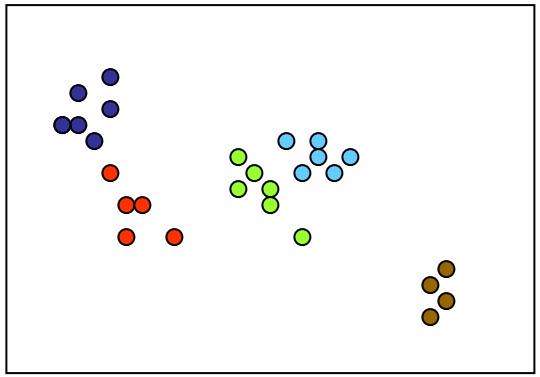
# Dirichlet Process Mixtures

Khalid El-Arini

November 28, 2005

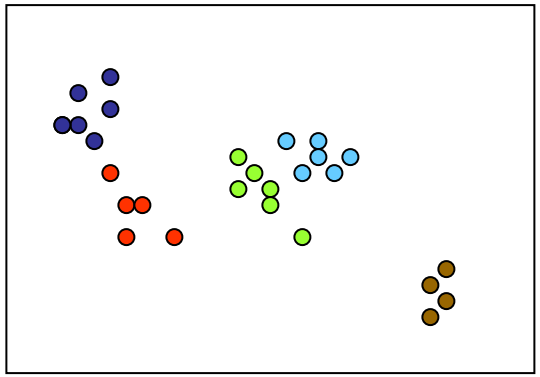
# Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussians.



# Motivation

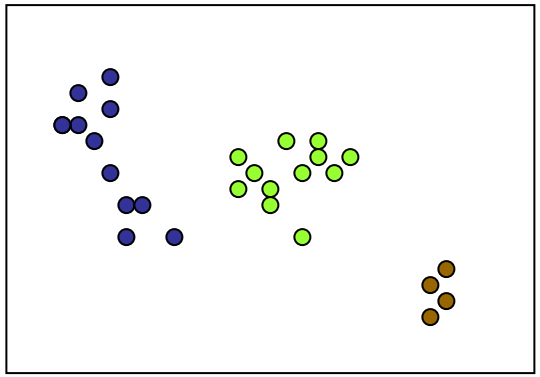
- We are given a data set, and are told that it was generated from a mixture of Gaussians.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

# Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussians.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

# What to do?

- We can guess the number of clusters, do EM for Gaussian Mixture Models, look at the results, and then try again…
- We can do hierarchical agglomerative clustering, and cut the tree at a visually appealing level…

# What to do?

- We can guess the number of clusters, do EM for Gaussian Mixture Models, look at the results, and then try again…
- We can do hierarchical agglomerative clustering, and cut the tree at a visually appealing level…

# What to do?

- We can guess the number of clusters, do EM for Gaussian Mixture Models, look at the results, and then try again…
- We can do hierarchical agglomerative clustering, and cut the tree at a visually appealing level…
- We want to cluster the data in a statistically principled manner, without resorting to hacks.

# Review: Dirichlet Distribution

- Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\Theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

- Distribution over possible parameter vectors for a multinomial distribution, and is in fact the conjugate prior for the multinomial.
- Beta distribution is the special case of a Dirichlet for 2 dimensions.
- Samples from the distribution lie in the *m* dimensional simplex
- Thus, it is in fact a “distribution over distributions.”

# Dirichlet Process

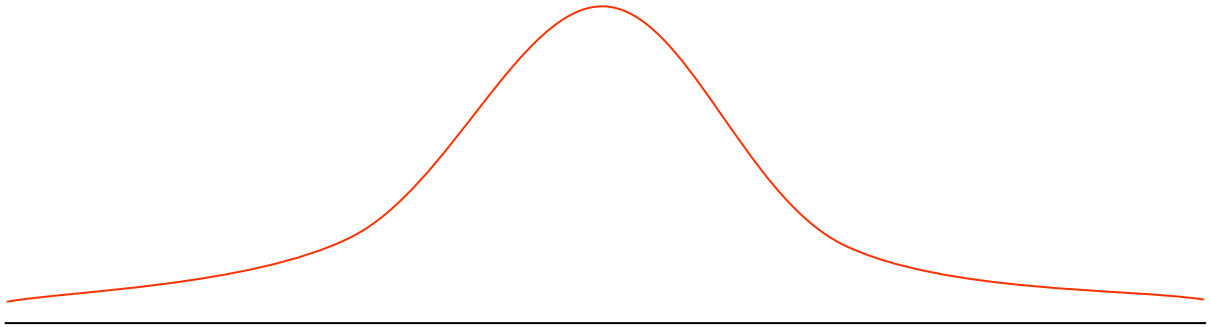
- A *Dirichlet Process* is also a distribution on distributions.
- We write:

$$G \sim DP(\alpha, G_0)$$

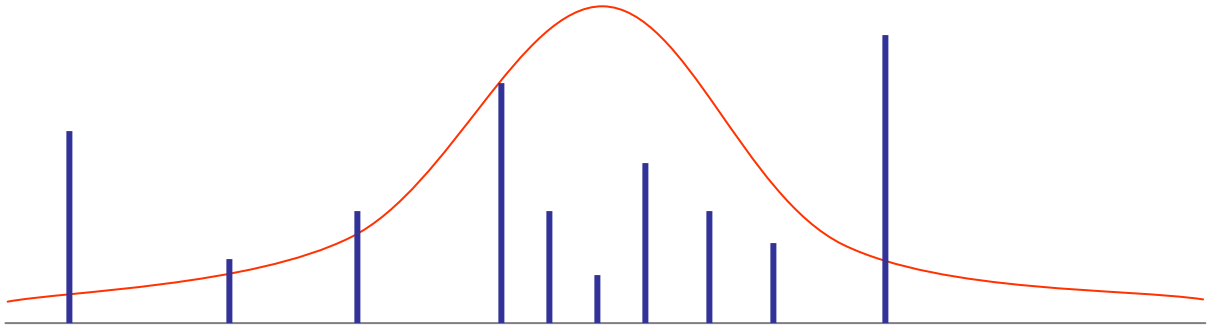
  - $G_0$ is a base distribution
  - $\alpha$ is a positive scaling parameter
- G has the same support as $G_0$

# Dirichlet Process

- Consider Gaussian $G_0$

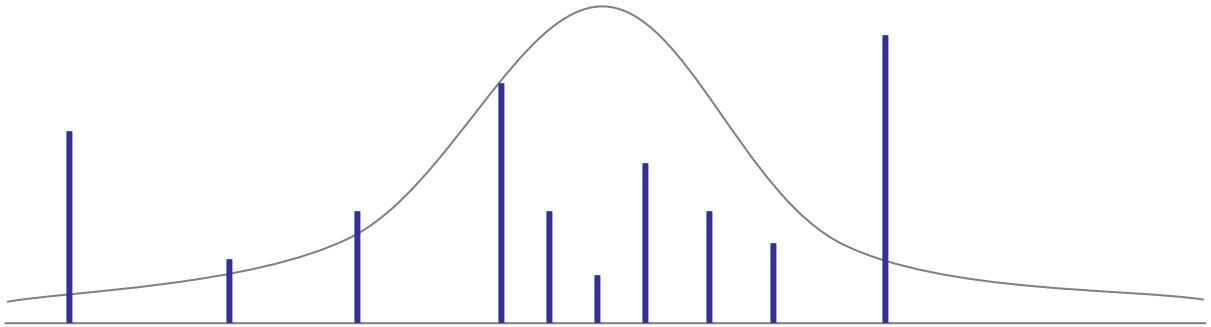


- $G \sim DP(\alpha, G_0)$



# Dirichlet Process

- $G \sim DP(\alpha, G_0)$

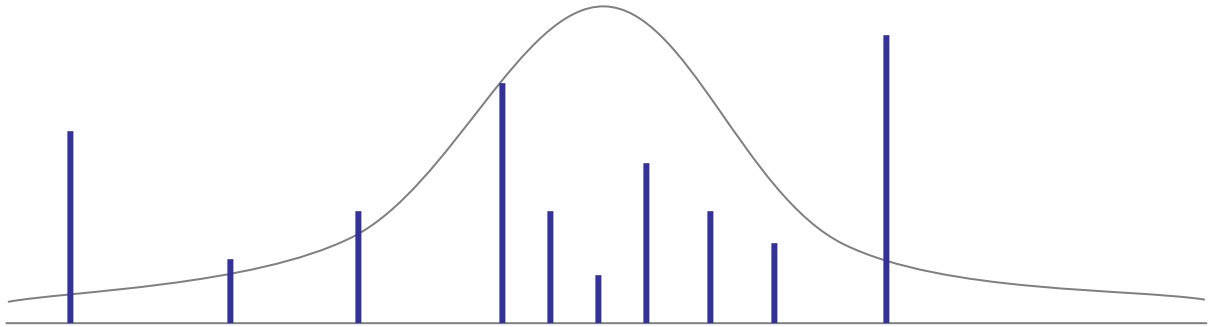


- $G_0$ is continuous, so the probability that any two samples are equal is precisely zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses [Blackwell]
  - Therefore, there is always a non-zero probability of two samples colliding

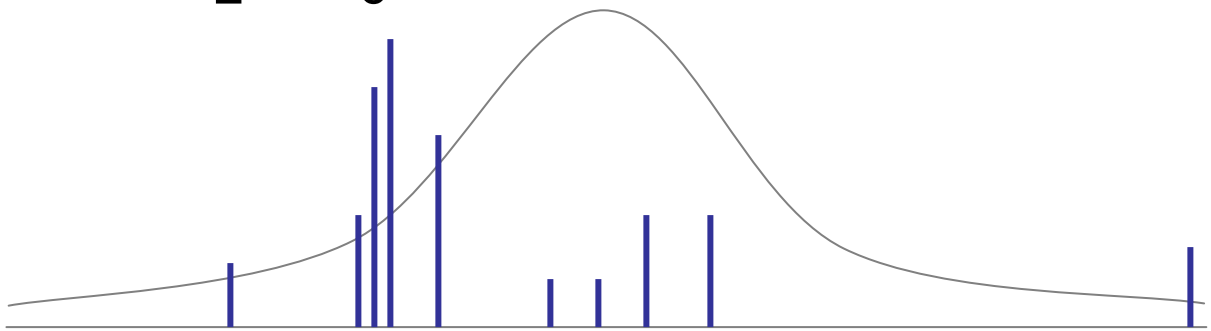
# Dirichlet Process

- $G \sim DP(\alpha_1, G_0)$

α values determine how close G is to $G_0$



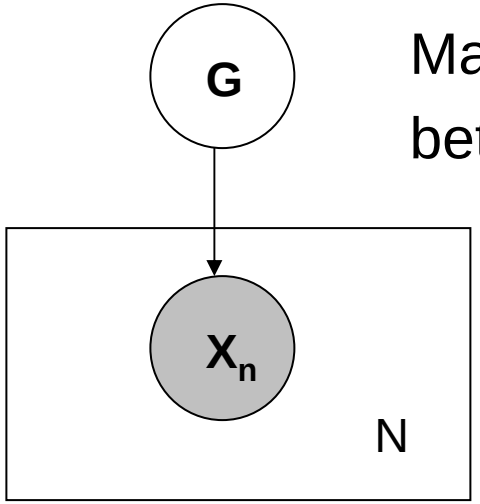
- $G \sim DP(\alpha_2, G_0)$



# Sampling from a DP

$G \sim DP(\alpha, G_0)$

$X_n \mid G \sim G$ for n = {1, …, N} (iid)




Marginalizing out G introduces dependencies between the $X_n$ variables

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

# Sampling from a DP

$$P(X_1, \ldots, X_N) = \int P(G) \prod_{n=1}^{N} P(X_n|G) dG$$

Assume we view these variables in a specific order, and are interested in the behavior of $X_n$ given the previous *n* - 1 observations.

$$X_n|X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be *K* unique values for the variables:

$$X_k^* \text{ for } k \in \{1, \ldots, K\}$$

# Sampling from a DP

$$X_n|X_1,\ldots,X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

$$P(X_1,\ldots,X_N) = P(X_1)P(X_2|X_1)\ldots P(X_N|X_1,\ldots,X_{N-1})$$

**Chain rule**

$$= \frac{\alpha^K}{\alpha(1+\alpha)\ldots(N-1+\alpha)} \prod_{k=1}^{K} G_0(X_k^*)$$

**P(partition)** **P(draws)**

Notice that above formulation of the joint does not depend on the order we consider the variables. We can use DeFinetti's Theorem to show that the variables are *exchangeable* under a Dirichlet Process model. This means we can consider them in any order.

# Chinese Restaurant Process

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_i & \text{with probability } \frac{1}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Let there be *K* unique values for the variables:

$$X_k^* \text{ for } k \in \{1, \ldots, K\}$$

Can rewrite as:

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_k^* & \text{with probability } \frac{\text{num}_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

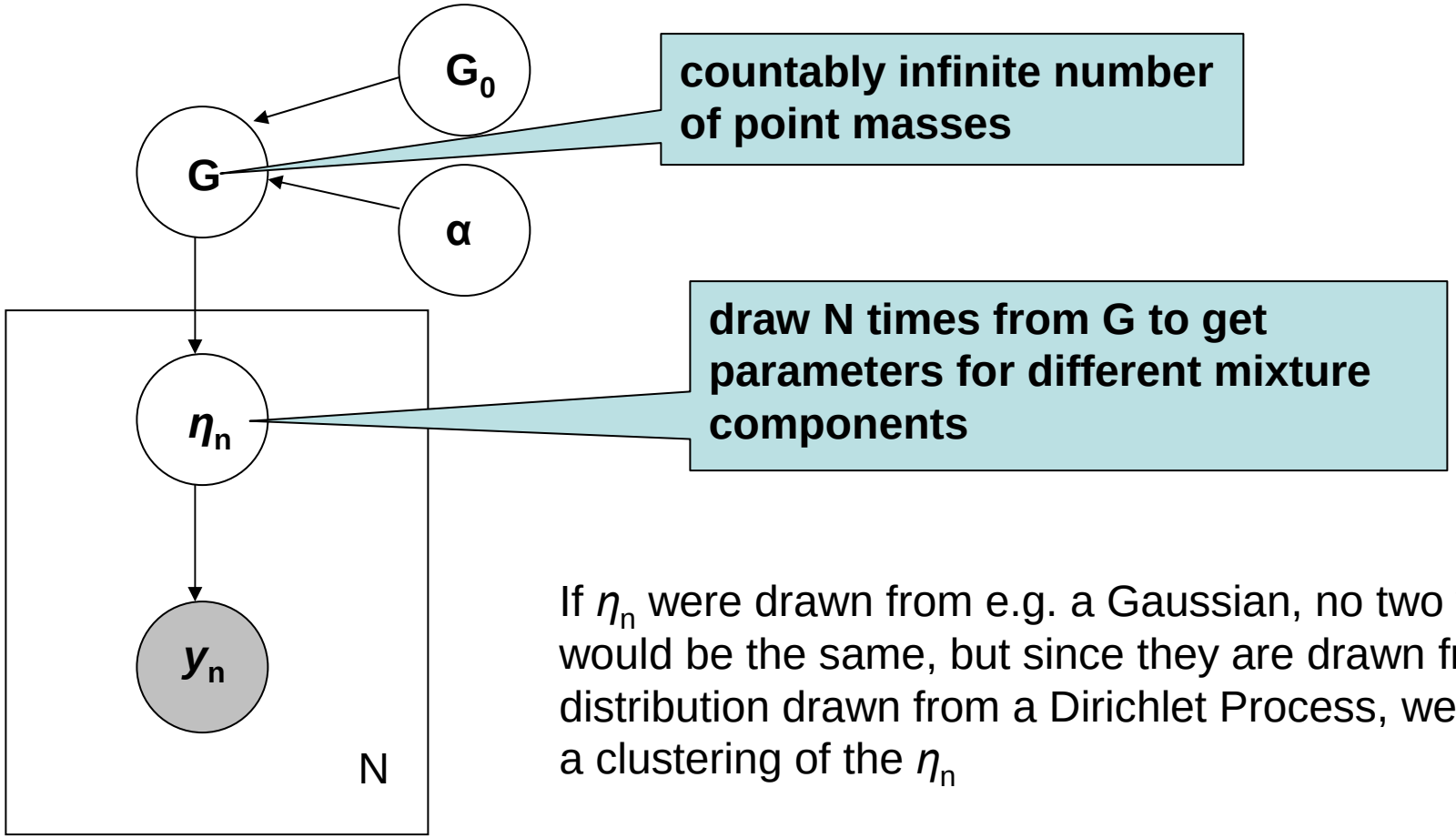
# Chinese Restaurant Process

$$X_n | X_1, \ldots, X_{n-1} = \begin{cases} X_k^* & \text{with probability } \frac{\text{num}_{n-1}(X_k^*)}{n-1+\alpha} \\ \text{new draw from } G_0 & \text{with probability } \frac{\alpha}{n-1+\alpha} \end{cases}$$

Consider a restaurant with infinitely many tables, where the $X_n$'s represent the patrons of the restaurant. From the above conditional probability distribution, we can see that a customer is more likely to sit at a table if there are already many people sitting there. However, with probability proportional to $\alpha$, the customer will sit at a new table.

Also known as the "clustering effect," and can be seen in the setting of social clubs. [Aldous]

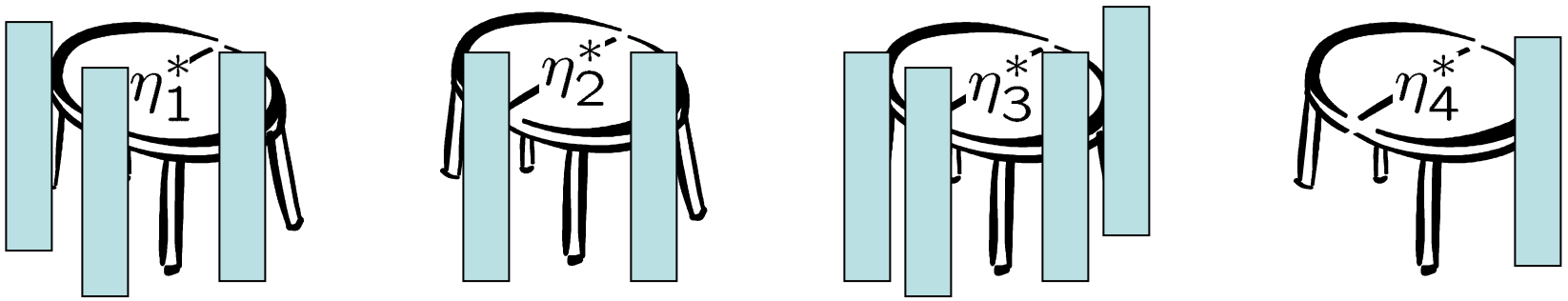
# Dirichlet Process Mixture

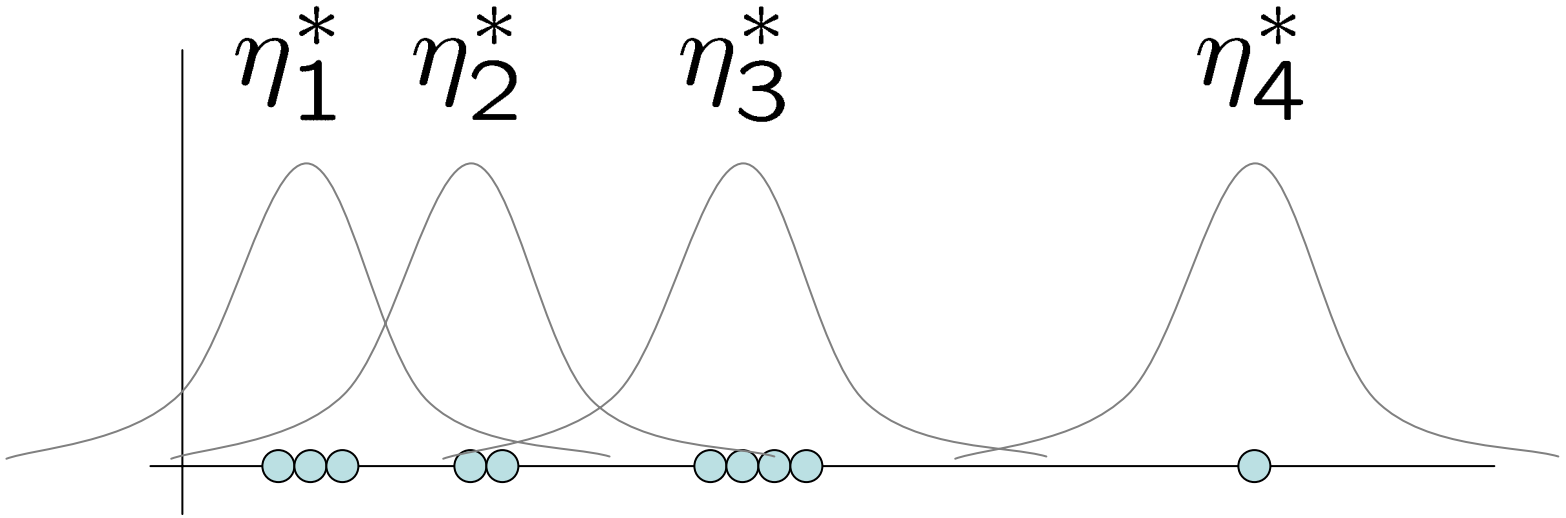



If $\eta_n$ were drawn from e.g. a Gaussian, no two values would be the same, but since they are drawn from a distribution drawn from a Dirichlet Process, we expect a clustering of the $\eta_n$

# unique values for $\eta_n$ = # mixture components

# CRP Mixture



# Stick Breaking

- So far, we've just mentioned properties of a distribution G drawn from a Dirichlet Process
- In 1994, Sethuraman developed a constructive way of forming G, known as "stick breaking"
- Who cares? Now we can perform variational inference [Blei and Jordan]

# Stick Breaking

$$V_1, V_2, \ldots, V_i, \ldots \sim \text{Beta}(1, \alpha)$$

$$f(V_i = v_i | \alpha) = \alpha(1 - v_i)^{\alpha - 1}$$

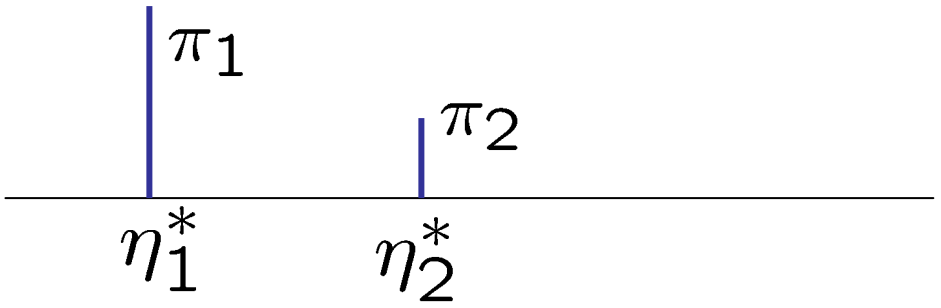
$$\eta_1^*, \eta_2^*, \ldots, \eta_i^*, \ldots \sim G_0$$

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v}) \delta_{\eta_i^*}$$

1. Draw $\eta_1^*$ from $G_0$
2. Draw $v_1$ from Beta(1, $\alpha$)
3. $\pi_1 = v_1$
4. Draw $\eta_2^*$ from $G_0$
5. Draw $v_2$ from Beta(1, $\alpha$)
6. $\pi_2 = v_2(1 - v_1)$

...



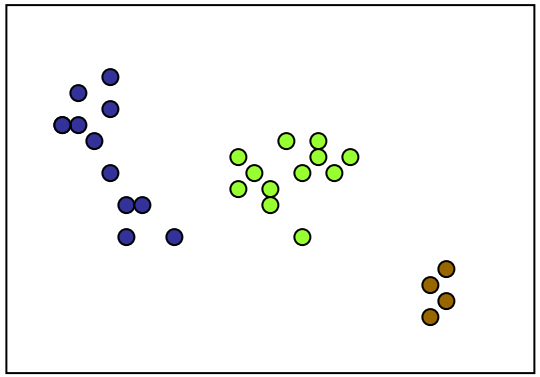
# Formal Definition

- Let $\alpha$ be a positive, real-valued scalar
- Let $G_0$ be a non-atomic probability distribution over support set A
- We say G ~ DP($\alpha$, $G_0$), if for all natural numbers k and k-partitions {$A_1$, …, $A_k$},

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$$

# Conclusion

- We now have a statistically principled mechanism for solving our original problem.



- This was intended as a general and fairly shallow overview of Dirichlet Processes.

# Acknowledgments


- Much thanks goes to David Blei for helping me understand the little I know about Dirichlet Processes.
- Some material for this presentation was inspired by slides from Teg Grenager and Zoubin Ghahramani.

# References


Blei, David M. and Michael I. Jordan. “Variational inference for Dirichlet process mixtures.” Bayesian Analysis 1(1), 2004.

Ghahramani, Zoubin. “Non-parametric Bayesian Methods.” UAI Tutorial July 2005.

Grenager, Teg. “Chinese Restaurants and Stick Breaking: An Introduction to the Dirichlet Process”

Blackwell, David and James B. MacQueen. “Ferguson Distributions via Polya Urn Schemes.” The Annals of Statistics 1(2), 1973, 353-355.

Ferguson, Thomas S. “A Bayesian Analysis of Some Nonparametric Problems” The Annals of Statistics 1(2), 1973, 209-230.