STATS305C: Applied Statistics III

Lecture 18: Dirichlet processes

Scott Linderman

May 25, 2022

Outline

- Collapsed Gibbs sampling for Bayesian Mixture Models
- Dirichlet process mixture models and random measures
- Poisson random measures

Finite Bayesian Mixture Models

1. Sample the proportions from a Dirichlet prior with $\boldsymbol{\alpha} \in \mathbb{R}_+^K$:

$$\pi \sim \mathrm{Dir}(\boldsymbol{\alpha}) \qquad [\boldsymbol{\alpha} = \alpha \mathbf{1}_K] \tag{1}$$

2. Sample the parameters for each component:

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu) \qquad \text{for } k = 1, \ldots, K \tag{2}$$

3. Sample the assignment of each data point:

$$z_n \overset{\text{iid}}{\sim} \pi \qquad \text{for } n = 1, \ldots, N \tag{3}$$

4. Sample data points given their assignments:

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_{z_n}) \qquad \text{for } n = 1, \ldots, N \tag{4}$$

Joint distribution

- This generative model corresponds to the following factorization of the joint distribution

$$p(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K [\pi_k p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)]^{\mathbb{I}[z_n=k]} \tag{5}$$

- Let's assume an **exponential family** likelihood,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = h(\boldsymbol{x}_n) \exp \left\{ \langle t(\boldsymbol{x}_n), \boldsymbol{\theta}_k \rangle - A(\boldsymbol{\theta}_k) \right\}. \tag{6}$$

- Then assume a **conjugate prior**,

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) = \frac{1}{Z(\boldsymbol{\phi}, \nu)} \exp \left\{ \langle \boldsymbol{\phi}, \boldsymbol{\theta}_k \rangle - \nu A(\boldsymbol{\theta}_k) \right\}. \tag{7}$$

where $Z_\theta(\boldsymbol{\phi}, \nu)$ is the normalizing function.

Joint distribution

- This generative model corresponds to the following factorization of the joint distribution

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K [\pi_k p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)]^{\mathbb{I}[z_n=k]} \tag{5}$$

- Let's assume an **exponential family** likelihood,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = h(\boldsymbol{x}_n) \exp\left\{\langle t(\boldsymbol{x}_n), \boldsymbol{\theta}_k\rangle - A(\boldsymbol{\theta}_k)\right\}. \tag{6}$$

- Then assume a **conjugate prior**,

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) = \frac{1}{Z(\boldsymbol{\phi}, \nu)} \exp\left\{\langle \boldsymbol{\phi}, \boldsymbol{\theta}_k\rangle - \nu A(\boldsymbol{\theta}_k)\right\}. \tag{7}$$

where $Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)$ is the normalizing function.

"Collapsing" out variables

In some models, we can marginalize (aka *collapse* or *integrate out*) some variables to work on a lower dimensional distribution.

Typically, this is possible in models constructed with conjugate exponential family distributions.

Collapsing out the parameters in a Bayesian mixture

Let's marginalize the parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ in the exponential family mixture model,

$$p(\boldsymbol{\pi}, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\int p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N [\pi_k \, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)]^{\mathbb{I}[z_n=k]} \, \mathrm{d}\boldsymbol{\theta}_k \right] \tag{8}$$

$$\propto \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\pi_k^{N_k} \int \frac{1}{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)} \exp \left\{ \left\langle \boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \boldsymbol{\theta}_k \right\rangle - (\nu + N_k) A(\boldsymbol{\theta}_k) \right\} \mathrm{d}\boldsymbol{\theta}_k \right] \tag{9}$$

$$= \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\pi_k^{N_k} \frac{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)} \right] \tag{10}$$

where $Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)$ is the normalizing function of the conjugate prior $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$.

Collapsing out the cluster probabilities in a Bayesian mixture

While we're at it, let's marginalize the mixture proportions $\boldsymbol{\pi}$, too. The Dirichlet density is,

$$\mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{1}{Z_\pi(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad \text{where} \quad Z_\pi(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{11}$$

Plugging this in and integrating over $\boldsymbol{\pi}$ yields,

$$p(\{(z_n, \boldsymbol{x}_n)\}_{n=1}^{N} \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \left[\int \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_k^{N_k} \, \mathrm{d}\boldsymbol{\pi}\right] \left[\prod_{k=1}^{K} \frac{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)}\right] \tag{12}$$

$$= \left[\frac{Z_\pi([\alpha_1 + N_1, \ldots, \alpha_K + N_K])}{Z_\pi(\boldsymbol{\alpha})}\right] \left[\prod_{k=1}^{K} \frac{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \nu)}\right] \tag{13}$$

The collapsed distribution in a Bayesian mixture model

We'll simplify the notation by writing,

$$p(\{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha})} \prod_{k=1}^K \frac{Z_\theta(\boldsymbol{\phi}_k', \nu_k')}{Z_\theta(\boldsymbol{\phi}, \nu)} \tag{14}$$

where

$$\boldsymbol{\alpha}' = [\alpha_1 + N_1, \ldots, \alpha_K + N_K] \tag{15}$$

$$\boldsymbol{\phi}_k' = \boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n) \tag{16}$$

$$\nu_k' = \nu + N_k. \tag{17}$$

This is a **general pattern**: in exponential families, marginal likelihoods are given by ratios of posterior and prior normalizing functions.

Exponential family posterior predictive distributions

Exercise: Consider an exponential family model with a conjugate prior,

$$\theta \sim p(\theta; \phi, \nu), \qquad x_n \overset{\text{iid}}{\sim} p(x \mid \theta) \tag{18}$$

Derive an expression for the posterior predictive distribution,

$$p(x_{N+1} \mid \{x_n\}_{n=1}^N; \phi, \nu) = \int p(x_{N+1} \mid \theta)\, \underbrace{p(\theta \mid \{x_n\}_{n=1}^N; \phi, \nu)}_{1}\, \mathrm{d}\theta \tag{19}$$

in terms of the log normalizing function of the conjugate prior.

$$1 = p(\theta \mid \phi_N, \nu_N) \;; \quad \phi_N = \phi + \sum_{n=1}^N t(x_n), \quad \nu_N = \nu + N$$

$$\int p(x_{N+1} \mid \theta)\, p(\theta \mid \phi_N, \nu_N)\, d\theta = \int h(x_{N+1})\, e^{\langle t(x_{N+1}), \theta \rangle - A(\theta)} \frac{1}{Z(\phi_N, \nu_N)} e^{\langle \phi_N, \theta \rangle - \nu_N A(\theta)}\, d\theta$$

$$= h(x_{N+1}) \cdot \frac{Z(\phi_{N+1}, \nu_{N+1})}{Z(\phi_N, \nu_N)} \;; \quad \phi_{N+1} = \phi_N + t(x_{N+1}), \quad \nu_{N+1} = \nu_N + 1$$

Collapsed Gibbs for Bayesian Mixtures

Now consider the conditional distribution of z_n, holding all the other assignments fixed,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto Z_\pi(\boldsymbol{\alpha}') \prod_{k=1}^{K} Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}'_k, \nu'_k) \tag{20}$$

where $\boldsymbol{\alpha}'$, $\boldsymbol{\phi}'_k$, and ν'_k are computed with $z_n = k$. To simplify, divide by a constant w.r.t. z_n,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha}'^{(\neg n)})} \prod_{k=1}^{K} \frac{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}'_k, \nu'_k)}{Z_{\boldsymbol{\theta}}(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} \tag{21}$$

where

$$\boldsymbol{\alpha}'^{(\neg n)} = [\alpha_1 + N_1^{(\neg n)}, \ldots, \alpha_K + N_K^{(\neg n)}] \qquad \boldsymbol{\phi}_k'^{(\neg n)} = \boldsymbol{\phi} + \sum_{n' \neq n} t(\boldsymbol{x}_{n'}) \mathbb{I}[z_{n'} = k] \tag{22}$$

$$\nu_k'^{(\neg n)} = \nu + N_k^{(\neg n)} \qquad N_k^{(\neg n)} = \sum_{n' \neq n} \mathbb{I}[z_{n'} = k] \tag{23}$$

Collapsed Gibbs for Bayesian Mixtures II

- Then many terms cancel. In the first ratio,

$$\frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha}'^{(\neg n)})} = \frac{\prod_{k=1}^K \Gamma(\alpha_k') \, \Gamma(\sum_{k=1}^K \alpha_k'^{(\neg n)})}{\prod_{k=1}^K \Gamma(\alpha_k'^{(\neg n)}) \, \Gamma(\sum_{k=1}^K \alpha_k')} \propto \alpha_k'^{(\neg n)} = \alpha + N_k^{(\neg n)} \tag{24}$$

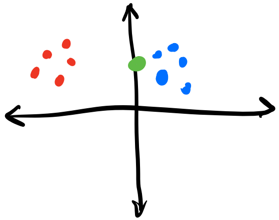
  In words, the first ratio is proportion to the size of cluster k before adding the n-th data point.

- In the second ratio, all but the k-th term in the product cancel to leave:

$$\prod_{k=1}^K \frac{Z_\theta(\boldsymbol{\phi}_k', \boldsymbol{\nu}_k')}{Z_\theta(\boldsymbol{\phi}_k'^{(\neg n)}, \boldsymbol{\nu}_k'^{(\neg n)})} = \frac{Z_\theta(\boldsymbol{\phi}_k', \boldsymbol{\nu}_k')}{Z_\theta(\boldsymbol{\phi}_k'^{(\neg n)}, \boldsymbol{\nu}_k'^{(\neg n)})} \propto p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \boldsymbol{\nu}). \tag{25}$$

  In other words, the second ratio is proportional to the *posterior predictive density*.

Collapsed Gibbs for Bayesian Mixtures II



- Then many terms cancel. In the first ratio,

$$\frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha}'^{(\neg n)})} = \frac{\prod_{k=1}^K \Gamma(\alpha_k') \, \Gamma(\sum_{k=1}^K \alpha_k'^{(\neg n)})}{\prod_{k=1}^K \Gamma(\alpha_k'^{(\neg n)}) \, \Gamma(\sum_{k=1}^K \alpha_k')} \propto \alpha_k'^{(\neg n)} = \alpha + N_k^{(\neg n)} \tag{24}$$

  In words, the first ratio is proportion to the size of cluster k before adding the n-th data point.

- In the second ratio, all but the k-th term in the product cancel to leave:

$$\prod_{k=1}^K \frac{Z_\theta(\boldsymbol{\phi}_k', \nu_k')}{Z_\theta(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} = \frac{Z_\theta(\boldsymbol{\phi}_k', \nu_k')}{Z_\theta(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} \propto p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{25}$$

  In other words, the second ratio is proportional to the *posterior predictive density*.

Collapsed Gibbs for Bayesian Mixtures III

Altogether, the conditional distribution of z_n is,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto (\alpha_k + N_k^{(\neg n)})\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu), \tag{26}$$

a function of the size of the cluster and the probability of $\boldsymbol{x}_n$ given other points in that cluster.

The infinite limit: informally speaking

- Now consider a special case where $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}_K$ and, loosely speaking, take $K \to \infty$. In this limit, we obtain a **Dirichlet process mixture model**.

- Note how the collapsed Gibbs sampling algorithm changes.

- The probability of assigning the n-th data point to a non-empty cluster is still,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \left(\frac{\alpha}{K} + N_k^{(\neg n)}\right) p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{27}$$

- But now there are only $K_{\text{used}} = \#\text{unique}(\{z_{n'}\}_{n' \neq n})$ non-empty clusters, and the remaining $K - K_{\text{used}}$ unoccupied clusters each have probability,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \frac{\alpha}{K} p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu). \tag{28}$$

The infinite limit: informally speaking

- Now consider a special case where $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}_K$ and, loosely speaking, take $K \to \infty$. In this limit, we obtain a **Dirichlet process mixture model**.
- Note how the collapsed Gibbs sampling algorithm changes.
- The probability of assigning the n-th data point to a non-empty cluster is still,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \left(\frac{\alpha}{K} + N_k^{(\neg n)}\right) p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{27}$$

- But now there are only $K_{\text{used}} = \#\mathrm{unique}(\{z_{n'}\}_{n' \neq n})$ non-empty clusters, and the remaining $K - K_{\text{used}}$ unoccupied clusters each have probability,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \tfrac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu). \tag{28}$$

The infinite limit: informally speaking

- Now consider a special case where $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}_K$ and, loosely speaking, take $K \to \infty$. In this limit, we obtain a **Dirichlet process mixture model**.
- Note how the collapsed Gibbs sampling algorithm changes.
- The probability of assigning the n-th data point to a non-empty cluster is still,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \left(\frac{\alpha}{K} + N_k^{(\neg n)}\right) p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{27}$$

- But now there are only $K_{\text{used}} = \#\mathrm{unique}(\{z_{n'}\}_{n' \neq n})$ non-empty clusters, and the remaining $K - K_{\text{used}}$ unoccupied clusters each have probability,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \tfrac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu). \tag{28}$$

The infinite limit: informally speaking II

- Since all the empty clusters are equivalent, we can combine them to get,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha})$$
$$\propto \begin{cases} \left(\frac{\alpha}{K} + N_k^{(\neg n)}\right) p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ (K - K_{\text{used}}) \frac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1, \end{cases} \tag{29}$$

where we assume that the cluster labels are permuted after each iteration so that only $k = 1, \ldots, K_{\text{used}}$ are non-empty.

- As $K \to \infty$, these updates simplify to the classic collapsed Gibbs updates for DPMMs,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha})$$
$$\propto \begin{cases} N_k^{(\neg n)}\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ \alpha\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1. \end{cases} \tag{30}$$

The infinite limit: informally speaking II

- Since all the empty clusters are equivalent, we can combine them to get,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \begin{cases} (\frac{\alpha}{K} + N_k^{(\neg n)})\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ (K - K_{\text{used}}) \frac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1, \end{cases} \tag{29}$$

  where we assume that the cluster labels are permuted after each iteration so that only $k = 1, \ldots, K_{\text{used}}$ are non-empty.

- As $K \to \infty$, these updates simplify to the classic collapsed Gibbs updates for DPMMs,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \begin{cases} N_k^{(\neg n)}\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ \alpha\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1. \end{cases} \tag{30}$$

The infinite limit: informally speaking III

As the Gibbs sampler runs, it has some probability of deleting a cluster (by removing its last data point) and some probability (determined by α) of creating a new cluster with one data point. In this sense, the model is **nonparametric**: it doesn't require you to specify K in advance.

These probabilities are *size-biased*, you're more likely to add a data point to a large cluster.

There are many other ways to arrive at the DPMM:

1. via an stochastic process on partitions called the **Chinese restaurant process (CRP)**
2. as a **random measure** on θ with a countably infinite number of weighted atoms, only a finite number of which are used.
3. via a **stick-breaking construction** to get the weights of the random measure.

Orbanz [2014] offers an accessible, book-length treatment of these important models.

Outline

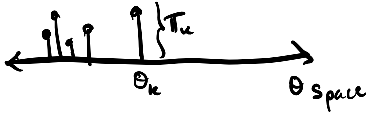
[Algorithm 8, Neal]

- Collapsed Gibbs sampling for Bayesian Mixture Models
- **Dirichlet process mixture models and random measures**
- Poisson random measures

Random measure perspective

- Another way to arrive at the DPMM is by thinking in terms of **random measures**,

$$\Theta = \sum_{k=1}^{\infty} \pi_k \, \delta_{\boldsymbol{\theta}_k} \tag{31}$$



  where $\pi_k \in \mathbb{R}_+$ are the weights and $\boldsymbol{\theta}_k$ are the locations. Since the weights and locations are random variables, Θ is a random measure.

- In particular, it's a random measure on the space of $\boldsymbol{\theta}$ with a countably infinite number of **atoms**.
- If the weights sum to one, it's a **random probability measure**.
- In Bayesian mixture models, Θ serves as the random **mixing measure** in,

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \, p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) \, \Theta(\mathrm{d}\boldsymbol{\theta}). \tag{32}$$

Random measure perspective

- Another way to arrive at the DPMM is by thinking in terms of **random measures**,

$$\Theta = \sum_{k=1}^{\infty} \pi_k \, \delta_{\boldsymbol{\theta}_k} \tag{31}$$

  where $\pi_k \in \mathbb{R}_+$ are the weights and $\boldsymbol{\theta}_k$ are the locations. Since the weights and locations are random variables, Θ is a random measure.

- In particular, it's a random measure on the space of $\boldsymbol{\theta}$ with a countably infinite number of **atoms**.

- If the weights sum to one, it's a **random probability measure**.

- In Bayesian mixture models, Θ serves as the random **mixing measure** in,

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \, p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) \, \Theta(\mathrm{d}\boldsymbol{\theta}). \tag{32}$$

Random measure perspective

- Another way to arrive at the DPMM is by thinking in terms of **random measures**,

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \tag{31}$$

  where $\pi_k \in \mathbb{R}_+$ are the weights and $\boldsymbol{\theta}_k$ are the locations. Since the weights and locations are random variables, Θ is a random measure.

- In particular, it's a random measure on the space of $\boldsymbol{\theta}$ with a countably infinite number of **atoms**.
- If the weights sum to one, it's a **random probability measure**.
- In Bayesian mixture models, Θ serves as the random **mixing measure** in,

$$p(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}) \Theta(\mathrm{d}\boldsymbol{\theta}). \tag{32}$$

Random measure perspective

- Another way to arrive at the DPMM is by thinking in terms of **random measures**,

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \qquad \Theta(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot) \tag{31}$$

  where $\pi_k \in \mathbb{R}_+$ are the weights and $\boldsymbol{\theta}_k$ are the locations. Since the weights and locations are random variables, Θ is a random measure.

- In particular, it's a random measure on the space of $\boldsymbol{\theta}$ with a countably infinite number of **atoms**.
- If the weights sum to one, it's a **random probability measure**.
- In Bayesian mixture models, Θ serves as the random **mixing measure** in,

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k p(\mathbf{x} \mid \boldsymbol{\theta}_k) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) \Theta(\mathrm{d}\boldsymbol{\theta}). \tag{32}$$

Constructing a random measure

- The simplest way to construct a random measure is to sample the locations independently,

$$\theta_k \overset{\text{iid}}{\sim} p(\theta \mid \phi, \nu). \quad = G \tag{33}$$

  Such a measure is called **homogeneous**.

- The weights cannot be independent if they're to sum to one. For finite mixtures, a simple alternative is to sample weights and then normalize them,

$$w_k \sim p(w), \qquad \pi_k = \frac{w_k}{\sum_{j=1}^{K} w_j}. \tag{34}$$

- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, 1)$, what distribution does this imply on π?

- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, \beta)$, what distribution does this imply on π?

Constructing a random measure

- The simplest way to construct a random measure is to sample the locations independently,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu). \tag{33}$$

  Such a measure is called **homogeneous**.

- The weights cannot be independent if they're to sum to one. For finite mixtures, a simple alternative is to sample weights and then normalize them,

$$w_k \sim p(w), \qquad \pi_k = \frac{w_k}{\sum_{j=1}^{K} w_j}. \tag{34}$$

- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, 1)$, what distribution does this imply on π?

- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, \beta)$, what distribution does this imply on π?

Constructing a random measure

- The simplest way to construct a random measure is to sample the locations independently,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu). \tag{33}$$

  Such a measure is called **homogeneous**.

- The weights cannot be independent if they're to sum to one. For finite mixtures, a simple alternative is to sample weights and then normalize them,

$$w_k \sim p(w), \qquad \pi_k = \frac{w_k}{\sum_{j=1}^K w_j}. \tag{34}$$

- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, 1)$, what distribution does this imply on π?
- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, \beta)$, what distribution does this imply on π?

Constructing a random measure

- The simplest way to construct a random measure is to sample the locations independently,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu). \tag{33}$$

  Such a measure is called **homogeneous**.

- The weights cannot be independent if they're to sum to one. For finite mixtures, a simple alternative is to sample weights and then normalize them,

$$w_k \sim p(w), \qquad \pi_k = \frac{w_k}{\sum_{j=1}^K w_j}. \tag{34}$$

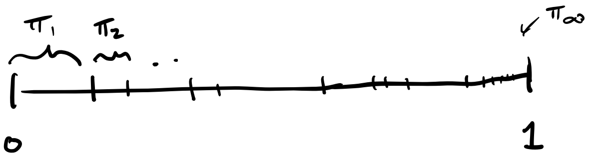
- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, 1)$, what distribution does this imply on π?
- **Question:** When $p(w) = \mathrm{Gamma}(w; \alpha, \beta)$, what distribution does this imply on π?

$= \beta^{-1} \tilde{w}_j \; ; \; \tilde{w} \sim \mathrm{Ga}(\alpha, 1)$

Constructing a random measure with an infinte number of atoms

This trick doesn't work for infinite mixtures; the sum of weights diverges almost surely.

Question: how else could you sample $\pi = (\pi_1, \pi_2, \ldots)$ so that $\sum_{k=1}^{\infty} \pi_k = 1$?



$l_1 \sim \mathrm{Beta}(1, \alpha)$

$\pi_1 = l_1$

$l_2 \sim \mathrm{Beta}(1, \alpha)$

$\pi_2 = (1 - l_1)\, l_2$

$l_3 \sim \mathrm{Beta}(1, \alpha)$

$\pi_3 = (1 - \pi_1)(1 - l_2)\, l_3$

$l_k \sim \mathrm{Beta}(1, \alpha)$

$\pi_k = l_k \prod_{j<k} (1 - l_j)$

$\pi_K = \prod_{j=1}^{K-1} (1 - l_j)$

$\vec{\pi} = (\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha \mathbb{1}_K)$

Stick breaking construction of the Dirichlet process

- **Stick breaking construction**: think of the interval $[0,1]$ as a unit-length "stick."
- Let ℓ_k denote the fraction of the remaining stick given to component k. Then sample,

$$\ell_k \sim p(\ell_k) \qquad \pi_k = \ell_k \prod_{j=1}^{k-1}(1-\ell_j). \tag{35}$$

- When $p(\ell_k) = \mathrm{Beta}(\ell_k; 1, \alpha)$, this yields a **Dirichlet process**.
- If we have finite K, setting $\pi_K = \prod_{j=1}^{K-1}(1-\ell_j)$ yields a finite Dirichlet distribution on π.
- We say $\Theta \sim \mathrm{DP}(\alpha, G)$ where G is the distribution with density $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \boldsymbol{\nu})$.

Stick breaking construction of the Dirichlet process

- **Stick breaking construction**: think of the interval $[0, 1]$ as a unit-length "stick."
- Let ℓ_k denote the fraction of the remaining stick given to component k. Then sample,

$$\ell_k \sim p(\ell_k) \qquad \pi_k = \ell_k \prod_{j=1}^{k-1}(1-\ell_j). \tag{35}$$

- When $p(\ell_k) = \mathrm{Beta}(\ell_k; 1, \alpha)$, this yields a **Dirichlet process**.
- If we have finite K, setting $\pi_K = \prod_{j=1}^{K-1}(1-\ell_j)$ yields a finite Dirichlet distribution on π.
- We say $\Theta \sim \mathrm{DP}(\alpha, G)$ where G is the distribution with density $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$.

Stick breaking construction of the Dirichlet process

- **Stick breaking construction**: think of the interval $[0,1]$ as a unit-length "stick."
- Let ℓ_k denote the fraction of the remaining stick given to component k. Then sample,

$$\ell_k \sim p(\ell_k) \qquad \pi_k = \ell_k \prod_{j=1}^{k-1}(1-\ell_j). \tag{35}$$

- When $p(\ell_k) = \mathrm{Beta}(\ell_k; 1, \alpha)$, this yields a **Dirichlet process**.
- If we have finite K, setting $\pi_K = \prod_{j=1}^{K-1}(1-\ell_j)$ yields a finite Dirichlet distribution on $\boldsymbol{\pi}$.
- We say $\Theta \sim \mathrm{DP}(\alpha, G)$ where G is the distribution with density $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$.

Stick breaking construction of the Dirichlet process

- **Stick breaking construction**: think of the interval $[0, 1]$ as a unit-length "stick."
- Let ℓ_k denote the fraction of the remaining stick given to component k. Then sample,

$$\ell_k \sim p(\ell_k) \qquad \pi_k = \ell_k \prod_{j=1}^{k-1}(1-\ell_j). \tag{35}$$

- When $p(\ell_k) = \mathrm{Beta}(\ell_k; 1, \alpha)$, this yields a **Dirichlet process**.
- If we have finite K, setting $\pi_K = \prod_{j=1}^{K-1}(1-\ell_j)$ yields a finite Dirichlet distribution on $\boldsymbol{\pi}$.
- We say $\Theta \sim \mathrm{DP}(\alpha, G)$ where G is the distribution with density $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$.

Naïve Gibbs sampling in the DPMM

- We can equivalently sample a Bayesian mixture model as,

$$\boldsymbol{\theta}_n \overset{\text{iid}}{\sim} \Theta \tag{36}$$

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \tag{37}$$

  for $n = 1, \ldots, N$

- Since Θ is an atomic measure, there is some probability that $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}$ for two different data points.

- Now we can run a Gibbs sampler on $\{\boldsymbol{\theta}_n\}_{n=1}^N$, sampling their conditionals,

$$p(\boldsymbol{\theta}_n \mid \{\boldsymbol{\theta}_{n'}\}_{n' \neq n}, \{\boldsymbol{x}_n\}_{n=1}^N) \propto \alpha p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n \mid \boldsymbol{\phi}, \nu) + \sum_{n' \neq n} p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_{n'}) \, \delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n), \tag{38}$$

  which is an uncollapsed Gibbs sampler.

- When $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is an exponential family distribution and $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$ is its conjugate prior, the first term is available in closed form.

Naïve Gibbs sampling in the DPMM

- We can equivalently sample a Bayesian mixture model as,

$$\boldsymbol{\theta}_n \overset{\text{iid}}{\sim} \Theta \tag{36}$$

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \tag{37}$$

  for $n = 1, \ldots, N$

- Since Θ is an atomic measure, there is some probability that $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}$ for two different data points.

- Now we can run a Gibbs sampler on $\{\boldsymbol{\theta}_n\}_{n=1}^N$, sampling their conditionals,

$$p(\boldsymbol{\theta}_n \mid \{\boldsymbol{\theta}_{n'}\}_{n' \neq n}, \{\boldsymbol{x}_n\}_{n=1}^N) \propto \alpha p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_n) p(\boldsymbol{\theta}_n \mid \boldsymbol{\phi}, \nu) + \sum_{n' \neq n} p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_{n'})\, \delta_{\boldsymbol{\theta}_{n'}}(\boldsymbol{\theta}_n), \tag{38}$$

  which is an uncollapsed Gibbs sampler.

- When $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is an exponential family distribution and $p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)$ is its conjugate prior, the first term is available in closed form.

Collapsed Gibbs sampling in the DPMM

- Unfortunately, the uncollapsed Gibbs sampler tends to mix slowly.
- As before, we can marginalize over ("**collapse out**") the cluster parameters $\boldsymbol{\theta}$.
- This is equivalent to performing **Bayesian inference over a partition** of indices $[N] \triangleq \{1, \ldots, N\}$.
- A **partition** is a set of disjoint, non empty sets whose union is $[N]$:

$$\mathscr{C} = \{\mathscr{C}_k : |\mathscr{C}_k| > 0\} \tag{39}$$

$$\text{where} \qquad \mathscr{C}_k = \{n : z_n = k\}. \tag{40}$$

- The Gibbs sampler over partitions reduces to a straightforward update,

$$p(z_n = k \mid \boldsymbol{X}, \{z_{n'}\}_{n' \neq n}) \propto \begin{cases} \frac{\alpha}{\alpha+N-1} p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k \text{ is in a new cluster} \\ \frac{N_k^{(\neg n)}}{\alpha+N-1} p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{o.w.} \end{cases} \tag{41}$$

Collapsed Gibbs sampling in the DPMM

- Unfortunately, the uncollapsed Gibbs sampler tends to mix slowly.
- As before, we can marginalize over (“**collapse out**”) the cluster parameters $\boldsymbol{\theta}$.
- This is equivalent to performing **Bayesian inference over a partition** of indices $[N] \triangleq \{1, \ldots, N\}$.
- A **partition** is a set of disjoint, non empty sets whose union is $[N]$:

$$\mathscr{C} = \{\mathscr{C}_k : |\mathscr{C}_k| > 0\} \tag{39}$$

$$\text{where} \qquad \mathscr{C}_k = \{n : z_n = k\}. \tag{40}$$

- The Gibbs sampler over partitions reduces to a straightforward update,

$$p(z_n = k \mid \boldsymbol{X}, \{z_{n'}\}_{n' \neq n}) \propto \begin{cases} \frac{\alpha}{\alpha+N-1} p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k \text{ is in a new cluster} \\ \frac{N_k^{(\neg n)}}{\alpha+N-1} p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{o.w.} \end{cases} \tag{41}$$

Collapsed Gibbs sampling in the DPMM

- Unfortunately, the uncollapsed Gibbs sampler tends to mix slowly.
- As before, we can marginalize over ("**collapse out**") the cluster parameters $\boldsymbol{\theta}$.
- This is equivalent to performing **Bayesian inference over a partition** of indices $[N] \triangleq \{1, \ldots, N\}$.
- A **partition** is a set of disjoint, non empty sets whose union is $[N]$:

$$\mathscr{C} = \{\mathscr{C}_k : |\mathscr{C}_k| > 0\} \tag{39}$$

$$\text{where} \qquad \mathscr{C}_k = \{n : z_n = k\}. \tag{40}$$

- The Gibbs sampler over partitions reduces to a straightforward update,

$$p(z_n = k \mid \boldsymbol{X}, \{z_{n'}\}_{n' \neq n}) \propto \begin{cases} \frac{\alpha}{\alpha+N-1} p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k \text{ is in a new cluster} \\ \frac{N_k^{(\neg n)}}{\alpha+N-1} p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{o.w.} \end{cases} \tag{41}$$

Collapsed Gibbs sampling in the DPMM

- Unfortunately, the uncollapsed Gibbs sampler tends to mix slowly.
- As before, we can marginalize over (**"collapse out"**) the cluster parameters $\boldsymbol{\theta}$.
- This is equivalent to performing **Bayesian inference over a partition** of indices $[N] \triangleq \{1, \ldots, N\}$.
- A **partition** is a set of disjoint, non empty sets whose union is $[N]$:

$$\mathscr{C} = \{\mathscr{C}_k : |\mathscr{C}_k| > 0\} \tag{39}$$

$$\text{where} \qquad \mathscr{C}_k = \{n : z_n = k\}. \tag{40}$$

- The Gibbs sampler over partitions reduces to a straightforward update,

$$p(z_n = k \mid \boldsymbol{X}, \{z_{n'}\}_{n' \neq n}) \propto \begin{cases} \frac{\alpha}{\alpha+N-1} p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k \text{ is in a new cluster} \\ \frac{N_k^{(\neg n)}}{\alpha+N-1} p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{o.w.} \end{cases} \tag{41}$$

The Chinese Restaurant Process (CRP)

- Another way to sample a DPMM is to first sample the partition of $[N]$,

$$\mathscr{C} \sim p(\mathscr{C}; N, \alpha) \tag{42}$$

and then for each $\mathscr{C}_k \in \mathscr{C}$ sample,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G \tag{43}$$

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) \qquad \text{for } n \in \mathscr{C}_k \tag{44}$$

- The prior distribution on partitions is called a **Chinese restaurant process** (CRP).

  Initialize $\mathscr{C} = \varnothing$. For each $n = 1, \ldots, N$:

  1. insert n into existing block $\mathscr{C}_k$ with probability $\frac{|\mathscr{C}_k|}{\alpha+n-1}$, or
  2. create a new block with probability $\frac{\alpha}{\alpha+n-1}$.

- **Question:** Why doesn't the CRP prior depend on G? (I.e. on the hyperparameters $\boldsymbol{\phi}$ and ν.)

The Chinese Restaurant Process (CRP)

- Another way to sample a DPMM is to first sample the partition of $[N]$,

$$\mathscr{C} \sim p(\mathscr{C}; N, \alpha) \tag{42}$$

and then for each $\mathscr{C}_k \in \mathscr{C}$ sample,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G \tag{43}$$

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) \qquad \text{for } n \in \mathscr{C}_k \tag{44}$$

- The prior distribution on partitions is called a **Chinese restaurant process** (CRP).

  Initialize $\mathscr{C} = \varnothing$. For each $n = 1, \ldots, N$:

  1. insert n into existing block $\mathscr{C}_k$ with probability $\frac{|\mathscr{C}_k|}{\alpha+n-1}$, or
  2. create a new block with probability $\frac{\alpha}{\alpha+n-1}$.

- **Question:** Why doesn't the CRP prior depend on G? (I.e. on the hyperparameters $\boldsymbol{\phi}$ and ν.)

The Chinese Restaurant Process (CRP)

- Another way to sample a DPMM is to first sample the partition of $[N]$,

$$\mathscr{C} \sim p(\mathscr{C}; N, \alpha) \tag{42}$$

and then for each $\mathscr{C}_k \in \mathscr{C}$ sample,

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G \tag{43}$$

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) \qquad \text{for } n \in \mathscr{C}_k \tag{44}$$

- The prior distribution on partitions is called a **Chinese restaurant process** (CRP).

  Initialize $\mathscr{C} = \varnothing$. For each $n = 1, \ldots, N$:

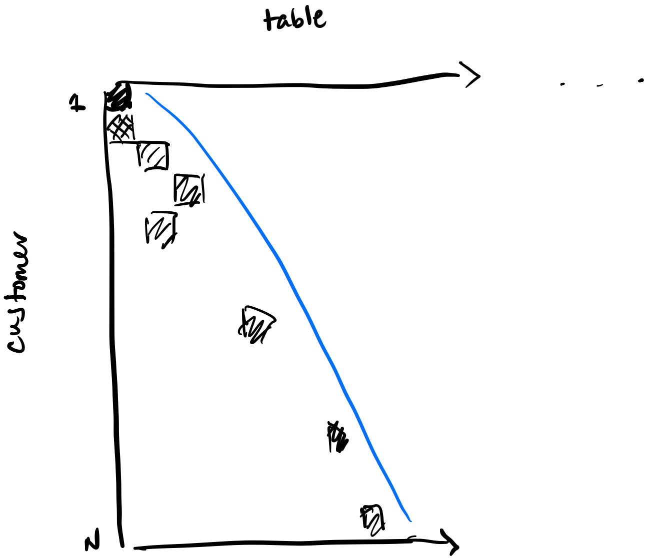
  1. insert n into existing block $\mathscr{C}_k$ with probability $\frac{|\mathscr{C}_k|}{\alpha+n-1}$, or

  2. create a new block with probability $\frac{\alpha}{\alpha+n-1}$.

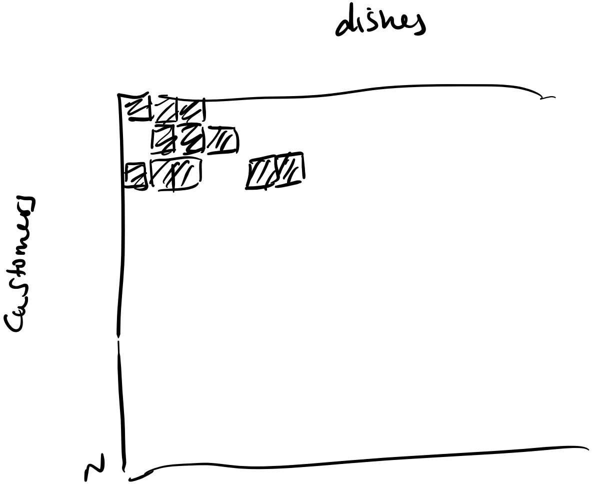
- **Question:** Why doesn't the CRP prior depend on G? (I.e. on the hyperparameters $\boldsymbol{\phi}$ and ν.)

The CRP suggests a way of sampling a DPMM one data point at a time

The CRP as a prior on binary matrices with one-hot rows



The Indian Buffet Process (IBP) as a prior on binary feature matrices



Pitman-Yor processes

The **Pitman-Yor process** (PYP) generalizes the DP to allow for more general distributions over cluster sizes.

We say $\Theta \sim \mathrm{PYP}(\alpha, d, G)$ is a Pitman-Yor process with **concentration** α, **discount** d, and **base measure** G if

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \tag{45}$$

$$\ell_k \sim \mathrm{Beta}(1-d, \alpha + kd) \tag{46}$$

$$\pi_k = \ell_k \prod_{j=1}^{k-1} (1-\ell_j) \tag{47}$$

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} G \tag{48}$$

When $d = 0$ we recover the DP; when $d > 0$ the PY produces a power law distribution over cluster sizes.

Mixture of finite mixture models

- DPMMs are often used to select the number of mixture components automatically, but they are actually misspecified for this task.

- The DP random measure has an infinite number of atoms almost surely. As $N \to \infty$, we get an infinite number of clusters with probability one.

- When we believe the data to have an unknown but finite number of clusters, **mixture of finite mixture models** (MFMMs) [Miller and Harrison, 2018] are more appropriate,

$$
\begin{aligned}
K &\sim p(K) && [\text{e.g. } K - 1 \sim \mathrm{Po}(\lambda)] && (49) \\
\boldsymbol{\pi} &\sim \mathrm{Dir}(\alpha \mathbf{1}_K) && && (50) \\
\boldsymbol{\theta}_k &\overset{\text{iid}}{\sim} G && \text{for } k = 1, \ldots, K && (51) \\
z_n &\overset{\text{iid}}{\sim} \boldsymbol{\pi} && \text{for } n = 1, \ldots, N && (52) \\
\boldsymbol{x}_n &\sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_{z_n}) && \text{for } n = 1, \ldots, N && (53)
\end{aligned}
$$

- Surprisingly, very similar collapsed Gibbs sampling algorithms can be derived for MFMMs.

Outline

- Collapsed Gibbs sampling for Bayesian Mixture Models
- Dirichlet process mixture models and random measures
- **Poisson random measures**

Poisson random measures

- Dirichlet processes and Poisson processes are closely related. In fact, DPs are instances of **Poisson random measures**.
- Consider the unnormalized weights and parameters to be a realization of a **marked point process**,

$$\{w_k, \boldsymbol{\theta}_k\}_{k=1}^{K} \sim \mathrm{PP}(\lambda(w, \boldsymbol{\theta})) \tag{54}$$

where $\lambda : \mathbb{R}_+ \times \mathbb{R}^D \to \mathbb{R}_+$, and define,

$$\mu = \sum_{k=1}^{K} w_k \delta_{\boldsymbol{\theta}_k}. \tag{55}$$

This is an unnormalized **random measure** on $\mathbb{R}^D$.

Poisson random measures II

- A Poisson random measure is **homogeneous** if the intensity factors as,

$$\lambda(w, \boldsymbol{\theta}) = \lambda(w) \cdot \lambda(\boldsymbol{\theta}). \tag{56}$$

- Now suppose the weight intensity is,

$$\lambda(w) = \alpha w^{-1} e^{-\beta w}. \tag{57}$$

  Then $\int_0^\infty \lambda(w) \, \mathrm{d}w = \infty$, so the random measure has infinitely many atoms almost surely.

- However, the measure assigned to any set $\mathscr{A} \subseteq \mathbb{R}^D$ is,

$$\mu(\mathscr{A}) = \sum_{k:\boldsymbol{\theta}_k \in \mathscr{A}} w_k \sim \mathrm{Ga}(\alpha G(\mathscr{A}), 1). \tag{58}$$

  and the total measure $W = \sum_{k=1}^\infty w_k \sim \mathrm{Ga}(\alpha, 1)$ is almost surely finite.

- We say $\mu = \sum_{k=1}^\infty w_k \delta_{\boldsymbol{\theta}_k}$ is a **gamma process** because $\lambda(w) \propto \mathrm{Ga}(w; 0, \beta)$.

Poisson random measures II

- A Poisson random measure is **homogeneous** if the intensity factors as,

$$\lambda(w, \boldsymbol{\theta}) = \lambda(w) \cdot \lambda(\boldsymbol{\theta}). \tag{56}$$

- Now suppose the weight intensity is,

$$\lambda(w) = \alpha w^{-1} e^{-\beta w}. \tag{57}$$

  Then $\int_0^\infty \lambda(w)\, \mathrm{d}w = \infty$, so the random measure has infinitely many atoms almost surely.

- However, the measure assigned to any set $\mathscr{A} \subseteq \mathbb{R}^D$ is,

$$\mu(\mathscr{A}) = \sum_{k:\boldsymbol{\theta}_k \in \mathscr{A}} w_k \sim \mathrm{Ga}(\alpha G(\mathscr{A}), 1). \tag{58}$$

  and the total measure $W = \sum_{k=1}^\infty w_k \sim \mathrm{Ga}(\alpha, 1)$ is almost surely finite.

- We say $\mu = \sum_{k=1}^\infty w_k \delta_{\boldsymbol{\theta}_k}$ is a **gamma process** because $\lambda(w) \propto \mathrm{Ga}(w; 0, \beta)$.

Poisson random measures II

- A Poisson random measure is **homogeneous** if the intensity factors as,

$$\lambda(w, \boldsymbol{\theta}) = \lambda(w) \cdot \lambda(\boldsymbol{\theta}). \tag{56}$$

- Now suppose the weight intensity is,

$$\lambda(w) = \alpha w^{-1} e^{-\beta w}. \qquad \propto \mathrm{Ga}(0, \beta) \tag{57}$$

  Then $\int_0^\infty \lambda(w)\, \mathrm{d}w = \infty$, so the random measure has infinitely many atoms almost surely.

- However, the measure assigned to any set $\mathscr{A} \subseteq \mathbb{R}^D$ is,

$$\mu(\mathscr{A}) = \sum_{k : \boldsymbol{\theta}_k \in \mathscr{A}} w_k \sim \mathrm{Ga}(\alpha G(\mathscr{A}), 1). \tag{58}$$

  and the total measure $W = \sum_{k=1}^\infty w_k \sim \mathrm{Ga}(\alpha, 1)$ is almost surely finite.

- We say $\mu = \sum_{k=1}^\infty w_k \delta_{\boldsymbol{\theta}_k}$ is a **gamma process** because $\lambda(w) \propto \mathrm{Ga}(w; 0, \beta)$.

Poisson random measures II

- A Poisson random measure is **homogeneous** if the intensity factors as,

$$\lambda(w, \boldsymbol{\theta}) = \lambda(w) \cdot \lambda(\boldsymbol{\theta}). \tag{56}$$

- Now suppose the weight intensity is,

$$\lambda(w) = \alpha w^{-1} e^{-\beta w}. \tag{57}$$

  Then $\int_0^\infty \lambda(w) \, \mathrm{d}w = \infty$, so the random measure has infinitely many atoms almost surely.

- However, the measure assigned to any set $\mathscr{A} \subseteq \mathbb{R}^D$ is,

$$\mu(\mathscr{A}) = \sum_{k : \boldsymbol{\theta}_k \in \mathscr{A}} w_k \sim \mathrm{Ga}(\alpha G(\mathscr{A}), 1). \tag{58}$$

  and the total measure $W = \sum_{k=1}^\infty w_k \sim \mathrm{Ga}(\alpha, 1)$ is almost surely finite.

- We say $\mu = \sum_{k=1}^\infty w_k \delta_{\boldsymbol{\theta}_k}$ is a **gamma process** because $\lambda(w) \propto \mathrm{Ga}(w; 0, \beta)$.

Dirichlet processes are normalized gamma processes

- If μ is a gamma process, the **normalized** random measure is a Dirichlet process,

$$\mu = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\theta}_k} \sim \mathrm{GaP}(\alpha, G) \quad \Rightarrow \quad \Theta = \sum_{k=1}^{\infty} \frac{w_k}{W} \delta_{\boldsymbol{\theta}_k} \sim \mathrm{DP}(\alpha, G). \tag{59}$$

- We can get other Poisson random measures by changing the weight intensity. E.g.
  - $\lambda(w) = \gamma w^{-(\alpha+1)}$ yields a *stable process*, and
  - $\lambda(w) = \gamma w^{-1}(1-w)^{\alpha-1}$ yields a *beta process*.
- **Completely random measures** further generalize Poisson random measures.
- If μ is a CRM, then $\Theta = \frac{\mu}{W}$ is independent of W iff μ is a gamma process; i.e. Θ is a DP.

References I

Peter Orbanz. Lecture notes on Bayesian nonparametrics. May 2014. URL http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf.

Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.