# CSci 8980: Advanced Topics in Graphical Models

## Dirichlet Processes

Instructor: Arindam Banerjee

October 4, 2007

## Measurable Space

- Given a set $X$, let $2^X$ be the power set

## Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if

# Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
    1. $\mathcal{A}$ contains $X$

# Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements

## Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions

# Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections

# Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections
- Examples

## Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections
- Examples
  - $X = \{a, b, c, d\}$, and $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$

# Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections
- Examples
  - $X = \{a, b, c, d\}$, and $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$
  - $X = \mathbb{R}$, and $\mathcal{A}$ is open intervals in $\mathbb{R}$

## Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections
- Examples
  - $X = \{a, b, c, d\}$, and $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$
  - $X = \mathbb{R}$, and $\mathcal{A}$ is open intervals in $\mathbb{R}$
- Tuple $(X, \mathcal{A})$ is called a measurable space

## Measurable Space

- Given a set $X$, let $2^X$ be the power set
- $\mathcal{A} \subseteq 2^X$ is called a $\sigma$-algebra if
  1. $\mathcal{A}$ contains $X$
  2. $\mathcal{A}$ is closed under complements
  3. $\mathcal{A}$ is closed under countable unions
- Hence, $\mathcal{A}$ is closed under countable intersections
- Examples
  - $X = \{a, b, c, d\}$, and $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$
  - $X = \mathbb{R}$, and $\mathcal{A}$ is open intervals in $\mathbb{R}$
- Tuple $(X, \mathcal{A})$ is called a measurable space
- One can define a *measure* $\mu$ on a measurable space

# Measurable Space (Contd.)

- Measurable function

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces

# Measurable Space (Contd.)

- Measurable function
    - Function between two measurable spaces
    - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$
- A measure is a function $\mu : \mathcal{A} \mapsto [0, \infty]$ such that

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$
- A measure is a function $\mu : \mathcal{A} \mapsto [0, \infty]$ such that
  - $\mu(\emptyset) = 0$, and

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$
- A measure is a function $\mu : \mathcal{A} \mapsto [0, \infty]$ such that
  - $\mu(\emptyset) = 0$, and
  - For a countable sequence of pairwise disjoint sets $E_1, E_2, \ldots$

$$\mu\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$
- A measure is a function $\mu : \mathcal{A} \mapsto [0, \infty]$ such that
  - $\mu(\emptyset) = 0$, and
  - For a countable sequence of pairwise disjoint sets $E_1, E_2, \ldots$

$$\mu\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

- A probability measure satisfies $P(X) = 1$

# Measurable Space (Contd.)

- Measurable function
  - Function between two measurable spaces
  - Consider two spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$
  - $f : X \mapsto Y$ is measurable if $\forall b \in \mathcal{B}, f^{-1}(b) \in A$
- Example
  - Random variables are measurable functions
  - For real-valued random variables, $Y = \mathbb{R}$
- A measure is a function $\mu : \mathcal{A} \mapsto [0, \infty]$ such that
  - $\mu(\emptyset) = 0$, and
  - For a countable sequence of pairwise disjoint sets $E_1, E_2, \ldots$

$$\mu\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

- A probability measure satisfies $P(X) = 1$
- $(X, \mathcal{A}, P)$ is called a probability space

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists$ $\mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists\ \mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - $\mathcal{P}$ yields the distributions $P$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists$ $\mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
    - $\mathcal{P}$ yields the distributions $P$
    - $[0,1]^{\mathcal{A}}$ is the space of all functions $P$ from $\mathcal{A} \mapsto [0,1]$

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists\ \mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - $\mathcal{P}$ yields the distributions $P$
  - $[0,1]^{\mathcal{A}}$ is the space of all functions $P$ from $\mathcal{A} \mapsto [0,1]$
  - With $P(X) = 1$ these functions are probability distributions

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists$ $\mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - $\mathcal{P}$ yields the distributions $P$
  - $[0,1]^{\mathcal{A}}$ is the space of all functions $P$ from $\mathcal{A} \mapsto [0,1]$
  - With $P(X) = 1$ these functions are probability distributions
- Goal: To construct such a $\mathcal{P}$ over probability distributions

# Distribution Over Distributions

- How to define random probability measures $P$ over $(X, \mathcal{A})$
- Consider any sequence of sets $A_1, \ldots, A_m (A_i \in \mathcal{A})$
- Define joint distribution $(P(A_1), \ldots, P(A_m))$
- Show $\exists\ \mathcal{P}$ on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - $\mathcal{P}$ yields the distributions $P$
  - $[0,1]^{\mathcal{A}}$ is the space of all functions $P$ from $\mathcal{A} \mapsto [0,1]$
  - With $P(X) = 1$ these functions are probability distributions
- Goal: To construct such a $\mathcal{P}$ over probability distributions
- Parametric vs non-parametric Bayes

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$

## Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \ldots, B_k)$ is a partition if

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \ldots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^{k} B_i = X$

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \dots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^{k} B_i = X$
- Define random probability $P$ as follows:

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \ldots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^{k} B_i = X$
- Define random probability $P$ as follows:
  - Define joint distribution $(P(B_1), \ldots, P(B_k))$

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \ldots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^{k} B_i = X$
- Define random probability $P$ as follows:
  - Define joint distribution $(P(B_1), \ldots, P(B_k))$
  - Use this to define joint distribution $(P(A_1), \ldots, P(A_m))$

## Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \ldots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^{k} B_i = X$
- Define random probability $P$ as follows:
  - Define joint distribution $(P(B_1), \ldots, P(B_k))$
  - Use this to define joint distribution $(P(A_1), \ldots, P(A_m))$
- For arbitrary sets $A_1, \ldots, A_m$, with $\gamma_j = 0$ or 1, define

$$B_{\gamma_1, \cdots, \gamma_m} = \cap_{j=1}^{m} A_j^{\gamma_j}$$

# Constructing $\mathcal{P}$

- It is convenient to work with a partition of $X$
- For any $k$, $(B_1, \dots, B_k)$ is a partition if
  - $B_i \in \mathcal{A}, \forall i; \quad B_i \cap B_j = \emptyset, \forall i \neq j; \quad \cup_{i=1}^k B_i = X$
- Define random probability $P$ as follows:
  - Define joint distribution $(P(B_1), \dots, P(B_k))$
  - Use this to define joint distribution $(P(A_1), \dots, P(A_m))$
- For arbitrary sets $A_1, \dots, A_m$, with $\gamma_j = 0$ or 1, define

$$B_{\gamma_1, \cdots, \gamma_m} = \cap_{j=1}^m A_j^{\gamma_j}$$

- Then $\{B_{\gamma_1, \cdots, \gamma_m}\}$ is a valid partition of $X$

# Constructing $\mathcal{P}$ (Contd.)

- We have a valid partition $\{B_{\gamma_1,\cdots,\gamma_m}\}$

# Constructing $\mathcal{P}$ (Contd.)

- We have a valid partition $\{B_{\gamma_1,\cdots,\gamma_m}\}$
- Now, define a joint distribution over partitions

$$\{P(B_{\gamma_1,\cdots,\gamma_m}); \gamma_j = 0 \text{ or } 1, j = 1, \ldots, m\}$$

## Constructing $\mathcal{P}$ (Contd.)

- We have a valid partition $\{B_{\gamma_1,\cdots,\gamma_m}\}$
- Now, define a joint distribution over partitions

$$\{P(B_{\gamma_1,\cdots,\gamma_m}); \gamma_j = 0 \text{ or } 1, j = 1, \ldots, m\}$$

- The joint distribution over $(P(A_1), \ldots, P(A_m))$

$$P(A_i) = \sum_{\substack{(\gamma_1,\cdots,\gamma_m) \\ \gamma_i = 1}} P(B_{\gamma_1,\cdots,\gamma_m})$$

## A Consistency Requirement

- There is one consistency requirement we need for $P(B_1, \cdots, B_k)$

# A Consistency Requirement

- There is one consistency requirement we need for $P(B_1, \cdots, B_k)$
- Consider two partitions $B' = (B'_1, \cdots, B'_{k'})$ and $B = (B_1, \cdots, B_k)$

## A Consistency Requirement

- There is one consistency requirement we need for $P(B_1, \cdots, B_k)$
- Consider two partitions $B' = (B'_1, \cdots, B'_{k'})$ and $B = (B_1, \cdots, B_k)$
- Let $B'$ be a refinement of $B$, i.e.,

$$B_1 = \cup_1^{r_1} B'_i, B_2 = \cup_{r_1+1}^{r_2} B'_i, \cdots, B_k = \cup_{r_{k-1}+1}^{k'} B'_i$$

## A Consistency Requirement

- There is one consistency requirement we need for $P(B_1, \cdots, B_k)$
- Consider two partitions $B' = (B'_1, \cdots, B'_{k'})$ and $B = (B_1, \cdots, B_k)$
- Let $B'$ be a refinement of $B$, i.e.,

$$B_1 = \cup_1^{r_1} B'_i, B_2 = \cup_{r_1+1}^{r_2} B'_i, \cdots, B_k = \cup_{r_{k-1}+1}^{k'} B'_i$$

- Then the distribution of $(P(B_1), \cdots, P(B_k))$ is identical to that of

$$\left( \sum_1^{r_1} P(B'_i), \sum_{r_1+1}^{r_2} P(B'_i), \cdots, \sum_{r_{k-1}+1}^{k'} P(B'_i) \right)$$

## A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes
- Based on the above construction

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes
- Based on the above construction
  - Sufficient to focus on partitions, rather than arbitrary sets

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes
- Based on the above construction
  - Sufficient to focus on partitions, rather than arbitrary sets
  - Can maintain distribution over distributions (non-parametric Bayes)

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes
- Based on the above construction
  - Sufficient to focus on partitions, rather than arbitrary sets
  - Can maintain distribution over distributions (non-parametric Bayes)
  - So far, we have only seen distribution over parameters

# A Key Lemma

- Lemma: If the joint distribution $(P(B_1), \cdots, P(B_k))$ satisfies the consistency condition, and, if for arbitrary sets $(A_1, \ldots, A_m)$, the joint distribution is constructed as outlined earlier, then there exists $\mathcal{P}$ which yields these distribution.
- Samples $P$ from $\mathcal{P}$ are distributions on $(X, \mathcal{A})$
- We will focus on a specific $\mathcal{P}$: Dirichlet processes
- Based on the above construction
    - Sufficient to focus on partitions, rather than arbitrary sets
    - Can maintain distribution over distributions (non-parametric Bayes)
    - So far, we have only seen distribution over parameters
- Can inference be tractably done over such models?

# Dirichlet Distribution

- Distribution over finite discrete distributions

# Dirichlet Distribution

- Distribution over finite discrete distributions
- The density function is given by

$$D(\alpha_1, \ldots, \alpha_k) = f(x_1, \ldots, x_k | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$$

# Dirichlet Distribution

- Distribution over finite discrete distributions
- The density function is given by

$$D(\alpha_1,\ldots,\alpha_k) = f(x_1,\ldots,x_k|\alpha_1,\ldots,\alpha_k) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$$

- Well defined on the unit simplex $\sum_{i=1}^k x_i = 1$

# Dirichlet Distribution

- Distribution over finite discrete distributions
- The density function is given by

$$D(\alpha_1,\ldots,\alpha_k) = f(x_1,\ldots,x_k|\alpha_1,\ldots,\alpha_k) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i-1}$$

- Well defined on the unit simplex $\sum_{i=1}^k x_i = 1$
- Key Property: If $(X_1,\ldots,X_k) \sim D(\alpha_1,\ldots,\alpha_k)$, and $r_1,\ldots,r_\ell$ are integers such that $0 < r_1 < \cdots < r_\ell$ then

$$\left(\sum_{1}^{r_1} X_i, \sum_{r_1+1}^{r_2} X_i, \cdots, \sum_{r_{\ell-1}+1}^{k} X_i\right) \sim D\left(\sum_{1}^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, \cdots, \sum_{r_\ell+1}^{k} \alpha_i\right)$$

# Dirichlet Distribution

- Distribution over finite discrete distributions
- The density function is given by

$$D(\alpha_1,\ldots,\alpha_k) = f(x_1,\ldots,x_k|\alpha_1,\ldots,\alpha_k) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$$

- Well defined on the unit simplex $\sum_{i=1}^k x_i = 1$
- Key Property: If $(X_1,\ldots,X_k) \sim D(\alpha_1,\ldots,\alpha_k)$, and $r_1,\ldots,r_\ell$ are integers such that $0 < r_1 < \cdots < r_\ell$ then

$$\left(\sum_1^{r_1} X_i, \sum_{r_1+1}^{r_2} X_i, \cdots, \sum_{r_{\ell-1}+1}^{k} X_i\right) \sim D\left(\sum_1^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, \cdots, \sum_{r_\ell+1}^{k} \alpha_i\right)$$

- In particular, the marginal distribution of $X_j \sim B(\alpha_j, \sum_1^k \alpha_i - \alpha_j)$ where

$$B(\alpha,\beta) = f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

# Gamma and Dirichlet

- Gamma distribution, with $x > 0, \alpha, \theta > 0$, is

$$\Gamma(\alpha, \theta) = f(x|\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}$$

## Gamma and Dirichlet

- Gamma distribution, with $x > 0, \alpha, \theta > 0$, is

$$\Gamma(\alpha, \theta) = f(x|\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1}$$

- Key property: If $X_i \sim \Gamma(\alpha_i, \theta), i = 1, \ldots, k$, then

$$\sum_{i=1}^{k} X_i \sim \Gamma\left(\sum_{i=1}^{k} \alpha_i, \theta\right)$$

# Gamma and Dirichlet

- Gamma distribution, with $x > 0, \alpha, \theta > 0$, is

$$\Gamma(\alpha, \theta) = f(x|\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1}$$

- Key property: If $X_i \sim \Gamma(\alpha_i, \theta), i = 1, \ldots, k$, then

$$\sum_{i=1}^{k} X_i \sim \Gamma\left(\sum_{i=1}^{k} \alpha_i, \theta\right)$$

- Let $Z_i = \frac{X_i}{\sum_{i=1} X_i}$, then

$$(Z_1, \ldots, Z_k) \sim D(\alpha_1, \ldots, \alpha_k)$$

# Gamma, Exponential, Geometric

- Recall Gamma distribution

$$\Gamma(\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}$$

# Gamma, Exponential, Geometric

- Recall Gamma distribution

$$\Gamma(\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1}$$

- With $\alpha = 1, \theta = 1/\lambda$, we get exponential distribution

$$f(x|\lambda) = G(1, 1/\lambda) = \lambda \exp(-\lambda x)$$

## Gamma, Exponential, Geometric

- Recall Gamma distribution

$$\Gamma(\alpha, \theta) = \frac{\exp(-x/\theta)}{\theta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}$$

- With $\alpha = 1, \theta = 1/\lambda$, we get exponential distribution

$$f(x|\lambda) = G(1, 1/\lambda) = \lambda \exp(-\lambda x)$$

- Discrete version of exponential is the geometric distribution

$$f(k|q) = (1-q)^{k-1} q$$

# Properties of Dirichlet Distribution

- $(X_1, \dots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$

## Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^k \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$

# Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$
  - Variance $E[X_i^2] = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha+1)}$

# Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$
  - Variance $E[X_i^2] = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha+1)}$
  - Covariance $E[X_i X_j] = \frac{-\alpha_i \alpha_j}{\alpha^2(\alpha+1)}, i \neq j$

# Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$
  - Variance $E[X_i^2] = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha+1)}$
  - Covariance $E[X_i X_j] = \frac{-\alpha_i \alpha_j}{\alpha^2(\alpha+1)}, i \neq j$
  - $X_1$ is independent of $X_2/(1 - X_1), \cdots, X_k/(1 - X_1)$

## Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$
  - Variance $E[X_i^2] = \frac{\alpha_i(\alpha-\alpha_i)}{\alpha^2(\alpha+1)}$
  - Covariance $E[X_i X_j] = \frac{-\alpha_i\alpha_j}{\alpha^2(\alpha+1)}, i \neq j$
  - $X_1$ is independent of $X_2/(1 - X_1), \cdots, X_k/(1 - X_1)$
  - Similarly for each $X_i$

# Properties of Dirichlet Distribution

- $(X_1, \ldots, X_k) \sim D(\alpha_1, \cdots, \alpha_k), \alpha = \sum_{i=1}^{k} \alpha_i$
  - Expectation $E[X_i] = \frac{\alpha_i}{\alpha}$
  - Variance $E[X_i^2] = \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha+1)}$
  - Covariance $E[X_i X_j] = \frac{-\alpha_i \alpha_j}{\alpha^2(\alpha+1)}, i \neq j$
  - $X_1$ is independent of $X_2/(1-X_1), \cdots, X_k/(1-X_1)$
  - Similarly for each $X_i$
- If prior distribution is $D(\alpha_1, \cdots, \alpha_k)$, then posterior

$$P(X_1, \ldots, X_k | X = j) = D(\alpha_1^{(j)}, \cdots, \alpha_k^{(j)})$$

where

$$\alpha_i^{(j)} = \begin{cases} \alpha_i & \text{if } i \neq j \\ \alpha_j + 1 & \text{if } i = j \end{cases}$$

# Dirichlet Processes

- Definition: Let $\alpha$ be a non-negative finite measure on $(X, \mathcal{A})$. Then $P$ is a Dirichlet Process on $(X, \mathcal{A})$ with parameter $\alpha$ if for every $k = 1, 2, \cdots$, and a partition $(B_1, \cdots, B_k)$ of $X$, the distribution of $(P(B_1), \cdots, P(B_k))$ is Dirichlet $D(\alpha(B_1), \cdots, \alpha(B_2))$.

# Dirichlet Processes

- Definition: Let $\alpha$ be a non-negative finite measure on $(X, \mathcal{A})$. Then $P$ is a Dirichlet Process on $(X, \mathcal{A})$ with parameter $\alpha$ if for every $k = 1, 2, \cdots$, and a partition $(B_1, \cdots, B_k)$ of $X$, the distribution of $(P(B_1), \cdots, P(B_k))$ is Dirichlet $D(\alpha(B_1), \cdots, \alpha(B_2))$.
- For any $A \in \mathcal{A}$, $E[P(A)] = \frac{\alpha(A)}{\alpha(X)}$

# Dirichlet Processes

- Definition: Let $\alpha$ be a non-negative finite measure on $(X, \mathcal{A})$. Then $P$ is a Dirichlet Process on $(X, \mathcal{A})$ with parameter $\alpha$ if for every $k = 1, 2, \cdots$, and a partition $(B_1, \cdots, B_k)$ of $X$, the distribution of $(P(B_1), \cdots, P(B_k))$ is Dirichlet $D(\alpha(B_1), \cdots, \alpha(B_2))$.
- For any $A \in \mathcal{A}$, $E[P(A)] = \frac{\alpha(A)}{\alpha(X)}$
- Let $Q$ be a fixed probability measure on $(X, A)$ with $Q \ll \alpha$. Then for any $m$, and any $A_1, \ldots, A_m$, and $\epsilon > 0$,

$$\mathcal{P}\{|P(A_i) - Q(A_i)| < \epsilon, i = 1, \ldots, m\} > 0$$

# Properties of Dirichlet Processes

- Three main properties for DPs

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
    - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$
  - Here $P$ acts as the "parameter," DP is the prior

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$
  - Here $P$ acts as the "parameter," DP is the prior
- Prop 2: DP gives probability 1 to discrete measures on $(X, \mathcal{A})$

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$
  - Here $P$ acts as the "parameter," DP is the prior
- Prop 2: DP gives probability 1 to discrete measures on $(X, \mathcal{A})$
  - Easy to show using a constructive definition of DP

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$
  - Here $P$ acts as the "parameter," DP is the prior
- Prop 2: DP gives probability 1 to discrete measures on $(X, \mathcal{A})$
  - Easy to show using a constructive definition of DP
- Prop 3: The posterior distribution given $X$ is the DP with parameter $\alpha + \delta_X$

# Properties of Dirichlet Processes

- Three main properties for DPs
- Prop 1: DP is a probability measure on $([0,1]^{\mathcal{A}}, \mathcal{F}^{\mathcal{A}})$
  - Samples from a DP are distributions $P$ on $(X, \mathcal{A})$
  - Here $P$ acts as the "parameter," DP is the prior
- Prop 2: DP gives probability 1 to discrete measures on $(X, \mathcal{A})$
  - Easy to show using a constructive definition of DP
- Prop 3: The posterior distribution given $X$ is the DP with parameter $\alpha + \delta_X$
  - Posterior given $X_1, \ldots, X_n$ is the DP with parameter $\alpha + \sum_{i=1}^{n} \delta_{X_i}$

# Stick Breaking Construction

- A constructive definition of DP

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$

## Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
    - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$

## Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0,1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0, 1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$
  - Recall that a r.v. is a measurable function

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0, 1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$
  - Recall that a r.v. is a measurable function
- The distribution of the r.v. is defined as follows

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0,1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$
  - Recall that a r.v. is a measurable function
- The distribution of the r.v. is defined as follows
  - $(\pi_1, \pi_2, \ldots)$ are i.i.d. with distribution $B(1, \alpha(X))$

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0,1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$
  - Recall that a r.v. is a measurable function
- The distribution of the r.v. is defined as follows
  - $(\pi_1, \pi_2, \ldots)$ are i.i.d. with distribution $B(1, \alpha(X))$
  - $(Y_1, Y_2, \ldots)$ are i.i.d. with distribution $\beta(A) = \alpha(A)/\alpha(X)$

# Stick Breaking Construction

- A constructive definition of DP
- Let $\alpha$ be a finite measure on $(X, \mathcal{A})$
- Let $N = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^N$
- Construct a probability space $(\Omega, \mathcal{S}, Q)$
  - Random variables $(\pi, Y, I) = ((\pi_j, Y_j), j = 1, 2, \ldots, I)$
  - Taking values in $(([0,1] \times \mathcal{X})^\infty \times N, (\mathcal{B} \times \mathcal{A})^\infty)$
  - Recall that a r.v. is a measurable function
- The distribution of the r.v. is defined as follows
  - $(\pi_1, \pi_2, \ldots)$ are i.i.d. with distribution $B(1, \alpha(X))$
  - $(Y_1, Y_2, \ldots)$ are i.i.d. with distribution $\beta(A) = \alpha(A)/\alpha(X)$
  - $Q(I = n | (\pi, Y)) = p_n = \pi_n \prod_{1 \le m \le (n-1)} (1 - \pi_m)$ so that

$$\sum_{1 \le m \le n} p_n = 1 - \prod_{1 \le m \le n} (1 - \pi_m) \to 1 \quad w.p.\ 1$$

# Stick Breaking Construction (Contd.)

- Now, we have a probability space $(\Omega, \mathcal{S}, Q)$

## Stick Breaking Construction (Contd.)

- Now, we have a probability space $(\Omega, \mathcal{S}, Q)$
- For any $A \in \mathcal{A}$, define

$$P_{(\theta, Y)}(A) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(A)$$

## Stick Breaking Construction (Contd.)

- Now, we have a probability space $(\Omega, \mathcal{S}, Q)$
- For any $A \in \mathcal{A}$, define

$$P_{(\theta,Y)}(A) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(A)$$

- $P$ is a random measure over $(X, \mathcal{A})$, due to $(\theta, Y)$

# Stick Breaking Construction (Contd.)

- Now, we have a probability space $(\Omega, \mathcal{S}, Q)$
- For any $A \in \mathcal{A}$, define

$$P_{(\theta,Y)}(A) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(A)$$

- $P$ is a random measure over $(X, \mathcal{A})$, due to $(\theta, Y)$
- $P$ is a sample from a Dirichlet process with parameter $\alpha$

# Stick Breaking Construction (Contd.)

- Now, we have a probability space $(\Omega, \mathcal{S}, Q)$
- For any $A \in \mathcal{A}$, define

$$P_{(\theta,Y)}(A) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(A)$$

- $P$ is a random measure over $(X, \mathcal{A})$, due to $(\theta, Y)$
- $P$ is a sample from a Dirichlet process with parameter $\alpha$
- By construction, clearly $P$ can only be discrete

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined
- Based on a fixed measure $\alpha$ on $\mathcal{A}$

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined
- Based on a fixed measure $\alpha$ on $\mathcal{A}$
- Consider a probability space $(U, \mathcal{B}, H)$

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined
- Based on a fixed measure $\alpha$ on $\mathcal{A}$
- Consider a probability space $(U, \mathcal{B}, H)$
- Define a transition measure $\alpha(u, A)$ on $U \times \mathcal{A}$

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined
- Based on a fixed measure $\alpha$ on $\mathcal{A}$
- Consider a probability space $(U, \mathcal{B}, H)$
- Define a transition measure $\alpha(u, A)$ on $U \times \mathcal{A}$
- For any $A_1, \ldots, A_m \in \mathcal{A}$, we have

$$(P(A_1), \ldots, P(A_m)) \sim \int_u D(\alpha(u, A_1), \ldots, D(u, A_m))dH(u)$$

# Dirichlet Process Mixtures

- $(X, \mathcal{A})$ is the space on which DP was defined
- Based on a fixed measure $\alpha$ on $\mathcal{A}$
- Consider a probability space $(U, \mathcal{B}, H)$
- Define a transition measure $\alpha(u, A)$ on $U \times \mathcal{A}$
- For any $A_1, \ldots, A_m \in \mathcal{A}$, we have

$$(P(A_1), \ldots, P(A_m)) \sim \int_u D(\alpha(u, A_1), \ldots, D(u, A_m))dH(u)$$

- In "practice" DPM is a infinite mixture model

# DPM (Contd.)

- Mike Jordan's NIPS'05 Tutorial

# Model-Based Clustering

- A generative approach to clustering:
  - pick one of $K$ clusters from a distribution $\pi = (\pi_1, \pi_2, \ldots \pi_K)$
  - generate a data point from a cluster-specific probability distribution

- This yields a finite mixture model:

$$p(x \mid \phi, \pi) = \sum_{k=1}^{K} \pi_k \, p(x \mid \phi_k),$$

  where $\pi$ and $\phi = (\phi_1, \phi_2, \ldots \phi_K)$ are the parameters, and where we've assumed the same parameterized family for each cluster (for simplicity)

- Data $\{x_i\}_{i=1}^{n}$ are assumed to be generated conditionally IID from this mixture

# Finite Mixture Models (cont)

- Another way to express this: define an underlying measure

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

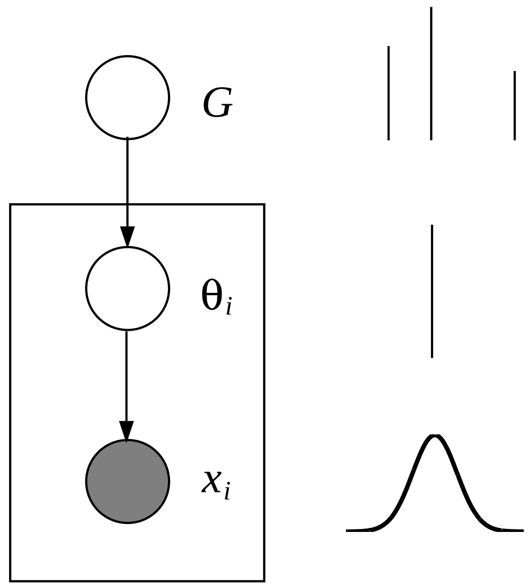
  where $\delta_{\phi_k}$ is an *atom* at $\phi_k$

- And define the process of obtaining a sample from a finite mixture model as follows. For $i = 1, \ldots, n$:

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(\cdot \mid \theta_i) \end{aligned}$$

- Note that each $\theta_i$ is equal to one of the underlying $\phi_k$
  - indeed, the subset of $\{\theta_i\}$ that maps to $\phi_k$ is exactly the $k$th cluster

# Finite Mixture Models (cont)




$$
\begin{aligned}
G &= \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot \mid \theta_i)
\end{aligned}
$$

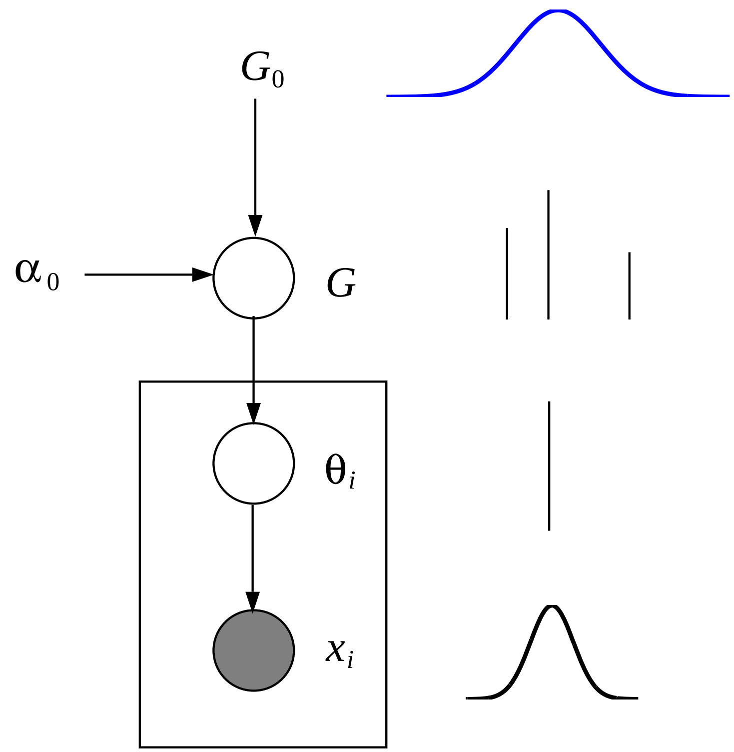
# Bayesian Finite Mixture Models

(e.g., Lo; Ferguson; Escobar & West; Robert; Green & Richardson; Neal; Ishawaran & Zarepour)

- Need to place priors on the parameters $\phi$ and $\pi$
- The choice of prior for $\phi$ is model-specific; e.g., we might use conjugate normal/inverse-gamma priors for a Gaussian mixture model
  - let's denote this prior as $G_0$
- Place a symmetric Dirichlet prior, $\mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K)$, on the mixing proportions $\pi$
  - the symmetry accords with the (usual) assumption that we could scramble the labels of the mixture components and not change the model
  - the scaling $(\alpha_0/K)$ gives $\alpha_0$ the semantics of a concentration parameter; the prior mean of $\phi_k$ is equal to $1/K$

# Bayesian Finite Mixture Models (cont)

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi_k &\sim \mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \\
G &= \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(\cdot \,|\, \theta_i)
\end{aligned}
$$




- Note that $G$ is now a *random measure*

# Going Nonparametric—A First Perspective

(e.g., Kingman; Waterson; Patil & Taillie; Liu; Ishwaran & Zarepour)

- Define a countably infinite mixture model by taking $K$ to infinity and hoping that "$G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k}$" means something, where

$$\begin{aligned} \phi_k &\sim G_0 \\ \pi_k &\sim \mathrm{Dir}(\alpha_0/K, \ldots, \alpha_0/K) \text{ as } K \to \infty \end{aligned}$$

- Several mathematical hurdles to overcome:
  - What is the distribution of any given $\pi_k$ as $K \to \infty$? Does it stabilize at some fixed distribution?
  - Is $\sum_{k=1}^{\infty} \pi_k = 1$ under some suitable notion of convergence?
  - Do we get a few large mixing proportions, or are they all of similar "size"?
  - Do we get any "clustering" at all?

- This seems hard; let's approach the problem from a different point of view

# A Second Perspective—Stick-Breaking

(e.g., Connor & Mosimann; Doksum; Freedman; Kingman; Pitman; Sethuraman)

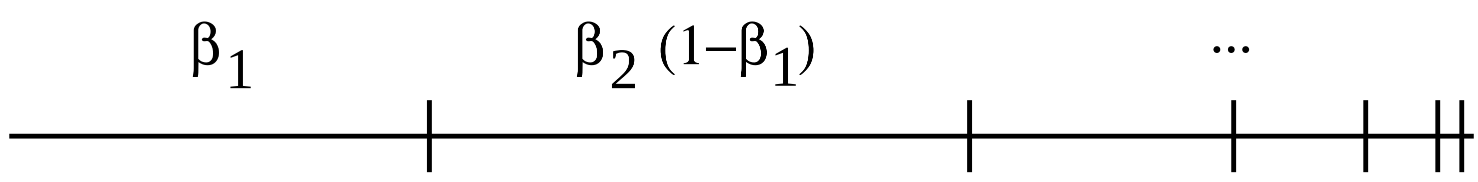
- Define an infinite sequence of Beta random variables:

$$\beta_k \sim \mathrm{Beta}(1, \alpha_0) \qquad\qquad k = 1, 2, \ldots$$

- And then define an infinite sequence of mixing proportions as:

$$\begin{aligned} \pi_1 &= \beta_1 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad\qquad k = 2, 3, \ldots \end{aligned}$$

- This can be viewed as breaking off portions of a stick:



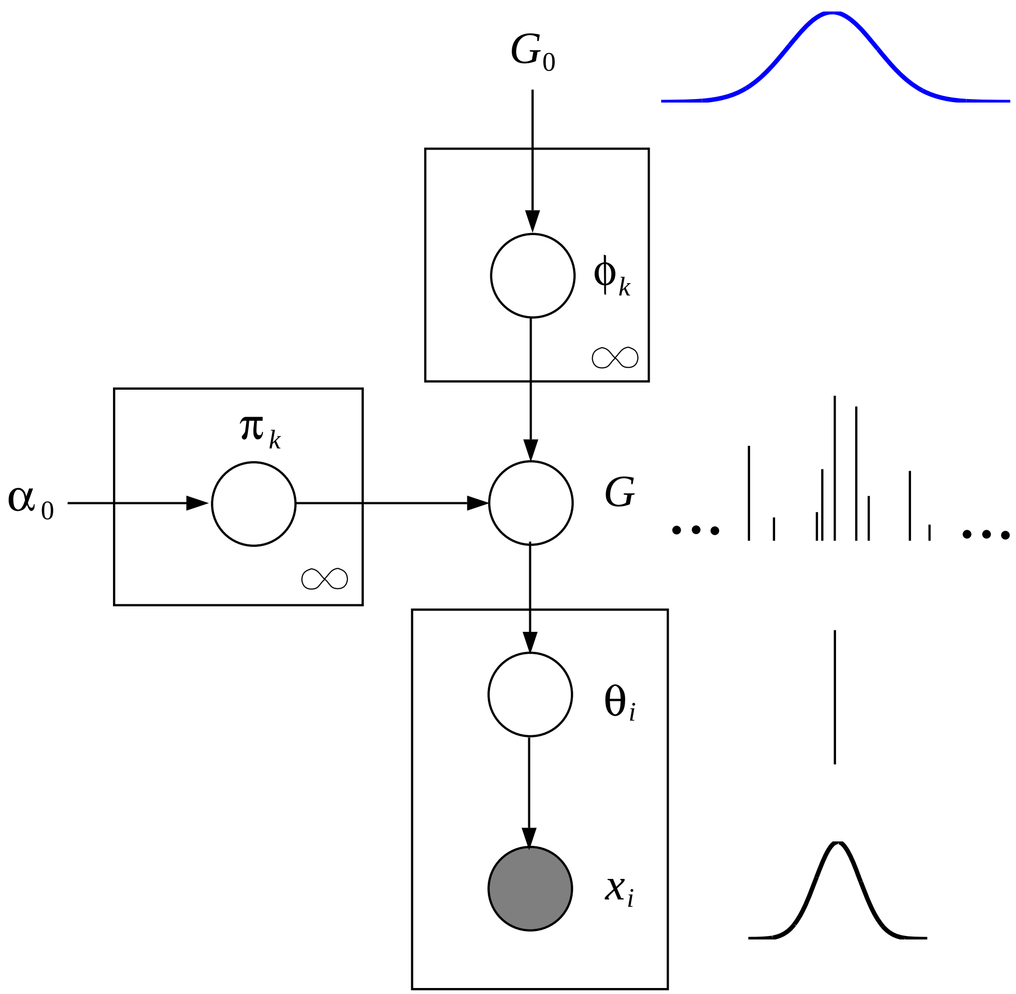
## Stick-Breaking (cont)

- We now have an explicit formula for each $\pi_k$: $\beta_k \prod_{l=1}^{k-1}(1-\beta_l)$
- We can also easily see that $\sum_{k=1}^{\infty} \pi_k = 1$ (wp1):

$$
\begin{aligned}
1 - \sum_{k=1}^{K} \pi_k &= 1 - \beta_1 - \beta_2(1-\beta_1) - \beta_3(1-\beta_1)(1-\beta_2) - \cdots \\
&= (1-\beta_1)(1-\beta_2-\beta_3(1-\beta_2) - \cdots) \\
&= \prod_{k=1}^{K}(1-\beta_k) \\
&\to 0 \qquad (\text{wp1 as } K \to \infty)
\end{aligned}
$$

- So now $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ has a clean definition as a random measure

# Graphical Model Representation



# The Posterior Dirichlet Process

- Suppose that we sample $G$ from a Dirichlet process and then sample $\theta_1$ from $G$. What is the posterior process?

- For a fixed partition, we get a standard Dirichlet update (for the cell that contains $\theta_1$ the exponent increases by one; stays the same for all other cells)
  - this is true for even the tiniest cell
  - suggests that the posterior is a Dirichlet process in which the base measure has an atom at $\theta_1$

- Indeed, we have (for a proof, see, e.g., Schervish, 1995):

$$G \mid \theta_1 \sim \mathrm{DP}(\alpha_0 G_0 + \delta_{\theta_1})$$

- Iterating the posterior update yields:

$$G \mid \theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha_0 G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$$

# Relationship to Stick-Breaking

- Recalling the formula for the expectation of a Dirichlet random variable, for any set $A \subseteq \Omega$, we have:

$$\mathbb{E}[G(A) \mid \theta_1, \ldots, \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{i=1}^n \delta_{\theta_i}(A)}{\alpha_0 + n} \rightarrow \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}(A)$$

  where $\phi_k$ are the unique values of the $\theta_i$, where $\pi_k = \lim_{n\rightarrow\infty} n_k/n$, and where $n_k$ is the number of repeats of $\phi_k$ in the sequence $(\theta_1, \ldots, \theta_n)$

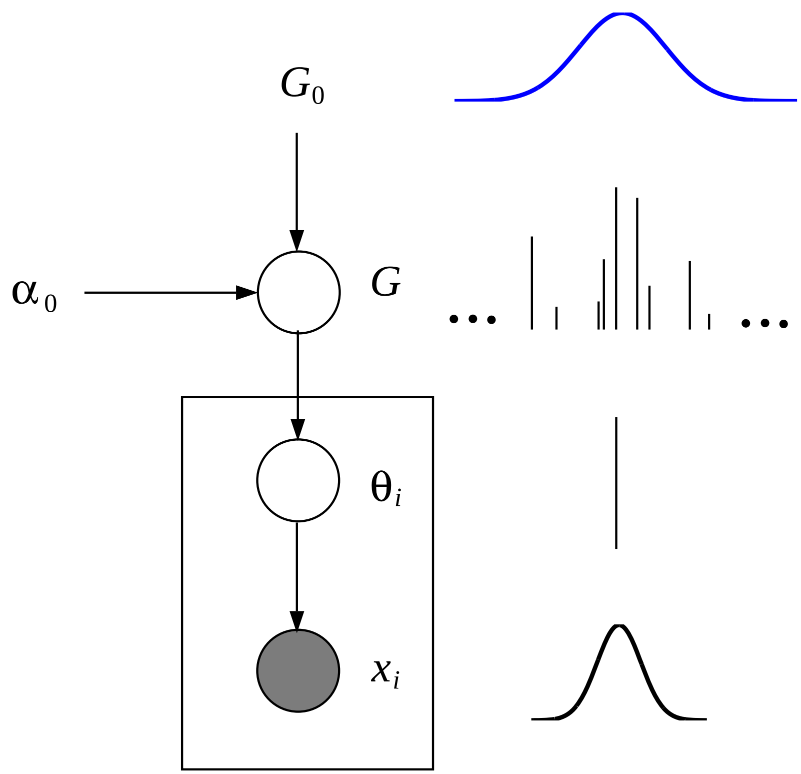
  – assuming that the posterior concentrates, this suggests that the random measures $G \sim \mathrm{DP}(\alpha_0 G_0)$ are discrete (wp1)

- Is there an infinite sum of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ that obeys the definition of the Dirichlet process?

  – yes, the stick-breaking random measure!

  – this important result is not hard to prove; it follows from elementary facts about the Dirichlet distribution (Sethuraman, 1994)

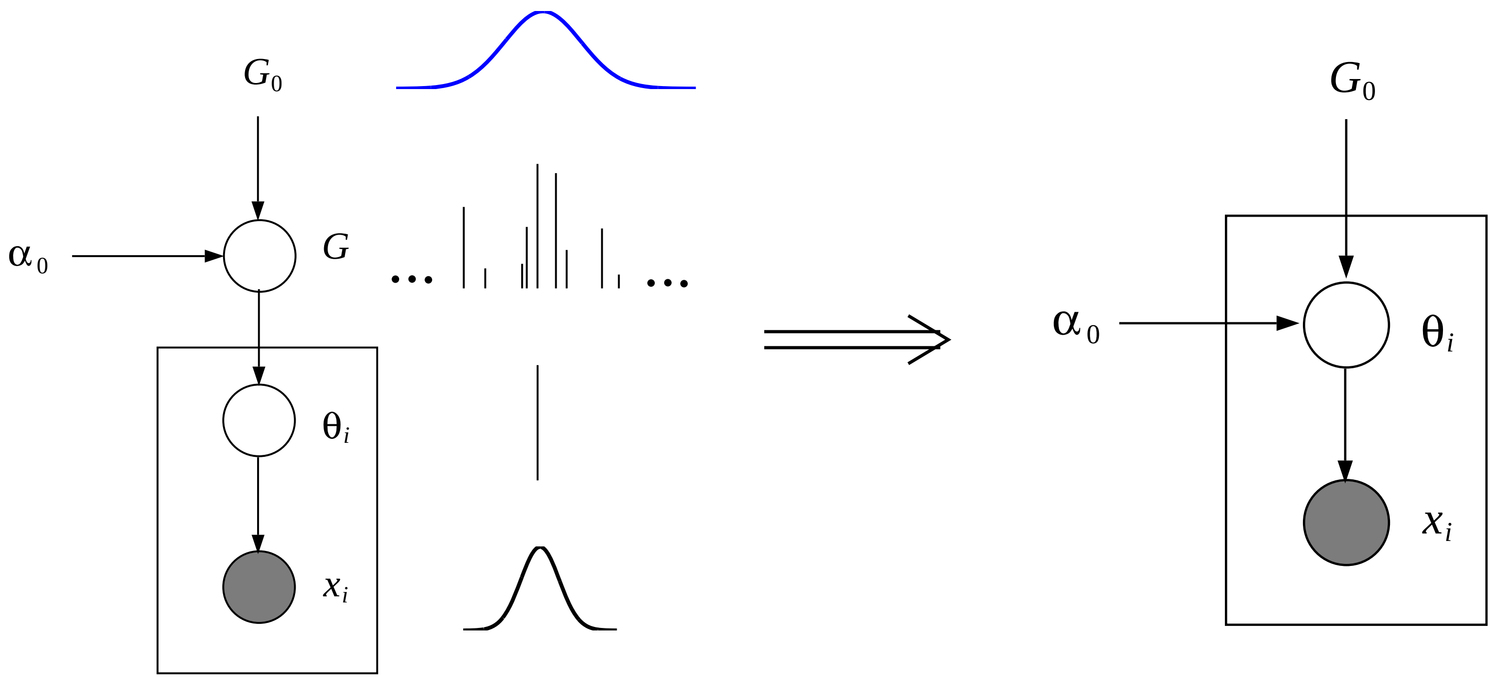
# Dirichlet Process Mixture Models




$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha_0 G_0) \\
\theta_i \,|\, G &\sim G \qquad\qquad\qquad i \in 1, \ldots, n \\
x_i \,|\, \theta_i &\sim F(x_i \,|\, \theta_i) \qquad i \in 1, \ldots, n
\end{aligned}
$$

# Marginal Probabilities

- To obtain the marginal probability of the parameters $\theta_1, \theta_2, \ldots$, we need to integrate out $G$



# Marginal Probabilities (cont)

- Recall the formula

$$\mathbb{E}[G(A) \mid \theta_1, \ldots, \theta_n] = \frac{\alpha_0 G_0(A) + \sum_{k=1}^{K} n_k \delta_{\phi_k}(A)}{\alpha_0 + n}$$

- Let $A$ be a singleton set equal to one of the $\phi_k$. The formula says that the marginal probability of observing $\phi_k$ again is proportional to $n_k$.

- And the marginal probability of observing a new $\phi$ vector is proportional to $\alpha_0$.

- This is just the Pólya urn scheme!

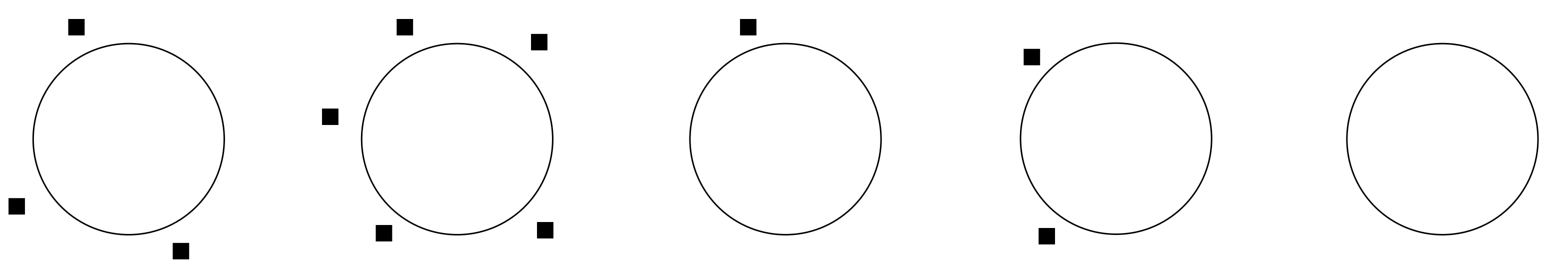
- I.e., integrating over the random measure $G$, where $G \sim \mathrm{DP}(\alpha_0 G_0)$, yields the Pólya urn

# Chinese Restaurant Process (CRP)

- A random process in which $n$ customers sit down in a Chinese restaurant with an infinite number of tables
  - first customer sits at the first table
  - $m$th subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} P(\text{previously occupied table } i \,|\, \mathcal{F}_{m-1}) &\propto n_i \\ P(\text{the next unoccupied table} \,|\, \mathcal{F}_{m-1}) &\propto \alpha_0 \end{aligned} \quad (1)$$

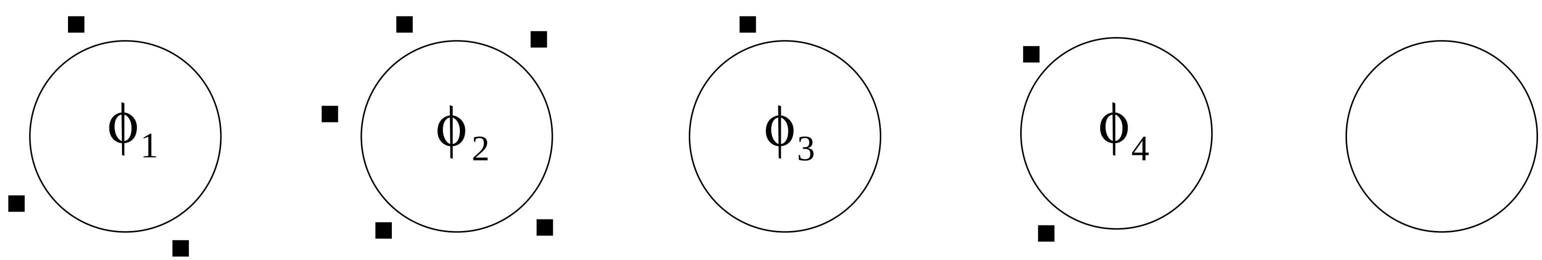
  where $n_i$ is the number of customers currently at table $i$ and where $\mathcal{F}_{m-1}$ denotes the state of the restaurant after $m-1$ customers have been seated



# The CRP and Clustering

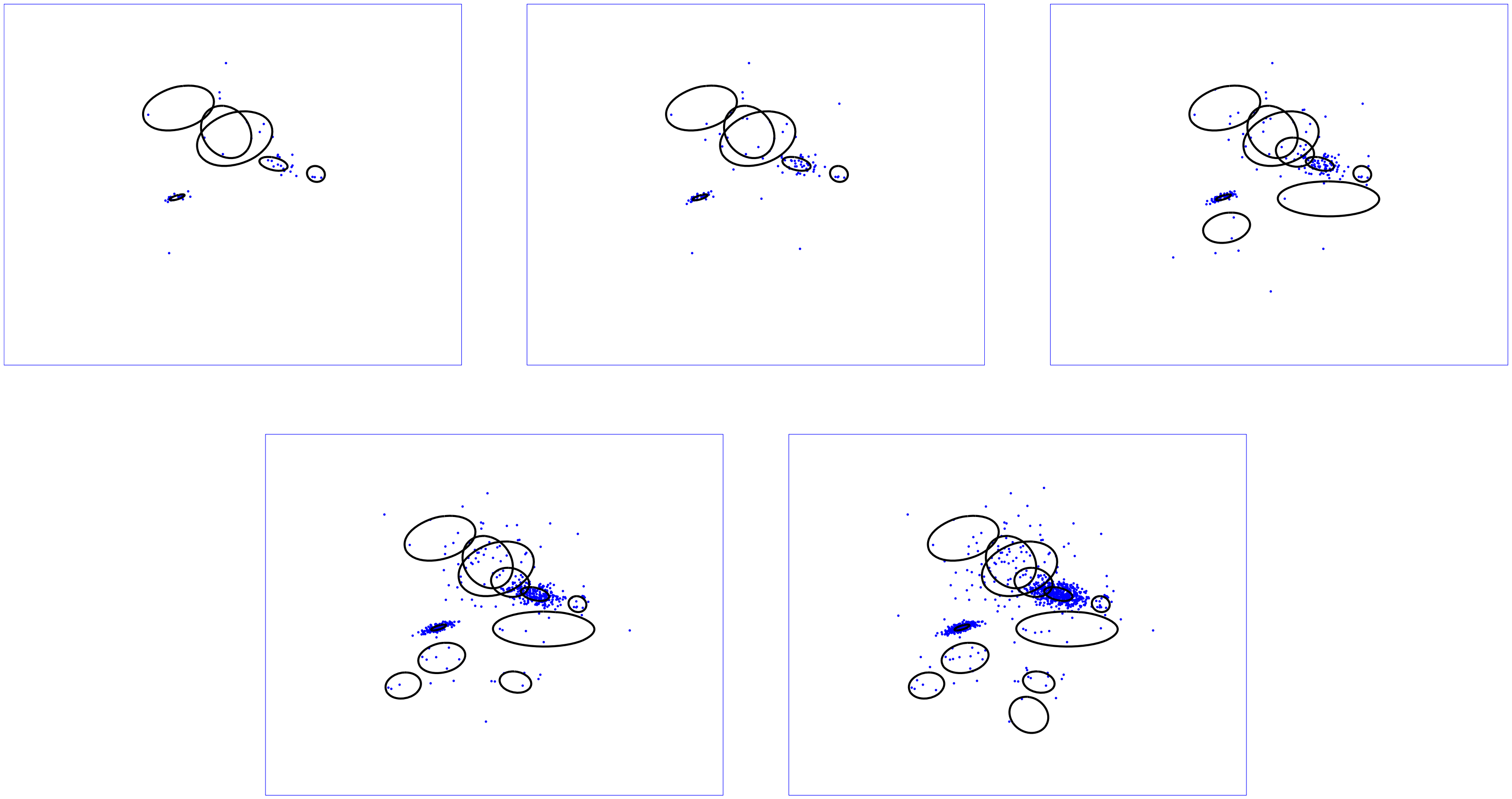
- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables

- This prior can be completed with:
  - a likelihood—e.g., associate a parameterized probability distribution with each table
  - a prior for the parameters—the first customer to sit at table $k$ chooses the parameter vector for that table ($\phi_k$) from the prior




- So we now have a distribution—or can obtain one—for any quantity that we might care about in the clustering setting

# CRP Prior, Gaussian Likelihood, Conjugate Prior



$$\phi_k = (\mu_k, \Sigma_k) \sim N(a,b) \otimes IW(\alpha, \beta)$$

$$x_i \sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k$$