Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 20, 2019

QUESTION 1

Introduction

Probabilisitc modeling follows a simple pipeline as shown in Figure 1. We have some questions, so we build a model to answer them, and use the available data to infer the parameters of our model. We then use our model to make predictions, or explore data further, and hence evaluate our model. The bottleneck of this pipeline is **inference**. More often than not, inference of real-world models is computationally intractable, and require some form of approximation. One such approximation technique is **variational** inference. Variational inference converts an inference problem to an optimization problem. The idea is to consider a variational distribution $q(\mathbf{z}|\lambda)$ over all the hidden variables and bring it as close as possible to the actual posterior distribution $p(\mathbf{z}|\theta)$. To do so, we minimize the KL divergence between the actual posterior and our variational approximation. Minimizing divergence can be shown to be the same as maximising the **ELBO** **E**vidence **L**ower **BO**und given by,

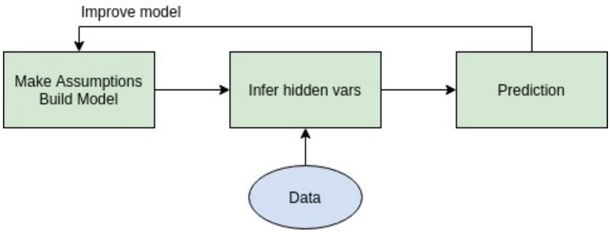

Figure 1: Probabilisitc pipeline

$$\text{ELBO} = L(\lambda) = \mathbb{E}_q[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z})]$$

The recipe for VI is as follows -

Build model $p(\mathbf{X},\mathbf{Z})$ → Choose approx. $q(\mathbf{z}|\lambda)$ → Compute ELBO → Take derivatives and optimize

This approach too has a bottleneck - the calculation of ELBO, which requires computing intractable expectations for most models. There are some model specific tricks (like Jaakola-Jordan's logistic likelihood trick) to make the computation easier, in addition to assuming simpler structures (like mean-field VI) for the variational distribution. We could also find tractable model-specific bounds for ELBO. However, these tricks don't generalize well across other models. We can't really go on deriving tricks for all models. Moreover, referring to our pipeline, we can't let our inference techniques determine what models to choose. Liabilities in computation of inference should in no way influence or restrict our choice of models. The paper addresses this problem by proposing a **model-agnostic** trick to make VI computations easier, thereby enabling us to freely explore complex models. The goal of the paper is to help users make easy iterations in the probabilistic pipeline - continuously trying new innovative models and improving upon them, instead of getting stuck in tedious inference calculations. The freedom that the paper provides to users is what I like the most about this paper.

The Black Box VI

The core idea is - What if we take derivatives first and then compute expectations? The Black-box VI recipe is as follows -

Build model $p(\mathbf{X}, \mathbf{Z})$ → Choose approx. $q(\mathbf{z}|\lambda)$ → Compute derivatives of ELBO → Compute expectations

Via some calculations, we can show that the gradient of ELBO with respect to variational parameters λ has the following form -

$$\nabla_\lambda L = \mathbb{E}_q \left[\nabla_\lambda \log q(\mathbf{z}|\lambda) \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}|\lambda)\right)\right]$$

$$\downarrow$$

$$\text{Monte-carlo approximation}$$

$$\downarrow$$

$$\nabla_\lambda L \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(\mathbf{z}_s|\lambda) \left(\log p(\mathbf{x}, \mathbf{z}_s) - \log q(\mathbf{z}_s|\lambda)\right) \quad \mathbf{z}_s \sim q(\mathbf{z}|\lambda)$$

There are no tedious assumptions on the model; It need not be continuous. It need not be differentiable. We just need to ensure that $\log p(\mathbf{x}, \mathbf{z})$ can be computed. Since $q(\mathbf{z}|\lambda)$ is in our hands, we could choose it to be such that we can sample from it and that it is differentiable. So this is not a problem too. One advantage is that we can actually **re-use** our variational distributions. Since the gradient computations do not involve any model-specific attributes (black box!), we could just collect some $q_i(\mathbf{z}|\lambda)$ distributions, compute gradients, sample from them, store in a library, and try it out on various models as and when needed.

Instead of computing expectations, we could use Monte carlo approximation and arrive at noisy unbiased gradients. Once we have all the gradients, we make gradient descent steps in the variational parameter space, ie, $\lambda^{(t+1)} = \lambda^{(t)} + \rho^{(t)} \nabla_\lambda L^{(t)}$

Issues with variance

Since VI is based on minimizing KL divergence, it is bound to have variance-related issues. Specifically, VI underestimates variance of true posterior. To see this, say true post. $p(z|\theta)$ is variadic, and shoots up in some region. Since $KL(q||p) = \int q \log(q/p)$, our optimization could do away with giving a small value to q, which also minimizes $KL(q||p)$, therein underestimating the variance of p.

In BBVI, since we are using Monte-carlo sampling, we end up with very noisy gradients. The steps we take are not so optimal more often than not, and this leads to very high variance. The time taken for convergence is also long, since the step size is small owing to the noisy nature of gradients. To put simply, basic BBVI fails if variance is not controlled. The paper proposes two methods to control variance of gradients. They are explained below.

Rao- Blackwellization

Suppose we want to compute the expectation of a random variable. If we could collect information that we already know of the random variable and condition the expectation on this information, we could possibly reduce the variance in the expectation. Mathematically, say, $\mathbf{T} = \mathbb{E}[J(\mathbf{X}, \mathbf{Y})]$ and $\hat{J}(\mathbf{X}) = \mathbb{E}[J(\mathbf{X}, \mathbf{Y})|\mathbf{X}]$. Then, using law of iterated expectations and law of total variance,

$$\mathbb{E}\left[\hat{J}(\mathbf{X})\right] = \mathbb{E}\left[\mathbb{E}\left[J(\mathbf{X}, \mathbf{Y})|\mathbf{X}\right]\right] = \mathbb{E}\left[J(\mathbf{X}, \mathbf{Y})\right] = T$$

$$\text{var}(J(\mathbf{X}, \mathbf{Y})) = \mathbb{E}\left[\text{var}(J(\mathbf{X}, \mathbf{Y})|\mathbf{X})\right] + \text{var}\left(\hat{J}(\mathbf{X})\right) \implies \text{var}(J(\mathbf{X}, \mathbf{Y})) > \text{var}\left(\hat{J}(\mathbf{X})\right)$$

Effectively, we have a proxy for computation of $\mathbf{T}$. Instead of computing expectation of $J(\mathbf{X},\mathbf{Y})$, use expectation of $\hat{J}(\mathbf{X})$, which is of lower variance. This is Rao-Blackwellization method.

In our case, using mean-field assumption, we can factorize and show that,

$$\mathbb{E}\left[J(\mathbf{X},\mathbf{Y})\,|\mathbf{X}\right]=\mathbb{E}_y\left[J(x,y)\right]$$

Essentially, we'd need to integrate out some variables to compute the conditional expectations, which would thereby reduce variance. So, in case we seek ELBO gradient with respect to λ_i, we could integrate out other factors by iteratively taking expectations, arriving at the final form -

$$\nabla_{\lambda_i}L=\mathbb{E}_{q(i)}\left[\nabla_{\lambda_i}\log q\left(\mathbf{z}_i|\lambda_i\right)\left(\log p_i\left(\mathbf{x},\mathbf{z}_{(i)}\right)-\log q\left(\mathbf{z}_i|\lambda_i\right)\right)\right]$$

Here, $z_{(i)}$ is the markov blanket of ith factor, p_i includes terms from ith factor. Importantly, this form is model-agnostic too. There are no model-specific conditional expectations! Experiments conducted by the authors showed great reduction in variance and time by incorporating Rao-Blackwellized gradients.

Control Variates

Again, the idea is to find a proxy for our computation, which gives the same expectation but with lesser variance. Control variates (Ross, 2002) are a family of functions that satisfy the conditions for this proxy. Define

$$\hat{f}(z)=f(z)-a\left(h(z)-\mathbb{E}\left[h(z)\right]\right)\implies\mathbb{E}\left[\hat{f}\right]=\mathbb{E}\left[f\right]\quad\text{and}$$
$$\operatorname{var}\left(\hat{f}\right)=\operatorname{var}(f)+a^2\left(\operatorname{var}(h)\right)-2a\operatorname{Cov}(f,h)$$

Here $\hat{f}$ are the control variates. We can minimize the variance of $\hat{f}$ by tweaking a. Note that its really important that we pick h that is correlated to f, thereby providing some additional information about f. Only then we can reduce variance of f.

For our ELBO gradient case, we could use $h=\nabla_\lambda\log q(\mathbf{z}|\lambda)$. Note that its expectation is zero. Also, we are in accordance with the model-agnostic flavor as h doesn't have any model-specific terms. We compute the optimum a^* that minimizes variance of $\hat{f}$ using this h, and applying Rao-Blackwellization too, we end up with

$$\nabla_{\lambda_i}L=\frac{1}{S}\sum_{s=1}^{S}\nabla_{\lambda_i}\log q_i\left(\mathbf{z}_s|\lambda_i\right)\left(\log p_i\left(\mathbf{x},\mathbf{z}_s\right)-\log q_i\left(\mathbf{z}_s|\lambda_i\right)-a_i^*\right)\quad\mathbf{z}_s\sim q_{(i)}(\mathbf{z}|\lambda)$$

Using the above Monte Carlo gradients, authors improve upon basic BBVI to arrive at **BBVI - II**. The authors achieved very good results, trying various complex models on longitudinal healthcare data. As expected, BBVI-II outperforms other standard methods such as Gibbs Sampling, mean-field VI etc. Note that the authors have used AdaGrad method for setting learning rates instead of Robbins-Monro rates. AdaGrad rates ensure that when variance in gradient is high, the learning rate is low and vice-versa.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

QUESTION 2

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 20, 2019

The generative story is as follows :

1. For each topic k in 1...K, draw $\boldsymbol{\phi}_k \sim \text{Dir}\left(\boldsymbol{\beta}\right)$
2. For each document d in 1...D, draw

$$\begin{aligned} \boldsymbol{\theta}_d &\sim \text{Dir}\left(\boldsymbol{\alpha}\right) \\ \mathbf{z}_{dn} &\sim \text{mult}\left(\boldsymbol{\theta}_d\right) \quad \forall \quad \text{n in 1...N}_d \\ \mathbf{w}_{dn} &\sim \text{mult}\left(\boldsymbol{\phi}_{\mathbf{z}_{dn}}\right) \forall \quad \text{n in 1...N}_d \\ \bar{\mathbf{z}_d} &= \frac{1}{N_d}\sum_{n=1}^{N_d}\mathbf{z}_{dn} \end{aligned}$$

3. For each pair of documents d, d' in 1...D,

$$A_{dd'} = \text{Bern}\left(\sigma\left(\bar{\mathbf{z}_d}^T\mathbf{H}\bar{\mathbf{z}_{d'}}\right)\right)$$

Here, we have extended the LDA model first to compute a size K vector $\bar{\mathbf{z}}_d$ wherein each z_{dk} is equal to the frequency of a particular topic in the document. Then, for each pair of documents (d, d'), we compute a similarity score $\bar{\mathbf{z}}_d^T\mathbf{H}\bar{\mathbf{z}}_{d'}$, and bring it to the range [0, 1] using sigmoid function, which can then be used as parameter of Bernoulli distribution to draw the binary link indicator $A_{dd'}$. Matrix $\mathbf{H}$ is a $K \times K$ matrix, wherein each element is represented by $\boldsymbol{\eta}_{a,b}$. Then,

$$\bar{\mathbf{z}}_d^T\mathbf{H}\bar{\mathbf{z}}_{d'} = \sum_{b=1}^{K}\sum_{a=1}^{K}\bar{z}_{da}\eta_{ab}\bar{z}_{d'b}$$

Each term $\bar{z}_{da}\eta_{ab}\bar{z}_{d'b}$ in the summation has the following meaning - if the frequency of topic a in document d is $\bar{z}_{da}$, and the frequency of topic b in document d' is $\bar{z}_{db}$, how much does this situation contribute to the formation of link between document d and d'. If η_{ab} is large, then situation is quite favourable for a link formation.

Note that in this model we assume asymmetric impact, ie, η_{ab} need not be same as η_{ba}. If binary links are symmetric, then we can take $\eta_{ab} = \eta_{ba}$, so lesser parameters to estimate. Also, we expect η_{ii} to take large values, as two documents containing same topics are more likely to be linked.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 20, 2019

QUESTION 3

Let us consider the complete likelihood of the model -

$$p\left(\mathbf{A},\mathbf{Z},\boldsymbol{\eta},\boldsymbol{\pi}|\boldsymbol{\alpha},a,b\right)=p\left(\mathbf{A}|\boldsymbol{\eta},\mathbf{z}\right)p\left(\mathbf{z}|\boldsymbol{\pi}\right)p\left(\boldsymbol{\eta}|a,b\right)p\left(\boldsymbol{\pi}|\boldsymbol{\alpha}\right)$$
$$=\prod_{n=1}^{N}\prod_{m=1}^{n}\left(\eta_{\mathbf{z}_n\mathbf{z}_m}\right)^{A_{nm}}\left(1-\eta_{\mathbf{z}_n\mathbf{z}_m}\right)^{1-A_{nm}}\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{\mathbb{I}[z_n=k]}\prod_{k=1}^{K}\prod_{l=1}^{k}\frac{\eta_{kl}^{a-1}\left(1-\eta_{kl}\right)^{b-1}}{\text{Beta}\left(a,b\right)}\text{Dir}\left(\boldsymbol{\pi}|\boldsymbol{\alpha}\right)$$

To compute the conditional posterior of various parameters, we consider the terms in the complete log likelihood where the particular parameters occur. First, let us compute the conditional posterior of $\boldsymbol{\pi}$.

$$\log p\left(\boldsymbol{\pi}|\mathbf{A},\mathbf{Z},\boldsymbol{\eta},\boldsymbol{\alpha},a,b\right)=\sum_{k=1}^{K}\left(\sum_{n=1}^{N}\mathbb{I}\left[\mathbf{z}_n=k\right]+\alpha-1\right)\log\pi_k+\text{const.}$$

This is of the form of a log of Dirichlet distribution. So, conditional posterior of $\boldsymbol{\pi}$ is a Dirichlet distribution $\text{Dir}\left(\left\{\sum_{n=1}^{N}\mathbb{I}\left[\mathbf{z}_n=k\right]+\alpha\right\}_{k=1}^{K}\right)$.

Now let us compute the conditional posterior of η_{kl}.

$$\log p\left(\eta_{kl}|\mathbf{A},\mathbf{Z},\boldsymbol{\eta}_{-kl},\boldsymbol{\alpha},a,b\right)=\left(a-1+N_{kl\mathbf{A}}\right)\log\eta_{kl}+\left(b-1+N_{kl}-N_{kl\mathbf{A}}\right)\log\left(1-\eta_{kl}\right)+\text{const.}$$

where $N_{kl}=\sum_{n=1}^{N}\sum_{m=1}^{n}\left(\mathbb{I}\left[\mathbf{z}_n=k,\mathbf{z}_m=l\right]+\mathbb{I}\left[\mathbf{z}_n=l,\mathbf{z}_m=k\right]\right)$ and
$N_{kl\mathbf{A}}=\sum_{n=1}^{N}\sum_{m=1}^{n}\left(\mathbb{I}\left[\mathbf{z}_n=k,\mathbf{z}_m=l\right]+\mathbb{I}\left[\mathbf{z}_n=l,\mathbf{z}_m=k\right]\right)A_{nm}$. Note that here we are using the fact that the matrix is symmetric. So, $\eta_{kl}=\eta_{lk}$. The above log likelihood is same as that of a Beta distribution. Hence the conditional posterior of η_{kl} is given by, Beta$(a+N_{kl\mathbf{A}},b+N_{kl}-N_{kl\mathbf{A}})$.

Let's compute the conditional posterior of $\mathbf{z}_n$.

$$\log p\left(\mathbf{z}_n=k|\mathbf{A},\mathbf{Z}_{-n},\boldsymbol{\eta},\boldsymbol{\alpha},a,b\right)=\sum_{m=1}^{N}\left(A_{nm}\right)\log\eta_{k\mathbf{z}_m}+\left(1-A_{nm}\right)\log\left(1-\eta_{k\mathbf{z}_m}\right)+\log\pi_k+\text{const.}$$

$$\implies p\left(\mathbf{z}_n=k|\mathbf{A},\mathbf{Z}_{-n},\boldsymbol{\eta},\boldsymbol{\alpha},a,b\right)\propto\pi_k\times\prod_{m=1}^{N}\left(\eta_{k\mathbf{z}_m}\right)^{A_{nm}}\left(1-\eta_{k\mathbf{z}_m}\right)^{1-A_{nm}}$$

$$\implies p\left(\mathbf{z}_n=k|\mathbf{A},\mathbf{Z}_{-n},\boldsymbol{\eta},\boldsymbol{\alpha},a,b\right)=\frac{\pi_k\prod_{m=1}^{N}\left(\eta_{k\mathbf{z}_m}\right)^{A_{nm}}\left(1-\eta_{k\mathbf{z}_m}\right)^{1-A_{nm}}}{\sum_{l=1}^{K}\pi_l\prod_{m=1}^{N}\left(\eta_{l\mathbf{z}_m}\right)^{A_{nm}}\left(1-\eta_{l\mathbf{z}_m}\right)^{1-A_{nm}}}=\pi_k'$$

Note that here we're taking m from 1 to N instead of m from 1 to n, because we have to consider $\eta_{\mathbf{z}_m k}$. As in, we need to pick up $\eta_{k\mathbf{z}_1},\eta_{k\mathbf{z}_2},\cdots,\eta_{k\mathbf{z}_n},$, and $\eta_{\mathbf{z}_{n+1}k},\eta_{\mathbf{z}_{n+2}k},\cdots,\eta_{\mathbf{z}_N k}$. But, $\eta_{\mathbf{z}_{n+1}k}=\eta_{k\mathbf{z}_{n+1}}$, assuming symmetry.

The Gibbs Sampler for this problem is as follows

1. Initialize $\boldsymbol{\pi}$, $\mathbf{Z}$, $\boldsymbol{\eta}$ randomly as $\boldsymbol{\pi}^{(0)}$, $\mathbf{Z}^{(0)}$, $\boldsymbol{\eta}^{(0)}$ respectively. Set t = 1.

2. $$\boldsymbol{\pi}^{(t)} \sim \text{Dir}\left(\left\{\sum_{n=1}^{N} \mathbb{I}\left[\mathbf{z}_n^{(t-1)} = k\right] + \alpha\right\}_{k=1}^{K}\right)$$

3. $\mathbf{z}_n^{(t)} \sim \text{mult}\left(\pi_1'^{(t)}, \pi_2'^{(t)}, \cdots, \pi_K'^{(t)}\right)$ for all n in 1...N

4. For k in 1..N,
   $\eta_{kl}^{(t)} \sim \text{Beta}\Big(a + N_{kl\mathbf{A}}^{(t)}, b + N_{kl} - N_{kl\mathbf{A}}^{(t)}\Big)$ for l in 1..k

5. Go to Step-2 if not converged.

QUESTION 4

Note that $p\left(\theta \in A_i | G\right) = G\left(A_i\right)$. Consider the fixed partition of space Ω as given in the problem. So,

$$\begin{aligned}
p\left(G|\theta_1, \theta_2, \cdots, \theta_N\right) \propto p\left(\theta_1, \theta_2, \cdots, \theta_N|G\right) p\left(G|G_0\right) &= \prod_{n=1}^{N} p\left(\theta_n|G\right) p\left(G|G_0\right) \\
&= \prod_{n=1}^{N}\prod_{i=1}^{K} G\left(A_i\right)^{\mathbb{I}\left[\theta_n \in A_i\right]} \operatorname{Dir}\left(\alpha G_0\left(A_1\right), \alpha G_0\left(A_2\right), \cdots, \alpha G_0\left(A_K\right)\right) \\
&= \prod_{i=1}^{K} G\left(A_i\right)^{\sum_{n=1}^{N}\mathbb{I}\left[\theta_n \in A_i\right]} \operatorname{Dir}\left(\alpha G_0\left(A_1\right), \alpha G_0\left(A_2\right), \cdots, \alpha G_0\left(A_K\right)\right)
\end{aligned}$$

Let $\delta_{\theta_n}\left(A_i\right) = \mathbb{I}\left[\theta_n \in A_i\right]$. Using the Dirichlet-Multinomial conjugacy, we get,

$$p\left(G|\theta_1, \theta_2, \cdots, \theta_N\right) = \operatorname{Dir}\left(\alpha G_0\left(A_1\right) + \sum_{n=1}^{N}\delta_{\theta_n}\left(A_1\right), \alpha G_0\left(A_2\right) + \sum_{n=1}^{N}\delta_{\theta_n}\left(A_2\right), \cdots, \alpha G_0\left(A_K\right) + \sum_{n=1}^{N}\delta_{\theta_n}\left(A_K\right)\right)$$

Since this is true for all finite partitions of Ω, we find that the posterior of G is also a Dirichlet process. To compute the base distribution, let us find the expectation of posterior.

$$\mathbb{E}\left[G\left(A\right)|\theta_1, \theta_2, \cdots, \theta_N\right] = \frac{\alpha G_0\left(A\right) + \sum_{n=1}^{N}\delta_{\theta_n}\left(A\right)}{\alpha + N} = \left(\frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{n=1}^{N}\delta_{\theta_n}\right)\left(A\right)$$

$$\implies \text{base distribution of posterior} = \frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{n=1}^{N}\delta_{\theta_n}$$

The concentration parameter of DP posterior is equal to $\alpha + N$. Hence, we get

$$G|\theta_1, \theta_2, \cdots, \theta_N \sim \operatorname{DP}\left(\alpha + N, \frac{\alpha}{\alpha+N}G_0 + \frac{1}{\alpha+N}\sum_{n=1}^{N}\delta_{\theta_n}\right)$$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Suryateja B.V.
Roll Number: 160729
Date: April 20, 2019

QUESTION

5

Introduction

Suppose we consider that every document discusses a set of topics. Our task is to learn the topics. We apriori do not know how many topics are there in a document. A salient approach to this problem is non-parametric Bayesian modeling, wherein we use a Dirichlet Process (DP) prior, which offers probabilities to infinite topics apriori. Using the document data, we could pick the required finite topics a posteriori.

Now, it is natural to expect that various documents share topics. For example, a document on Beta Processes and a document on Gaussian Processes might share a topic like 'Stochastic'. Learning topics for each document seperately denies us an opportunity to share data across documents. We could've learned about the topic 'Stochastic' better if we **jointly** learned on both documents. Note that we don't want to club both documents into one! This destroys the individual document structure. A solution in statistics which enables joint learning while preserving the document structure is the hierarchical modeling technique.

Let's take 2 documents, and put up $G_1, G_2 \sim DP\left(\alpha_0, G_0\right)$ priors where $G_1 = \sum_{k=1}^{\infty} \pi_{1k} \delta_{\phi_{1k}}$ and $G_2 = \sum_{k=1}^{\infty} \pi_{2k} \delta_{\phi_{2k}}$. Since G_0 is continuous, none of ϕ_{1k}'s and ϕ_{2k}'s are equal, implying that there is no sharing of topics across both documents. So, we need a *discrete* G_0 with broad support. The paper **Hierarchical Dirichlet Process** proposes a straightforward solution - draw G_0 from another dirichlet process. An example generative story using HDP is given below, wherein H is a Normal-Wilshart distribution.

$$G_0 | \gamma, H \sim DP\left(\gamma, H\right) \qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

$$G_j | \alpha_0, G_0 \sim DP\left(\alpha_0, G_0\right) \quad \forall \quad j \qquad G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

$$\theta_{ji} | G_j \sim G_j$$

$$x_{ji} | \theta_{ji} \sim \mathcal{N}\left(x_{ji} | \mu_{ji}, \Sigma_{ji}\right)$$

One appealing aspect of the model is its recursiveness, as in, we could easily extend the model to another level of hierarchy by sampling H from another DP. This can be used to learn topics jointly from multiple corpora. For example, it is likely that documents from Statistics corpora and Computer Science corpora share a topic like 'Python'.

Stick Breaking Construction

How are β_k and π_{jk} related? Sethuraman suggested a stick breaking construction to compute β_k's. The target is to find a sequence of β_k's such that $\sum_{k=1}^{\infty} \beta_k = 1$. The construction is as follows - we recursively break a stick of unit length. Call this the parent stick.

$$\beta_k' \sim \text{Beta}\,(1, \gamma) \qquad\qquad \beta_k = \beta_k' \prod_{l=1}^{k-1} \left(1 - \beta_l'\right)$$

For our case with J documents (groups), we can think of recursively breaking J sticks of unit length such that $\sum_{k=1}^{\infty} \pi_{jk} = 1$ for all j. The catch is that each piece length of jth stick depends on the piece lengths in parent stick, thereby establishing a hierarchical flavor.

$$\pi_{jk}' \sim \text{Beta}\left(\alpha_0 \beta_k, \alpha_0 \left(\sum_{l=k+1}^{\infty} \beta_l\right)\right) \qquad\qquad \pi_{jk} = \pi_{jk}' \prod_{l=1}^{k-1} \left(1 - \pi_{jl}'\right)$$

The stick breaking process works for π_{jk}'s because it can be shown that $\boldsymbol{\pi}_j \sim \text{DP}\,(\alpha_0, \boldsymbol{\beta})$. Note the contrast with respect to a normal DP. We'd have drawn π_{jk}' from Beta$(1, \alpha_0)$, in which there is no sharing of values.

Chinese Restaurant Franchise

The CRP metaphor for Dirichlet Process can be extended to the Chine Restaurant Franchise metaphor to explain HDP intuitively. The core idea is that, in addition to customers going to restaurants and picking tables, we now have *tables* taking up a customer role and going to a parent restaurant. For simplicity's sake, let's take the case of 2 restaurants X and Y. A, B, C are tables in restaurant X which has 4 customers. D and E are tables in restaurant Y with 5 customers. The parent restaurant has tables p, q, and r, with the 5 tables from two restaurants being customers. The customer seating arrangement is as shown in Figure 2. Since A and D sit at table p, it means the dishes served at table A and D would be the same as p.

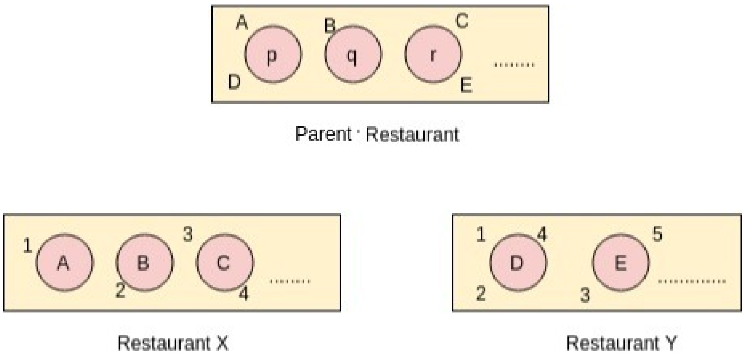

Figure 2: Chinese Restaurant Franchise when J = 2

Suppose a new customer 6 comes in restaurant X. She would sit at table A with probability $\frac{1}{4+\alpha_0}$, and eat dish p (since A sits at p in parent restaurant). Similarly, she'd sit at table B with probability $\frac{1}{4+\alpha_0}$ and eat dish q; at table C with probability $\frac{2}{4+\alpha_0}$ and eat dish p. Note that the numerator is the number of customers in restaurant X sitting at that particular table, and the denominator is α_0 plus the number of customers in the restaurant before current customer arrived. Customer 6 chooses to sit at a new table with probability $\frac{\alpha_0}{4+\alpha_0}$. Now, we could think of a waiter at the new table going to the parent restaurant, and a CRP starts at parent restaurant with the waiter as a new customer. The probabilities are as shown in Table 1

Table/Dish	Probability
p	$\frac{2}{5+\gamma}$
q	$\frac{1}{5+\gamma}$
r	$\frac{2}{5+\gamma}$
New dish	$\frac{\gamma}{5+\gamma}$

Table 1: CRP in parent restaurant

In case the waiter sits at table p, she'd call restaurant X and tell the chef to serve dish p at the new table that customer 6 chose. If the waiter chooses a new dish, say s, then this new dish s will be served to customer 6.

The process in similar in Restaurant Y too. A new customer would either sit at one of the tables, or choose to sit at a new table. If she prefers a new table, a waiter would go to the parent restaurant and choose a dish according to another CRP. Notice that there is a global set of dishes; a subset of those dishes is being served in each restaurant.

The dishes p, q, r are analogous to ϕ_k's, ie, the support of G_0. A, B, C, D, E are analogous to $\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}$ respectively. Note that θ_{11} and θ_{21} are the same, as in, the latent variables are being shared across the 2 groups. Due to such sharing of latent variables, it makes sense that we use CRF as a prior for jointly modeling multiple related datasets. The posterior predictive distribution would pool the information across multiple restaurants before assigning a table to a new customer in a restaurant.