Applied Nonparametric Hierarchical Bayes

Michael I. Jordan

Department of Statistics
Department of Electrical Engineering and Computer Science
University of California, Berkeley

http://www.cs.berkeley.edu/~jordan

Acknowledgments: David Blei, Yee Whye Teh

Themes

- Intelligent systems are based on certain architectural and algorithmic choices
- What are the general principles underlying these architectural and algorithmic choices?
- How do those principles allow us to go from a problem specification to an algorithmic solution?

Applied Nonparametric Hierarchical Bayes

- nonparametric: the number of parameters (degrees of freedom) of a system should be allowed to grow as more data are observed

- Bayes: probability theory and decision theory provide a solid foundation on which to understand learning, perception, reasoning and action

- hierarchical: we often have multiple, related streams of data, and we want to share information among those streams

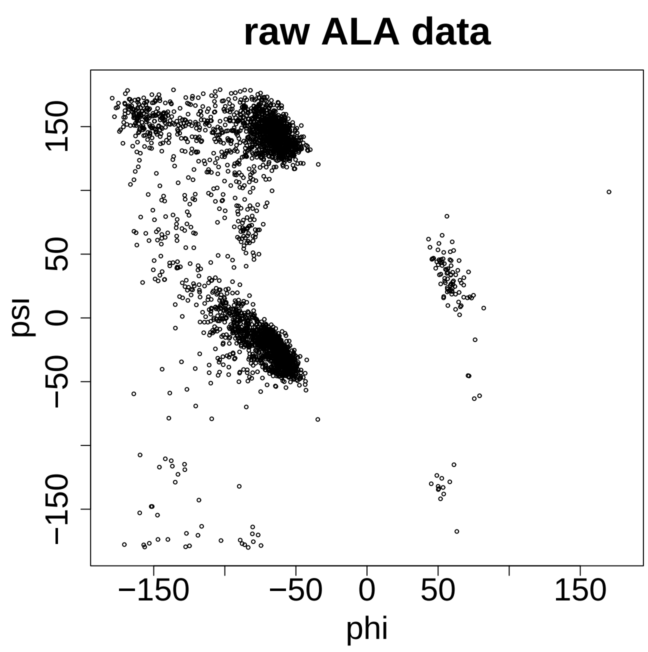
- applied: we want to solve real-world problems

Document and Image Modeling

- Define a topic to be a probability distribution across *words* in some vocabulary
- Define a document to be a probability distribution across topics
- Given a corpus of documents, find the topics and find the patterns of usage of topics across documents
- Each document is a clustering problem; we must link multiple clusterings across a corpus
- Note that a "document" can be an image, where a "word" is a local image feature

Protein Folding

- A protein is a folded chain of amino acids
- The backbone of the chain has two degrees of freedom per amino acid (phi and psi angles)
- Empirical plots of phi and psi angles are called *Ramachandran diagrams*



Protein Folding (cont.)

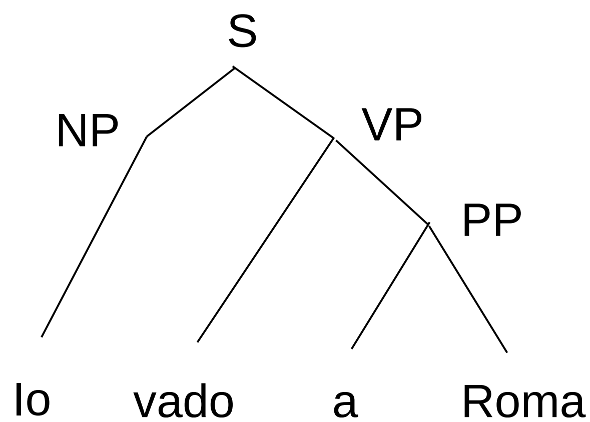
- We want to model the density in the Ramachandran diagram to provide an energy term for protein folding algorithms

- We actually have a linked set of Ramachandran diagrams, one for each amino acid neighborhood

- We thus have a linked set of density estimation problems

Natural Language Parsing

- Given a corpus of sentences, some of which have been parsed by humans, find a grammar that can be used to parse future sentences



- Much progress over the past decade; state-of-the-art methods are all statistical

Natural Language Parsing (cont.)

- Key idea: *lexicalization* of context-free grammars
  - the grammatical rules (S $\rightarrow$ NP VP) are conditioned on the specific lexical items (words) that they derive

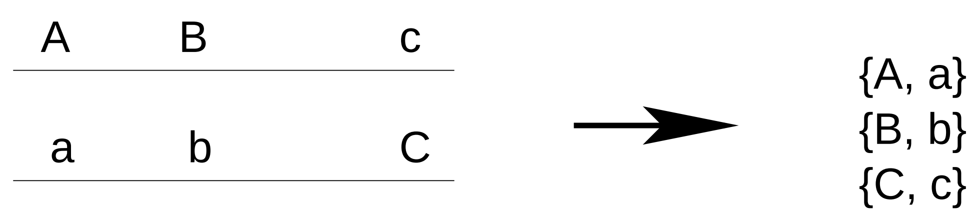
- This leads to huge numbers of potential rules, and (adhoc) shrinkage methods are used to control the counts

- Need to control the numbers of clusters (model selection) in a setting in which many tens of thousands of clusters are needed

- Need to consider related groups of clustering problems (one group for each grammatical context)

Haplotype Modeling

- Consider M binary markers in a genomic region
- There are 2^M possible *haplotypes*—i.e., states of a single chromosome
  - but in fact, far fewer are seen in human populations
- A *genotype* is a set of unordered pairs of markers (from one individual)

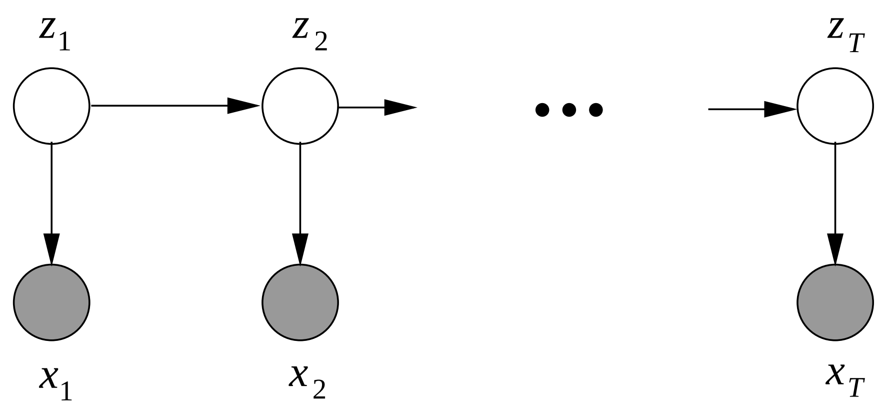


- Given a set of genotypes (multiple individuals), estimate the underlying haplotypes
- This is a clustering problem

Haplotype Modeling (cont.)

- A key problem is inference for the number of clusters
- Consider now the case of multiple groups of genotype data (e.g., ethnic groups)
- Geneticists would like to find clusters within each group but they would also like to share clusters between the groups

Nonparametric Hidden Markov Models



- An open problem—how to work with HMMs and state space models that have an unknown and unbounded number of states?

- Each row of a transition matrix is a probability distribution across "next states"

- We need to estimation these transitions in a way that links them across rows

Outline

- Dirichlet Processes (clusters)
- Hierarchical Dirichlet Processes (tied clusters)
- Beta Processes (features)
- Hierarchical Beta Processes (tied features)

Clustering—How to Choose K?

Clustering—How to Choose K?

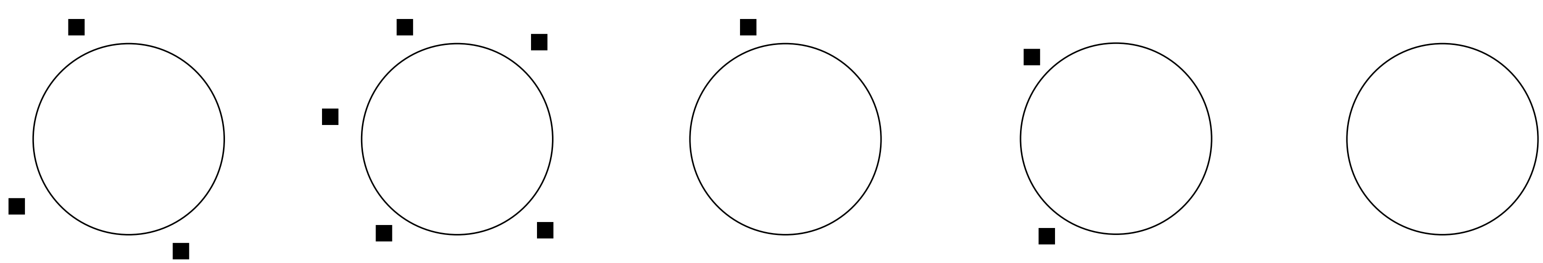
- Adhoc approaches (e.g., hierarchical clustering)
  - they do often yield a data-driven choice of K
  - but there is little understanding of how good these choices are
- Methods based on objective functions (M-estimators)
  - e.g., K-means, spectral clustering
  - do come with some frequentist guarantees
  - but it's hard to turn these into data-driven choices of K
- Parametric likelihood-based approaches
  - finite mixture models, Bayesian variants thereof
  - various model choice methods: hypothesis testing, cross-validation, bootstrap, AIC, BIC, DIC, Laplace, bridge sampling, reversible jump, etc
  - but do the assumptions underlying the method really apply to this setting? (not often)
- Let's try something different...

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
  - mth subsequent customer sits at a table drawn from the following distribution:

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

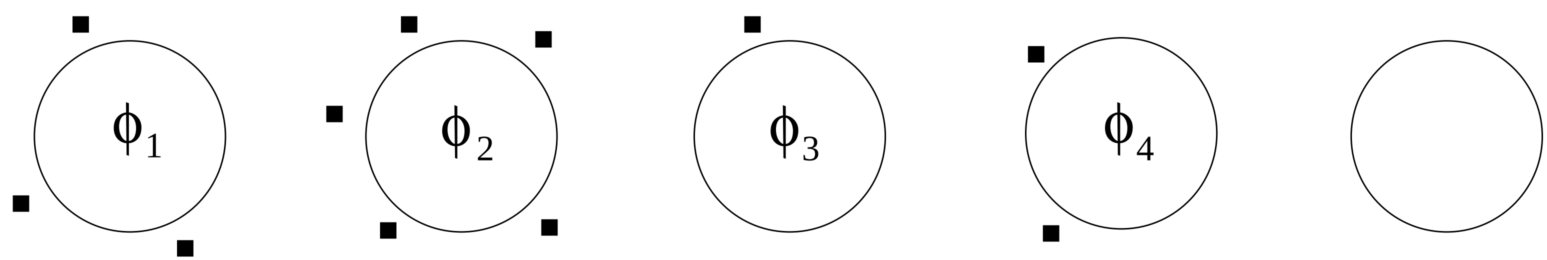
where n_i is the number of customers currently at table i and where $\mathcal{F}_{m-1}$ denotes the state of the restaurant after $m-1$ customers have been seated



The CRP and Clustering

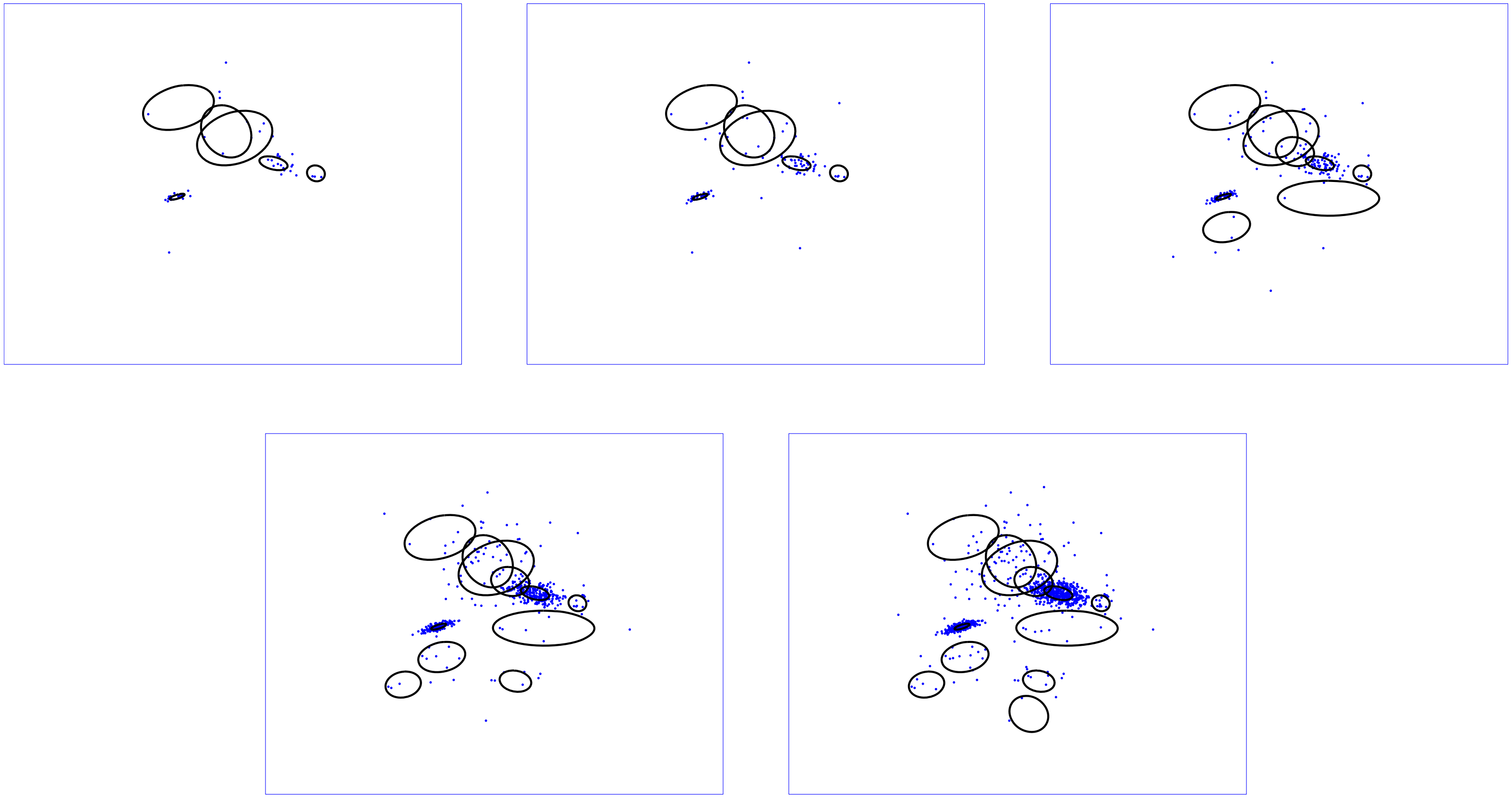
- Data points are customers; tables are clusters
  - the CRP defines a prior distribution on the partitioning of the data and on the number of tables

- 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 (ϕ_k) from a prior G_0



- 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



$$\begin{aligned} \phi_k &= (\mu_k, \Sigma_k) \sim N(a, b) \otimes IW(\alpha, \beta) \\ x_i &\sim N(\phi_k) \qquad \text{for a data point } i \text{ sitting at table } k \end{aligned}$$

Inference for the CRP

- We've described how to generate data from the model; how do we go backwards and generate a model from data?

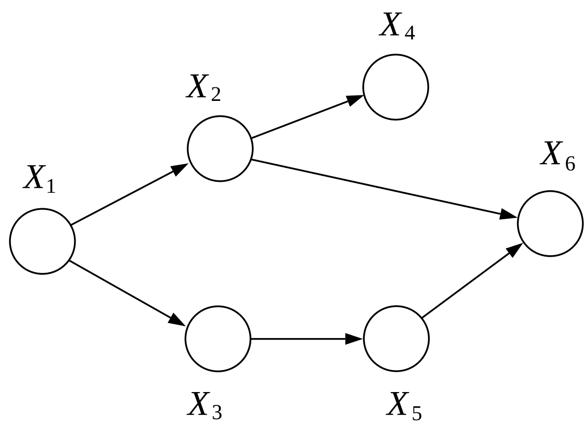
- A wide variety of variational, combinatorial and MCMC algorithms have been developed

- E.g., a Gibbs sampler is readily developed by using the (deep) fact that the Chinese restaurant process is exchangeable
  - to sample the table assignment for a given customer given the seating of all other customers, simply treat that customer as the last customer to arrive
  - in which case, the assignment is made proportional to the number of customers already at each table (cf. preferential attachment)
  - parameters are sampled at each table based on the customers at that table (cf. K means)

- (This isn't the state of the art, but it's easy to explain on one slide)

Directed Graphical Models

- Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ is associated with a random variable X_v:

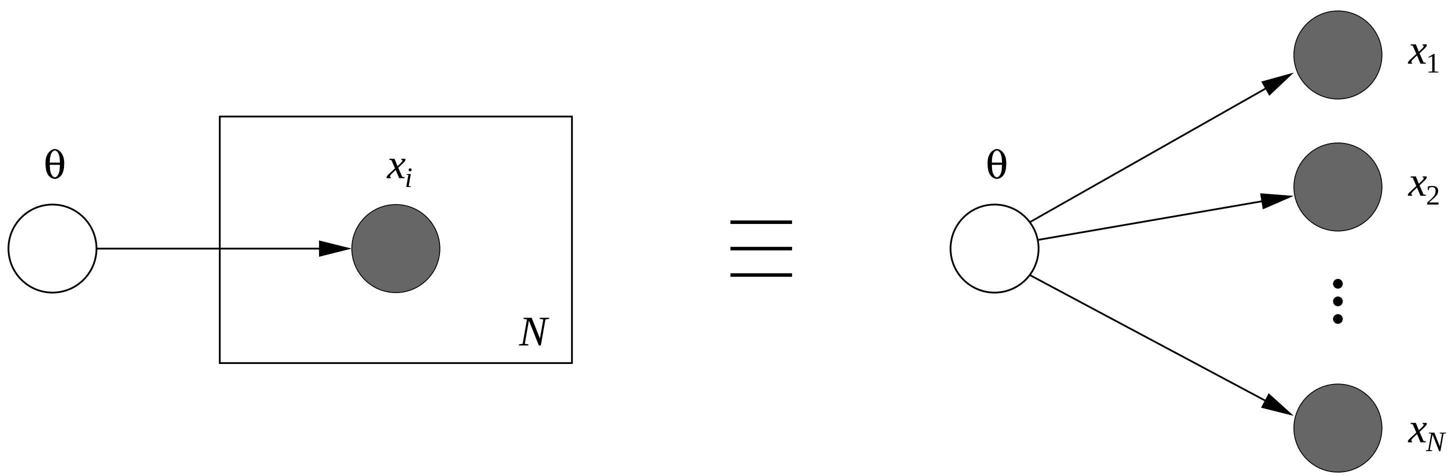


- The joint distribution on $(X_1, X_2, \ldots, X_N)$ factorizes according to the "parent-of" relation defined by the edges $\mathcal{E}$:

$$p(x_1, x_2, x_3, x_4, x_5, x_6; \theta) = p(x_1; \theta_1)\; p(x_2 \,|\, x_1; \theta_2)$$
$$p(x_3 \,|\, x_1; \theta_3)\; p(x_4 \,|\, x_2; \theta_4)\; p(x_5 \,|\, x_3; \theta_5)\; p(x_6 \,|\, x_2, x_5; \theta_6)$$

Plates

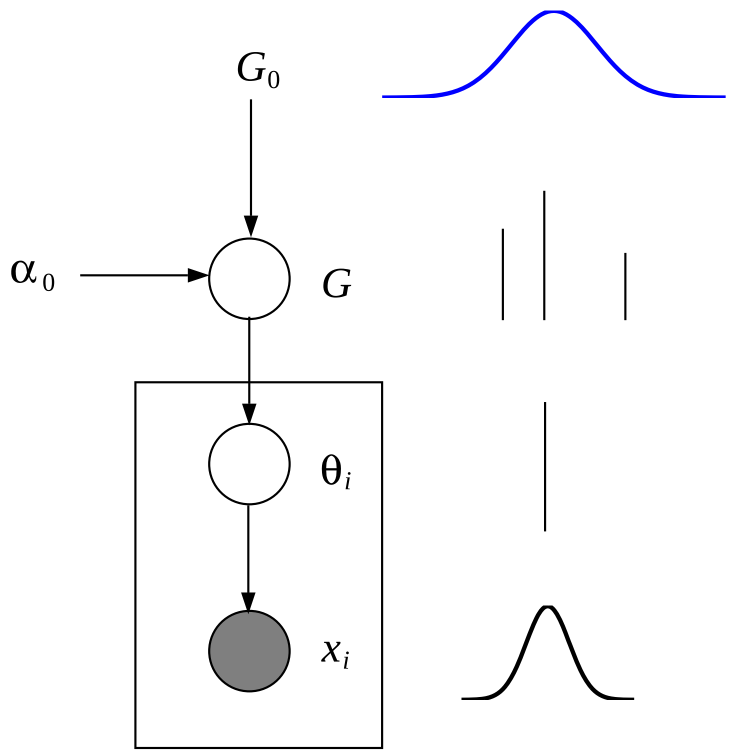
- A *plate* is a "macro" that allows subgraphs to be replicated:



- Shading denotes conditioning

Finite Mixture Models

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



- Note that G is a *random measure*

Going Nonparametric—A First Attempt

- 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 π_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

Stick-Breaking

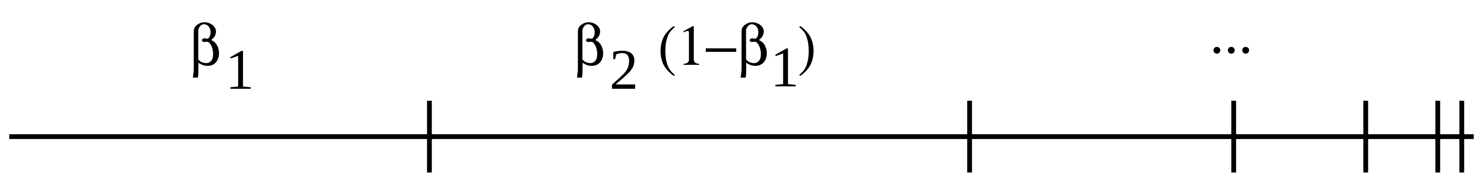
- Define an infinite sequence of Beta random variables:

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

- 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 π_k:

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

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

- The distribution of G is known as a Dirichlet process
  - it can be shown that for any finite partition $(A_1, \ldots, A_r)$ of the sample space, the random vector $(G(A_1), \ldots, G(A_r))$ is distributed as a finite-dimensional Dirichlet distribution

- We write this as

$$G \sim \mathrm{DP}(\alpha_0, G_0),$$

where α_0 is known as the concentration parameter and G_0 is known as the base measure

Dirichlet Process Mixture Models