

# **STATS305C: Applied Statistics III**

## **Lecture 18: Dirichlet processes**

Scott Linderman

June 1, 2023

# 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  $\alpha \in \mathbb{R}_+^K$ :

$$\pi \sim \text{Dir}(\alpha) \tag{1}$$

2. Sample the parameters for each component:

$$\theta_k \stackrel{\text{iid}}{\sim} p(\theta \mid \phi, \nu) \quad \text{for } k = 1, \dots, K \tag{2}$$

3. Sample the assignment of each data point:

$$z_n \stackrel{\text{iid}}{\sim} \pi \quad \text{for } n = 1, \dots, N \tag{3}$$

4. Sample data points given their assignments:

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \theta_{z_n}) \quad \text{for } n = 1, \dots, N \tag{4}$$

# Joint distribution

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

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

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

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

- Then assume a **conjugate prior**,

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

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

## Joint distribution

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

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

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

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

- Then assume a **conjugate prior**,

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

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,

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

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

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

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  $\pi$ , too. The Dirichlet density is,

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

Plugging this in and integrating over  $\pi$  yields,

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

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



# The collapsed distribution in a Bayesian mixture model

We'll simplify the notation by writing,

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

where

$$\boldsymbol{\alpha}' = [\alpha_1 + N_1, \dots, \alpha_K + N_K] \quad (15)$$

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

$$\nu'_k = \nu + N_k. \quad (17)$$

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,

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}; \boldsymbol{\phi}, \nu), \quad \mathbf{x}_n \stackrel{\text{iid}}{\sim} p(\mathbf{x} | \boldsymbol{\theta}) \quad (18)$$

Derive an expression for the posterior predictive distribution,

$$p(\mathbf{x}_{N+1} | \{\mathbf{x}_n\}_{n=1}^N; \boldsymbol{\phi}, \nu) = \int p(\mathbf{x}_{N+1} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \{\mathbf{x}_n\}_{n=1}^N; \boldsymbol{\phi}, \nu) d\boldsymbol{\theta} \quad (19)$$

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

## Collapsed Gibbs for Bayesian Mixtures

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

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

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

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

where

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

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

## Collapsed Gibbs for Bayesian Mixtures II

- Then many terms cancel. In the first ratio,

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

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:

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

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,

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

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:

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

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,

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

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

## The infinite limit: informally speaking

- ▶ Now consider a special case where  $\alpha = \frac{\alpha}{K} \mathbf{1}_K$  and, loosely speaking, take  $K \rightarrow \infty$ . 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,

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

- ▶ 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,

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

## The infinite limit: informally speaking

- ▶ Now consider a special case where  $\alpha = \frac{\alpha}{K} \mathbf{1}_K$  and, loosely speaking, take  $K \rightarrow \infty$ . 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,

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

- ▶ 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,

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



## The infinite limit: informally speaking

- ▶ Now consider a special case where  $\alpha = \frac{\alpha}{K} \mathbf{1}_K$  and, loosely speaking, take  $K \rightarrow \infty$ . 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,

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

- ▶ 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,

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

## The infinite limit: informally speaking II

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

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

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

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

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

## The infinite limit: informally speaking II

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

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

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

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

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

## 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  $\alpha$ ) 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  $\theta$  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

- ▶ 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**,

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (31)$$

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

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

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k p(\mathbf{x} \mid \theta_k) = \int p(\mathbf{x} \mid \theta) \Theta(d\theta). \quad (32)$$

## Random measure perspective

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

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (31)$$

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

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

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k p(\mathbf{x} \mid \theta_k) = \int p(\mathbf{x} \mid \theta) \Theta(d\theta). \quad (32)$$

## Random measure perspective

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

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (31)$$

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

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

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k p(\mathbf{x} \mid \theta_k) = \int p(\mathbf{x} \mid \theta) \Theta(d\theta). \quad (32)$$



## Random measure perspective

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

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (31)$$

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

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

$$p(\mathbf{x}) = \sum_{k=1}^{\infty} \pi_k p(\mathbf{x} \mid \theta_k) = \int p(\mathbf{x} \mid \theta) \Theta(d\theta). \quad (32)$$

## Constructing a random measure

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

$$\theta_k \stackrel{\text{iid}}{\sim} p(\theta \mid \phi, \nu). \quad (33)$$

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,

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

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

## Constructing a random measure

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

$$\theta_k \stackrel{\text{iid}}{\sim} p(\theta \mid \phi, \nu). \quad (33)$$

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,

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

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

## Constructing a random measure

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

$$\theta_k \stackrel{\text{iid}}{\sim} p(\theta \mid \phi, \nu). \quad (33)$$

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,

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

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

## Constructing a random measure

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

$$\theta_k \stackrel{\text{iid}}{\sim} p(\theta \mid \phi, \nu). \quad (33)$$

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,

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

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

## Constructing a random measure with an infinite number of atoms

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

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

## Stick breaking construction of the Dirichlet process

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

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

- ▶ When  $p(\ell_k) = \text{Beta}(\ell_k; 1, \alpha)$ , 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  $\pi$ .
- ▶ We say  $\Theta \sim \text{DP}(\alpha, G)$  where  $G$  is the distribution with density  $p(\theta \mid \phi, \nu)$ .

## Stick breaking construction of the Dirichlet process

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

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

- ▶ When  $p(\ell_k) = \text{Beta}(\ell_k; 1, \alpha)$ , 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  $\pi$ .
- ▶ We say  $\Theta \sim \text{DP}(\alpha, G)$  where  $G$  is the distribution with density  $p(\theta \mid \phi, \nu)$ .



# Stick breaking construction of the Dirichlet process

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

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

- ▶ When  $p(\ell_k) = \text{Beta}(\ell_k; 1, \alpha)$ , 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  $\pi$ .
- ▶ We say  $\Theta \sim \text{DP}(\alpha, G)$  where  $G$  is the distribution with density  $p(\theta \mid \phi, \nu)$ .

# Stick breaking construction of the Dirichlet process

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

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

- ▶ When  $p(\ell_k) = \text{Beta}(\ell_k; 1, \alpha)$ , 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  $\pi$ .
- ▶ We say  $\Theta \sim \text{DP}(\alpha, G)$  where  $G$  is the distribution with density  $p(\theta \mid \phi, \nu)$ .

## Naïve Gibbs sampling in the DPMM

- We can equivalently sample a Bayesian mixture model as,

$$\theta_n \stackrel{\text{iid}}{\sim} \Theta \quad (36)$$

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \theta_n) \quad (37)$$

for  $n = 1, \dots, N$

- Since  $\Theta$  is an atomic measure, there is some probability that  $\theta_n = \theta_{n'}$  for two different data points.
- Now we can run a Gibbs sampler on  $\{\theta_n\}_{n=1}^N$ , sampling their conditionals,

$$p(\theta_n \mid \{\theta_{n'}\}_{n' \neq n}, \{\mathbf{x}_n\}_{n=1}^N) \propto \alpha p(\mathbf{x}_n \mid \theta_n) p(\theta_n \mid \phi, \nu) + \sum_{n' \neq n} p(\mathbf{x}_n \mid \theta_{n'}) \delta_{\theta_{n'}}(\theta_n), \quad (38)$$

which is an uncollapsed Gibbs sampler.

- When  $p(\mathbf{x} \mid \theta)$  is an exponential family distribution and  $p(\theta \mid \phi, \nu)$  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,

$$\boldsymbol{\theta}_n \stackrel{\text{iid}}{\sim} \Theta \quad (36)$$

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

for  $n = 1, \dots, N$

- Since  $\Theta$  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,

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

which is an uncollapsed Gibbs sampler.

- When  $p(\mathbf{x} \mid \boldsymbol{\theta})$  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  $\theta$ .
- ▶ This is equivalent to performing **Bayesian inference over a partition** of indices  $[N] \triangleq \{1, \dots, N\}$ :
- ▶ A **partition** is a set of disjoint, non empty sets whose union is  $[N]$ :

$$\mathcal{C} = \{\mathcal{C}_k : |\mathcal{C}_k| > 0\} \quad (39)$$

$$\text{where } \mathcal{C}_k = \{n : z_n = k\}. \quad (40)$$

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

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

## 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  $\theta$ .
- ▶ This is equivalent to performing **Bayesian inference over a partition** of indices  $[N] \triangleq \{1, \dots, N\}$ :
- ▶ A **partition** is a set of disjoint, non empty sets whose union is  $[N]$ :

$$\mathcal{C} = \{\mathcal{C}_k : |\mathcal{C}_k| > 0\} \quad (39)$$

$$\text{where } \mathcal{C}_k = \{n : z_n = k\}. \quad (40)$$

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

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

## 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  $\theta$ .
- ▶ This is equivalent to performing **Bayesian inference over a partition** of indices  $[N] \triangleq \{1, \dots, N\}$ .
- ▶ A **partition** is a set of disjoint, non empty sets whose union is  $[N]$ :

$$\mathcal{C} = \{\mathcal{C}_k : |\mathcal{C}_k| > 0\} \quad (39)$$

$$\text{where } \mathcal{C}_k = \{n : z_n = k\}. \quad (40)$$

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

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

## 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  $\theta$ .
- ▶ This is equivalent to performing **Bayesian inference over a partition** of indices  $[N] \triangleq \{1, \dots, N\}$ .
- ▶ A **partition** is a set of disjoint, non empty sets whose union is  $[N]$ :

$$\mathcal{C} = \{\mathcal{C}_k : |\mathcal{C}_k| > 0\} \quad (39)$$

$$\text{where } \mathcal{C}_k = \{n : z_n = k\}. \quad (40)$$

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

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



# The Chinese Restaurant Process (CRP)

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

$$\mathcal{C} \sim p(\mathcal{C}; N, \alpha) \quad (42)$$

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

$$\boldsymbol{\theta}_k \stackrel{\text{iid}}{\sim} G \quad (43)$$

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \boldsymbol{\theta}_k) \quad \text{for } n \in \mathcal{C}_k \quad (44)$$

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

Initialize  $\mathcal{C} = \emptyset$ . For each  $n = 1, \dots, N$ :

1. insert  $n$  into existing block  $\mathcal{C}_k$  with probability  $\frac{|\mathcal{C}_k|}{\alpha + n - 1}$ , 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  $\phi$  and  $\nu$ .)

# The Chinese Restaurant Process (CRP)

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

$$\mathcal{C} \sim p(\mathcal{C}; N, \alpha) \quad (42)$$

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

$$\boldsymbol{\theta}_k \stackrel{\text{iid}}{\sim} G \quad (43)$$

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \boldsymbol{\theta}_k) \quad \text{for } n \in \mathcal{C}_k \quad (44)$$

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

Initialize  $\mathcal{C} = \emptyset$ . For each  $n = 1, \dots, N$ :

1. insert  $n$  into existing block  $\mathcal{C}_k$  with probability  $\frac{|\mathcal{C}_k|}{\alpha + n - 1}$ , 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  $\phi$  and  $\nu$ .)

# The Chinese Restaurant Process (CRP)

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

$$\mathcal{C} \sim p(\mathcal{C}; N, \alpha) \quad (42)$$

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

$$\boldsymbol{\theta}_k \stackrel{\text{iid}}{\sim} G \quad (43)$$

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \boldsymbol{\theta}_k) \quad \text{for } n \in \mathcal{C}_k \quad (44)$$

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

Initialize  $\mathcal{C} = \emptyset$ . For each  $n = 1, \dots, N$ :

1. insert  $n$  into existing block  $\mathcal{C}_k$  with probability  $\frac{|\mathcal{C}_k|}{\alpha + n - 1}$ , 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  $\phi$  and  $\nu$ .)

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

$$\Theta = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad (45)$$

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

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

$$\theta_k \stackrel{\text{iid}}{\sim} G \quad (48)$$

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  $N \rightarrow \infty$ , 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,

$$K \sim p(K) \quad [\text{e.g. } K - 1 \sim \text{Po}(\lambda)] \quad (49)$$

$$\pi \sim \text{Dir}(\alpha \mathbf{1}_K) \quad (50)$$

$$\theta_k \stackrel{\text{iid}}{\sim} G \quad \text{for } k = 1, \dots, K \quad (51)$$

$$z_n \stackrel{\text{iid}}{\sim} \pi \quad \text{for } n = 1, \dots, N \quad (52)$$

$$\mathbf{x}_n \sim p(\mathbf{x} \mid \theta_{z_n}) \quad \text{for } n = 1, \dots, N \quad (53)$$

- ▶ 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**,

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

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

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

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,

$$\lambda(w, \theta) = \lambda(w) \cdot \lambda(\theta). \quad (56)$$

- Now suppose the weight intensity is,

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

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

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

$$\mu(\mathcal{A}) = \sum_{k: \theta_k \in \mathcal{A}} w_k \sim \text{Ga}(\alpha G(\mathcal{A}), 1). \quad (58)$$

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

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

## Poisson random measures II

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

$$\lambda(w, \theta) = \lambda(w) \cdot \lambda(\theta). \quad (56)$$

- Now suppose the weight intensity is,

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

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

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

$$\mu(\mathcal{A}) = \sum_{k: \theta_k \in \mathcal{A}} w_k \sim \text{Ga}(\alpha G(\mathcal{A}), 1). \quad (58)$$

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

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

## Poisson random measures II

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

$$\lambda(w, \theta) = \lambda(w) \cdot \lambda(\theta). \quad (56)$$

- ▶ Now suppose the weight intensity is,

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

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

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

$$\mu(\mathcal{A}) = \sum_{k: \theta_k \in \mathcal{A}} w_k \sim \text{Ga}(\alpha G(\mathcal{A}), 1). \quad (58)$$

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

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

## Poisson random measures II

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

$$\lambda(w, \theta) = \lambda(w) \cdot \lambda(\theta). \quad (56)$$

- ▶ Now suppose the weight intensity is,

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

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

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

$$\mu(\mathcal{A}) = \sum_{k: \theta_k \in \mathcal{A}} w_k \sim \text{Ga}(\alpha G(\mathcal{A}), 1). \quad (58)$$

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

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

# Dirichlet processes are normalized gamma processes

- ▶ If  $\mu$  is a gamma process, the **normalized** random measure is a Dirichlet process,

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

- ▶ 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  $\mu$  is a CRM, then  $\Theta = \frac{\mu}{W}$  is independent of  $W$  iff  $\mu$  is a gamma process; i.e.  $\Theta$  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](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.