# Advanced Machine Learning
# Lecture 12

## Dirichlet Processes II

03.12.2012

**Bastian Leibe**
**RWTH Aachen**
http://www.vision.rwth-aachen.de/

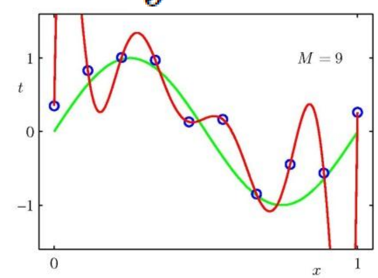
**leibe@vision.rwth-aachen.de**

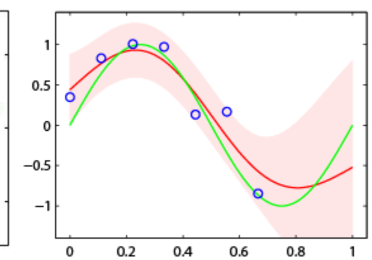
# This Lecture: *Advanced Machine Learning*

- **Regression Approaches**

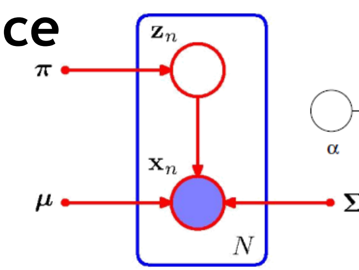
$$f : \mathcal{X} \to \mathbb{R}$$

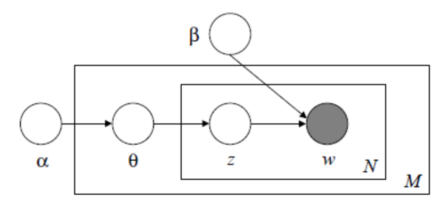
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes




- **Bayesian Estimation & Bayesian Non-Parametrics**
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes




- **SVMs and Structured Output Learning**
  - SV Regression, SVDD
  - Large-margin Learning

$$f : \mathcal{X} \to \mathcal{Y}$$

# Topics of This Lecture

- Dirichlet Processes
  - Recap: Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

# Recap: Dirichlet Processes

- **Gaussian Processes**
  - Gaussian Processes (GP) define a distribution over functions

$$f \sim \mathrm{GP}(\cdot|\mu, c)$$

where $\mu$ is the mean function and $c$ is the covariance function.

$\Rightarrow$ We can think of GPs as "infinite-dimensional" Gaussians.

- **Dirichlet Processes**
  - Dirichlet Processes (DP) define a distribution over distributions (a measure on measures)

$$G \sim \mathrm{DP}(\cdot|G_0, \alpha)$$

  - Where $\alpha > 0$ is a scaling parameter and $G_0$ is the base measure.

$\Rightarrow$ We can think of DPs as "infinite-dimensional" Dirichlet distributions.

# Sidenote: Bayesian Nonparametric Methods

- Bayesian Nonparametric Methods (BNPs)
  - Both Gaussian Processes and Dirichlet Processes are examples of BNPs.

- *What does that mean?*
  - Nonparametric: does NOT mean there are no parameters!
  - It means (very roughly) that the number of parameters grows with the number of data points.

- Parametric methods:
  - Get data → build model → predict using model

- Nonparametric methods
  - Get data → predict directly based on data

# Recap: Dirichlet Processes

- **Definition** [Ferguson, 1973]
  - Let $\Theta$ be a measurable space, $G_0$ be a probability measure on $\Theta$, and $\alpha$ a positive real number.
  - For all $(A_1,\ldots,A_K)$ finite partitions of $\Theta$,

$$G \sim \mathrm{DP}(\cdot|G_0, \alpha)$$

  means that

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

- **Translation**
  - *A random probability distribution $G$ on $\Theta$ is drawn from a Dirichlet Process if its measure on every finite partition follows a Dirichlet distribution.*

# Recap: Dirichlet Processes

- **Important property** **[Blackwell]**
  - Draws from a DP will always place all their mass on a countable set of points, the so-called atoms $\delta_{\theta k}$.

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

where $\delta_{\theta k}$ is a Dirac delta at $\theta_k$, and $\theta_k \sim G_0(\cdot)$.

$\Rightarrow$ Samples from DP are discrete with probability one.

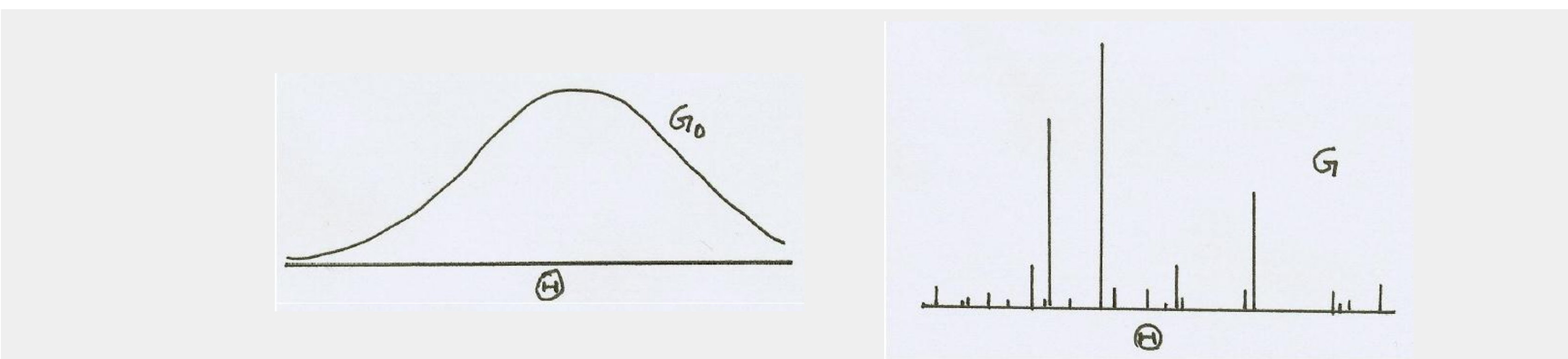
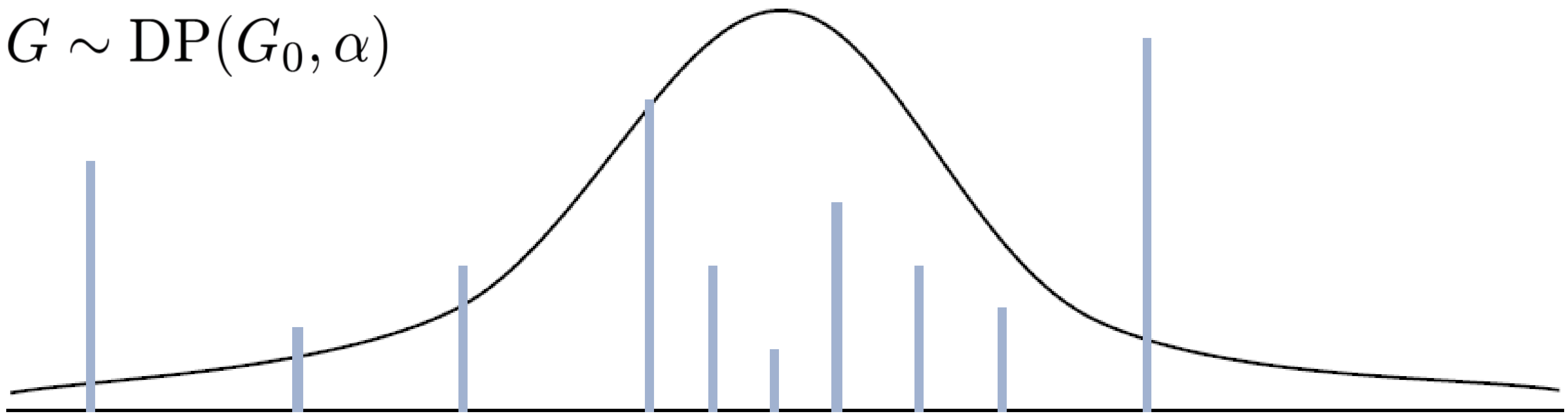





Image source: Zoubin Gharamani

# Recap: Dirichlet Processes




- **Consider a DP with a Gaussian as base measure $G_0$**
  - $G_0$ 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.
  
  $\Rightarrow$ There is always a non-zero probability of two samples colliding.
  
  $\Rightarrow$ *This is what allows us to use DPs for clustering!*

# Recap: Dirichlet Process Properties

- **Sampling**
  - Since $G$ is a probability measure, we can draw samples from it

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

$$\theta_1, \ldots, \theta_N | G \sim G$$

- **Posterior of $G$ given observations $\theta_1,\ldots,\theta_N$?**
  - The usual Dirichlet-multinomial conjugacy carries over to the nonparametric DP as well.

  $\Rightarrow$ Posterior is again a DP.

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

# Existence of Dirichlet Processes

- Summary so far
  - A probability measure is a function from subsets of a space $\Theta$ to [0,1] satisfying certain properties.
  - A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.

- *How do we know that such an object exists?*
  - Kolmogorov Consistency Theorem: If we can prescribe consistent finite dimensional distributions, then a distribution over functions exists.
  - De Finetti's Theorem: If we have an infinite exchangeable sequence of random variables, then a distribution over measures exists making them independent.

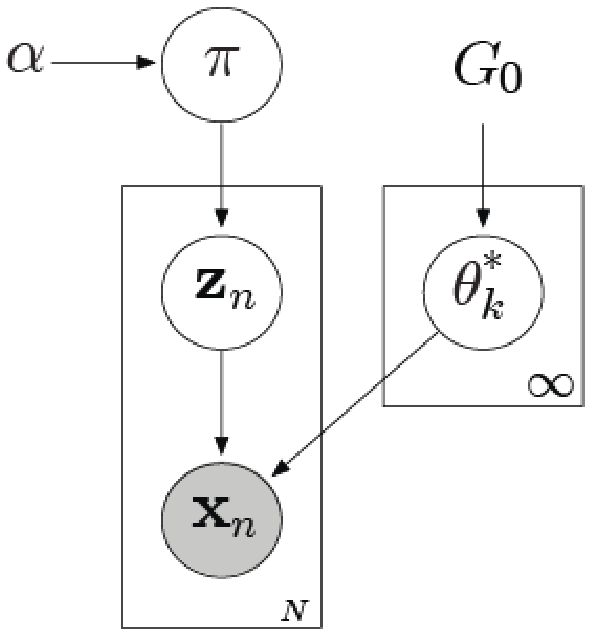
  $\Rightarrow$ Pólya's urn, Chinese Restaurant Process
  - Stick-breaking Construction: just construct it.

# Topics of This Lecture

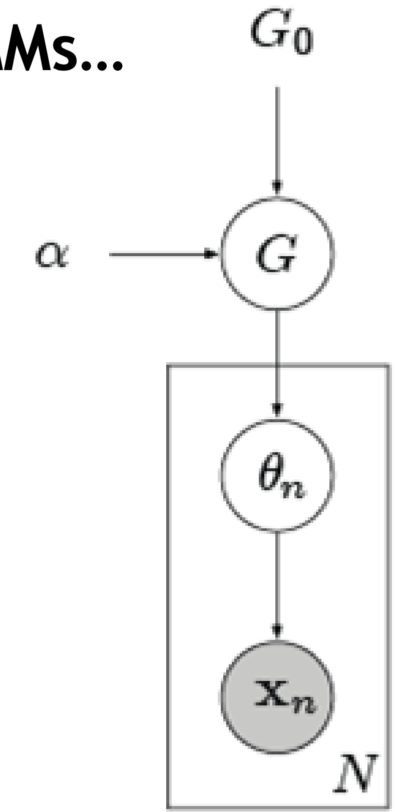
- Dirichlet Processes
  - Recap: Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

# Dirichlet Process Mixture Models

- During this lecture, we will use the following two forms for DPMMs…




"Indicator variable representation"

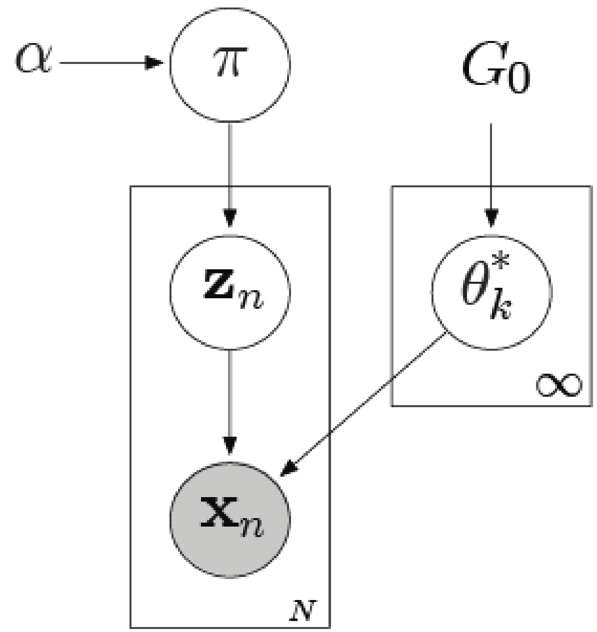



"Distributional form"



Image sources: Yee Whye The, Kurt Miller

# Dirichlet Process Mixture Models

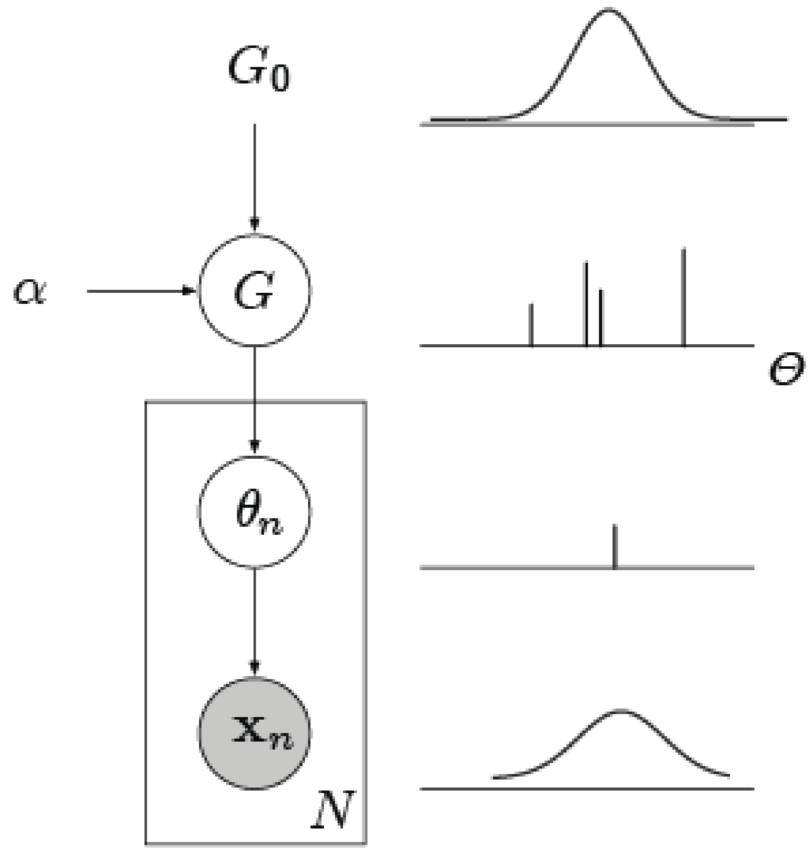



- Indicator variable representation
  - Form of an infinite mixture model
  - The DP is implicit through the choice of priors
  - We will use this form whenever we want to make the assignment of points to clusters explicit (⇒ use for clustering).



Image sources: Yee Whye The

# Dirichlet Process Mixture Models




Base distribution $G_0$

Infinite discrete distribution on $\Theta$, defines the clusters

Parameters of the cluster that generates $\mathbf{x}_n$

Likelihood of $\mathbf{x}_n$ given the cluster

- **Distributional form**
  - Explicit representation of the DP through the node $G$.
  - Useful when we want to use the DPMM's predictive distribution.



Image sources: Yee Whye The

# Topics of This Lecture

- Dirichlet Processes
  - Recap: Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

# Recap: Pólya's Urns [Blackwell & MacQueen, 1973]

- *Can we sample observations without constructing $G$?*

$$G \sim \mathrm{DP}(G_0, \alpha) \qquad \bar{\theta}_n \sim G$$

- **Yes, by a variation of the classical balls-in-urns analogy**
  - **Assume that $G_0$ is a distribution over colors, and that each $\theta_n$ represents the color of a single ball placed in the urn.**
  - **Start with an empty urn. Repeat for $N$ steps:**

1. **With probability proportional to α, draw $\theta_n \sim G_0$ and add a ball of that color to the urn.**
2. **With probability proportional to $n-1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as $\theta_n$ and return the ball into the urn, along with a new one of the same color.**

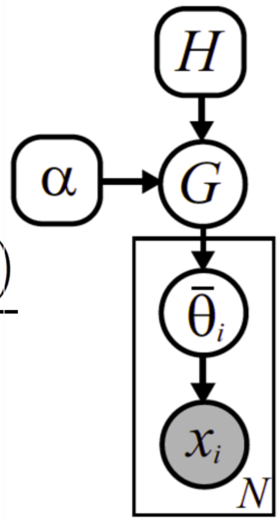


Image source: Yee Whye Teh

# Pólya's Urns: Discussion

- Pólya Urn scheme
  - Simple generative process for the predictive distribution of a DP
  - Consider a set of $N$ observations $\bar{\theta}_n \sim G$ taking $K$ distinct values $\{\theta_k\}_{k=1}^{K}$. The predictive distribution of the next observation is then

$$p(\bar{\theta}_N = \theta | \bar{\theta}_{1:N-1}, \alpha, H) = \frac{\alpha H(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k)}{N - 1 + \alpha}$$




- Remarks
  - This procedure can be used to sample observations from a DP without explicitly constructing the underlying mixture.

  ⇒ DPs lead to simple predictive distributions that can be evaluated by caching the number of previous observations taking each distinct value.

# De Finetti's Theorem

[De Finetti, 1935]

- Theorem
  - *For any **infinitely exchangeable** sequence of random variables $\{\mathbf{x}_i\}^{1:\infty}$, $\mathbf{x}_i \in \mathcal{X}$, there exists some space $\Theta$ of probability measures and corresponding distribution $P(\theta)$ such that the joint probability of any $N$ observations has a **mixture representation***

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \int_\Theta \prod_{n=1}^{N} p(\mathbf{x}_n|\theta)\mathrm{d}P(\theta)$$

- Interpretation
  - If you assert exchangeability, it is reasonable to act as if there is an underlying parameter, there is a prior on this parameter, and the data are i.i.d. given that parameter.
  - In order for this to work, we need to allow $\theta$ to range over measures, in which case $P(\theta)$ is a distribution over measures.
    - As we know, the Dirichlet Process is a distribution on measures!

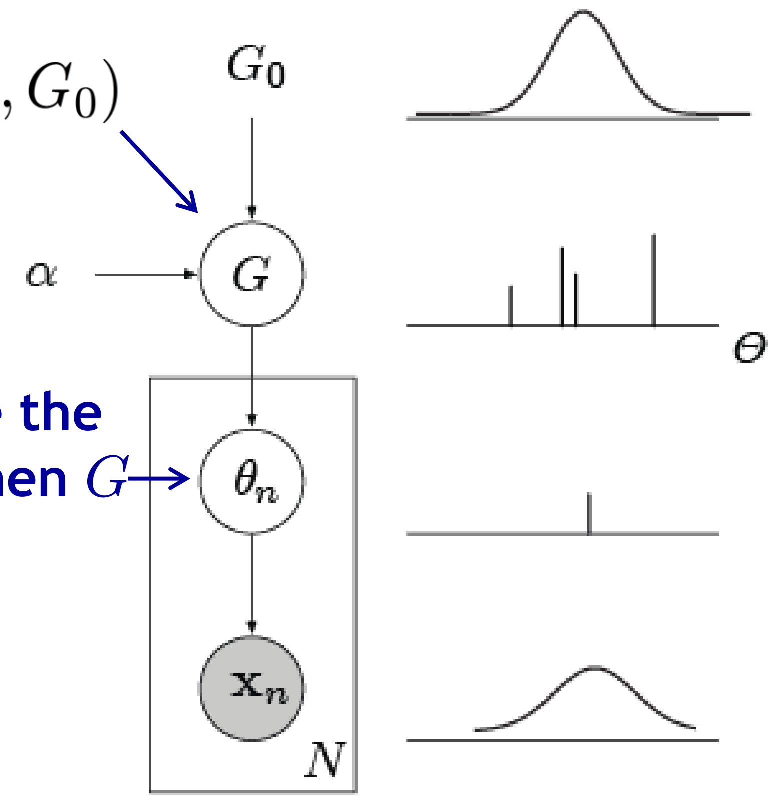
# Pólya Urn Scheme

- Existence proof for DP
  - Starting with a DP, we constructed Pólya's urn scheme.
  - The reverse is possible using De Finetti's theorem:
  - Since the $\theta_n$ are i.i.d. $\sim G$, their joint distribution is invariant to permutations, thus $\theta_1, \theta_2, \ldots$ are exchangeable.
  - Thus a distribution over measures must exist making them i.i.d.
  - This is the DP.

- We have just (informally) proven that DPs exist
  - Hooray!
  - Now, let's move on to see how we can use them…

# Big Picture: Pólya Urns and the DP

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




**Pólya urns describe the distribution of $\theta$ when $G$ is marginalized out**

# Topics of This Lecture

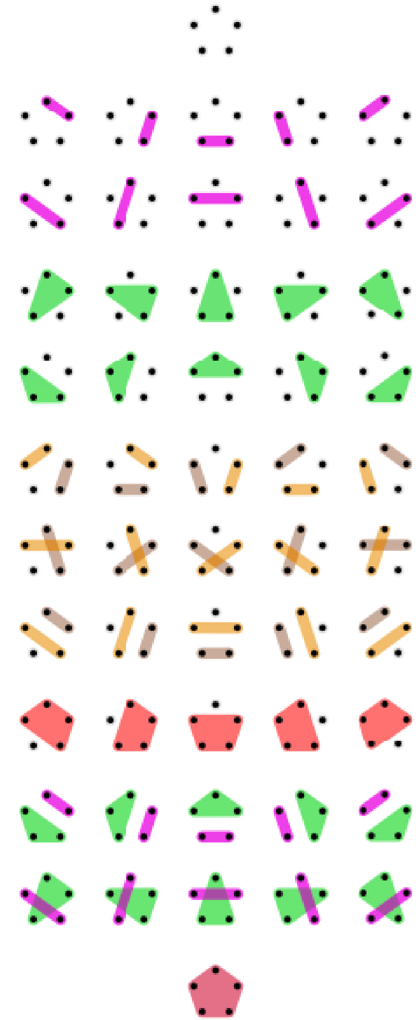
- Dirichlet Processes
  - Recap: Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

# Sidenote on Partitions

- Problem with partitions
  - If our goal is clustering, the output grouping is defined by an assignment of indicator variables

$$\left.\begin{array}{l} \mathbf{z}_n \sim \mathrm{Mult}(\boldsymbol{\pi}) \\ \mathbf{z}_n \sim \mathrm{Cat}(\boldsymbol{\pi}) \end{array}\right\} \boldsymbol{\pi} \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

  - The number of ways of assigning $N$ data points to $K$ mixtures is $K^N$.
  - If $K \geq N$, this is much larger than the number of ways of partitioning the data!
  - Example: $N$ = 5: 52 partitions vs. $5^5$ = 3125



⇒ ***Need representation that is invariant to relabeling!***


Image source: Wikipedia

# Chinese Restaurant Process (CRP)

- *How can DPs support clustering?*
- Chinese Restaurant Process
  - Visualize clustering as a sequential process of customers sitting at tables in an (infinitely large) restaurant.

    Customers ⇔ observed data to be clustered

    Tables ⇔ distinct blocks of partition, or clusters

  - This will help us see the clustering effect of DPs explicitly
- Relation to the clustering problem
  - We typically don't know the number of clusters and want to learn it from data
  - CRPs address this problem by assuming that there is an infinite number of latent clusters, but that only a finite number of them is used to generate the observed data.

# Chinese Restaurant Process (CRP)

- Procedure
  - Imagine a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers.
  - The 1st customer enters and sits at the first table.
  - The $N^{\text{th}}$ customer enters and sits at table

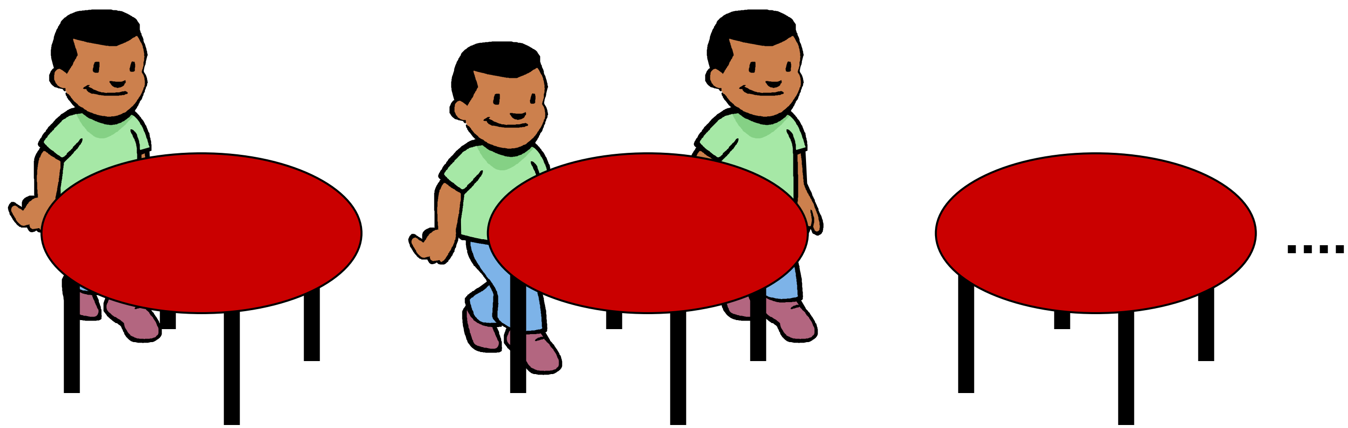
$$\begin{cases} k & \text{with prob } \dfrac{N_k}{N-1+\alpha} \quad \text{for } k = 1,\ldots,K \\ K+1 & \text{with prob } \dfrac{\alpha}{N-1+\alpha} \quad \text{(new table)} \end{cases}$$

where $N_k$ is the number of customers already sitting at table $k$.

- Remark
  - Metaphor was motivated by the seemingly infinite seating capability of Chinese restaurants in San Francisco…

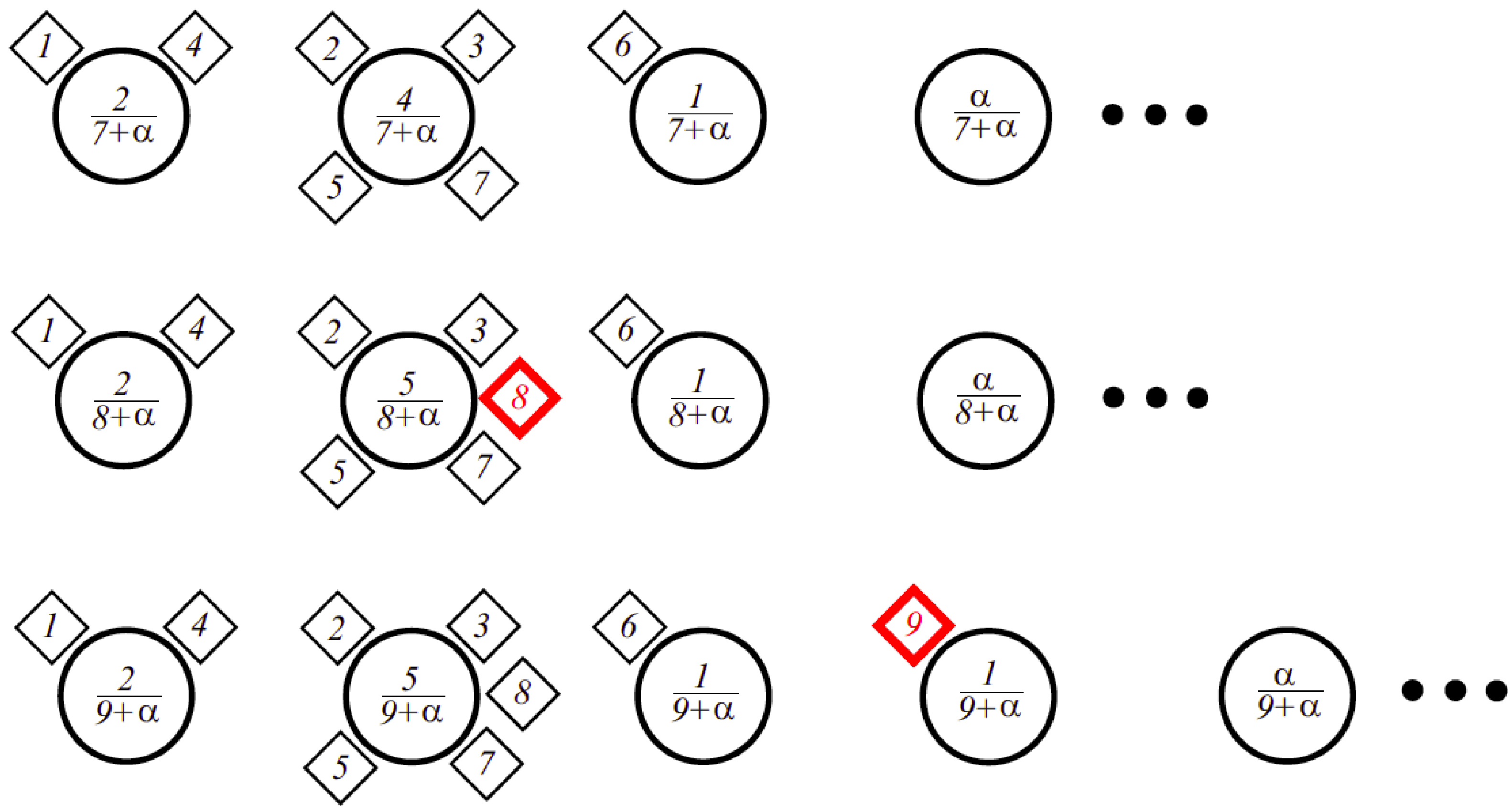
# Chinese Restaurant Process (CRP)

- **Visualization**




| $p(\mathbf{z}_n = \mathbf{z} \mid \mathbf{z}_{-n}) =$ | | | |
|---|---|---|---|
| | $1$ | $0$ | $0$ |
| | $\frac{1}{1+\alpha}$ | $\frac{\alpha}{1+\alpha}$ | $0$ |
| | $\frac{1}{2+\alpha}$ | $\frac{1}{2+\alpha}$ | $\frac{\alpha}{2+\alpha}$ |
| | $\frac{1}{3+\alpha}$ | $\frac{2}{3+\alpha}$ | $\frac{\alpha}{3+\alpha}$ |

# Chinese Restaurant Process (CRP)



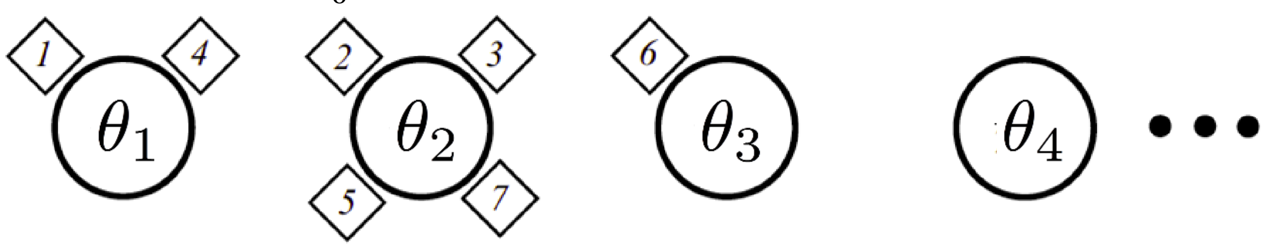

- **Resulting conditional distribution**

$$p(\mathbf{z}_N = \mathbf{z} | \mathbf{z}_1, \ldots, \mathbf{z}_{N-1}, \alpha) = \frac{1}{N-1+\alpha} \left( \sum_{k=1}^{K} N_k \delta(\mathbf{z}, k) + \alpha \delta(\mathbf{z}, \bar{k}) \right)$$


Image source: Erik Sudderth

# Relationship between CRPs and DPs

- Discussion
  - DP is a distribution over distributions.
  - DP results in discrete distributions, so if you draw $N$ points, you are likely to get repeated values.
  - A DP therefore induces a partitioning of the $N$ points.
  - The CRP is the corresponding distribution over partitions.
  - We can easily get back from the CRP to the Pólya urn scheme by the following extension:
    - When the first customer sits down at an empty table, he independently chooses a dish $\theta_k$ for the entire table from a prior distribution $G_0$.




    - Dish ⇔ parameters of the cluster

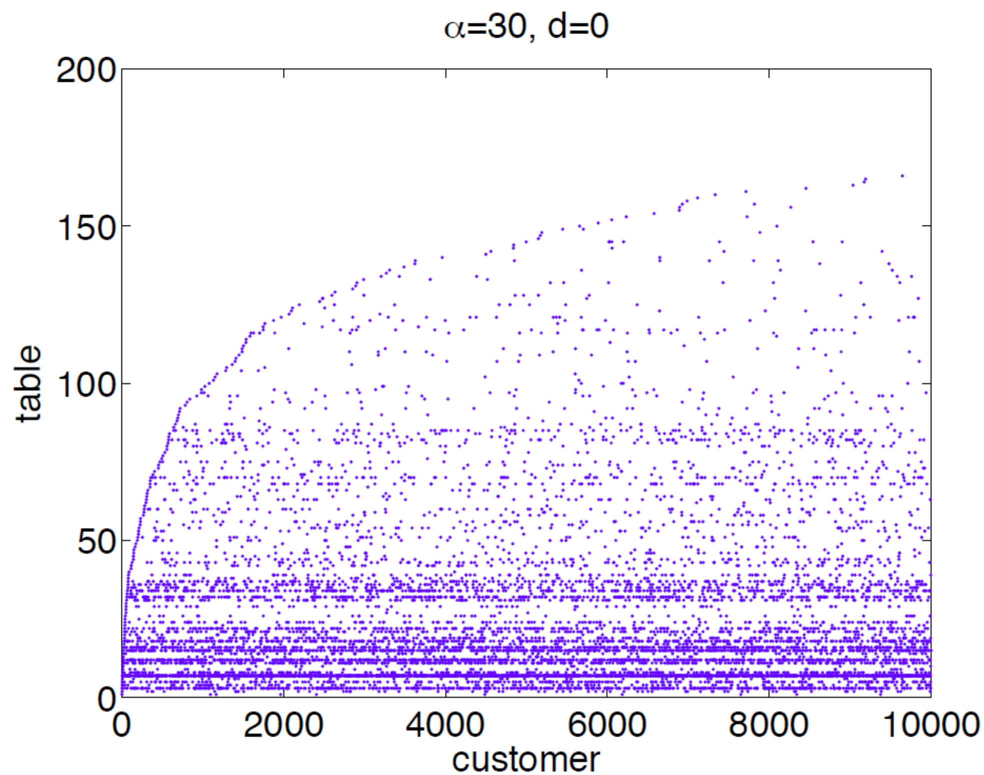




Slide inspired by: Zoubin Gharamani, Yee Whye Teh

Image source: Erik Sudderth

# Chinese Restaurant Process (CRP)




- **The CRP exhibits the clustering property of the DP.**
  - **Rich-gets-richer effect implies small number of large clusters.**
  - **Expected number of clusters is $K = \mathcal{O}(\alpha \log N)$.**

# CRPs & Exchangeable Partitions

$$p(\mathbf{z}_N = \mathbf{z} | \mathbf{z}_1, ..., \mathbf{z}_{N-1}, \alpha) = \frac{1}{N-1+\alpha} \left( \sum_{k=1}^{K} N_k \delta(\mathbf{z}, k) + \alpha \delta(\mathbf{z}, \bar{k}) \right)$$

- **Closer analysis**
  - Consider the probability of a certain seating arrangement:

$$p(\mathbf{z}_1, ..., \mathbf{z}_N | \alpha) = p(\mathbf{z}_1 | \alpha) p(\mathbf{z}_2 | \mathbf{z}_1, \alpha) \dots p(\mathbf{z}_N | \mathbf{z}_{N-1}, ..., \mathbf{z}_1, \alpha)$$

$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k)$$

  - Derivation of the terms

$$\alpha$$

First customer to sit at each table

$$1 \cdot 2 \cdots (N_k - 1)! = \Gamma(N_k)$$

Other customers joining each table

$$\frac{1}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdots \frac{1}{N-1+\alpha} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}$$

Normalization constants

# CRPs & Exchangeable Partitions

- Probability of a seating arrangement

$$p(\mathbf{z}_1, ..., \mathbf{z}_N | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k)$$

- Exchangeability property
  - The probability of a seating arrangement of $N$ customers is *independent* of the order they enter the restaurant!
  - The CRP is thus a prior on infinitely exchangeable partitions.
  - (Definition exchangeability: The joint probability underlying the data is invariant to permutation.)

- Why is this of importance?
  - Two reasons…

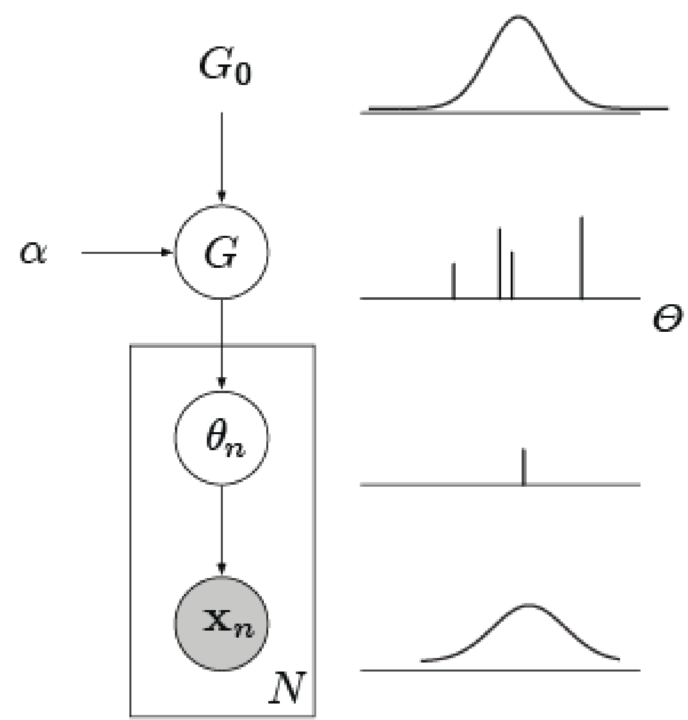
# Reason 1: De Finetti's Theorem

- **Putting all of this together…**
  - **De Finetti's theorem tells us that the CRP has an underlying mixture distribution with a prior distribution over measures.**
  - **The Dirichlet Process is the De Finetti mixing distribution for the CRP.**

- **Graphical model visualization**
  - **This means, when we integrate out $G$, we get the CRP:**

$$p(\theta_1, \ldots, \theta_N) = \int \prod_{n=1}^{N} p(\theta_n | G) \mathrm{d}P(G)$$

$\Rightarrow$ ***If the DP is the prior on $G$, then the CRP defines how points are assigned to clusters when we integrate out $G$.***








Slide adapted from Kurt Miller, Mike Jordan

Image source: Kurt Miller

# Reason 2: Efficient Inference

- **Taking advantage of exchangeability…**
  - In clustering applications, we are ultimately interested in the cluster assignments $\mathbf{z}_1,\ldots,\mathbf{z}_\mathrm{N}$.
  - Equivalent question in the CRP: Where should customer $n$ sit, conditioned on the seating choices of all the other customers?
    - This is easy when customer $n$ is the last customer to arrive:

$$p(\mathbf{z}_N = \mathbf{z}|\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}, \alpha) = \frac{1}{N-1+\alpha}\left(\sum_{k=1}^{K} N_k \delta(\mathbf{z}, k) + \alpha\delta(\mathbf{z}, \bar{k})\right)$$

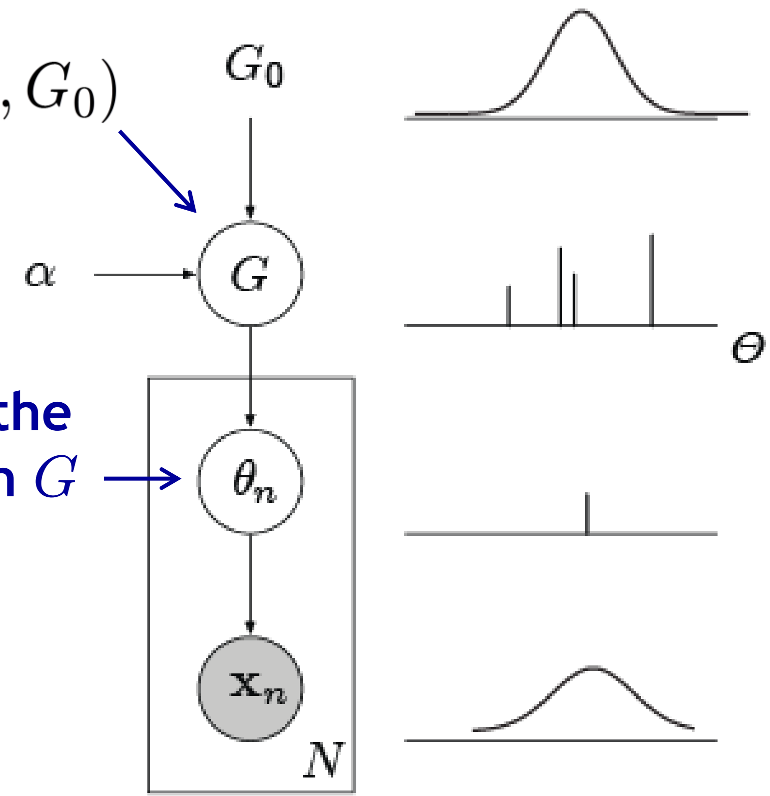
    - (Seemingly) hard otherwise…

$\Rightarrow$ ***Because of exchangeability, we can always swap customer $n$ with the final customer and use the above formula!***

$\Rightarrow$ **We'll use this for efficient Gibbs sampling later on…**

# Big Picture: CRPs and the DP

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




**The CRP describes the partitions of $\theta$ when $G$ is marginalized out**

# Topics of This Lecture

- Dirichlet Processes
  - Recap: Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

# Stick-Breaking Construction [Sethuraman, 1994]

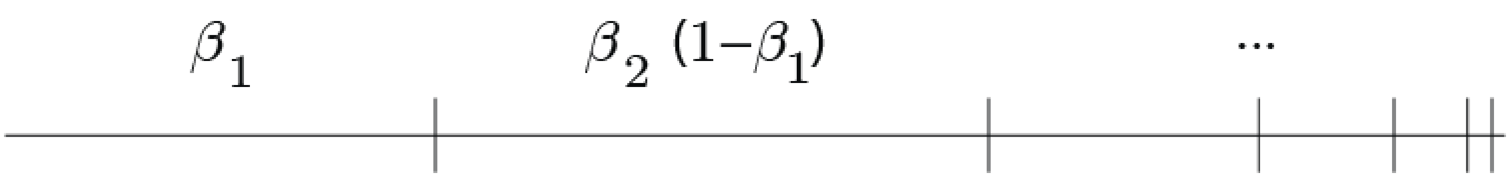
- **Explicit construction for the weights in DP realizations**
  - **Define an infinite sequence of random variables**

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

  - **Then define an infinite sequence of mixing proportions as**

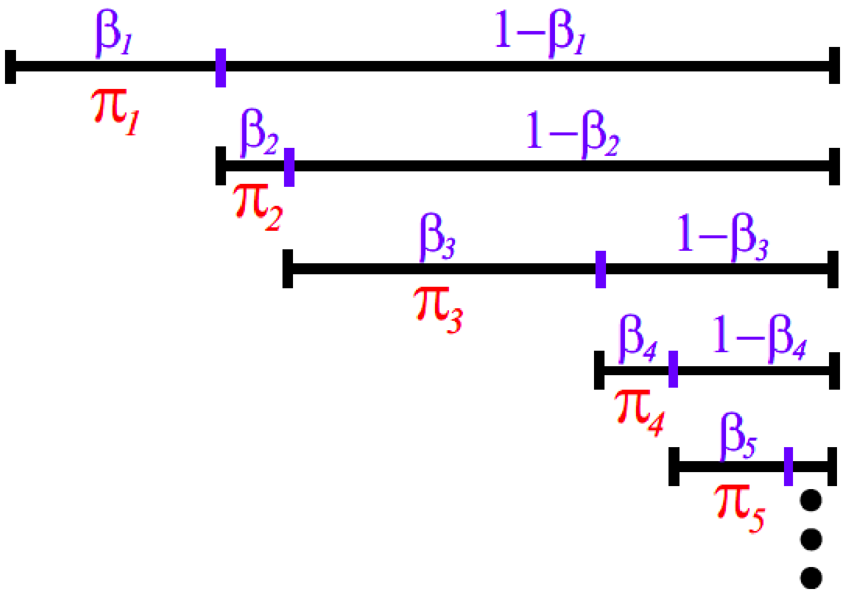
$$\begin{aligned} \pi_1 &= \beta_1 \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad\qquad k = 2, 3, \ldots \end{aligned}$$

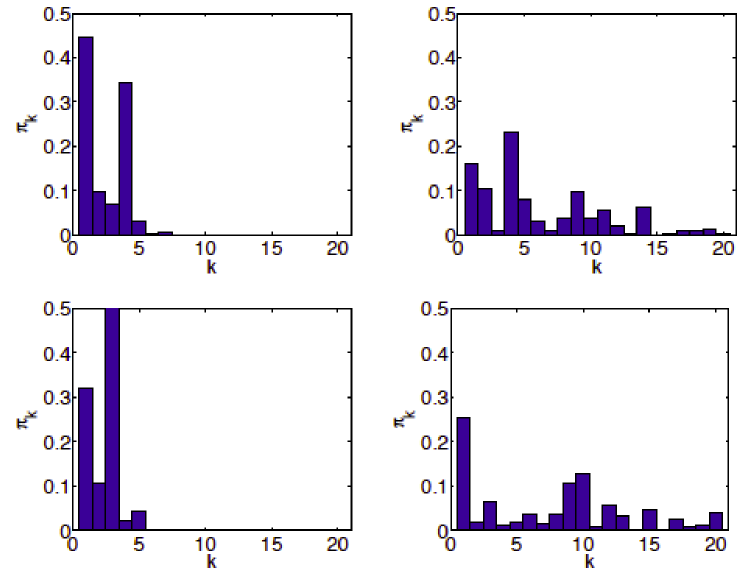
  - **This can be viewed as breaking off portions of a stick**




  - **When the $\pi_k$ are drawn this way, we can write $\pi \sim \mathrm{GEM}(\alpha)$.**
**(where $\mathrm{GEM}$ stands for Griffiths, Engen, McCloskey)**

# Stick-Breaking Example




$\alpha = 1$ $\alpha = 5$

- **Interpretation**
  - **Mixture weights $\pi_k$ partition a unit-length "stick" of probability mass among an infinite set of random parameters.**
  - **Note: The weights do not decrease monotonically!**

# Stick-Breaking Construction

- We now have an explicit formula for each $\pi_k$:

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

- We can also easily see that $\sum_{k=1}^{\infty} \pi_k = 1$:

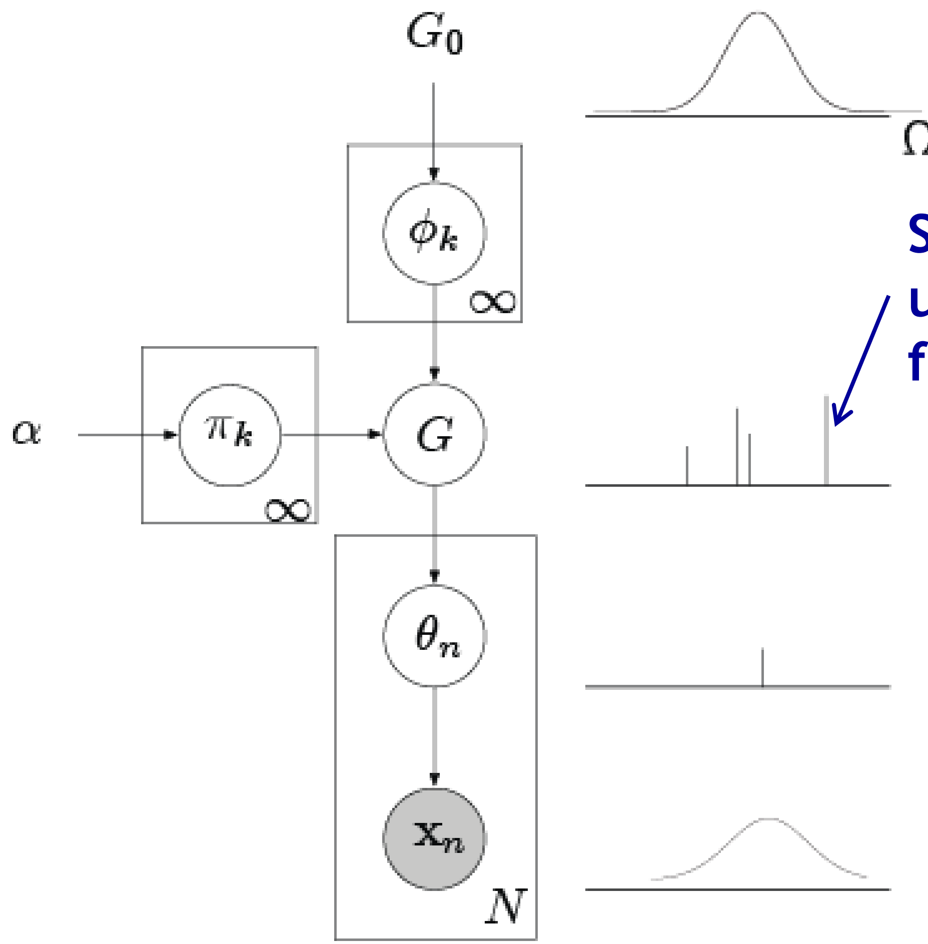
$$\begin{aligned} 1 - \sum_{k=1}^{K} \pi_k &= 1 - \beta_1 - \beta_2(1-\beta_1) - \beta_3(1-\beta_1)(1-\beta_2) - \ldots \\ &= (1-\beta_1)(1-\beta_2-\beta_3(1-\beta_2)-\ldots) \\ &= \prod_{k=1}^{K}(1-\beta_k) \\ &\to 0 \qquad \text{as } K \to \infty \end{aligned}$$

  - This shows that Dirichlet measures are discrete with probability one (as we already noted before).

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

# Big Picture: Stick-Breaking and the DP

- **Graphical Model representation**

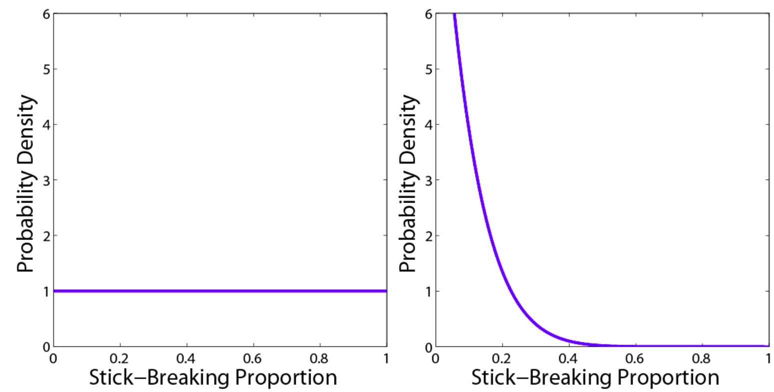



**Stick-Breaking allows us to sample directly from the weights**

# Dirichlet Stick-Breaking

- Sidenote
  - The Stick-Breaking representation provides another interpretation of the concentration parameter $\alpha$.
  - Since $\beta_k \sim \mathrm{Beta}(1,\alpha)$, we can apply standard moment formulas and find

$$\mathbb{E}[\beta_k] = \frac{1}{1+\alpha}$$



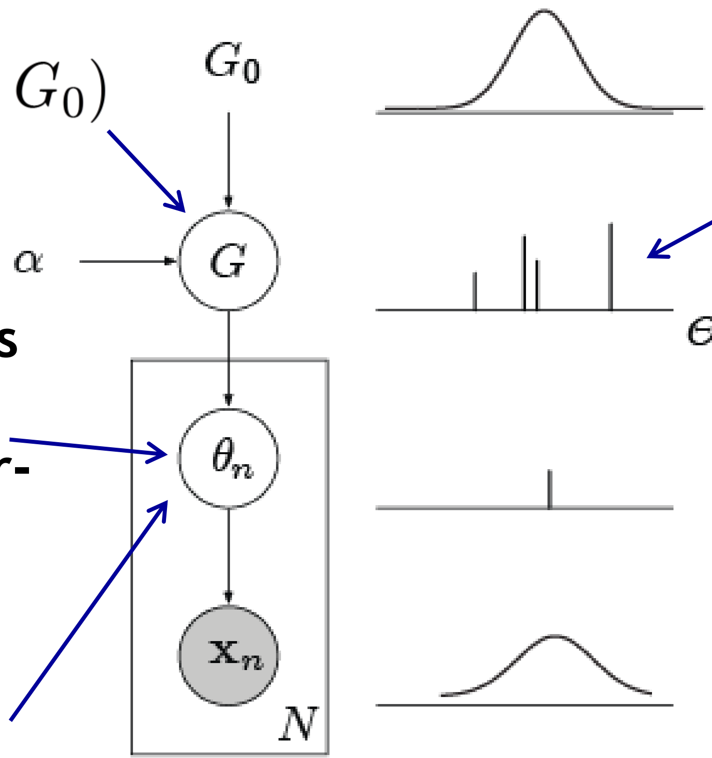

$\alpha = 1$ $\alpha = 10$

$\Rightarrow$ For small $\alpha$, the first few mixture components are typically assigned the majority of the probability mass.

$\Rightarrow$ For $\alpha \to \infty$, samples $G \sim \mathrm{DP}(\alpha, G_0)$ approach the base measure $G_0$ by assigning small, roughly uniform weights to a densely sampled set of discrete parameters.

# Summary: Pólya Urns, CRPs, and Stick-Breaking

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




The Pólya urn describes the predictive distribution of $\theta$ when $G$ is marginalized out

The CRP describes the partitions of $\theta$ when $G$ is marginalized out

The Stick-Breaking Process describes the partition weights

# References and Further Reading



- Unfortunately, there are currently no good introductory textbooks on the Dirichlet Process. We will therefore post a number of tutorial papers on their different aspects.
  - One of the best available general introductions
    - E.B. Sudderth, "Graphical Models for Visual Object Recognition and Tracking", PhD thesis, Chapter 2, Section 2.5, 2006.
  - A gentle introductory tutorial (recommended 1st read)
    - S.J. Gershman, D.M. Blei, „A Tutorial on Bayesian Nonparametric Methods", In Journal of Mathematical Psychology, Vol. 56, 2012.
  - Good overview of MCMC methods for DPMMs
    - R. Neal, Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, Vol. 9(2), p. 249-265, 2000.