87}$$

where

$$\gamma \,|\, \mathbf{r}, t \sim \mathcal{G}(a_\gamma + r_\cdot, b_\gamma - \log(t)) \tag{1.88}$$

$$\alpha \,|\, \mathbf{u}, \mathbf{M} \sim \mathcal{G}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1 + u_j)) \tag{1.89}$$

$$\boldsymbol{\beta} \,|\, \gamma, \mathbf{M} \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \ldots, \frac{\gamma}{J} + m_{\cdot J}) \tag{1.90}$$

$$\pi_{jj'} \,|\, \alpha, \beta_{j'}, \mathbf{z}, \mathbf{u}, \mathbf{Q} \overset{ind}{\sim} \mathcal{G}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j) \tag{1.91}$$

1.3.2 Sampling z and the auxiliary variables

The hidden state sequence, $\mathbf{z}$, is sampled jointly with the auxiliary variables, which consist of $\mathbf{u}$, $\mathbf{M}$, $\mathbf{Q}$, $\mathbf{r}$ and t. The joint conditional distribution of these variables is defined directly by the generative model:

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q}, \mathbf{M}, \mathbf{r}, t \,|\, \boldsymbol{\pi}, \boldsymbol{\beta}, \alpha, \gamma, \boldsymbol{\theta}) = p(\mathbf{z} \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{u} \,|\, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{Q} \,|\, \mathbf{u}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\mathbf{M} \,|\, \mathbf{z}, \mathbf{Q}, \alpha, \boldsymbol{\beta}) \tag{1.92}$$

$$\times\, p(\mathbf{r} \,|\, \gamma, \mathbf{M}) p(t \,|\, \gamma, \mathbf{M}) \tag{1.93}$$

Since we are representing the transition matrix explicitly, we can sample the entire sequence $\mathbf{z}$ at once with the forward-backward algorithm, as in an ordinary HMM (or, if we are employing the HSMM variant described in Sec. 1.2.4, then we can use the modified message passing scheme for HSMMs described by Johnson and Willsky (2013)). Having done this, we can sample $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$ and t from their forward distributions. To summarize, we have

$$u_j \,|\, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{G}(n_{j\cdot}, \sum_{j'} \pi_{jj'} \phi_{jj'}) \tag{1.94}$$

$$q_{jj'} \,|\, u_j, \pi_{jj'}, \phi_{jj'} \overset{ind}{\sim} \mathcal{P}\text{ois}(u_j (1 - \phi_{jj'}) \pi_{jj'}) \tag{1.95}$$

$$m_{jj'} \,|\, n_{jj'}, q_{jj'}, \beta_{j'}, \alpha \overset{ind}{\sim} \frac{\Gamma(\alpha\beta_j)}{\Gamma(\alpha\beta_j + n_{jj'} + q_{jj'})} \sum_{m=1}^{n_{jj'}+q_{jj'}} s(n_{jj'} + q_{jj'}, m) \alpha^m \beta_{j'}^m \delta_m \tag{1.96}$$

$$r_j \,|\, m_{\cdot j}, \gamma \overset{ind}{\sim} \frac{\Gamma(\frac{\gamma}{J})}{\Gamma(\frac{\gamma}{J} + m_{\cdot j})} \sum_{r=1}^{m_{j\cdot}} s(m_{\cdot j}, r) \left(\frac{\gamma}{J}\right)^r \delta_r \tag{1.97}$$

$$t \,|\, \gamma, \mathbf{M} \sim \mathcal{B}\text{eta}(\gamma, m_{\cdot\cdot}) \tag{1.98}$$

1.3.3 Sampling state and emission parameters

The state parameters, $\boldsymbol{\theta}$, influence the transition matrix, $\boldsymbol{\pi}$ and the auxiliary vector q through the similarity matrix matrix $\boldsymbol{\phi}$, and also control the emission distributions. We have likelihood factors

$$p(\mathbf{z}, \mathbf{Q} \,|\, \boldsymbol{\theta}) \propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} \tag{1.99}$$

$$p(\mathbf{Y} \,|\, \mathbf{z}, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t; \theta_{z_t}) \tag{1.100}$$

where proportionality is with respect to variation in $\boldsymbol{\theta}$.

The parameter space for the hidden states, the associated prior H on $\boldsymbol{\theta}$, and the similarity function Φ, is application-specific, but we consider here the case where a state, θ_j, consists of a finite-length binary vector, motivated by the application of inferring the set of relevant entities in each sentence of a text document.

Let $\theta_j = (\theta_{j1}, \ldots, \theta_{jD})$, with $\theta_{jd} = 1$ indicating presence of feature d in context state j, and $\theta_{jd} = 0$ indicating absence. Of course, in this case, the set of possible states is finite, and so on its face it may seem that a nonparametric model is unnecessary. However, if D is reasonably large, it is likely that most of the 2^D possible states are vanishingly unlikely (and, in fact, the number of observations may well be less than 2^D), and so we would like a model that encourages the selection of a sparse set of states. Moreover, there may be more than one state with the same θ, but with different transition dynamics.

Sampling $\boldsymbol{\theta}$

In principle, H can be 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.

We require a similarity function, $\Phi(\theta_j, \theta_{j'})$, which varies between 0 to 1, and is equal to 1 if and only if $\theta_j = \theta_{j'}$. A natural choice in this setting is the Laplacian kernel:

$$\phi_{jj'} = \Phi(\theta_j, \theta_{j'}) = \exp(-\lambda \Delta_{jj'}) \tag{1.101}$$

where $\Delta_{jj'd} = |\theta_{jd} - \theta_{j'd}|$, $\Delta_{jj'} = \sum_{d=1}^{D} \Delta_{jj'}$ is the Hamming distance between θ_j and $\theta_{j'}$, 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

$$\phi_{jj'-d} = \exp(-\lambda(\Delta_{jj'} - \Delta_{jj'd})) \tag{1.102}$$

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

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

$$\frac{p(\mathbf{z}, \mathbf{Q} \,|\, \theta_{jd} = 1, \boldsymbol{\theta} \setminus \theta_{jd})}{p(\mathbf{z}, \mathbf{Q} \,|\, \theta_{jd} = 0, \boldsymbol{\theta} \setminus \theta_{jd})} \propto \prod_{j' \neq j} \frac{e^{-\lambda(n_{jj'} + n_{j'j})\left|1 - \theta_{j'd}\right|} \left(1 - \phi_{jj'-d} e^{-\lambda \left|1 - \theta_{j'd}\right|}\right)^{q_{jj'} + q_{j'j}}}{e^{-\lambda(n_{jj'} + n_{j'j})\left|\theta_{j'd}\right|} \left(1 - \phi_{jj'-d} e^{-\lambda \left|\theta_{j'd}\right|}\right)^{q_{jj'} + q_{j'j}}} \tag{1.103}$$

$$= e^{-\lambda(c_{jd0} - c_{jd1})} \prod_{j' \neq j} \left(\frac{1 - \phi_{jj'-d} e^{-\lambda}}{1 - \phi_{jj'-d}} \right)^{(-1)^{\theta_{j'd}} (q_{jj'} + q_{j'j})} \tag{1.104}$$

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,

$$c_{jd0} = \sum_{\{j' \,|\, \theta_{j'd} = 0\}} n_{jj'} + n_{j'j} \qquad c_{jd1} = \sum_{\{j' \,|\, \theta_{j'd} = 1\}} n_{jj'} + n_{j'j} \tag{1.105}$$

Therefore, we can Gibbs sample θ_{jd} from its conditional posterior Bernoulli distribution

given the rest of $\boldsymbol{\theta}$, where we compute the Bernoulli parameter via the log-odds

$$\log\left(\frac{p(\theta_{jd}=1\,|\,\mathbf{Y},\mathbf{z},\mathbf{Q},\boldsymbol{\theta}\setminus\theta_{jd})}{p(\theta_{jd}=0\,|\,\mathbf{Y},\mathbf{z},\mathbf{Q},\boldsymbol{\theta}\setminus\theta_{jd})}\right)=\log\left(\frac{p(\theta_{jd}=1)p(\mathbf{z},\mathbf{Q}\,|\,\theta_{jd}=1,\boldsymbol{\theta}\setminus\theta_{jd})p(\mathbf{Y}\,|\,\mathbf{z},\theta_{jd}=1,\boldsymbol{\theta}\setminus\theta_{jd})}{p(\theta_{jd}=0)p(\mathbf{z},\mathbf{Q}\,|\,\theta_{jd}=0,\boldsymbol{\theta}\setminus\theta_{jd})p(\mathbf{Y}\,|\,\mathbf{z},\theta_{jd}=0,\boldsymbol{\theta}\setminus\theta_{jd})}\right) \tag{1.106}$$

$$=\log\left(\frac{\mu_d}{1-\mu_d}\right)+(c_{jd1}-c_{jd0})\lambda+\sum_{j'\neq j}(-1)^{\theta_{j'd}}(q_{jj'}+q_{j'j})\log\left(\frac{1-\phi_{jj'}^{(-d)}e^{-\lambda}}{1-\phi_{jj'}^{(-d)}}\right) \tag{1.107}$$

$$+\sum_{\{t\,|\,z_t=j\}}\log\left(\frac{f(\mathbf{y}_t;\theta_{jd}=1,\theta_j\setminus\theta_{jd})}{f(\mathbf{y}_t;\theta_{jd}=0,\theta_j\setminus\theta_{jd})}\right) \tag{1.108}$$

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

$$f(\mathbf{y}_t;\theta_j)=g(\mathbf{y}_t;\mathbf{W}^{\mathsf{T}}\theta_j) \tag{1.109}$$

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

$$g(\mathbf{y}_t;\mathbf{W}^{\mathsf{T}}\theta_j)=\prod_{k=1}^{K}g_k(y_{tk};\mathbf{w}_k\cdot\theta_j) \tag{1.110}$$

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

$$\log\left(\frac{f(\mathbf{y}_t;\theta_{jd}=1,\theta_j\setminus\theta_{jd})}{f(\mathbf{y}_t;\theta_{jd}=0,\theta_j\setminus\theta_{jd})}\right)=\sum_{k=1}^{K}\log\left(\frac{g_k(y_{tk};x_{tk}^{(-d)}+w_{dk})}{g_k(y_{tk};x_{tk}^{(-d)})}\right) \tag{1.111}$$

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

$$\log\left(\frac{g_k(y_{tk};x_{tk}^{(-d)}+w_{dk})}{g_k(y_{tk};x_{tk}^{(-d)})}\right)=-w_{dk}(y_{tk}-x_{tk}^{(-d)}+\frac{1}{2}w_{dk}) \tag{1.112}$$

Sampling $\boldsymbol{\mu}$

Sampling the μ_d is straightforward with a Beta prior. Suppose

$$\mu_d \overset{ind}{\sim} \mathcal{B}\text{eta}(a_\mu, b_\mu) \tag{1.113}$$

Then, conditioned on $\boldsymbol{\theta}$ the μ_d are independent with

$$\mu_d \,|\, \boldsymbol{\theta} \sim \mathcal{B}\text{eta}(a_\mu + \sum_j \theta_{jd}, b_\mu + \sum_j (1 - \theta_{jd})) \tag{1.114}$$

Sampling λ

The parameter λ governs the connection between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Writing (1.99) in terms of λ and the difference matrix $\boldsymbol{\Delta} = (\Delta_{jj'})_{1 \le j,j' \le J}$ gives

$$p(\mathbf{z}, \mathbf{Q} \,|\, \lambda, \boldsymbol{\theta}) \propto \prod_j \prod_{j'} e^{-\lambda \Delta_{jj'} n_{jj'}} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.115}$$

Put an $\mathcal{E}\text{xp}(b_\lambda)$ prior on λ, so that

$$p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}) \propto e^{-(b_\lambda + \sum_j \sum_{j'} \Delta_{jj'} n_{jj'})\lambda} \prod_j \prod_{j'} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.116}$$

This density is log-concave, with

$$-\frac{d^2 \log(p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}))}{d\lambda^2} = \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \frac{\Delta_{jj'}^2 q_{jj'} e^{\lambda \Delta_{jj'}}}{(e^{\lambda \Delta_{jj'}} - 1)^2} > 0 \tag{1.117}$$

and so we can use Adaptive Rejection Sampling (Gilks and Wild, 1992) to sample from it. The relevant h and h', representing the log density and its first derivative, respectively, are

$$h(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'})\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} q_{jj'} \log(1 - e^{-\lambda \Delta_{jj'}}) \tag{1.118}$$

$$h'(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'}) + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \frac{q_{jj'} \Delta_{jj'}}{e^{\lambda \Delta_{jj'}} - 1} \tag{1.119}$$

Sampling W

Conditioned on the state matrix $\boldsymbol{\theta}$ and the data matrix $\mathbf{Y}$, the weight matrix $\mathbf{W}$ can be sampled as well using standard methods for Bayesian regression problems. For example, suppose that the weights are *a priori* i.i.d. Normal:

$$p(\mathbf{W}) = \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(w_{dk} \,|\, 0, \sigma_0^2) \tag{1.120}$$

and the likelihood is

$$g_k(y; x) = \mathcal{N}(y \,|\, x, 1) \tag{1.121}$$

Then it is a standard result from Bayesian linear modeling that

$$p(\mathbf{W} \,|\, \boldsymbol{\theta}, \mathbf{Y}) = \prod_{k=1}^{K} \mathcal{N}\left(\left(\sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\theta} \right)^{-1} \boldsymbol{\theta}^\mathsf{T} \mathbf{y}_k, \sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\theta} \right) \tag{1.122}$$

If one or more output features, say $\mathbf{y}_k$, is binary, we can adopt a probit model where we introduce a latent data vector $\mathbf{y}_k^*$ for each such k, and assume

$$p(\mathbf{y}_k^* \,|\, \mathbf{x}_k) = \prod_t \mathcal{N}(y_{tk}^* \,|\, x_{tk}, 1) \tag{1.123}$$

and

$$y_{tk} = \begin{cases} 0, & y^*_{tk} \leq 0 \\ 1, & y^*_{tk} > 0 \end{cases} \tag{1.124}$$

And so, after marginalizing over $\mathbf{y}^*_k$

$$p(\mathbf{y}_k \,|\, \mathbf{x}_k) = \prod_{t=1}^{T} F(x_{tk})^{y_{tk}} (1 - F(x_{tk}))^{1-y_{tk}} \tag{1.125}$$

where F is the standard Normal CDF, since

$$\int_0^\infty dy^*_{tk} \mathcal{N}(y^*_{tk} \,|\, x_{tk}, 1) = \int_{-x_{tk}}^\infty dy^*_{tk} \mathcal{N}(y^*_{tk} \,|\, 0, 1) = 1 - F(-x_{tk}) = F(x_{tk}) \tag{1.126}$$

Then, conditioned on x_{tk} and y_{tk}, we can sample y^*_{tk} from a Normal distribution left- or right-truncated at 0:

$$p(y^*_{tk} \,|\, x_{tk}, y_{tk}) = \begin{cases} \mathcal{N}(x_{tk}, 1) I(y^*_{tk} \leq 0), & y_{tk} = 0 \\ \mathcal{N}(x_{tk}, 1) I(y^*_{tk} > 0), & y_{tk} = 1 \end{cases} \tag{1.127}$$

Conditioned on the y^*_{tk} and $\boldsymbol{\theta}$, the weights are distributed as in (1.122).

Summary

I have made the following assumptions about the representation of the hidden states and observed data in this subsection: (1) $\boldsymbol{\theta}$ consists of D binary features (2) the similarity function Φ is the Laplacian kernel with respect to Hamming distance with decay parameter λ, and (3) $\mathbf{Y}$ consists of K continuous or binary features associated with each time step t.

In addition, we make the following distributional assumptions:

$$\mu_d \overset{i.i.d}{\sim} \mathcal{B}\text{eta}(a_\mu, b_\mu) \tag{1.128}$$

$$\lambda \sim \mathcal{E}\text{xp}(b_\lambda) \tag{1.129}$$

$$\theta_{jd} \,|\, \boldsymbol{\mu} \overset{ind}{\sim} \mathcal{B}\text{ern}\,(\mu_d) \tag{1.130}$$

$$\mathbf{W} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}) \tag{1.131}$$

$$y^*_{tk} \,|\, \mathbf{W}, \mathbf{z}, \boldsymbol{\theta} \overset{ind}{\sim} \mathcal{N}(x_{tk}, 1) \tag{1.132}$$

$$y_{tk} = \begin{cases} y^*_{tk}, & \text{if } k \text{ is a continuous feature} \\ \mathbb{I}(y^*_{tk} > 0) & \text{if } k \text{ is a binary feature} \end{cases} \tag{1.133}$$

where we have defined

$$x_{tk} = \mathbf{w}_k \cdot \theta_{z_t} \tag{1.134}$$

I introduce Gibbs blocks corresponding to (1) each θ_{jd} individually, (2) the vector $\boldsymbol{\mu}$, (3) the decay parameter λ, (4) the weight matrix $\mathbf{W}$, and (5) the latent data $\mathbf{Y}^*$ associated with binary features. We have

$$\theta_{jd} \,|\, \boldsymbol{\theta} \setminus \theta_{jd}, \mathbf{z}, \mathbf{Q}, \boldsymbol{\mu}, \lambda, \mathbf{W}, \mathbf{Y}^* \sim \mathcal{B}\text{ern}\left(\frac{e^{\zeta_{jd}}}{1 + e^{\zeta_{jd}}}\right) \tag{1.135}$$

$$\mu_d \,|\, \boldsymbol{\theta}, \ldots \overset{ind}{\sim} \mathcal{B}\text{eta}(a_\mu + \sum_j \theta_{jd}, b_\mu + \sum_j (1 - \theta_{jd})) \tag{1.136}$$

$$p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}, \ldots) \propto e^{-(b_\lambda + \sum_j \sum_{j'} \Delta_{jj'} n_{jj'})\lambda} \prod_j \prod_{j'} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{1.137}$$

$$\mathbf{w}_k \,|\, \boldsymbol{\theta}, \mathbf{Y}^*, \ldots \overset{ind}{\sim} \mathcal{N}((\sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta})^{-1} \boldsymbol{\theta}^\mathsf{T} \mathbf{y}^*_k, \sigma_0^2 \mathbf{I} + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta}) \tag{1.138}$$

$$\mathbf{y}^*_{tk} \,|\, \mathbf{X}, \mathbf{Y}, \ldots \overset{ind}{\sim} \begin{cases} \mathcal{N}(x_{tk}, 1)\mathbb{I}(y^*_{tk} \leq 0), & y_{tk} = 0 \\ \mathcal{N}(x_{tk}, 1)\mathbb{I}(y^*_{tk} > 0), & y_{tk} = 1 \end{cases} \tag{1.139}$$

where $\Delta_{jj'} = \left|\left|\theta_j - \theta'_j\right|\right|_{L_1}$ and

$$\zeta_{jd} = \log\left(\frac{\mu_d}{1-\mu_d}\right) + (c_{jd1} - c_{jd0})\lambda + \sum_{j' \neq j}(-1)^{\theta_{j'd}}(q_{jj'} + q_{j'j}) \log\left(\frac{1 - \phi_{jj'}^{(-d)} e^{-\lambda}}{1 - \phi_{jj'}^{(-d)}}\right)$$
$$- \sum_{\{t \,|\, z_t = j\}} \sum_{k=1}^{K} w_{dk}(y^*_{tk} - x^{(-d)}_{tk} + \frac{1}{2} w_{dk}) \tag{1.140}$$

All distributions can be sampled from directly except for λ, which requires Adaptive Rejection Sampling, with the equations

$$h(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'})\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} q_{jj'} \log(1 - e^{-\lambda \Delta_{jj'}}) \tag{1.141}$$

$$h'(\lambda) = -(b_\lambda + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \Delta_{jj'} n_{jj'}) + \sum_{\{(j,j') \,|\, \Delta_{jj'} > 0\}} \frac{q_{jj'} \Delta_{jj'}}{e^{\lambda \Delta_{jj'}} - 1} \tag{1.142}$$

1.3.4 Generalization to Categorical-Valued θ

If we relax the assumption that the θ_j are binary vectors and instead allow each θ_{jd} to take on one in an arbitrary set of discrete values — i.e., to be categorically distributed rather than Bernoulli distributed — then the inference steps involving θ change somewhat.

I will continue to assume that the Φ function decays exponentially as a function of the number of component-wise differences between a pair of states; that is, $\Delta_{jj'd} = 0$ iff $\theta_{jd} = \theta_{j'd}$ and $\Delta jj'd = 1$ otherwise, $\Delta_{jj'} = \sum_d \Delta_{jj'd}$, and $\Phi(\theta_j \theta_{jj'}) = e^{-\lambda \Delta_{jj'}}$. Therefore the process of sampling λ is unchanged.

In place of Beta priors on each θ_{jd}, we use Chinese Restaurant Process priors, with concentration parameter $\alpha_d^{(\theta)}$. In the case that the number of categorical values is known in advance, this can be replaced by a Dirichlet prior, but I present the more flexible case here.

The weight matrix $\mathbf{W}$ must be expanded to allow for distinct weights associated with each possible value of the θ_{jd}. Rather than the matrix given by $(w_{dk})_{d=1,\ldots,D,k=1,\ldots,K}$, we now

need a set of weights, (w_{sdk}), where s indexes the categorical values that θ_{jd} can take. With a CRP prior, there are infinitely many s. We can use a "dummy variable" representation of $\boldsymbol{\theta}_j$, where we define S_d to be the number of realized states for dimension d and a 1 in position $\sum_{d'<d} S_{d'} + s$ indicates that $\theta_{jd} = s$. There will thus be D entries equal to 1, with the remaining entries equal to zero. We can then represent the weight matrix $\mathbf{W}$ as stacked block matrix, where each block is $S_d \times K$, and there is one block for each d. In practice we only need to instantiate a new block when some θ_{jd} is assigned to a "new table" in the CRP metaphor, so that the dimension of $\mathbf{W}$ is $\sum_d S_d \times K$. Then we have

$$\mathbf{y}_t^* \sim \mathcal{N}(\mathbf{w}^{(b)} + \boldsymbol{\theta}_{z_t}\mathbf{W}, \boldsymbol{\Sigma}) \tag{1.143}$$

where $\mathbf{w}^{(b)}$ is a K-dimensional bias vector with a separate Normal prior, $\mathbf{W}$ is the weight matrix as defined just above, z_t is the state indicator for time t, and $\boldsymbol{\Sigma}$ is a $K \times K$ noise covariance matrix.

Sampling $\boldsymbol{\theta}$

As before, the conditional posterior for θ_{jd} is proportional to the product of three terms: the prior (now a CRP), the likelihood of all successful and failed transitions to and from state j, and the likelihood of the observation sequence.

Under the CRP, the prior probability $P(\theta_{jd} = s)$ is proportional to the number of other $j' \neq j$ such that $\theta_{j'd} = s$ where this count is positive; and proportional to $\alpha_j^{(\theta)}$ otherwise. Let

$$\tilde{n}_{ds}^{-j} = \sum_{j' \neq j} I(\theta_{jd} = s), \quad s = 1, \ldots, S_d \tag{1.144}$$

be these counts, where we assume that there are S_d distinct values taken by the θ_{jd} for a

particular d. Then

$$p(\theta_{jd} = s \,|\, \boldsymbol{\theta}_d^{-j}) \propto \begin{cases} \tilde{n}_{ds}^{-j} & s = 1, \ldots, S_d \\ \alpha_d^{(\theta)} & s = S_d + 1 \end{cases} \tag{1.145}$$

The transition component of the likelihood is as in the binary case:

$$p(\mathbf{z}, \mathbf{Q} \,|\, \theta_{jd}, \boldsymbol{\theta}_d^{-j}) \propto e^{-\lambda \sum_{j'} \Delta_{jj'}(n_{jj'}+n_{j'j})} \prod_{j' \neq j} (1 - e^{-\lambda \Delta_{jj'}(q_{jj'}+q_{j'j})}) \tag{1.146}$$

$$\propto e^{-\lambda \sum_{j'} I(\theta_{jd} \neq \theta_{j'd})(n_{jj'}+n_{j'j})} \prod_{j' \neq j} (1 - a \cdot e^{-\lambda I(\theta_{jd} \neq \theta_{j'd})(q_{jj'}+q_{j'j})}) \tag{1.147}$$

where a is a constant in θ_{jd}, defined as $e^{-\lambda \Delta_{jj'-d}(q_{jj'd}+q_{j'jd})}$. Taking a log yields

$$\log p(\mathbf{z}, \mathbf{Q} \,|\, \theta_{jd}, \boldsymbol{\theta}_d^{-j}) = -\lambda \sum_{\{j':\ \theta_{jd} \neq \theta_{j'd}\}} (n_{jj'}+n_{j'j}) + \sum_{j' \neq j} \log(1 - a \cdot e^{-\lambda I(\theta_{jd} \neq \theta_{j'd})(q_{jj'}+q_{j'j})}) \tag{1.148}$$

The emission component of the likelihood is given for each t by

$$p(\mathbf{y}_t^* \,|\, \boldsymbol{\theta}_{z_t}, \mathbf{W}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp(-\frac{1}{2}(\mathbf{y}_t^* - \mathbf{w}^{(b)} - \boldsymbol{\theta}_{z_t}\mathbf{W})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t^* - \mathbf{w}^{(b)} - \boldsymbol{\theta}_{z_t}\mathbf{W})) \tag{1.149}$$

Assuming a diagonal covariance matrix, isolating $\theta_{j,d}$, and taking a log, this becomes, for each k and t,

$$\log p(y_{tk}^* \,|\, \theta_{z_t,d}, \boldsymbol{\theta}_{z_t}^{-d}, w_k^{(b)}, \sigma_k^2) = -\frac{1}{2\sigma_k^2} \left(y_{tk}^* - w_k^{(b)} - \boldsymbol{\theta}_{z_t} \mathbf{w}_k \right)^2 \tag{1.150}$$

$$\propto -\frac{1}{2\sigma_k^2} (\sum_{d'} w_{\theta_{z_t d'}, d', k} - (y_{tk}^* - w_k^{(b)}))^2 \tag{1.151}$$

$$\propto -\frac{1}{2\sigma_k^2} (w_{\theta_{z_t d}, d, k} - (y_{tk}^* - w_k^{(b)} - \sum_{d' \neq d} w_{\theta_{z_t, d'}, d', k}))^2 \tag{1.152}$$

For a particular j and d, the full emission log likelihood is a sum of terms like the above, over all t such that $z_t = j$.

Sampling W

Having expanded $\boldsymbol{\theta}$ to a dummy variable representation, and having constructed a stacked block form of $\mathbf{W}$, we can sample each column of $\mathbf{W}$ from a conditional posterior multivariate Normal just as in the binary case.

Bibliography

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2001). The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348.

Johnson, M. J. and Willsky, A. S. (2013). Bayesian nonparametric hidden semi-markov models. *The Journal of Machine Learning Research*, 14(1):673–701.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Paisley, J., Wang, C., and Blei, D. (2011). The discrete infinite logistic normal distribution. *arXiv preprint arXiv:1103.4789*.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. In Solla, S., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 554–560. MIT Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).