Nonparametric Bayesian Modeling (Contd)

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

March 27, 2019
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi^{(0)}_k\}_{k=1}^K$.

For $t = 1, \ldots, T$

For each observation $i = 1, \ldots, N$,

- Sample the cluster id $z_i$:
  
  $$p(z_i = k | Z^{(t-1)}_{i-1}, \phi^{(t-1)}_k, X) \propto n^{(t-1)}_k \times p(x_i | \phi^{(t-1)}_k) = \hat{\pi}_{ik}$$

- Set $K = K + 1$ (if $x_i$ assigned to a new cluster).

- Sample the mixture component parameters $\{\phi^{(t)}_k\}_{k=1}^K$ and $\alpha^{(t)}$ from the respective CPs.

Note: "Markov Chain Sampling Methods for Dirichlet Process Mixture Models" (Neal, 2000) is an excellent reference for various MCMC sampling algorithms for nonparametric Bayesian mixture models (including collapsed versions that don't require sampling for $\{\phi_k\}_{k=1}^K$).
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$

Note: "Markov Chain Sampling Methods for Dirichlet Process Mixture Models" (Neal, 2000) is an excellent reference for various MCMC sampling algorithms for nonparametric Bayesian mixture models (including collapsed versions that don't require sampling for $\{\phi_k\}_{k=1}^K$)
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$’s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$’s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
    
    $$
    p(z_i = k|Z_{-i}^{(t-1)}, \phi^{(t)}, X) \propto n_k^{(t-1)} \times p(x_i|\phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K)
    $$
  - Set $K = K + 1$ (if $x_i$ assigned to a new cluster)
A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$’s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
    
    $$p(z_i = k|Z_{\sim i}^{(t-1)}, \phi^{(t)}, X) \propto n_k^{(t-1)} \times p(x_i|\phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K)$$

    $$p(z_i = k_{new}|Z_{\sim i}^{(t-1)}, \phi^{(t-1)}, X) \propto \alpha^{(t-1)} \times p(x_i|G_0) = \hat{\pi}_{ik_{new}}$$
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $\mathbf{Z}$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(\mathbf{z}_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $\mathbf{Z}^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$

$$
p(z_i = k|\mathbf{Z}_{-i}^{(t-1)}, \phi^{(t)}, \mathbf{X}) \propto n_k^{(t-1)} \times p(x_i|\phi_k^{(t-1)}) = \hat{\lambda}_{ik} \quad (k = 1, \ldots, K)
$$

$$
p(z_i = k_{new}|\mathbf{Z}_{-i}^{(t-1)}, \phi^{(t-1)}, \mathbf{X}) \propto \alpha^{(t-1)} \times p(x_i|\mathbf{G}_0) = \hat{\lambda}_{ik_{new}}
$$

$$
z_i^{(t)} \sim \text{multinoulli}(\hat{\lambda}_{i1}, \hat{\lambda}_{i2}, \ldots, \hat{\lambda}_{ik_{new}})
$$

Note: "Markov Chain Sampling Methods for Dirichlet Process Mixture Models" (Neal, 2000) is an excellent reference for various MCMC sampling algorithms for nonparametric Bayesian mixture models (including collapsed versions that don't require sampling for $\{\phi_k\}_{k=1}^K$).
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i | \pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
    \[
    p(z_i = k | Z_{-i}^{(t-1)}, \phi^{(t)}, X) \propto n_k^{(t-1)} \times p(x_i | \phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K)
    \]
    \[
    p(z_i = k_{\text{new}} | Z_{-i}^{(t-1)}, \phi^{(t-1)}, X) \propto \alpha^{(t-1)} \times p(x_i | G_0) = \hat{\pi}_{ik_{\text{new}}}
    \]
    \[
    z_i^{(t)} \sim \text{multinoulli}(\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{ik_{\text{new}}})
    \]
  - set $K = K + 1$ (if $x_i$ assigned to a new cluster)
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i|\pi)$)

Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
    \[
p(z_i = k|Z_{-i}^{(t-1)}, \phi^{(t)}, X) \propto n_k^{(t-1)} \times p(x_i|\phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K)
    \]
    \[
p(z_i = k_{\text{new}}|Z_{-i}^{(t-1)}, \phi^{(t-1)}, X) \propto \alpha^{(t-1)} \times p(x_i|G_0) = \hat{\pi}_{ik_{\text{new}}}
    \]
    \[
z_i^{(t)} \sim \text{multinoulli}(\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{ik_{\text{new}}})
    \]
    set $K = K + 1$ (if $x_i$ assigned to a new cluster)
- Sample the mixture component parameters $\{\phi_k^{(t)}\}_{k=1}^K$ and $\alpha^{(t)}$ from the respective CPs
Recap: A Nonparametric Bayesian Mixture Model

A brief sketch of a basic Gibbs sampler (samples $Z$ and $\{\phi_k\}_{k=1}^K$) for this model with unbounded $K$ (note: The mixing proportions $\pi_k$'s were collapsed from the prior $p(z_i|\pi)$).

### Gibbs Sampler for NPBayes Mixture Model

- Set an initial $K$. Initialize $Z^{(0)}$ and $\{\phi_k^{(0)}\}_{k=1}^K$.
- For $t = 1, \ldots, T$
  - For each observation $i = 1, \ldots, N$, sample the cluster id $z_i$
    
    $$
    p(z_i = k|Z_{-i}^{(t-1)}, \phi^{(t)}, X) \propto n_k^{(t-1)} \times p(x_i|\phi_k^{(t-1)}) = \hat{\pi}_{ik} \quad (k = 1, \ldots, K)
    $$
    
    $$
    p(z_i = k_{new}|Z_{-i}^{(t-1)}, \phi^{(t-1)}, X) \propto \alpha^{(t-1)} \times p(x_i|G_0) = \hat{\pi}_{ik_{new}}
    $$
    
    $$
    z_i^{(t)} \sim \text{multinoulli}(\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{ik_{new}})
    $$
    
    set $K = K + 1$ (if $x_i$ assigned to a new cluster)

- Sample the mixture component parameters $\{\phi_k^{(t)}\}_{k=1}^K$ and $\alpha^{(t)}$ from the respective CPs.

Note: “Markov Chain Sampling Methods for Dirichlet Process Mixture Models” (Neal, 2000) is an excellent reference for various MCMC sampling algorithms for nonparametric Bayesian mixture models (including collapsed versions that don’t require sampling for $\{\phi_k\}_{k=1}^K$).
Recap: An Alternate View of Mixture Models

- Can represent a mixture model as a **discrete distribution** as \( G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \)

\[
\begin{array}{cccc}
\pi_1 & \pi_2 & \pi_3 & \pi_4 \\
\phi_1 & \phi_2 & \phi_3 & \phi_4
\end{array}
\]
Recap: An Alternate View of Mixture Models

- Can represent a mixture model as a discrete distribution as \( G = \sum_{k=1}^{K} \pi_k \delta \phi_k \)

- Assume \( \{\pi_k\}_{k=1}^{K} \) drawn from Dirichlet and parameters \( \{\phi_k\}_{k=1}^{K} \) from some base distribution \( G_0 \)

\[
(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
\phi_k \sim G_0 \quad k = 1, \ldots, K
\]
Can represent a mixture model as a **discrete distribution** as

\[ G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \]

Assume \( \{\pi_k\}_{k=1}^{K} \) drawn from Dirichlet and parameters \( \{\phi_k\}_{k=1}^{K} \) from some **base distribution** \( G_0 \)

\[(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)\]

\[\phi_k \sim G_0 \quad k = 1, \ldots, K\]

The mixture model defined by \( G \) would generate observations \( x_i \) \((i = 1, \ldots, N)\) as follows

\[\theta_i \sim G\]

\[x_i \sim p(x|\theta_i)\]
Can represent a mixture model as a discrete distribution as 
\[ G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k} \]

Assume \( \{\pi_k\}_{k=1}^{K} \) drawn from Dirichlet and parameters \( \{\phi_k\}_{k=1}^{K} \) from some base distribution \( G_0 \)

\[ (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \]
\[ \phi_k \sim G_0 \quad k = 1, \ldots, K \]

The mixture model defined by \( G \) would generate observations \( x_i \) (\( i = 1, \ldots, N \)) as follows

\[ \theta_i \sim G \]
\[ x_i \sim p(x|\theta_i) \]

Discrete \( G \) implies that many params \( \theta_i \)'s will be identical (leading to clustered observations)
Recap: An Alternate View of Mixture Models

\[ \pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \]

\[ \phi_k \sim G_0 \]

\[ G = \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \]

\[ \theta_i \sim G \]

\[ x_i \sim p(x|\theta_i) \]
Recap: An Alternate View of Mixture Models

\[ \pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \]
\[ \phi_k \sim G_0 \]
\[ G = \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \]
\[ \theta_i \sim G \]
\[ x_i \sim p(x|\theta_i) \]

- This representation doesn’t have an explicit cluster id \( z_i \) for each observation \( x_i \). In this representation, clustering is implicit (non-uniqueness of \( \theta_i \)'s implies clustering of \( x_i \)'s)
Recap: An Alternate View of Mixture Models

- This representation doesn’t have an explicit cluster id $z_i$ for each observation $x_i$. In this representation, clustering is implicit (non-uniqueness of $\theta_i$’s implies clustering of $x_i$’s)

- $G$ defines a prior on a $K$ comp. mixture model with mix. prop. $\{\pi_k\}_{k=1}^K$ and params $\{\phi_k\}_{k=1}^K$
Nonparametric Bayesian Mixture Model (Contd)

- Also known as an “infinite mixture model”
- Can have an unbounded number of components (limited only by the data size)
Nonparametric Bayesian Mixture Model (Contd)

- Also known as an “infinite mixture model”
- Can have an unbounded number of components (limited only by the data size)
- Can think of these models in two equivalent ways
Nonparametric Bayesian Mixture Model (Contd)

- Also known as an “infinite mixture model”
- Can have an unbounded number of components (limited only by the data size)
- Can think of these models in two equivalent ways
  - An infinite mixture model can be obtained using an \( \text{infinite-dim Dirichlet} \) on its mixing proportions
Nonparametric Bayesian Mixture Model (Contd)

- Also known as an “infinite mixture model”
- Can have an unbounded number of components (limited only by the data size)
- Can think of these models in two equivalent ways
  - An infinite mixture model can be obtained using an infinite-dim Dirichlet on its mixing proportions
  - An infinite mixture model is as a discrete distribution $G$ of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$
Also known as an “infinite mixture model”

Can have an unbounded number of components (limited only by the data size)

Can think of these models in two equivalent ways

- An infinite mixture model can be obtained using an infinite-dim Dirichlet on its mixing proportions
- An infinite mixture model is as a discrete distribution $G$ of the form

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

Can view/define such infinite mixture models using various equivalent ways

- Stick-breaking Process
- Dirichlet Process
- Chinese Restaurant Process
- Pólya-Urn Scheme
Stick-Breaking Process

- Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$
Stick-Breaking Process

- Sethuraman (1994) showed how to construct \( G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \).
- Sethuraman’s stick-breaking construction provides a sequential way to generate \( \pi_k \)’s.
Sethuraman (1994) showed how to construct \( G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \).

Sethuraman’s stick-breaking construction provides a sequential way to generate \( \pi_k \)'s.

We basically need to generate a sequence \( \{\pi_k\}_{k=1}^{\infty} \) s.t. \( \pi_k \in (0, 1) \) and \( \sum_{k=1}^{\infty} \pi_k = 1 \).
Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$.

Sethuraman’s stick-breaking construction provides a sequential way to generate $\pi_k$’s.

We basically need to generate a sequence $\{ \pi_k \}_{k=1}^{\infty}$ s.t. $\pi_k \in (0, 1)$ and $\sum_{k=1}^{\infty} \pi_k = 1$.

Can be done using a stick-breaking construction for $\{ \pi_k \}_{k=1}^{\infty}$ as follows.
Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

Sethuraman’s stick-breaking construction provides a sequential way to generate $\pi_k$’s

We basically need to generate a sequence $\{\pi_k\}_{k=1}^{\infty}$ s.t. $\pi_k \in (0, 1)$ and $\sum_{k=1}^{\infty} \pi_k = 1$

Can be done using a stick-breaking construction for $\{\pi_k\}_{k=1}^{\infty}$ as follows

$$\beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, \ldots, \infty$$
Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

Sethuraman’s stick-breaking construction provides a sequential way to generate $\pi_k$’s.

We basically need to generate a sequence $\{\pi_k\}_{k=1}^{\infty}$ s.t. $\pi_k \in (0, 1)$ and $\sum_{k=1}^{\infty} \pi_k = 1$

Can be done using a stick-breaking construction for $\{\pi_k\}_{k=1}^{\infty}$ as follows

$$\beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, \ldots, \infty$$

$$\pi_1 = \beta_1$$
Sethuraman (1994) showed how to construct $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

- Sethuraman’s stick-breaking construction provides a sequential way to generate $\pi_k$’s
- We basically need to generate a sequence $\{\pi_k\}_{k=1}^{\infty}$ s.t. $\pi_k \in (0,1)$ and $\sum_{k=1}^{\infty} \pi_k = 1$
- Can be done using a stick-breaking construction for $\{\pi_k\}_{k=1}^{\infty}$ as follows

\[
\beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, \ldots, \infty \\
\pi_1 = \beta_1 \\
\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_{\ell-1}) \quad k = 2, \ldots, \infty
\]
The Stick-Breaking Construction: Pictorial Illustration

- Assume a stick of length 1 to begin with. Now recursively break it as follows:
  - Choose a random location $\beta_k \in (0, 1)$ drawn from Beta$(1, \alpha)$ at which to break the stick
  - Record $\pi_k$ as “$\beta_k$ times the length of the remaining stick”
The Stick-Breaking Construction: Pictorial Illustration

- Assume a stick of length 1 to begin with. Now recursively break it as follows:
  - Choose a random location $\beta_k \in (0, 1)$ drawn from Beta$(1, \alpha)$ at which to break the stick
  - Record $\pi_k$ as “$\beta_k$ times the length of the remaining stick”

- It is also very popular in deriving inference algorithms for nonparametric Bayesian mixture models
A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a \textit{distribution over distributions}

**Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed $[G(A_1),...,G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1),...,\alpha G_0(A_K))$ for any finite partition $A_1,...,A_K$ of the space $\Omega$ (Ferguson, 1973)

**Fact 2:** Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

**Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)

Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)
Dirichlet Process (DP)

- A Dirichlet Process DP$(\alpha, G_0)$ defines a distribution over distributions

So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution

Fact 1: If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed $[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$ for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

Fact 2: Any $G$ drawn from DP$(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

Fact 3: $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)

Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)
Dirichlet Process (DP)

- A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
  - $\alpha$: concentration param, $G_0$: base distribution (=mean of DP)

Fact 1: If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed
\[
G(A_1), \ldots, G(A_K) \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))
\] for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

Fact 2: Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form
\[
G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(\cdot)
\] (Sethuraman, 1994)

Fact 3: $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)

Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)
A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions

- So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
- $\alpha$: concentration param, $G_0$: base distribution (=mean of DP)
- Large $\alpha$ means $G \to G_0$
Dirichlet Process (DP)

A Dirichlet Process DP(\(\alpha, G_0\)) defines a distribution over distributions

- So \(G \sim DP(\alpha, G_0)\) will give us a distribution
- \(\alpha\) : concentration param, \(G_0\): base distribution (=mean of DP)
- Large \(\alpha\) means \(G \rightarrow G_0\)

Fact 1: If \(G \sim DP(\alpha, G_0)\) then any finite dim. marginal of \(G\) is Dirichlet distributed

Fact 2: Any \(G\) drawn from \(DP(\alpha, G_0)\) will be of the form \(G = \sum_{\infty}^{\infty} \pi_k \delta \phi_k(Sethuraman, 1994)\)

Fact 3: \(G\) is a discrete dist, i.e., only a few \(\pi_k\)'s will be significant (an informal proof shown next)

Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)
Dirichlet Process (DP)

- A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
  - $\alpha$: concentration param, $G_0$: base distribution (= mean of DP)
  - Large $\alpha$ means $G \rightarrow G_0$

- **Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed
  $$[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$
  for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

- **Fact 2:** Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

- **Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)

Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name).
Dirichlet Process (DP)

A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions.

- So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution.
- $\alpha$: concentration param, $G_0$: base distribution ($= \text{mean of DP}$).
- Large $\alpha$ means $G \rightarrow G_0$.

**Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed.

$$[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973).

**Fact 2:** Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994).
Dirichlet Process (DP)

- A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
  - $\alpha$: concentration param, $G_0$: base distribution (=mean of DP)
  - Large $\alpha$ means $G \rightarrow G_0$

- **Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed
  \[
  [G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))
  \]
  for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

- **Fact 2:** Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

- **Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)
Dirichlet Process (DP)

- A Dirichlet Process $DP(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim DP(\alpha, G_0)$ will give us a distribution
  - $\alpha$: concentration param, $G_0$: base distribution (=mean of DP)
  - Large $\alpha$ means $G \rightarrow G_0$

- **Fact 1:** If $G \sim DP(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed
  \[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet} (\alpha G_0(A_1), \ldots, \alpha G_0(A_K))\]
  for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

- **Fact 2:** Any $G$ drawn from $DP(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

- **Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant (an informal proof shown next)
  - Intuitively, can think of DP as an infinite-dim generalization of a Dirichlet (hence the name)
Detour: Some Properties of Dirichlet Distribution

- Aggregation: If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) then

\[
\left(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K \right) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
\]
Detour: Some Properties of Dirichlet Distribution

- **Aggregation:** If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) then
  \[
  \left(\frac{\pi_1 + \pi_2, \pi_3, \ldots, \pi_K}{K-1 \text{ dim}}\right) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
  \]

- **Expansion:** If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) and \(\hat{\pi} \sim \text{Beta}(\alpha_1b, \alpha_1(1-b))\)
Detour: Some Properties of Dirichlet Distribution

- Aggregation: If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) then
  \[
  (\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
  \]
  \(K-1\) dim

- Expansion: If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) and \(\hat{\pi} \sim \text{Beta}(\alpha_1 b, \alpha_1(1 - b))\) then
  \[
  (\pi_1 \hat{\pi}, \pi_1(1 - \hat{\pi}), \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1 - b), \alpha_2, \ldots, \alpha_K)
  \]
  \(K + 1\) dim
Detour: Some Properties of Dirichlet Distribution

- Aggregation: If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) then
  \[
  \left(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
  \]
  \(K-1\) dim

- Expansion: If \((\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)\) and \(\hat{\pi} \sim \text{Beta}(\alpha_1 b, \alpha_1(1 - b))\) then
  \[
  \left(\pi_1 \hat{\pi}, \pi_1(1 - \hat{\pi}), \pi_2, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1 - b), \alpha_2, \ldots, \alpha_K)
  \]
  \(K + 1\) dim

- Expansion: If \((\hat{\pi}_1, \ldots, \hat{\pi}_M) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_M)\) with \(\sum_{m=1}^{M} b_m = 1\)
• Aggregation: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ then

$$\left(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

• Expansion: If $(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ and $\hat{\pi} \sim \text{Beta}(\alpha_1 b, \alpha_1(1 - b))$ then

$$\left(\pi_1 \hat{\pi}, \pi_1(1 - \hat{\pi}), \pi_2, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1 b, \alpha_1(1 - b), \alpha_2, \ldots, \alpha_K)$$

• Expansion: If $(\hat{\pi}_1, \ldots, \hat{\pi}_M) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_M)$ with $\sum_{m=1}^{M} b_m = 1$ then

$$\left(\pi_1 \hat{\pi}_1, \ldots, \pi_1 \hat{\pi}_M, \pi_2, \ldots, \pi_K\right) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_M, \alpha_2, \ldots, \alpha_K)$$
An Informal Proof: Discreteness of DP Draws

- Note: \((x, 1 - x) \sim \text{Dirichlet}(\alpha, \alpha)\) is equivalent to \(x \sim \text{Beta}(\alpha, 1)\).
An Informal Proof: Discreteness of DP Draws

Note: \((x, 1 - x) \sim \text{Dirichlet}(\alpha, \alpha)\) is equivalent to \(x \sim \text{Beta}(\alpha, 1)\)

If \(\alpha\) is very small, \(x\) will be close to 0 or close to 1 (thus \((x, 1 - x)\) will be skewed)
An Informal Proof: Discreteness of DP Draws

- Note: \((x, 1 - x) \sim \text{Dirichlet}(\alpha, \alpha)\) is equivalent to \(x \sim \text{Beta}(\alpha, 1)\)

- If \(\alpha\) is very small, \(x\) will be close to 0 or close to 1 (thus \((x, 1 - x)\) will be skewed)

- Therefore, if we recursively keep expanding a Dirichlet, it will eventually become discrete
DP Posterior Distribution

Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete.
Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete.

Assume $N$ i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete $G$.
DP Posterior Distribution

- Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete.
- Assume $N$ i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete $G$.
- What will be the posterior distribution of $G$?
DP Posterior Distribution

- Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete
- Assume $N$ i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete $G$
- What will be the posterior distribution of $G$?
- For the finite-dimensional marginal of $G$, due to Dirichlet-multinoulli conjugacy, we will have
  $$[G(A_1), \ldots, G(A_K)]|\theta_1, \ldots, \theta_N \sim \text{Dirichlet}((\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_K) + n_K)$$
  .. where $n_k = \#\{i : \theta_i \in A_k\}$
Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete.

Assume $N$ i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete $G$.

What will be the posterior distribution of $G$?

For the finite-dimensional marginal of $G$, due to Dirichlet-multinoulli conjugacy, we will have

$$[G(A_1), \ldots, G(A_K)]|\theta_1, \ldots, \theta_N \sim \text{Dirichlet}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_K) + n_K)$$

.. where $n_k = \#\{i : \theta_i \in A_k\}$

This implies that the posterior of $G$ will also be a DP (a nice property!)

$$G|\theta_1, \ldots, \theta_N \sim \text{DP}(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i})$$

.. note also that $n_k = \sum_{i=1}^{N} \delta_{\theta_i}(A_k)$
Assume $G \sim \text{DP}(\alpha, G_0)$. Note that $G$ is discrete.

Assume $N$ i.i.d. (non-unique) draws $\theta_1, \ldots, \theta_N$ from the discrete $G$.

What will be the posterior distribution of $G$?

For the finite-dimensional marginal of $G$, due to Dirichlet-multinoulli conjugacy, we will have

$$[G(A_1), \ldots, G(A_K)]|\theta_1, \ldots, \theta_N \sim \text{Dirichlet}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_K) + n_K)$$

where $n_k = \#\{i : \theta_i \in A_k\}$

This implies that the posterior of $G$ will also be a DP (a nice property!)

$$G|\theta_1, \ldots, \theta_N \sim \text{DP}(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i})$$

Note also that $n_k = \sum_{i=1}^{N} \delta_{\theta_i}(A_k)$

Note that the posterior’s base dist. is a weighted avg. of prior base dist. and an empirical dist.

$$\frac{\alpha}{\alpha + N} G_0 + \frac{N}{\alpha + N} \frac{\sum_{i=1}^{N} \delta_{\theta_i}}{N}$$
We saw that the posterior of $G$ is another DP

$$G|\theta_1, \ldots, \theta_N \sim \text{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}\right)$$

where $n_k = \sum_{i=1}^{N} \delta_i(A_k)$ or $n_k = \#\{i: \theta_i \in A_k\}$
DP Predictive Distribution

- We saw that the posterior of $G$ is another DP
  \[ G|\theta_1, \ldots, \theta_N \sim \text{DP} \left( \alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i} \right) \]
  where $n_k = \sum_{i=1}^{N} \delta_i(A_k)$ or $n_k = \#\{i : \theta_i \in A_k\}$

- What will be the predictive posterior $p(\theta_{N+1}|\theta_1, \ldots, \theta_N)$?
  \[
p(\theta_{N+1}|\theta_1, \ldots, \theta_N) = \int p(\theta_{N+1}|G, \theta_1, \ldots, \theta_N) p(G|\theta_1, \ldots, \theta_N) dG = \int p(\theta_{N+1}|G) p(G|\theta_1, \ldots, \theta_N) dG
  \]
DP Predictive Distribution

- We saw that the posterior of $G$ is another DP:

$$G|\theta_1, \ldots, \theta_N \sim DP\left(\alpha + N, \frac{\alpha}{\alpha + N}G_0 + \frac{1}{\alpha + N}\sum_{i=1}^{N} \delta_{\theta_i}\right)$$

where $n_k = \sum_{i=1}^{N} \delta_i(A_k)$ or $n_k = \#\{i : \theta_i \in A_k\}$

- What will be the predictive posterior $p(\theta_{N+1}|\theta_1, \ldots, \theta_N)$?

$$p(\theta_{N+1}|\theta_1, \ldots, \theta_N) = \int p(\theta_{N+1}|G, \theta_1, \ldots, \theta_N)p(G|\theta_1, \ldots, \theta_N)dG = \int p(\theta_{N+1}|G)p(G|\theta_1, \ldots, \theta_N)dG$$

- Intuitively, due to the discreteness of the DP posterior, this would simply be the mean of the DP posterior ($= \text{the posterior base distribution}$)

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N}G_0 + \frac{1}{\alpha + N}\sum_{i=1}^{N} \delta_{\theta_i}$$
DP Predictive Distribution

We saw that the posterior of $G$ is another DP

$$G|\theta_1, \ldots, \theta_N \sim \text{DP} \left( \alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i} \right)$$

where $n_k = \sum_{i=1}^{N} \delta_i(A_k)$ or $n_k = \#\{i: \theta_i \in A_k\}$

What will be the predictive posterior $p(\theta_{N+1}|\theta_1, \ldots, \theta_N)$?

$$p(\theta_{N+1}|\theta_1, \ldots, \theta_N) = \int p(\theta_{N+1}|G, \theta_1, \ldots, \theta_N) p(G|\theta_1, \ldots, \theta_N) dG = \int p(\theta_{N+1}|G) p(G|\theta_1, \ldots, \theta_N) dG$$

Intuitively, due to the discreteness of the DP posterior, this would simply be the mean of the DP posterior (= the posterior base distribution)

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}$$

Thus $\theta_{N+1}$ will be equal to a previous $\theta_i$ with probability proportional to $\sum_{j=1}^{N} \delta_{\theta_j=\theta_i}$, and will be a new value with probability proportional to $\alpha$
A Sequential Generative Scheme

The form of the DP predictive distribution

\[ \theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i} \]

suggests the following scheme to generate a sequence of parameters \( \theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots \)
A Sequential Generative Scheme

- The form of the DP predictive distribution

\[ \theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i} \]

suggests the following scheme to generate a sequence of parameters \( \theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots \)

\[ \theta_1 \sim G_0 \]

Note that \( \theta_1, \ldots, \theta_{n-1}, \theta_n \) is an "exchangeable sequence" (joint probability invariant to ordering)

\[ p(\theta_1, \theta_2, \ldots, \theta_n) = n! \prod_{i=1}^{n} p(\theta_i|\theta_1, \ldots, \theta_{i-1}) \]

Related to de-Finetti's theorem (next class)
A Sequential Generative Scheme

- The form of the DP predictive distribution

\[ \theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i} \]

suggests the following scheme to generate a sequence of parameters \( \theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots \)

\[ \theta_1 \sim G_0 \]

\[ \theta_2|\theta_1 \sim \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1} \]

\[ \vdots \]

Note that \( \theta_1, \ldots, \theta_{N-1}, \theta_N \) is an "exchangeable sequence" (joint probability invariant to ordering)

\[ p(\theta_1, \theta_2, \ldots, \theta_N) = \prod_{i=1}^{N} p(\theta_i|\theta_1, \ldots, \theta_{i-1}) \]

Related to de-Finetti's theorem (next class)
A Sequential Generative Scheme

- The form of the DP predictive distribution

\[
\theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}
\]

suggests the following scheme to generate a sequence of parameters \(\theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots\)

- \(\theta_1 \sim G_0\)
- \(\theta_2|\theta_1 \sim \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1}\)
- \(\vdots\)
- \(\theta_n|\theta_1, \ldots, \theta_{n-1} \sim \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}\)

Note that \(\theta_1, \ldots, \theta_{n-1}, \theta_n\) is an "exchangeable sequence" (joint probability invariant to ordering)

\[p(\theta_1, \theta_2, \ldots, \theta_n) = n \prod_{i=1}^{n} p(\theta_i|\theta_1, \ldots, \theta_{i-1})\]

Related to de-Finetti's theorem (next class)
A Sequential Generative Scheme

The form of the DP predictive distribution

$$\theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}$$

suggests the following scheme to generate a sequence of parameters $\theta_1, \ldots, \theta_N, \theta_{N+1}, \ldots$

$$\theta_1 \sim G_0$$

$$\theta_2|\theta_1 \sim \frac{\alpha}{\alpha + 1} G_0 + \frac{1}{\alpha + 1} \delta_{\theta_1}$$

$$\vdots$$

$$\theta_n|\theta_1, \ldots, \theta_{n-1} \sim \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1}$$

Note that $\theta_1, \ldots, \theta_{n-1}, \theta_n$ is an “exchangeable sequence” (joint probability invariant to ordering)

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \prod_{i=1}^{n} p(\theta_i|\theta_1, \ldots, \theta_{i-1})$$

Related to de-Finetti’s theorem (next class)
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.
- Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

Imagine table $k$ is associated with a unique $\phi_k$. Then the arrangement would look like...

The table assignment distribution is the same as the DP predictive distribution.
The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated. Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

- Probability of sitting at an already occupied table $\propto n^k$ ($n^k$: # of people sitting at table $k$)
- Probability of sitting at an unoccupied table $\propto \alpha$ (where $\alpha$ is a novelty hyperparameter)

Imagine table $k$ is associated with a unique $\phi_k$. Then the arrangement would look like...

The table assignment distribution is the same as the DP predictive distribution.
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.

- Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored” /labeled identical.

Probability of sitting at an already occupied table $\propto n^k$ ($n^k$: # of people sitting at table $k$)

Probability of sitting at an unoccupied table $\propto \alpha$ (where $\alpha$ is a novelty hyperparameter)

Imagine table $k$ is associated with a unique $\phi_k$. Then the arrangement would look like...

The table assignment distribution is the same as the DP predictive distribution.
The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated. Think of the $\theta_i$'s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$'s sitting at the same table will be “colored” /labeled identical.
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.
- Think of the $\theta_i$'s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$'s sitting at the same table will be “colored” / labeled identical.

![Diagram of tables with fractions and symbols corresponding to probabilities.]
The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.

Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

The table assignment distribution is the same as the DP predictive distribution.

Probability of sitting at an already occupied table

$\Pr(\text{table } k) \propto \frac{1}{n^k}$

$\Pr(\text{table } k)$: # of people sitting at table $k$

Probability of sitting at an unoccupied table

$\Pr(\text{table is empty}) \propto \frac{1}{\alpha}$

(Where $\alpha$ is a novelty hyperparameter.)
The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.

Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

- Probability of sitting at an already occupied table $k \propto n_k$ ($n_k$: # of people sitting at table $k$)
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.
- Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

- Probability of sitting at an already occupied table $k \propto n_k$ ($n_k$: # of people sitting at table $k$)
- Probability of sitting at an unoccupied table $\propto \alpha$ (where $\alpha$ is a novelty hyperparameter)
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way \( \theta_1, \ldots, \theta_n \) are sequentially generated.
- Think of the \( \theta_i \)'s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All \( \theta_i \)'s sitting at the same table will be “colored”/labeled identical.

- Probability of sitting at an already occupied table \( k \propto n_k \) (\( n_k \): # of people sitting at table \( k \))
- Probability of sitting at an unoccupied table \( \propto \alpha \) (where \( \alpha \) is a novelty hyperparameter)
- Imagine table \( k \) is associated with a unique \( \phi_k \). Then the arrangement would look like..
Chinese Restaurant Process (CRP)

- The CRP is another (culinary) metaphor to describe the way $\theta_1, \ldots, \theta_n$ are sequentially generated.
- Think of the $\theta_i$’s as customers who sequentially enter a restaurant (need not be Chinese!) and decide which table to sit at. All $\theta_i$’s sitting at the same table will be “colored”/labeled identical.

- Probability of sitting at an already occupied table $k \propto n_k$ ($n_k$: # of people sitting at table $k$)
- Probability of sitting at an unoccupied table $\propto \alpha$ (where $\alpha$ is a novelty hyperparameter)
- Imagine table $k$ is associated with a unique $\phi_k$. Then the arrangement would look like..

- The table assignment distribution is the same as the DP predictive distribution.
Hierarchical Dirichlet Process (HDP)

- Defines a DP whose base distribution \( G_0 \) itself is drawn from another DP

- Can be used if we would like to cluster \( m \) data sets, each using a DP mixture model

\[
\begin{align*}
G_0 & \sim \text{DP}(\gamma, H) \\
G_i & \sim \text{DP}(\alpha, G_0) \\
\theta_{ij} | G_i & \sim G_i \\
x_{ij} | \theta_{ij} & \sim p(x_{ij} | \theta_{ij})
\end{align*}
\]
Hierarchical Dirichlet Process (HDP)

- Defines a DP whose base distribution $G_0$ itself is drawn from another DP

- Can be used if we would like to cluster $m$ data sets, each using a DP mixture model

- The discreteness of the shared base distribution $G_0$ enables sharing information across the $m$ clustering problems (reason: because the discreteness allows sharing clusters/atoms)

Hierarchical Dirichlet Processes (Teh et al, 2006)
Hierarchical Dirichlet Process (HDP)

- Defines a DP whose base distribution $G_0$ itself is drawn from another DP

![Diagram showing the structure of HDP]

- Can be used if we would like to cluster $m$ data sets, each using a DP mixture model
- The discreteness of the shared base distribution $G_0$ enables sharing information across the $m$ clustering problems (reason: because the discreteness allows sharing clusters/atoms)
- Important: If $G_0$ were a continuous distribution, we won’t be able to share atoms (probability of $G_i$ and $G_j$ sharing any atoms will be zero if $G_0$ is a continuous distribution)

Hierarchical Dirichlet Processes (Teh et al, 2006)
Hierarchical Dirichlet Process (HDP)

- Defines a DP whose base distribution $G_0$ itself is drawn from another DP

- Can be used if we would like to cluster $m$ data sets, each using a DP mixture model

- The discreteness of the shared base distribution $G_0$ enables sharing information across the $m$ clustering problems (reason: because the discreteness allows sharing clusters/atoms)

- Important: If $G_0$ were a continuous distribution, we won’t be able to share atoms (probability of $G_i$ and $G_j$ sharing any atoms will be zero if $G_0$ is a continuous distribution)

- HDP used in nonparametric Bayesian version of LDA topic model
Next Class

- Some other aspects of NPBayes mixture models
- Other examples of NPBayes models and wrap-up of discussion on NPBayes
- On to next topic: Deep Probabilistic Modeling