Advanced Machine Learning
Lecture 11

Dirichlet Processes
28.11.2012

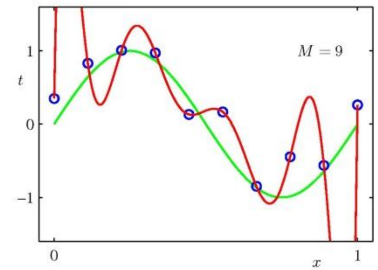
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

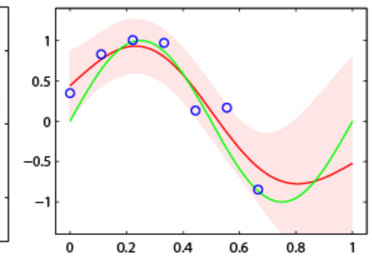
leibe@vision.rwth-aachen.de

This Lecture: *Advanced Machine Learning*

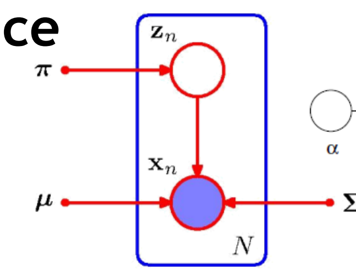
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

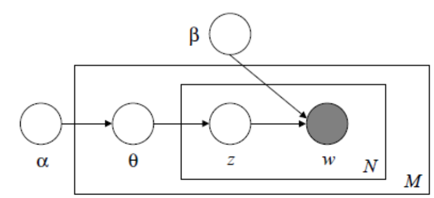
$$f : \mathcal{X} \rightarrow \mathbb{R}$$





- 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

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

Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Polya Urn Process
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Discussion
- Dirichlet Process Mixture Models
  - Comparison to finite mixture models
  - Efficient sampling
  - Applications

Recap: Bayesian Mixture Models

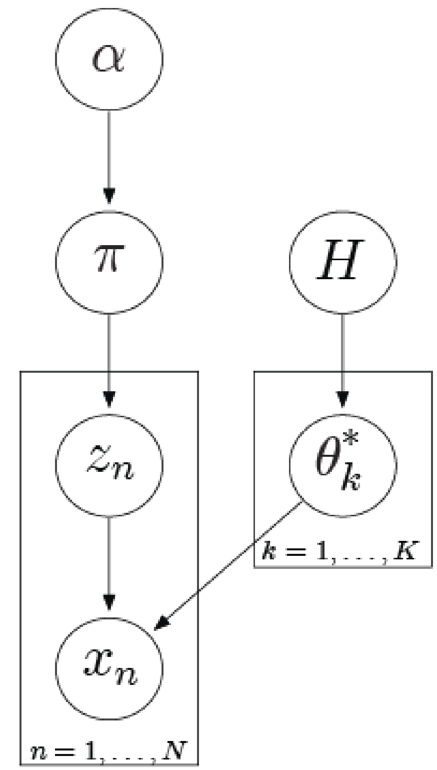
- Let's be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable $\mathbf{z}_n$ as indicator which component data point $\mathbf{x}_n$ belongs to.

$$\mathbf{z}_n|\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\mathbf{x}_n|\mathbf{z}_n = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$$

  - Introduce conjugate priors over parameters

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$
$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim H = \mathcal{N} - \mathcal{IW}(0, s, d, \phi)$$

"Normal - Inverse Wishart"



Recap: Bayesian Mixture Models

- **Full Bayesian Treatment**
  - Given a dataset, we are interested in the cluster assignments

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{\sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}$$

where the likelihood is obtained by marginalizing over the parameters θ

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z}) &= \int p(\mathbf{X}|\mathbf{Z},\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \int \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|z_{nk},\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|H)\mathrm{d}\boldsymbol{\theta} \end{aligned}$$

- **The posterior over assignments is intractable!**
  - Denominator requires summing over all possible partitions of the data into K groups!

$\Rightarrow$ We will see efficient approximate inference methods later on...

Recap: Mixture Models with Dirichlet Priors

- **Integrating out the mixing proportions π**

$$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi}$$

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

- **Conditional probabilities**
  - Examine the conditional of $\mathbf{z}_n$ given all other variables $\mathbf{z}_{-n}$

$$p(z_{nk}=1|\mathbf{z}_{-n},\alpha) = \frac{p(z_{nk}=1,\mathbf{z}_{-n}|\alpha)}{p(\mathbf{z}_{-n}|\alpha)}$$

$$= \frac{N_{-n,k} + \alpha/K}{N-1+\alpha} \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1,i\neq n}^{N} z_{ik}$$

$\Rightarrow$ **The more populous a class is, the more likely it is to be joined!**

Recap: Infinite Dirichlet Mixture Models

- Conditional probabilities: Finite K

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \frac{N_{-n,k} + \alpha/K}{N - 1 + \alpha}, \qquad N_{-n,k} \stackrel{\text{def}}{=} \sum_{i=1, i \neq n}^{N} z_{ik}$$

- Conditional probabilities: Infinite K
  - Taking the limit as $K \to \infty$ yields the conditionals

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \begin{cases} \frac{N_{-n,k}}{N-1+\alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N-1+\alpha} & \text{if all } k \text{ not represented} \end{cases}$$

  - Left-over mass α $\Rightarrow$ countably infinite number of indicator settings

Note

- **Why this term if all k are not represented?**

$$p(z_{nk} = 1 | \mathbf{z}_{-n}, \alpha) = \begin{cases} \frac{N_{-n,k}}{N-1+\alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N-1+\alpha} & \text{if all } k \text{ not represented} \end{cases}$$

  - **The total probability assigned to all unoccupied clusters is determined by the complement of existing cluster weights:**

$$\lim_{K \to \infty} p(\mathbf{z}_n \neq \mathbf{z}_m \text{ for all } n \neq m | \mathbf{z}_{-n}, \alpha) = 1 - \sum_{k=1}^{K} \frac{N_{-n,k}}{N-1+\alpha}$$

$$= \frac{N-1+\alpha \; - \; (N-1)}{N-1+\alpha}$$

$$= \frac{\alpha}{N-1+\alpha}$$

Topics of This Lecture

- **Finite Bayesian Mixture Models**
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Polya Urn Process
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Discussion
- Dirichlet Process Mixture Models
  - Comparison to finite mixture models
  - Efficient sampling
  - Applications

Gibbs Sampling for Finite Mixtures

- We need approximate inference here
  - Gibbs Sampling: Conditionals are simple to compute

$$p(\mathbf{z}_n = k|\text{others}) \propto \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\pi} \mid \mathbf{z} \sim \text{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K)$$

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\text{others} \sim \mathcal{N} - \mathcal{IW}(v', s', d', \phi')$$

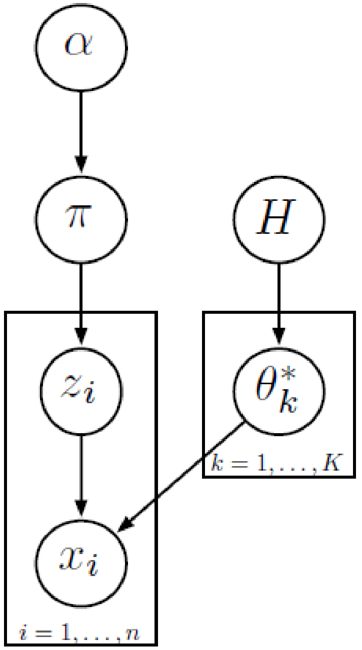
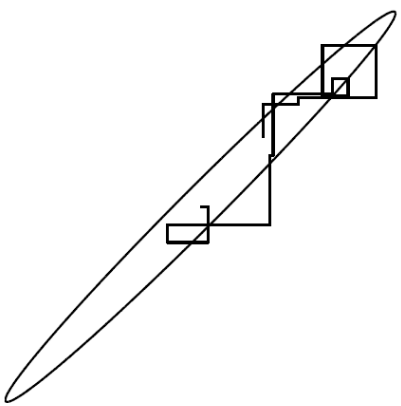


Image source: Yee Whye Teh

Recap: Gibbs Sampling

- Approach
  - MCMC-algorithm that is simple and widely applicable.
  - May be seen as a special case of Metropolis-Hastings.
- Idea
  - Sample variable-wise: replace $\mathbf{z}_i$ by a value drawn from the distribution $p(z_i|\mathbf{z}_{\setminus i})$.
    - This means we update one coordinate at a time.
  - Repeat procedure either by cycling through all variables or by choosing the next variable.
- Properties
  - The algorithm always accepts!
  - Completely parameter free.
  - Can also be applied to subsets of variables.



Gibbs Sampling for Finite Mixtures

- Standard finite mixture sampler
  - Given mixture weights $\boldsymbol{\pi}^{(t\text{-}1)}$ and cluster parameters $\left\{\boldsymbol{\theta}_k^{(t-1)}\right\}_{k=1}^{K}$ from the previous iteration, sample new parameters as follows

1. Independently assign each point $\mathbf{x}_n$ to one of the K clusters by sampling the variables $\mathbf{z}_n$ from the multinomial distributions

$$\mathbf{z}_n^{(t)} \sim \frac{1}{Z_n} \sum_{k=1}^{K} z_{nk}^{(t-1)} \pi_k^{(t-1)} p(\mathbf{x}_n | \theta_k^{(t-1)}) \qquad Z_n = \sum_{k=1}^{K} \pi_k^{(t-1)} p(\mathbf{x}_n | \theta_k^{(t-1)})$$

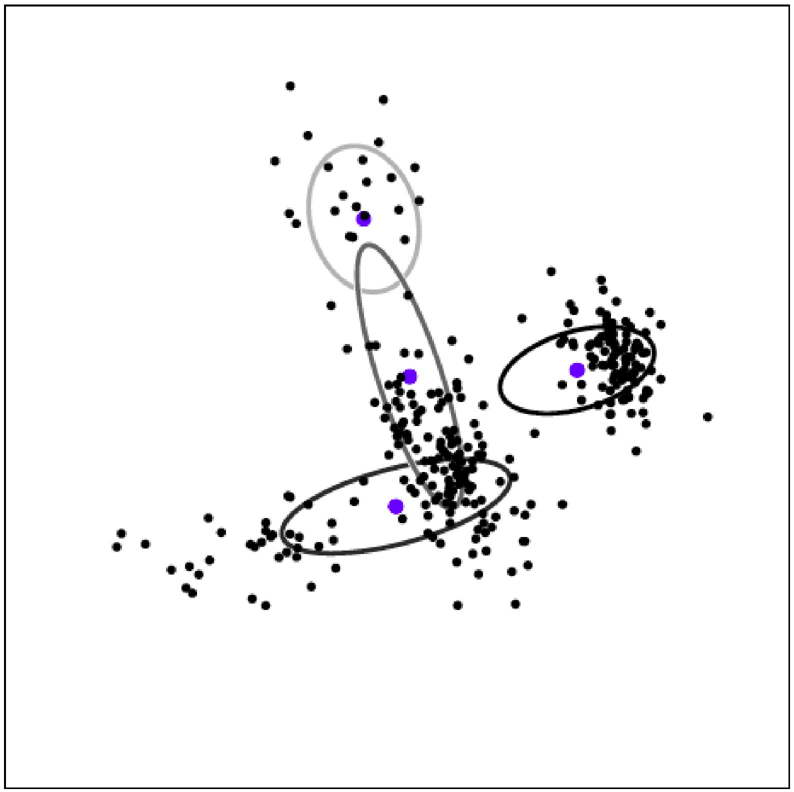
2. Sample new mixture weights from the Dirichlet distribution

$$\boldsymbol{\pi}^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{n=1}^{N} z_{nk}^{(t)}$$

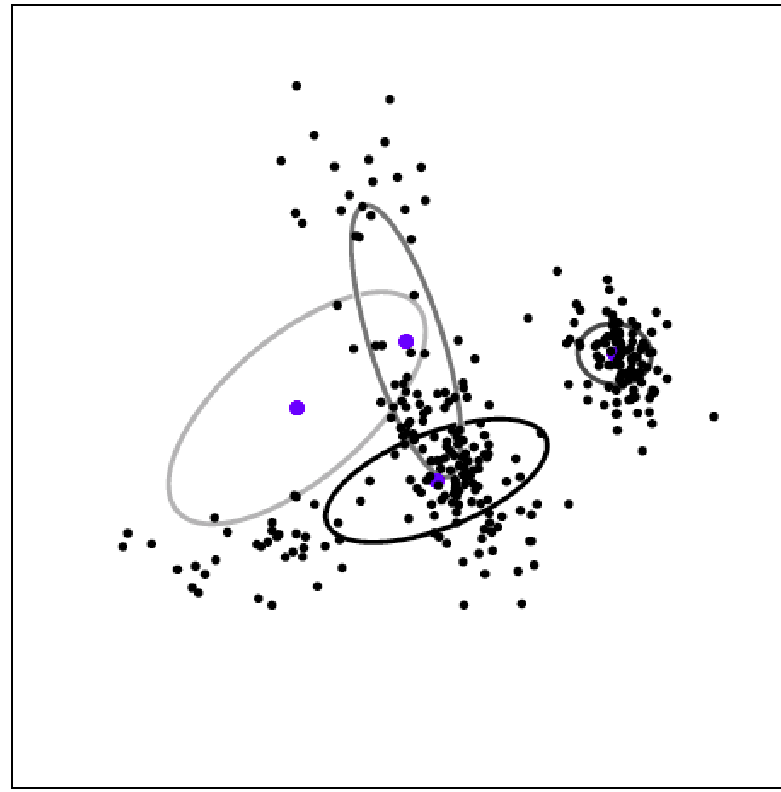
3. For each of the K clusters, independently sample new parameters from the conditional of the assigned observations

$$\boldsymbol{\theta}_k^{(t)} \sim p(\boldsymbol{\theta}_k | \left\{\mathbf{x}_n | z_{nk} = 1\right\}, H)$$

Standard Sampler: 2 Iterations

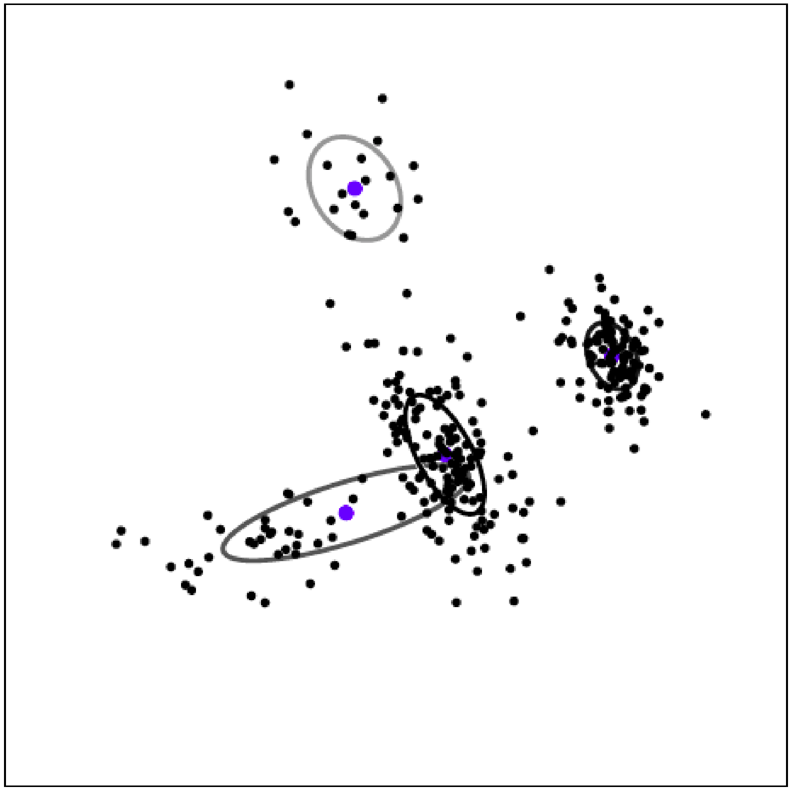


log p(x | π, θ) = –539.17

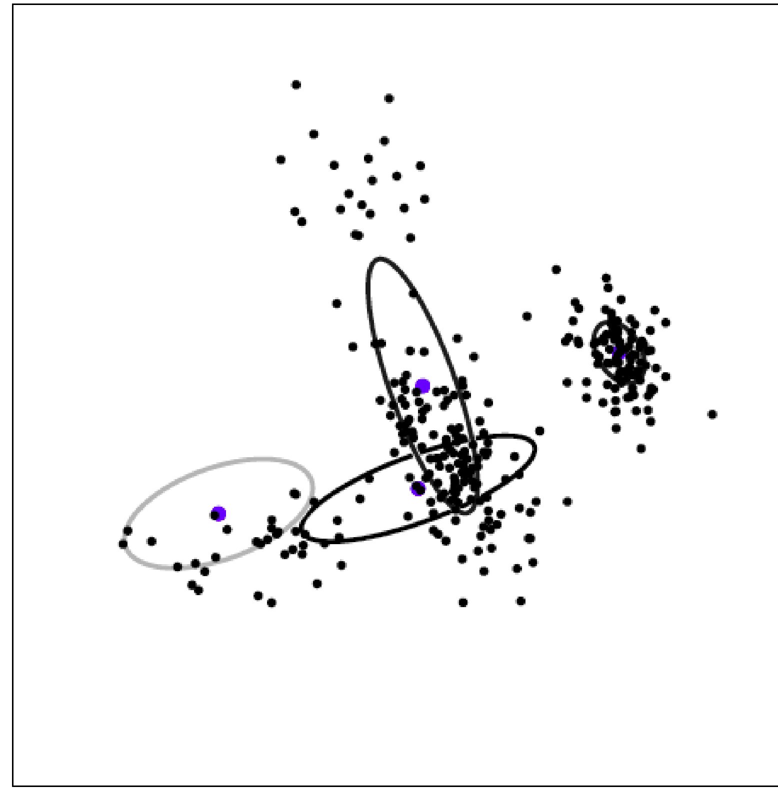


log p(x | π, θ) = –497.77

Standard Sampler: 10 Iterations

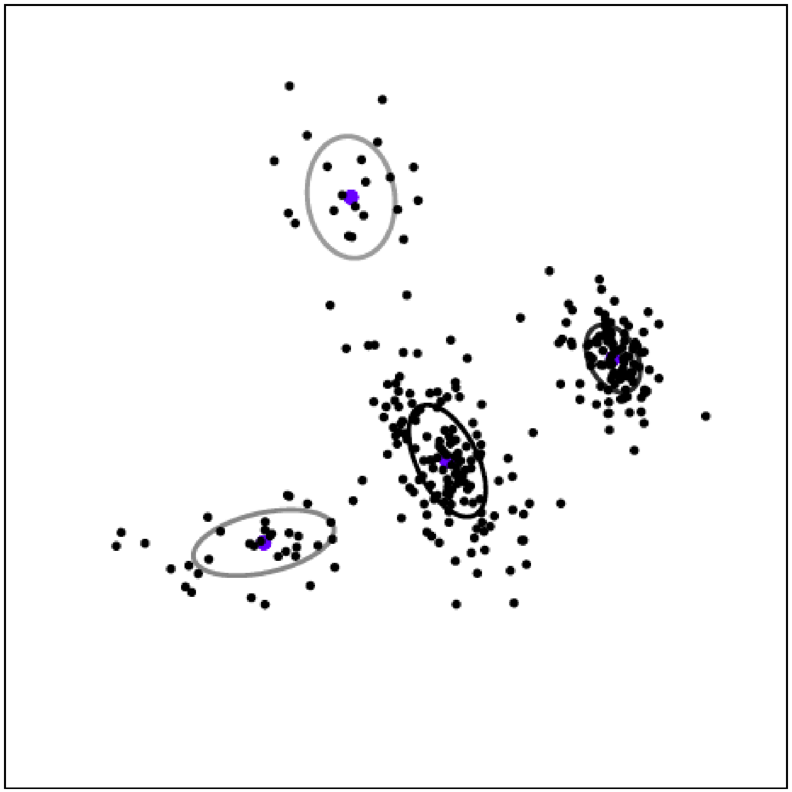


log p(x | π, θ) = –404.18

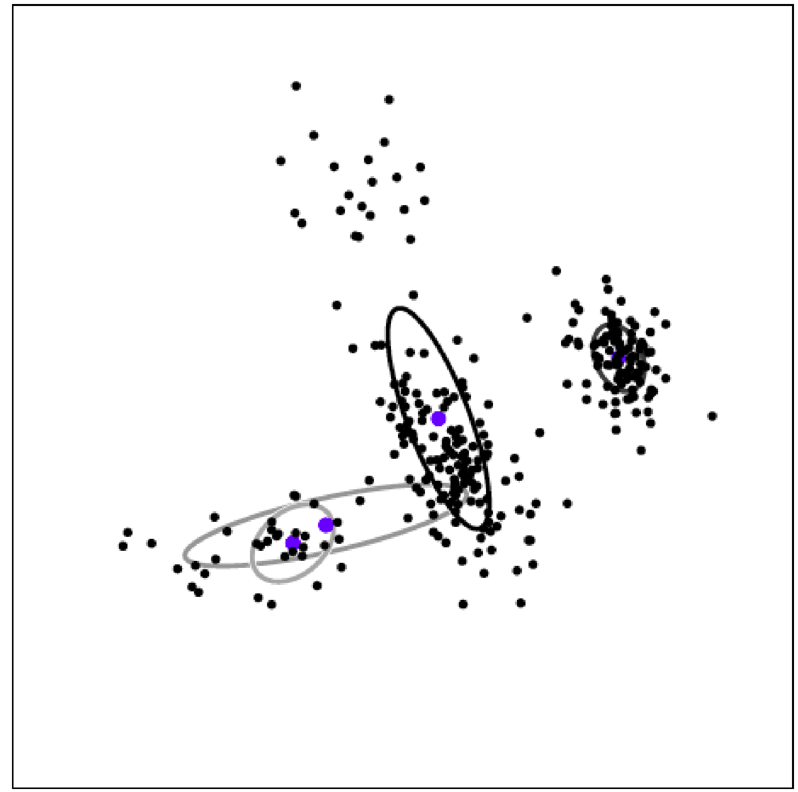


log p(x | π, θ) = –454.15

Standard Sampler: 50 Iterations



log p(x | π, θ) = –397.40



log p(x | π, θ) = –442.89

Slide credit: Erik Sudderth

Image source: Erik Sudderth

Gibbs Sampling for Finite Mixtures

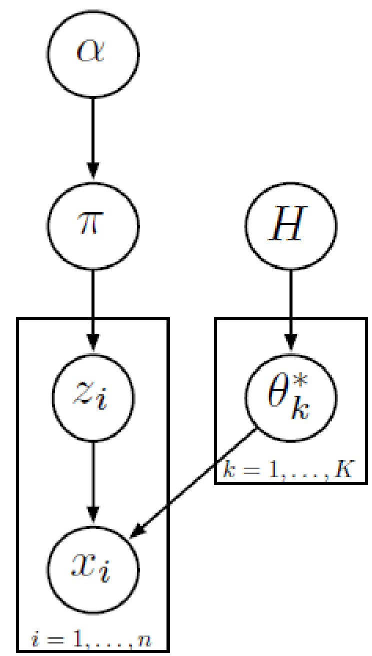
- **We need approximate inference here**
  - Gibbs Sampling: Conditionals are simple to compute

$$p(\mathbf{z}_n = k|\text{others}) \propto \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{\pi} \mid \mathbf{z} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K)$$

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\text{others} \sim \mathcal{N} - \mathcal{IW}(v', s', d', \phi')$$

- **However, this will be rather inefficient…**
  - In each iteration, algorithm can only change the assignment for individual data points.
  - There are often groups of data points that are associated with high probability to the same component. $\Rightarrow$ Unlikely that group is moved.
  - Better performance by collapsed Gibbs sampling which integrates out the parameters π, μ, Σ.



Collapsed Finite Bayesian Mixture

- More efficient algorithm
  - Conjugate priors allow analytic integration of some parameters
  - Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)
- Necessary steps
  - The model implies the factorization

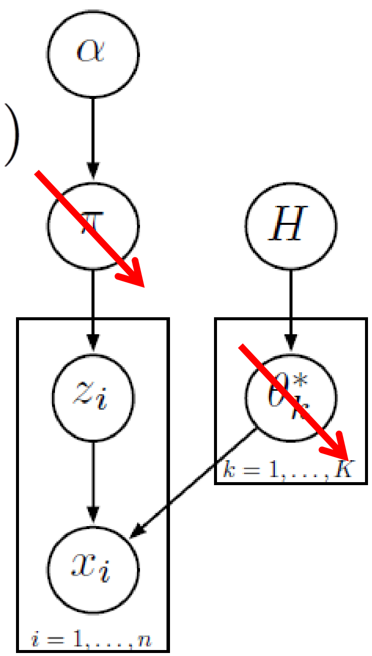
$$p(\mathbf{z}_n|\mathbf{z}_{-n},\mathbf{x},\alpha,H) \propto p(\mathbf{z}_n|\mathbf{z}_{-n},\alpha)p(\mathbf{x}_n|\mathbf{z},\mathbf{x}_{-n},H)$$

  - Derive

$$p(\mathbf{z}|\alpha) = \int p(\mathbf{z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi}$$

$$p(\mathbf{x}_n|\mathbf{z}_n,H) = \int \sum_{k=1}^{K} z_{nk}p(\mathbf{x}_n|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|H)\mathrm{d}\boldsymbol{\theta}$$

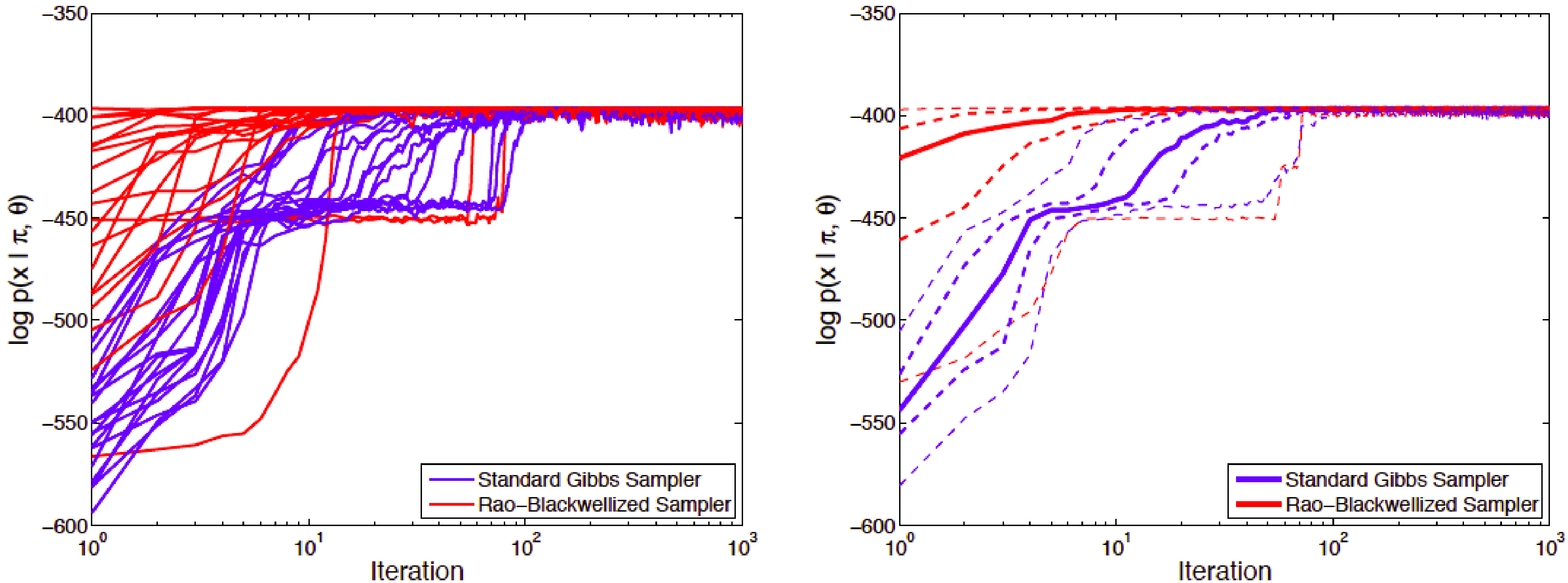
$\Rightarrow$ Conjugate prior, Normal - Inverse Wishart



Collapsed Finite Mixture Sampler

- **Algorithm**
  1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1,\ldots,N\}$.
  2. Set $\mathbf{z} = \mathbf{z}^{(t\text{-}1)}$. For each $n \in \{\tau(1),\ldots,\tau(N)\}$, sequentially resample $\mathbf{z}_n$ as follows
     a) For each of the K clusters, determine the predictive likelihood (this can be computed from cached sufficient statistics)
     $$p_k(\mathbf{x}_n|\mathbf{z}_{-n},H) = p(\mathbf{x}_n|\{\mathbf{x}_m|z_{mk}=1, m \neq n\}, H)$$
     b) Sample a new assignment $\mathbf{z}_n$ from the multinomial distribution
     $$\mathbf{z}_n \sim \sum_{k=1}^{K} \frac{z_{nk}(N_{-n,k}+\alpha/K)p_k(\mathbf{x}_n|\mathbf{z}_{-n},H)}{\sum_{j=1}^{K}(N_{-n,j}+\alpha/K)p_j(\mathbf{x}_n|\mathbf{z}_{-n},H)}$$
     c) Update cached sufficient statistics to reflect assignment z_{nk}.
  3. Set $\mathbf{z}^{(t)} = \mathbf{z}$. Optionally, mixture parameters may be sampled via steps 2-3 of the standard finite mixture sampler.

Standard vs. Collapsed Samplers



⇒ **Collapsed sampler converges much more quickly.**

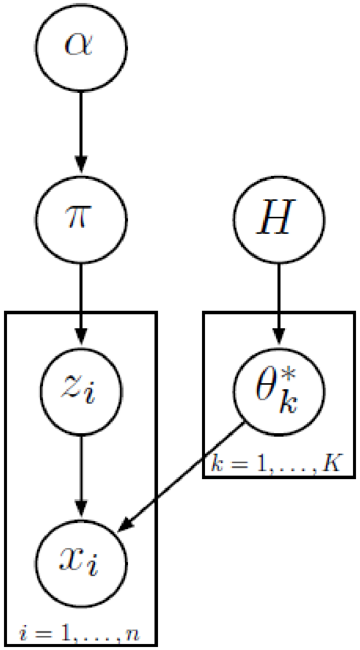
- **Theorem (Rao-Blackwell)**

  "Analytical marginalization of some variables from a joint distribution always reduces the variance of later estimates."

Discussion

- **Collapsed Gibbs sampling**
  - Integrates out the parameters π, μ, Σ.

$$p(z_{nk} = 1 | \text{others}) \propto \frac{(N_{-n,k} + \alpha/K)}{N - 1 + \alpha} p_k(\mathbf{x}_n | \mathbf{z}_{-n}, H)$$



- **Properties**
  - Can change all assignments in each iteration.
  - $\Rightarrow$ Able to move entire groups between clusters.
  - $\Rightarrow$ Faster convergence.
  - However, similar worst-case performance as standard sampler, may get stuck in local optima for many iterations.

Image source: Yee Whye Teh

Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Polya Urn Process
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Discussion
- Dirichlet Process Mixture Models
  - Comparison to finite mixture models
  - Efficient sampling
  - Applications

Dirichlet Processes

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

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

where μ 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.

Dirichlet Processes

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

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

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 Θ is drawn from a Dirichlet Process if its measure on every finite partition follows a Dirichlet distribution.*

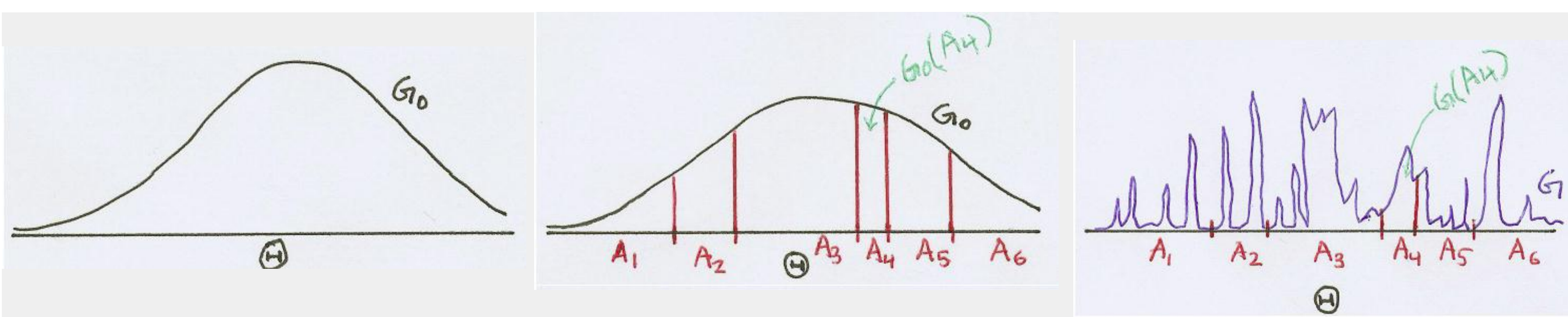
Dirichlet Processes

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

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

means that

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



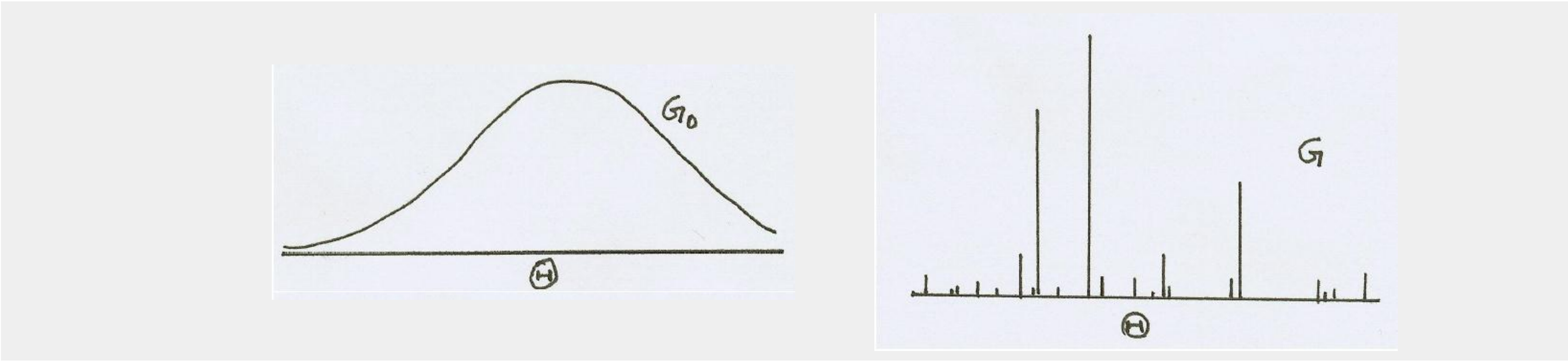
Dirichlet Processes

- **Important property** **[Blackwell]**
  - **Draws from a DP will always place all their mass on a countable set of points.**

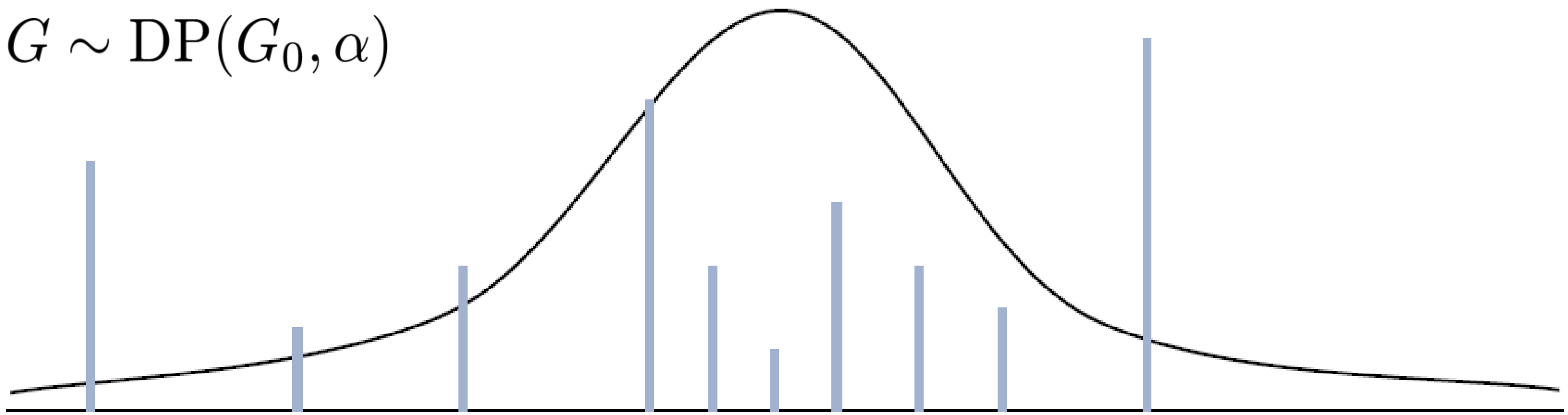
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

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

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



Dirichlet Processes: Discussion



- **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!*

Dirichlet Processes: Properties

- **Moments**

$$\mathbb{E}[G(A)] = G_0(A) \qquad \mathrm{var}[G(A)] = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1}$$

- **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. ⇒ Posterior is again a DP.

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

Properties

- Summary so far
  - We have seen some of the formal properties of DPs.
  - But how can we use them?
  - How can we sample from them?

  - In the following, we will characterize DPs through several different constructions in order to highlight key properties…
- Constructions
  - Polya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking Construction

Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Polya Urn Scheme
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Discussion
- Dirichlet Process Mixture Models
  - Comparison to finite mixture models
  - Efficient sampling
  - Applications

Polya's Urns [Blackwell & MacQueen, 1973]

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

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

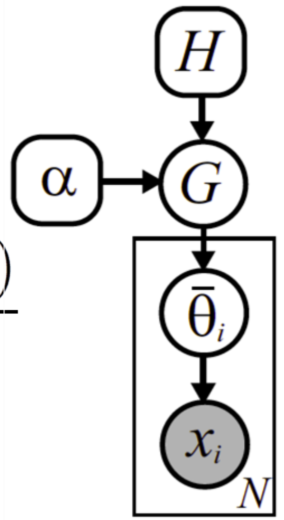
- **Yes, by a variation of the classical balls-in-urns analogy**
  - **Assume that G_0 is a distribution over colors, and that each θ_n represents the color of a single ball placed in the urn.**
  - **Start with an empty urn. Repeat for N steps:**

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 θ_n and return the ball into the urn, along with a new one of the same color.**

Image source: Yee Whye Teh

Polya's Urns: Discussion

- Polya 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.

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

Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Polya Urn Scheme
  - Chinese Restaurant Process
  - Stick-breaking construction
  - Discussion
- Dirichlet Process Mixture Models
  - Comparison to finite mixture models
  - Efficient sampling
  - Applications

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

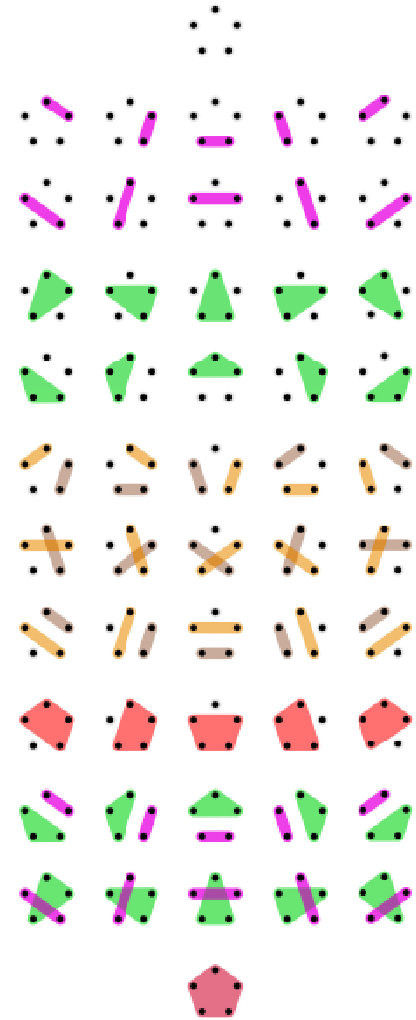
  - This will help us see the clustering effect of DPs explicitly
- Relation to the clustering problem
  - We typically don't know the number of clusters and want to learn it from data
  - CRPs address this problem by assuming that there is an infinite number of latent clusters, but that only a finite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

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!

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^{th} customer enters and sits at table

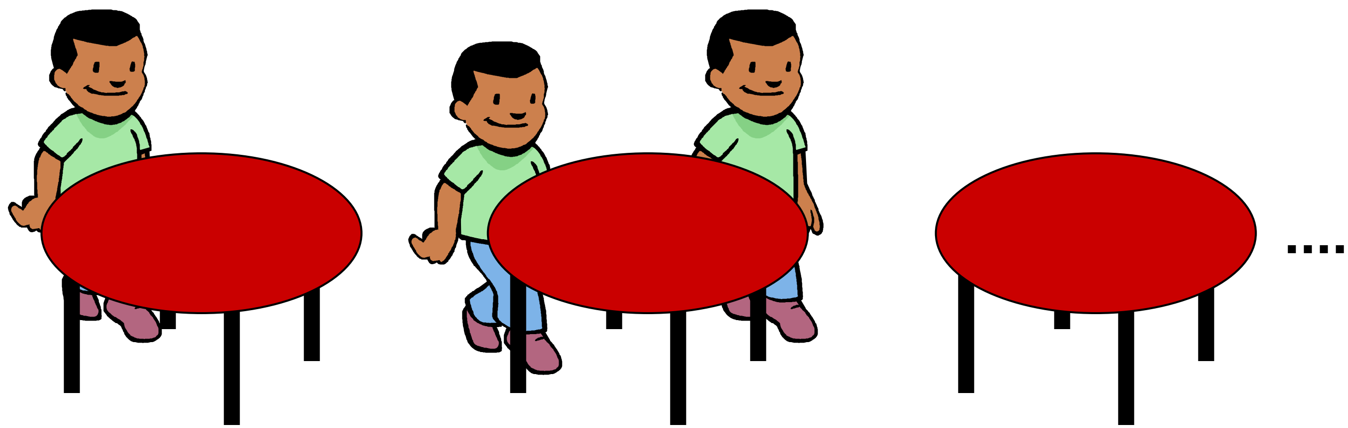
$$\begin{cases} k & \text{with prob } \dfrac{N_k}{N-1+\alpha} \text{ for } k = 1,\ldots,K \\ K+1 & \text{with prob } \dfrac{\alpha}{N-1+\alpha} \quad \text{(new table)} \end{cases}$$

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...

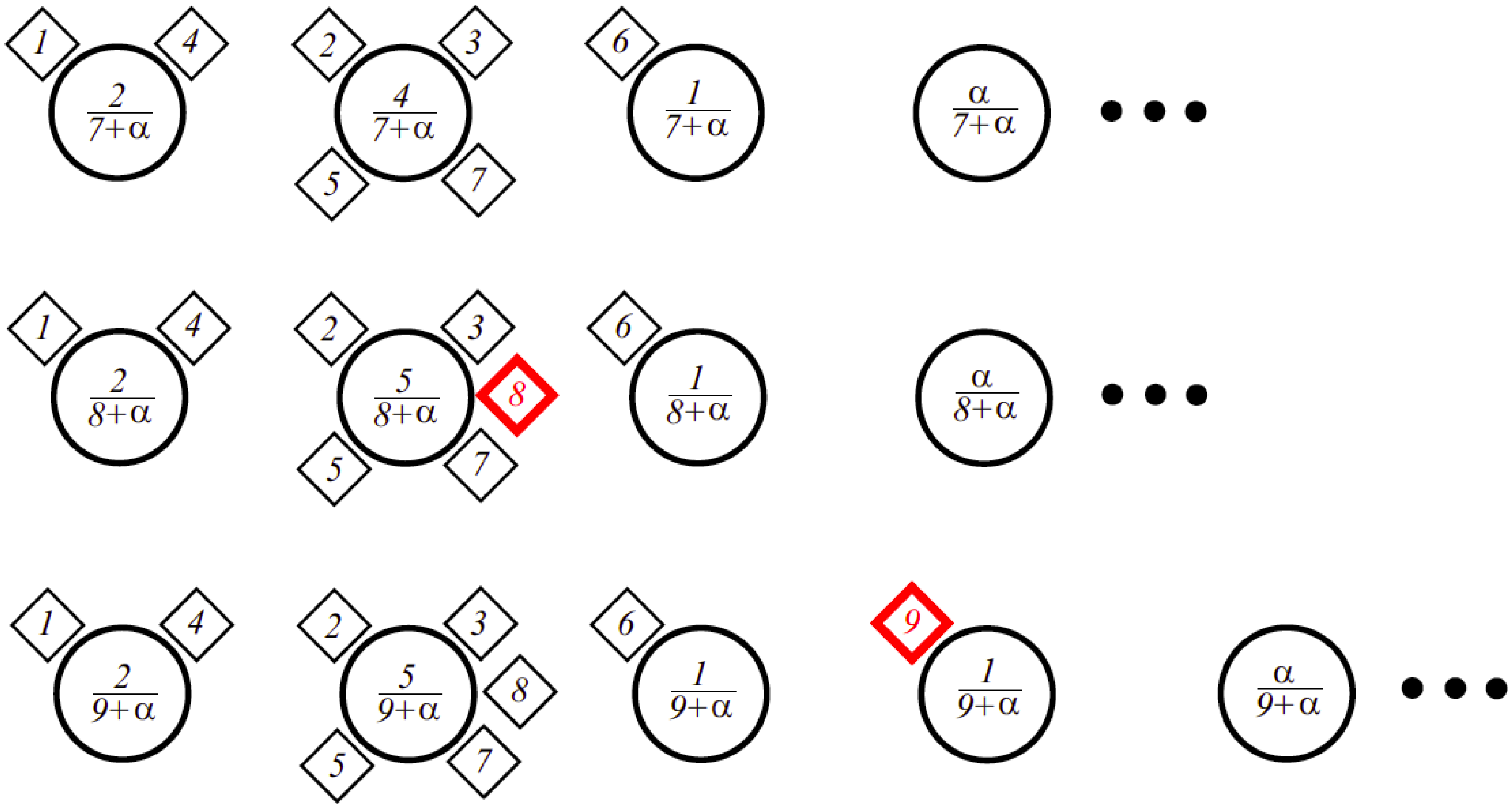
Chinese Restaurant Process (CRP)

- **Visualization**



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

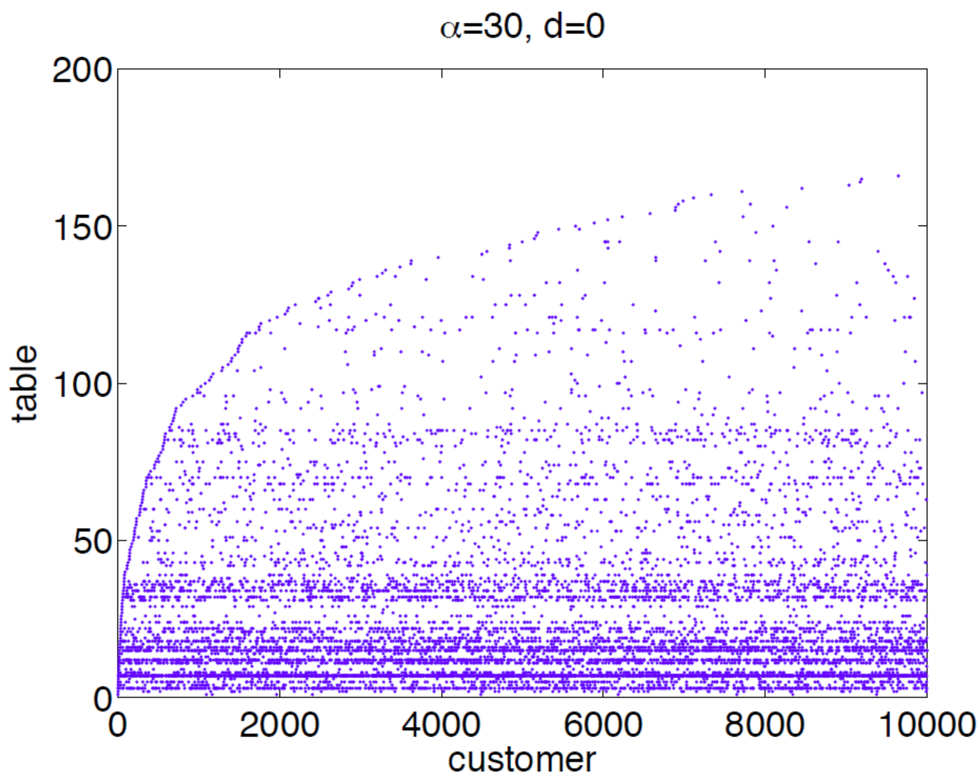
Chinese Restaurant Process (CRP)



- **Resulting conditional distribution**

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

Chinese Restaurant Process



- 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

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

- **Exchangeability property**
  - **The probability of a seating arrangement of N customers is *independent* of the order they enter the restaurant:**

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

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

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

normalization constants

$$\alpha$$

first customer to sit at each table

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

other customers joining each table

Discussion

- Relationship between CRPs and DPs
  - 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 induces a partitioning of the N points

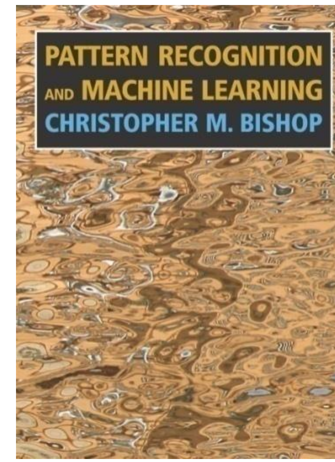
    e.g., $(1\ 3\ 4)\ (2\ 5)$, $\mathbf{z}_1 = \mathbf{z}_3 = \mathbf{z}_4 \neq \mathbf{z}6 = \mathbf{z}_2 = \mathbf{z}_5$
  - CRP is the corresponding distribution over partitions.

RWTH AACHEN UNIVERSITY

References and Further Reading

- More information about EM estimation is available in Chapter 9 of Bishop's book (recommendable to read).

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



- Additional information
  - Original EM paper:
    - A.P. Dempster, N.M. Laird, D.B. Rubin, „Maximum-Likelihood from incomplete data via EM algorithm", In Journal Royal Statistical Society, Series B. Vol 39, 1977
  - EM tutorial:
    - J.A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", TR-97-021, ICSI, U.C. Berkeley, CA,USA