School of Computer Science
Carnegie Mellon

10-708 Probabilistic Graphical Models

The Dirichlet Process (DP) and DP Mixture Models

Matt Gormley
Lecture 18
March 21, 2016

Readings:
Teh (2010)

Reminders

- Midway Project Report
  - Due March 23, 12:00 noon
- Course Survey #1
- Today: wrap up Topic Modeling

Outline

- **Motivation / Applications**
- **Background**
  - de Finetti Theorem
  - Exchangeability
  - Aglommerative and decimative properties of Dirichlet distribution
- **CRP and CRP Mixture Model**
  - Chinese Restaurant Process (CRP) definition
  - Gibbs sampling for CRP-MM
  - Expected number of clusters
- **DP and DP Mixture Model**
  - Ferguson definition of Dirichlet process (DP)
  - Stick breaking construction of DP
  - Uncollapsed blocked Gibbs sampler for DP-MM
  - Truncated variational inference for DP-MM
- **DP Properties**
- **Related Models**
  - Hierarchical Dirichlet process Mixture Models (HDP-MM)
  - Infinite HMM
  - Infinite PCFG

Parametric vs. Nonparametric

- **Parametric models:**
  - **Finite** and **fixed** number of parameters
  - Number of parameters is **independent of the dataset**
- **Nonparametric models:**
  - **Have** parameters ("**infinite dimensional**" would be a better name)
  - Can be understood as having an **infinite** number of parameters
  - Can be understood as having a **random** number of parameters
  - Number of parameters can **grow with the dataset**
- **Semiparametric models:**
  - Have a **parametric** component and a **nonparametric** component

Parametric vs. Nonparametric

	Frequentist	Bayesian
Parametric	Logistic regression, ANOVA, Fisher discriminant analysis, ARMA, etc.	Conjugate analysis, hierarchical models, conditional random fields
Semiparametric	Independent component analysis, Cox model, nonmetric MDS, etc.	[Hybrids of the above and below cells]
Nonparametric	Nearest neighbor, kernel methods, boostrap, decision trees, etc.	Gaussian processes, Dirichlet processes, Pitman-Yor processes, etc.

Table adapted from Ghahramani 2015

Parametric vs. Nonparametric

Application	Parametric	Nonparametric
function approximation	polynomial regression	Gaussian processes
classification	logistic regression	Gaussian process classifiers
clustering	mixture model, k-means	Dirichlet process mixture model
time series	hidden Markov model	infinite HMM
feature discovery	factor analysis, pPCA, PMF	infinite latent factor models

Parametric vs. Nonparametric

- **Def**: a *model* is a collection of distributions

$$\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$$

- *parametric model*: the parameter vector is finite dimensional

$$\Theta \subset \mathcal{R}^k$$

- *nonparametric model*: the parameters are from a possibly infinite dimensional space, $\mathcal{F}$

$$\Theta \subset \mathcal{F}$$

Motivation #1

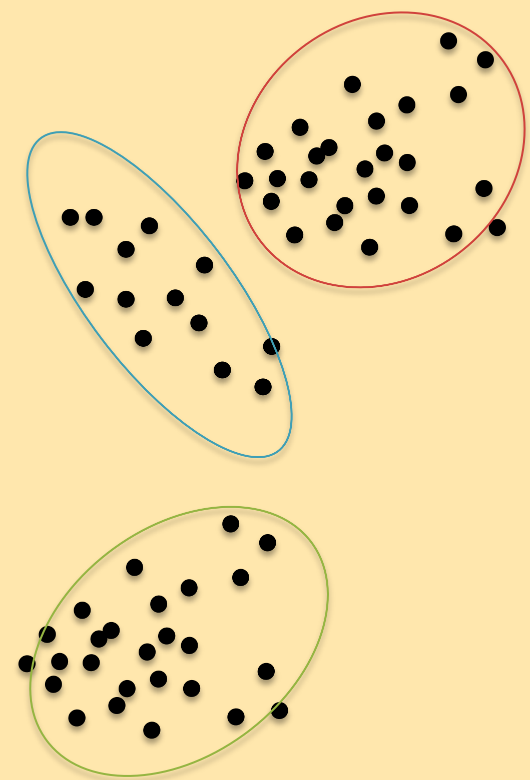
Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?

Motivation #1

Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?



Motivation #1

Model Selection

- For clustering:
  How many clusters in a **mixture model**?
- For topic modeling:
  How many topics in **LDA**?
- For grammar induction:
  How many non-terminals in a **PCFG**?
- For visual scene analysis:
  How many objects, parts, features?

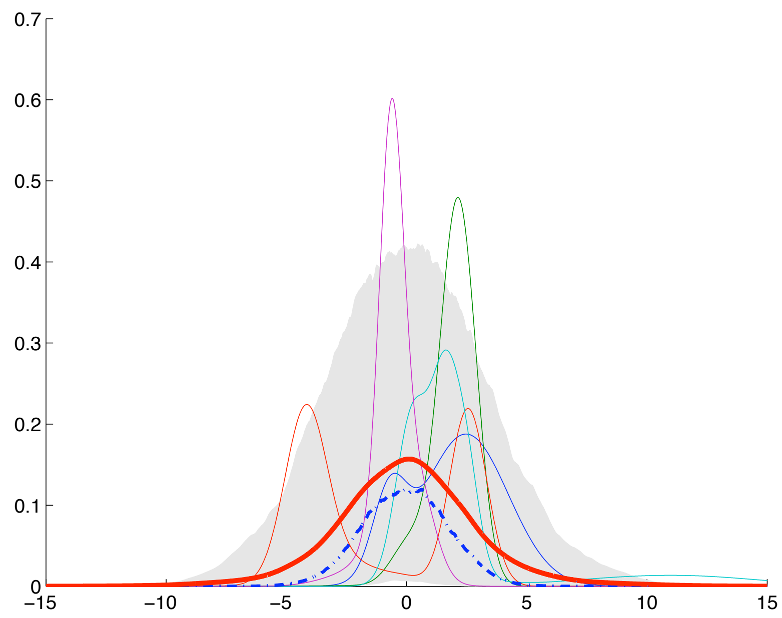
1. **Parametric approaches:** cross-validation, bootstrap, AIC, BIC, DIC, MDL, Laplace, bridge sampling, etc.
2. **Nonparametric approach:** average of an infinite set of models

Motivation #2

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Prior:



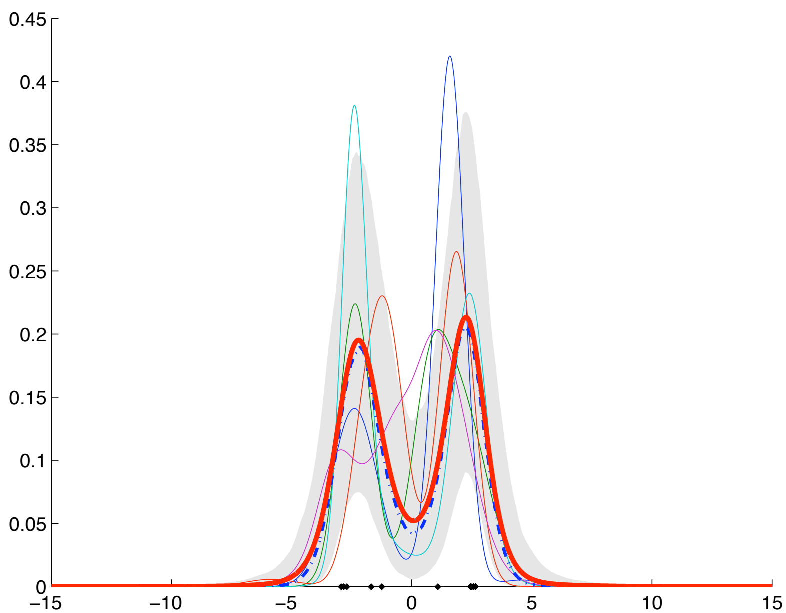
Red: mean density. Blue: median density. Grey: 5-95 quantile.
Others: draws.

Motivation #2

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Posterior:

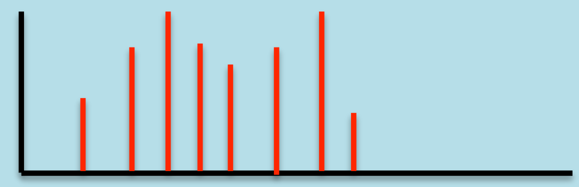


Red: mean density. Blue: median density. Grey: 5-95 quantile.
Black: data. Others: draws.

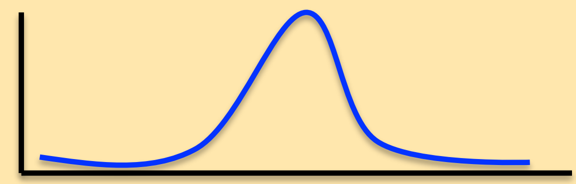
Background

Suppose we have a random variable X drawn from some distribution $P_\theta(X)$ and X ranges over a set $\mathcal{S}$.

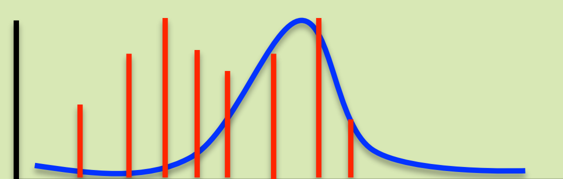
- Discrete distribution:
  $\mathcal{S}$ is a countable set.



- Continuous distribution:
  $P_\theta(X = x) = 0$ for all $x \in \mathcal{S}$



- Mixed distribution:
  $\mathcal{S}$ can be partitioned into two disjoint sets $\mathcal{D}$ and $\mathcal{C}$ s.t.
  1. $\mathcal{A}$ is countable and $0 < P_\theta(X \in D) < 1$
  2. $P_\theta(X = x) = 0$ for all $x \in \mathcal{C}$



Exchangability and de Finetti's Theorem

Exchangeability:

- **Def #1:** a joint probability distribution is **exchangeable** if it is invariant to permutation
- **Def #2:** The possibly infinite sequence of random variables $(X_1, X_2, X_3, \ldots)$ is **exchangeable** if for any finite permutation s of the indices $(1, 2, \ldots n)$:

$$P(X_1, X_2, \ldots, X_n) = P(X_{s(1)}, X_{s(2)}, \ldots, X_{s(n)})$$

Notes:

- *i.i.d.* and *exchangeable* are not the same!
- the latter says that if our data are reordered it doesn't matter

Exchangability and de Finetti's Theorem

Slide from Jordan ICML 2005

Theorem (De Finetti, 1935). *If* $(x_1, x_2, \ldots)$ *are infinitely exchangeable, then the joint probability* $p(x_1, x_2, \ldots, x_N)$ *has a representation as a mixture:*

$$p(x_1, x_2, \ldots, x_N) = \int \left(\prod_{i=1}^{N} p(x_i \,|\, \theta) \right) dP(\theta)$$

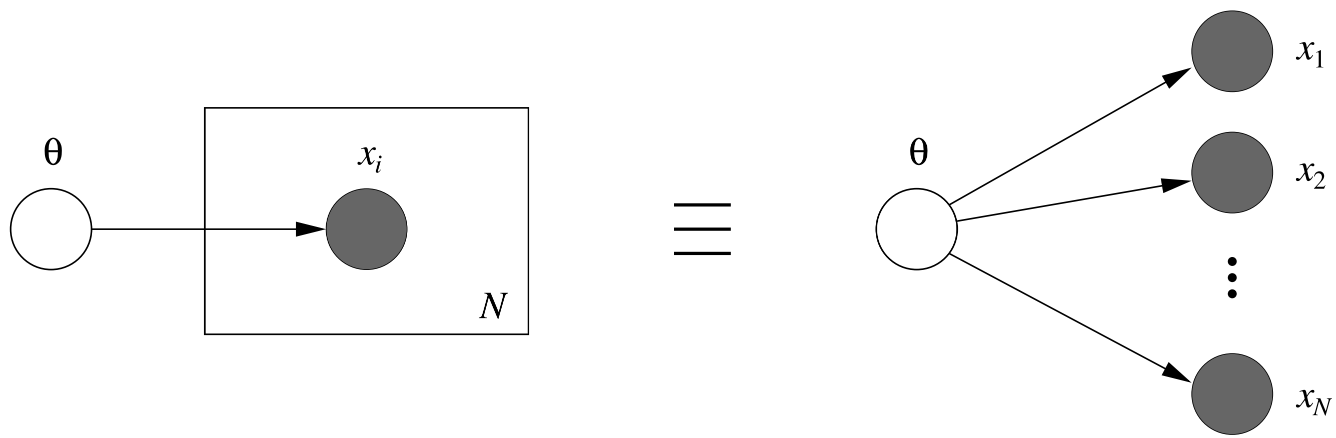
for some random variable θ.

- The theorem wouldn't be true if we limited ourselves to parameters θ ranging over Euclidean vector spaces
- In particular, we need to allow θ to range over measures, in which case $P(\theta)$ is a measure on measures
  - the Dirichlet process is an example of a measure on measures...

Actually, this is the Hewitt-Savage generalization of the de Finetti theorem.
The original version was given for the Bernoulli distribution

Exchangability and de Finetti's Theorem

- A *plate* is a "macro" that allows subgraphs to be replicated:



- Note that this is a graphical representation of the De Finetti theorem

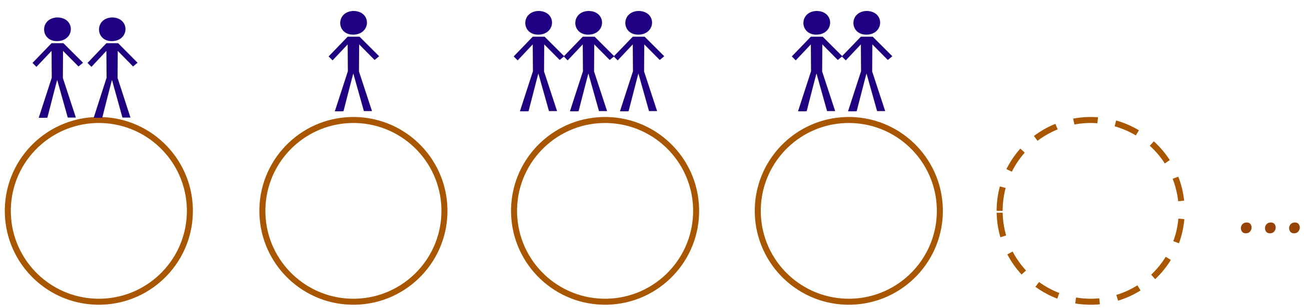
$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \left(\prod_{i=1}^{N} p(x_i \mid \theta) \right) d\theta$$

Chinese Restaurant Process

- Imagine a Chinese restaurant with an infinite number of tables
- Each customer enters and sits down at a table
  - The first customer sits at the first unoccupied table
  - Each subsequent customer chooses a table according to the following probability distribution:

    $p(kth\ occupied\ table) \propto n_k$
    $p(next\ unoccupied\ table) \propto \alpha$

    where n_k is the number of people sitting at the table k



$$\frac{2}{8+\alpha} \qquad \frac{1}{8+\alpha} \qquad \frac{3}{8+\alpha} \qquad \frac{2}{8+\alpha} \qquad \frac{\alpha}{8+\alpha}$$

Chinese Restaurant Process

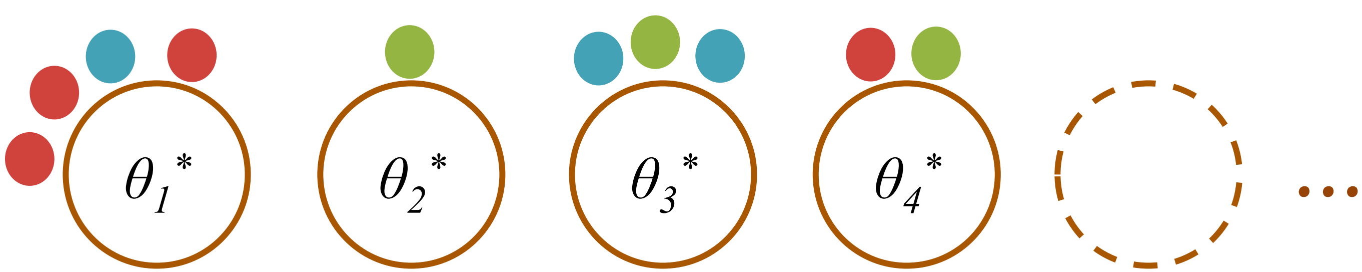
Properties:

- CRP defines a **distribution over clusterings** (i.e. partitions) of the indices $1, \ldots, n$
  - customer = index
  - table = cluster
- **Expected number of clusters** given n customers (i.e. observations) is $O(\alpha \log(n))$
  - *rich-get-richer* effect on clusters: popular tables tend to get more crowded
- Behavior of CRP with α:
  - As α goes to 0, the number of clusters goes to 1
  - As α goes to $+\infty$, the number of clusters goes to n
- The CRP is an **exchangeable process**
- We write $z_1, z_2, \ldots, z_n \sim CRP(\alpha)$ to denote a **sequence of cluster indices** drawn from a Chinese Restaurant Process

CRP Mixture Model

- Draw n cluster indices from a CRP:
  $z_1, z_2, \ldots, z_n \sim CRP(\alpha)$
- For each of the resulting K clusters:
  $\theta_k^* \sim H$
  where H is a base distribution
- Draw n observations:
  $$x_i \sim p(x_i \mid \theta^*_{z_i})$$

Customer i orders a dish x_i (observation) from a table-specific distribution over dishes θ_k^* (cluster parameters)



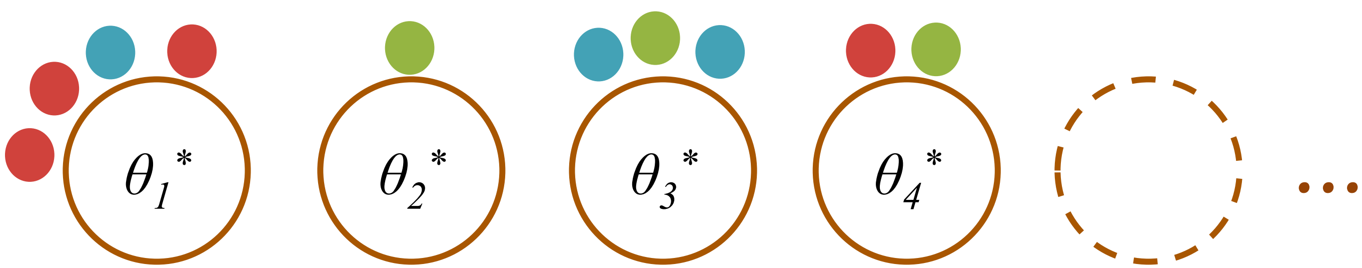
(color denotes different values of x_i)

CRP Mixture Model

- Draw n cluster indices from a CRP:
  $z_1, z_2, \ldots, z_n \sim CRP(\alpha)$
- For each of the resulting K clusters:
  $\theta_k^* \sim H$
  where H is a base distribution
- Draw n observations:
  $$x_i \sim p(x_i \mid \theta^*_{z_i})$$

- The Gibbs sampler is easy thanks to **exchangeability**
- For each observation, we remove the customer / dish from the restaurant and resample as if they were the **last to enter**
- If we **collapse out the parameters,** the Gibbs sampler draws from the conditionals:

  $z_i \sim p(z_i \mid \boldsymbol{z}_{-i}, \boldsymbol{x})$



(color denotes different values of x_i)

CRP Mixture Model

Overview of 3 Gibbs Samplers for Conjugate Priors

- Alg. 1: (uncollapsed)
  - Markov chain state: per-customer parameters $\theta_1, \ldots, \theta_n$
  - For $i = 1, \ldots, n$: Draw $\theta_i \sim p(\theta_i \mid \boldsymbol{\theta}_{-i}, \boldsymbol{x})$

All the thetas except θ_i

- Alg. 2: (uncollapsed)
  - Markov chain state: per-customer cluster indices $z_1, \ldots, z_n$ and per-cluster parameters $\theta_1^*, \ldots, \theta_k^*$
  - For $i = 1, \ldots, n$: Draw $z_i \sim p(z_i \mid \boldsymbol{z}_{-i}, \boldsymbol{x}, \boldsymbol{\theta}^*)$
  - Set K = number of clusters in $\boldsymbol{z}$
  - For $k = 1, \ldots, K$: Draw $\theta_k^* \sim p(\theta_k^* \mid \{x_i : z_i = k\})$
- Alg. 3: (collapsed)
  - Markov chain state: per-customer cluster indices $z_1, \ldots, z_n$
  - For $i = 1, \ldots, n$: Draw $z_i \sim p(z_i \mid \boldsymbol{z}_{-i}, \boldsymbol{x})$

CRP Mixture Model

- Q: How can the Alg. 2 Gibbs samplers permit an infinite set of clusters in finite space?
- A: Easy!
  - We are only representing a finite number of clusters at a time – those to which the data have been assigned
  - We can always bring back the parameters for the "next unoccupied table" if we need them

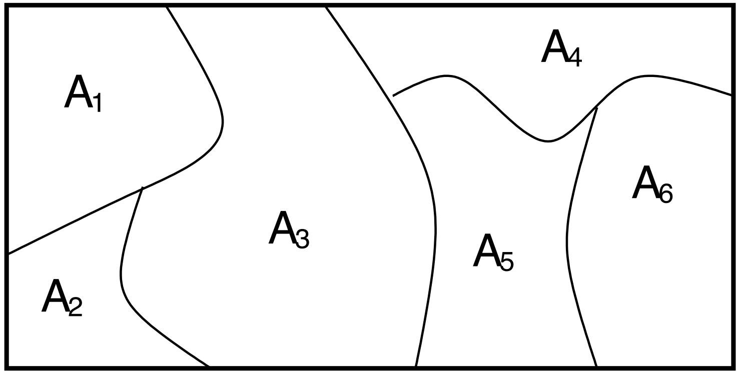
Dirichlet Process

Ferguson Definition

- Parameters of a DP:
  1. Base distribution, H, is a probability distribution over Θ
  2. Strength parameter, $\alpha \in \mathcal{R}$
- We say $G \sim \mathrm{DP}(\alpha, H)$
  if for any partition $A_1 \cup A_2 \cup \ldots \cup A_K = \Theta$
  we have:
  $(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$

In English: the DP is a distribution over probability measures s.t. marginals on finite partitions are Dirichlet distributed

A partition of the space Θ



Whiteboard

- Stick-breaking construction of the DP

Properties of the DP

1. Base distribution is the "mean" of the DP:

$$\mathbb{E}[G(A)] = H(A) \text{ for any } A \subset \Theta$$

2. Strength parameter is like "inverse variance"

$$V[G(A)] = H(A)(1 - H(A))/(\alpha + 1)$$

3. Samples from a DP are discrete distributions
   (stick-breaking construction of $G \sim \mathrm{DP}(\alpha, H)$
   makes this clear)

4. Posterior distribution of $G \sim \mathrm{DP}(\alpha, H)$
   given samples $\theta_1, \ldots, \theta_n$ from G is a DP

$$G|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right)$$

Whiteboard

- Dirichlet Process Mixture Model (stick-breaking version)

CRP-MM vs. DP-MM

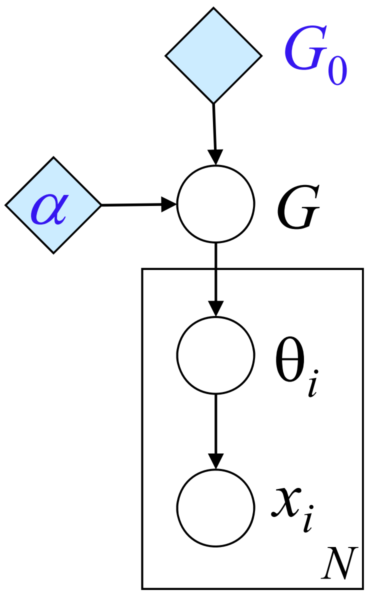
For both the CRP and stick-breaking constructions, if we marginalize out G, we have the following predictive distribution:

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i} \right)$$

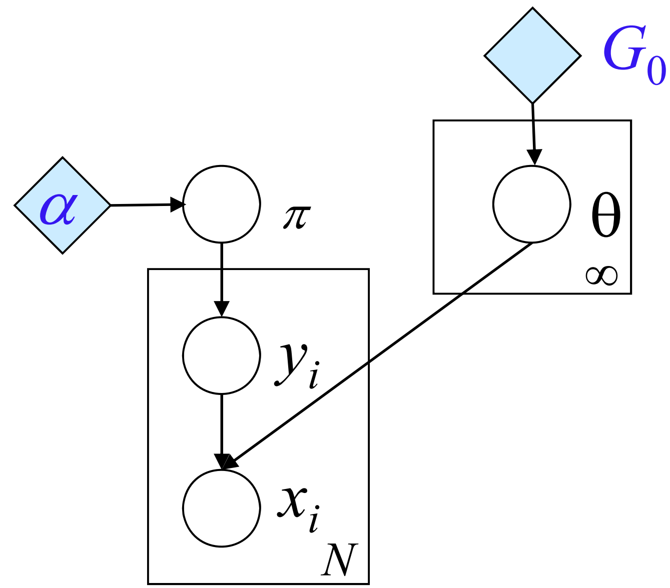
(Blackwell-MacQueen Urn Scheme)

The Chinese Restaurant Process Mixture Model is just a different construction of the Dirichlet Process Mixture Model where we have marginalized out G

Graphical Models for DPs



The Pólya urn construction



The Stick-breaking construction

Figure from Blei & Jordan (2006)

Example: DP Gaussian Mixture Model

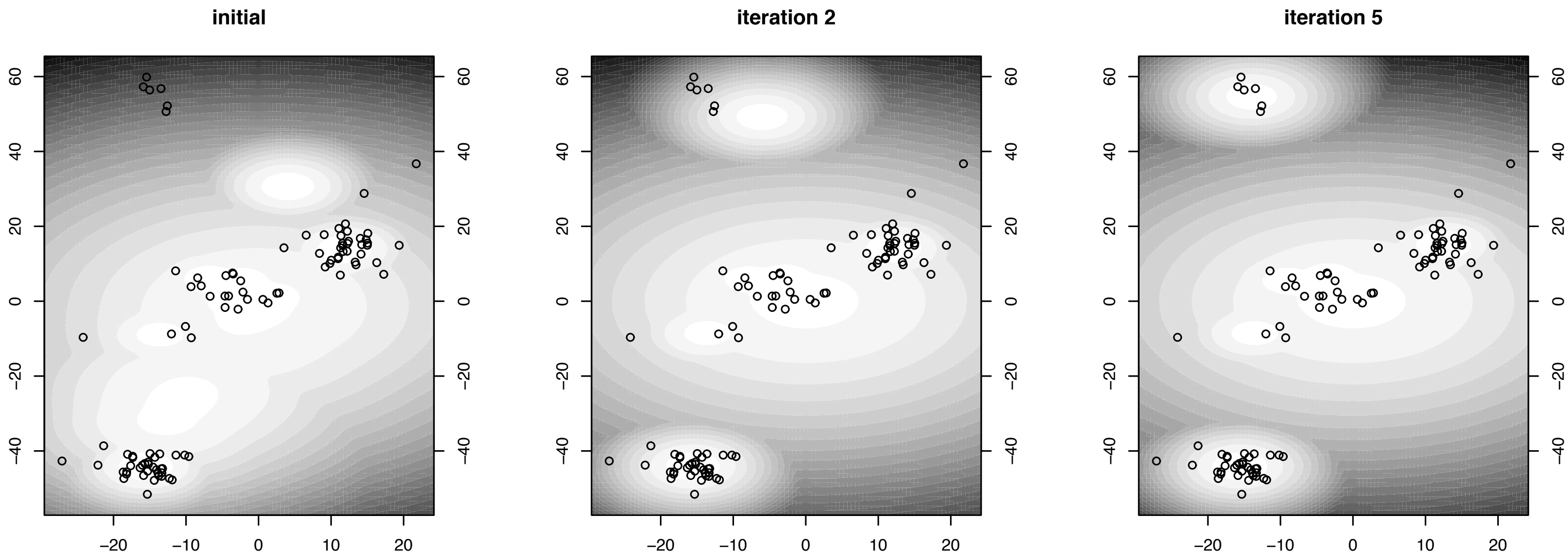


Figure 2: The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

Figure from Blei & Jordan (2006)

Example: DP Gaussian Mixture Model

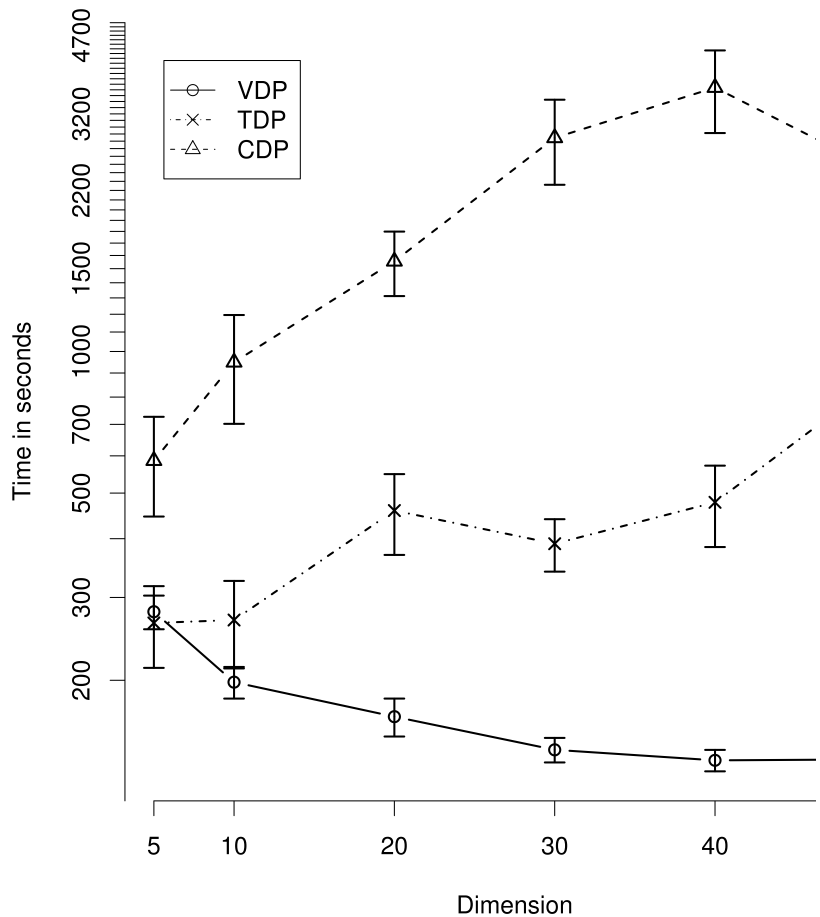


Figure 3: Mean convergence time and standard error across ten data sets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.

Summary of DP and DP-MM

- **DP** has many **different representations:**
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Blackwell-MacQueen Urn Scheme
  - Limit of finite mixtures
  - etc.
- These representations give rise to a variety of **inference techniques** for the **DP-MM** and related models
  - Gibbs sampler (CRP)
  - Gibbs sampler (stick-breaking)
  - Variational inference (stick-breaking)
  - etc.

Related Models

- Hierarchical Dirichlet Process Mixture Model (HDP-MM)
- Infinite HMM
- Infinite PCFG

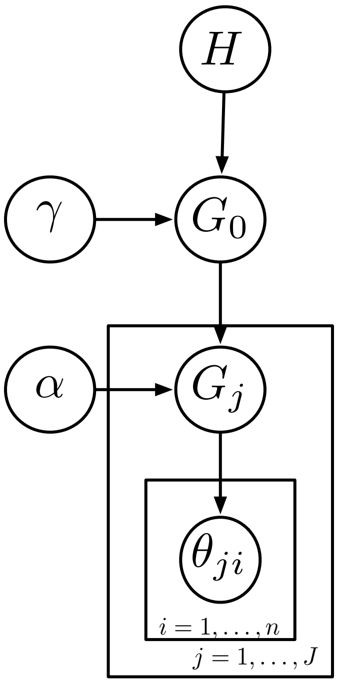
HDP-MM

- In LDA, we have *M* independent samples from a Dirichlet distribution.
- The weights are different, but the topics are fixed to be the same.
- If we replace the Dirichlet distributions with Dirichlet processes, each atom of each Dirichlet process will pick a topic *independently* of the other topics.

- Because the base measure is *continuous*, we have zero probability of picking the same topic twice.
- If we want to pick the same topic twice, we need to use a *discrete* base measure.
- For example, if we chose the base measure to be $H = \sum_{k=1}^{K} \alpha_k \delta_{\beta_k}$ then we would have LDA again.

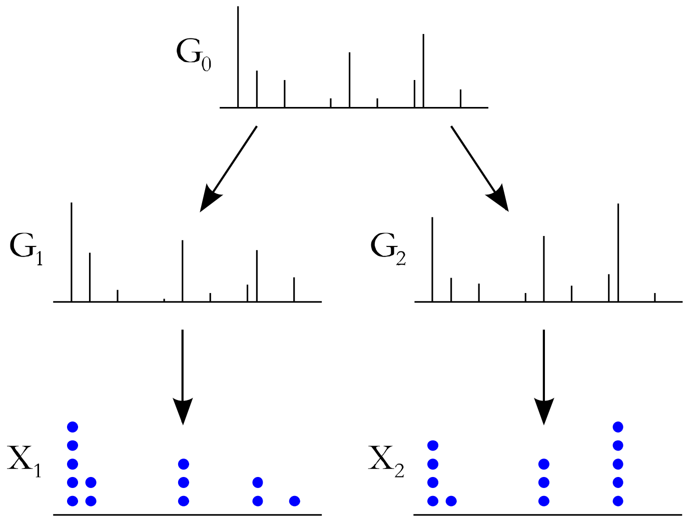
- We want there to be an infinite number of topics, so we want an *infinite, discrete* base measure.
- We want the location of the topics to be random, so we want an *infinite, discrete, random* base measure.

HDP-MM



Hierarchical Dirichlet process:

$$G_0 | \gamma, H \sim \mathrm{DP}(\gamma, H)$$
$$G_j | \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$$
$$\theta_{ji} | G_j \sim G_j$$



HDP-PCFG (Infinite PCFG)

HDP-PCFG

$\boldsymbol{\beta} \sim \mathrm{GEM}(\alpha)$ [draw top-level symbol weights]

For each grammar symbol $z \in \{1, 2, \dots\}$:

- $\phi_z^T \sim \mathrm{Dirichlet}(\alpha^T)$ [draw rule type parameters]
- $\phi_z^E \sim \mathrm{Dirichlet}(\alpha^E)$ [draw emission parameters]
- $\phi_z^B \sim \mathrm{DP}(\alpha^B, \boldsymbol{\beta}\boldsymbol{\beta}^T)$ [draw binary production parameters]

For each node i in the parse tree:

- $t_i \sim \mathrm{Multinomial}(\phi_{z_i}^T)$ [choose rule type]
- If $t_i = \textsc{Emission}$:
  - $x_i \sim \mathrm{Multinomial}(\phi_{z_i}^E)$ [emit terminal symbol]
- If $t_i = \textsc{Binary-Production}$:
  - $(z_{L(i)}, z_{R(i)}) \sim \mathrm{Multinomial}(\phi_{z_i}^B)$ [generate children symbols]