Applied Bayesian Nonparametrics

1. Models & Inference

Tutorial at CVPR 2012

Erik Sudderth

Brown University

Additional detail & citations in background chapter:
E. B. Sudderth, Graphical Models for Visual Object Recognition and Tracking, PhD Thesis, MIT, 2006.

Applied

Focus on those models which are most useful in practice.
To understand those models, we'll start with theory…

Bayesian

Not no parameters! Models with infinitely many parameters.
Distributions on uncertain functions, distributions, …

Nonparametric

Complex data motivates models of unbounded complexity.
We often need to learn the structure of the model itself.

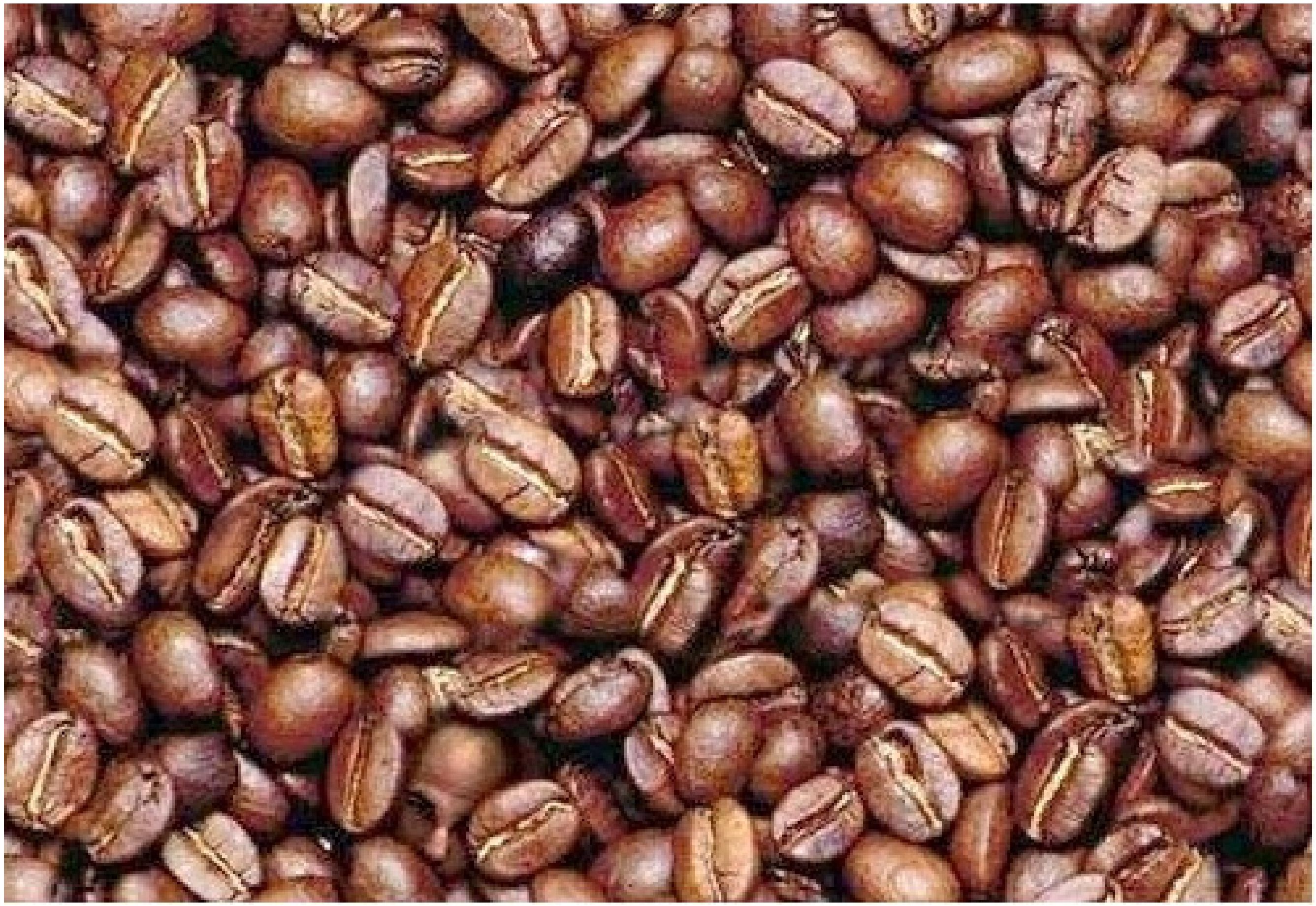
Statistics

Learning probabilistic models of visual data.
Clustering & features, space & time, mostly unsupervised.

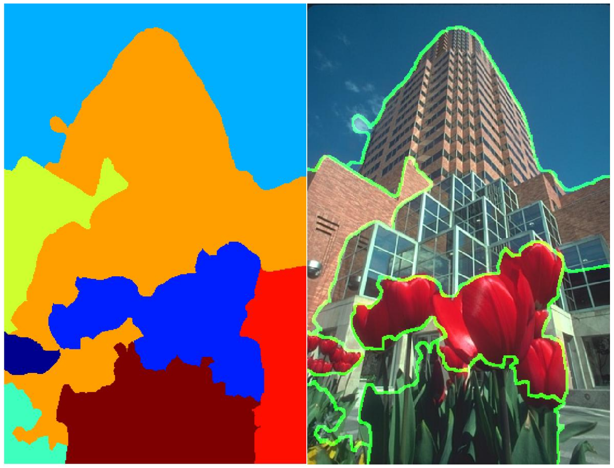
Applied BNP: Part I

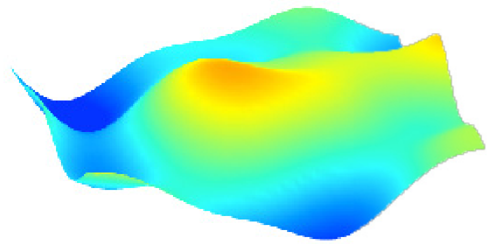
- *Review of parametric Bayesian models*
  - Finite mixture models
  - Beta and Dirichlet distributions
- *Canonical Bayesian nonparametric (BNP) model families*
  - Dirichlet & Pitman-Yor processes for infinite clustering
  - Beta processes for infinite feature induction
- *Key representations for BNP learning*
  - Infinite-dimensional stochastic processes
  - Stick-breaking constructions
  - Partitions and Chinese restaurant processes
  - Infinite limits of finite, parametric Bayesian models
- *Learning and inference algorithms*
  - Representation and truncation of infinite models
  - MCMC methods and Gibbs samplers
  - Variational methods and mean field

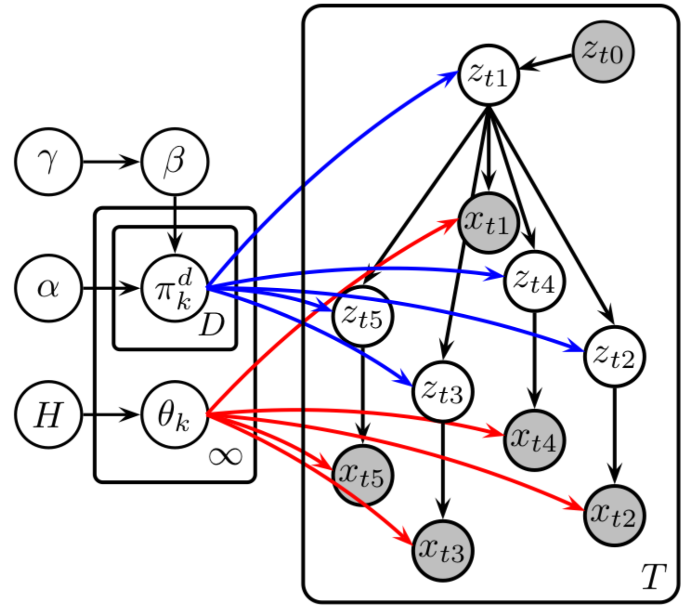
Coffee Break

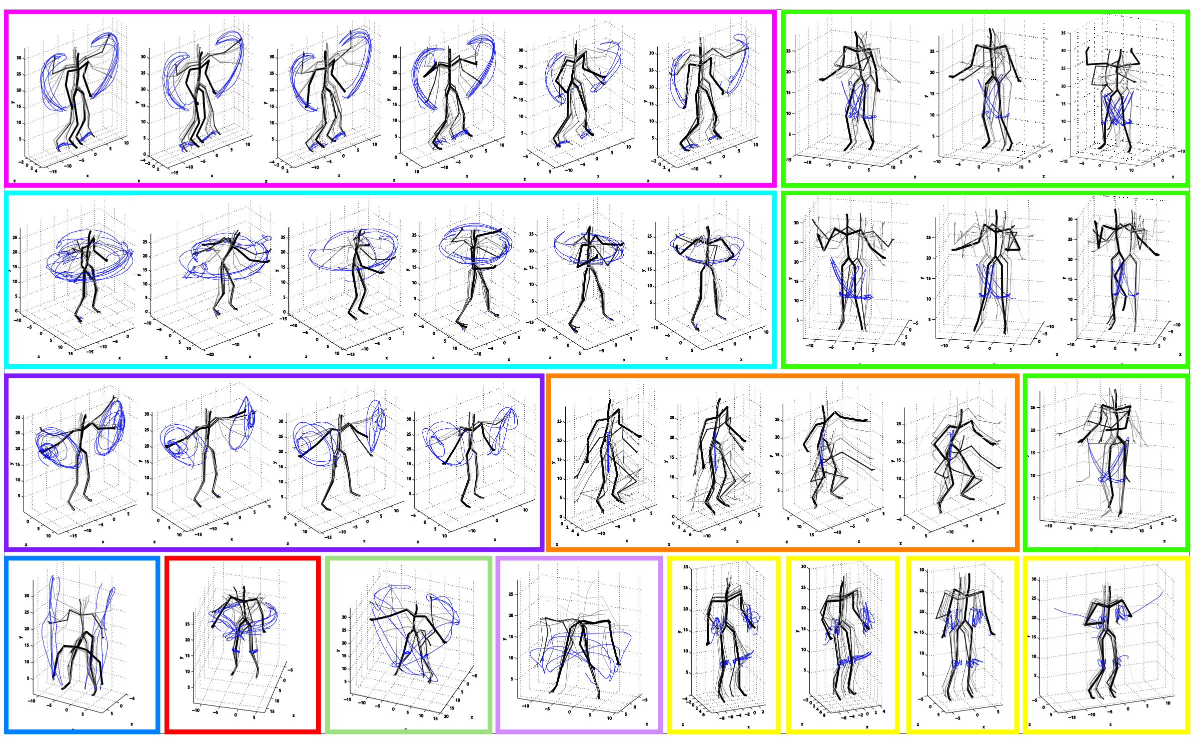


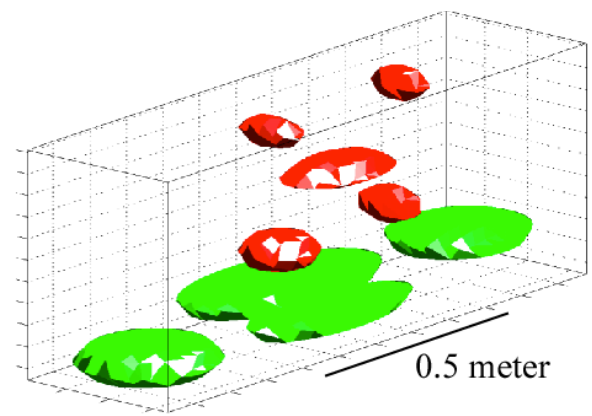
Applied BNP: Part II











Bayes Rule (Bayes Theorem)

$\theta \longrightarrow$ unknown parameters (many possible models)

$\mathcal{D} \longrightarrow$ observed data available for learning

$p(\theta) \longrightarrow$ prior distribution (domain knowledge)

$p(\mathcal{D} \mid \theta) \longrightarrow$ likelihood function (measurement model)

$p(\theta \mid \mathcal{D}) \longrightarrow$ posterior distribution (learned information)

$$p(\theta, \mathcal{D}) = p(\theta)p(\mathcal{D} \mid \theta) = p(\mathcal{D})p(\theta \mid \mathcal{D})$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\sum_{\theta' \in \Theta} p(\mathcal{D} \mid \theta')p(\theta')}$$

$$\propto p(\mathcal{D} \mid \theta)p(\theta)$$

Gaussian Mixture Models

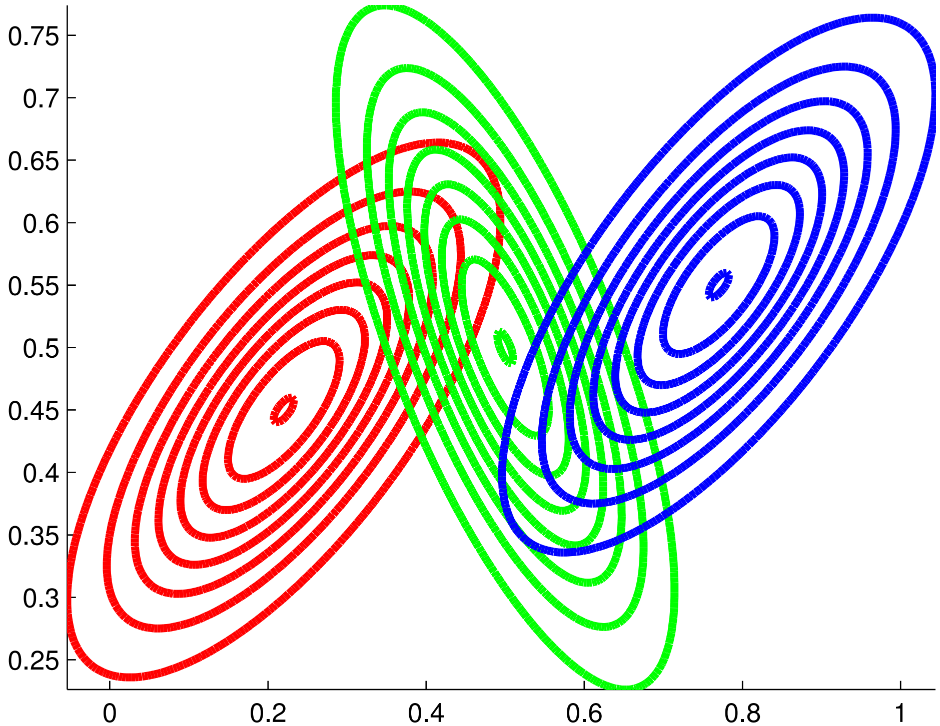
- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances: $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities: $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$
- Gaussian mixture marginal likelihood:

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

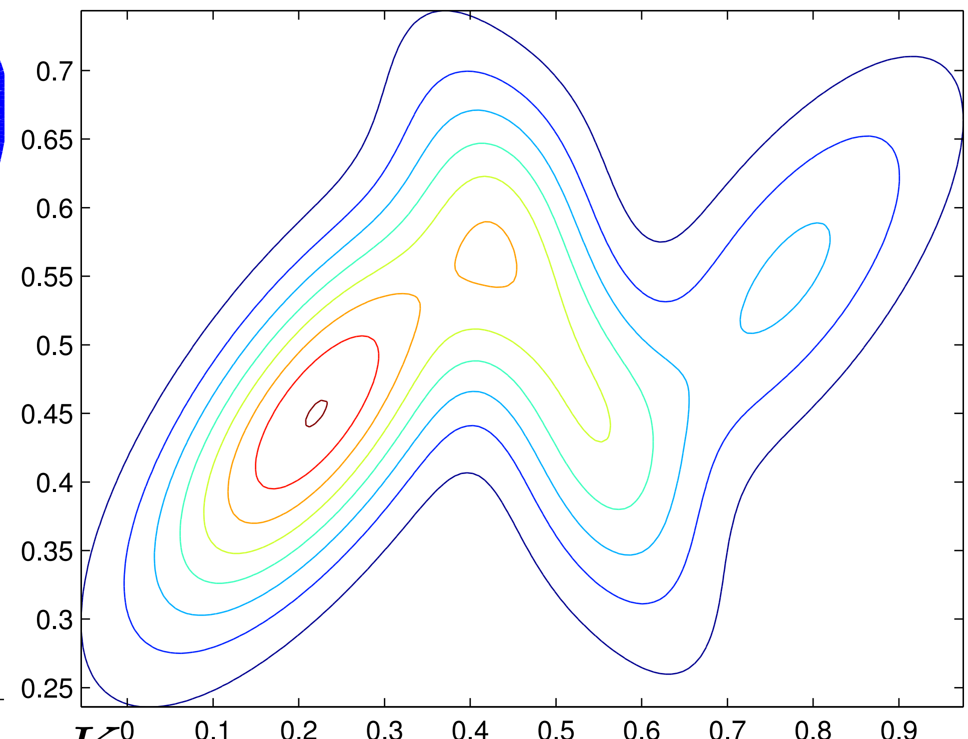
$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Gaussian Mixture Models

Mixture of 3 Gaussian Distributions in 2D



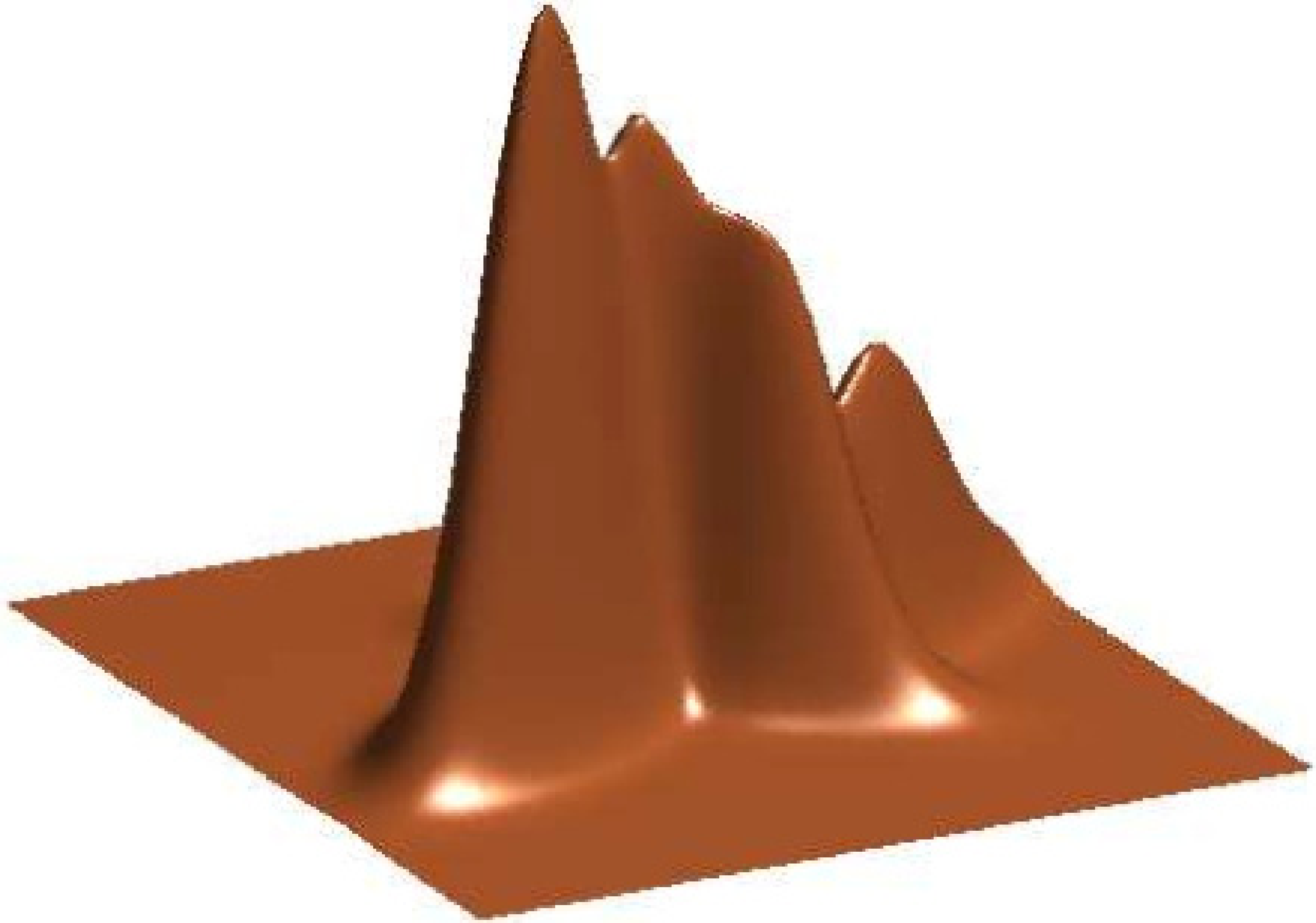
Contour Plot of Joint Density, Marginalizing Cluster Assignments



$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

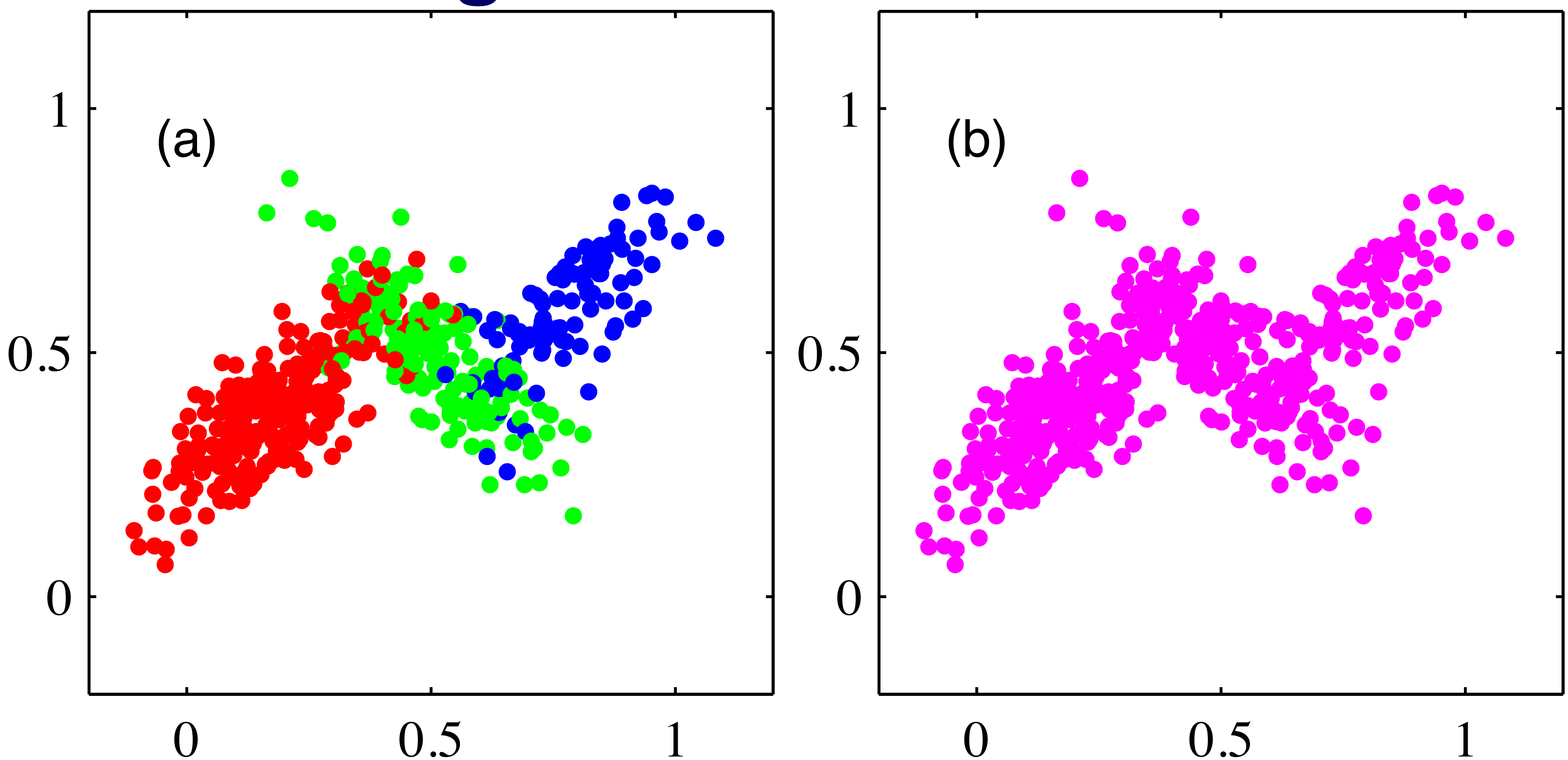
$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Gaussian Mixture Models



Surface Plot of Joint Density,
Marginalizing Cluster Assignments

Clustering with Gaussian Mixtures

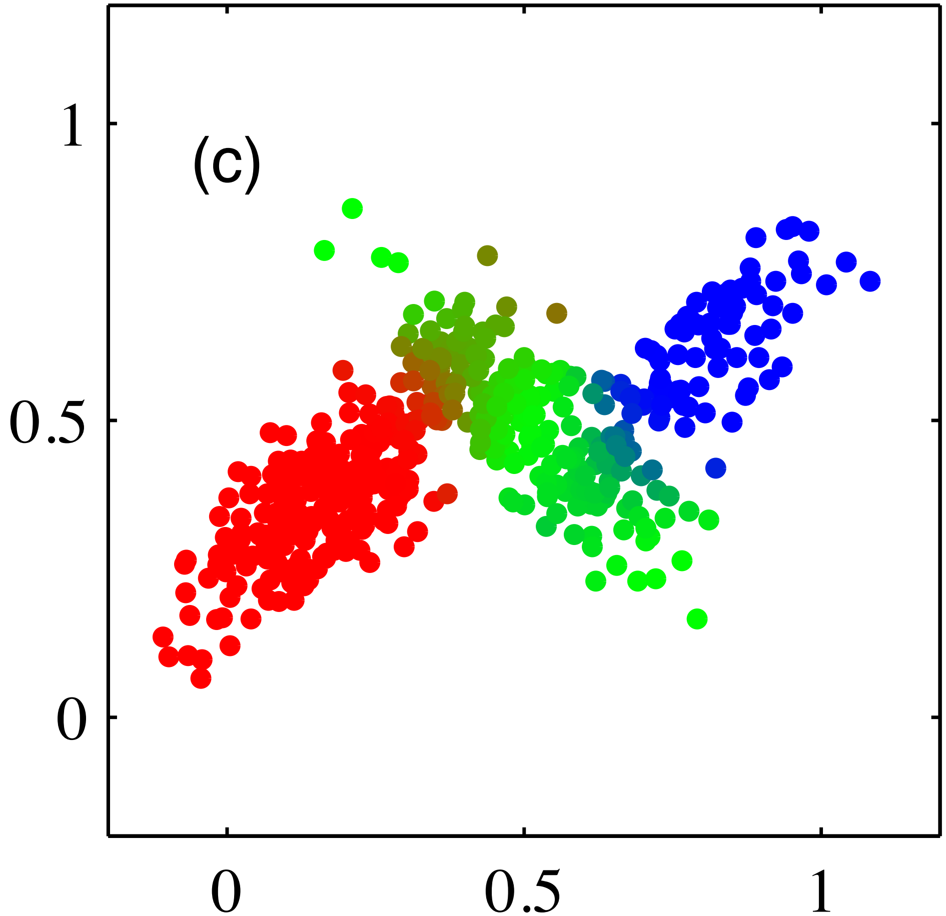


Complete Data Labeled by True Cluster Assignments

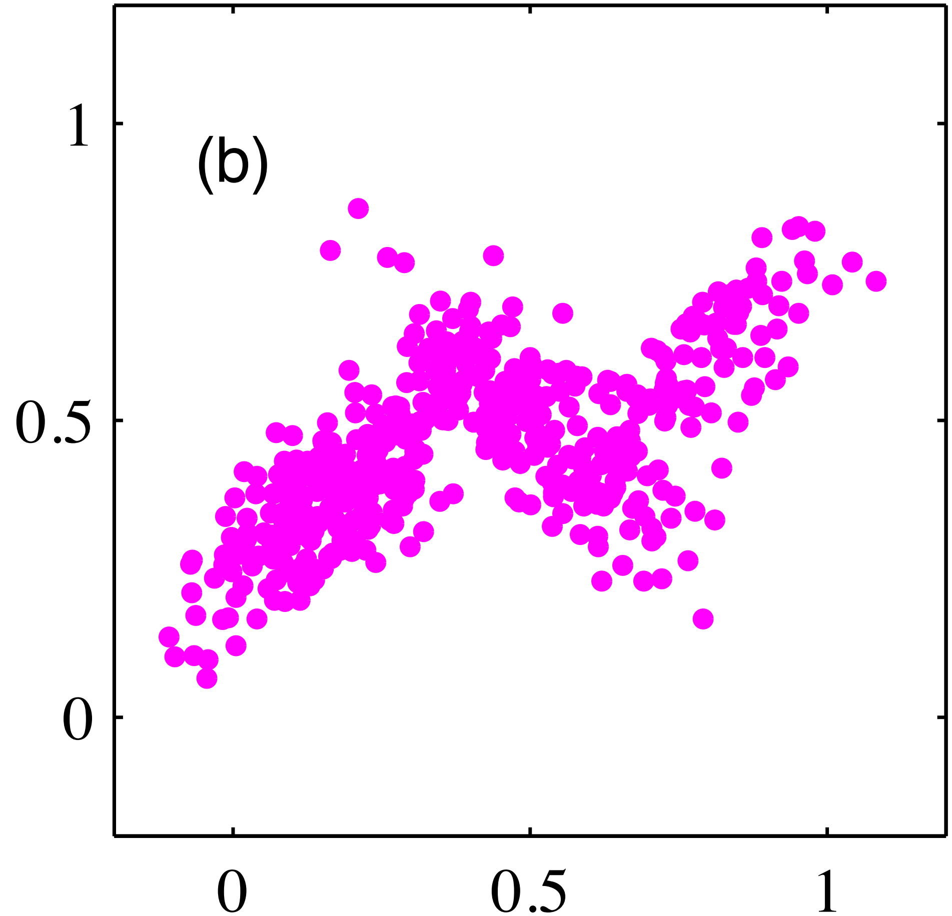
Incomplete Data: Points to be Clustered

C. Bishop, Pattern Recognition & Machine Learning

Inference Given Cluster Parameters



Posterior Probabilities of Assignment to Each Cluster



Incomplete Data: Points to be Clustered

$$r_{ik} = p(z_i = k \mid x_i, \pi, \theta) = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{\ell=1}^{K} \pi_\ell p(x_i \mid \theta_\ell)}$$

Learning Binary Probabilities

Bernoulli Distribution: Single toss of a (possibly biased) coin

$$\mathrm{Ber}(x \mid \theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \qquad 0 \leq \theta \leq 1$$

- Suppose we observe N samples from a Bernoulli distribution with unknown mean:

$$X_i \sim \mathrm{Ber}(\theta), i = 1, \ldots, N$$

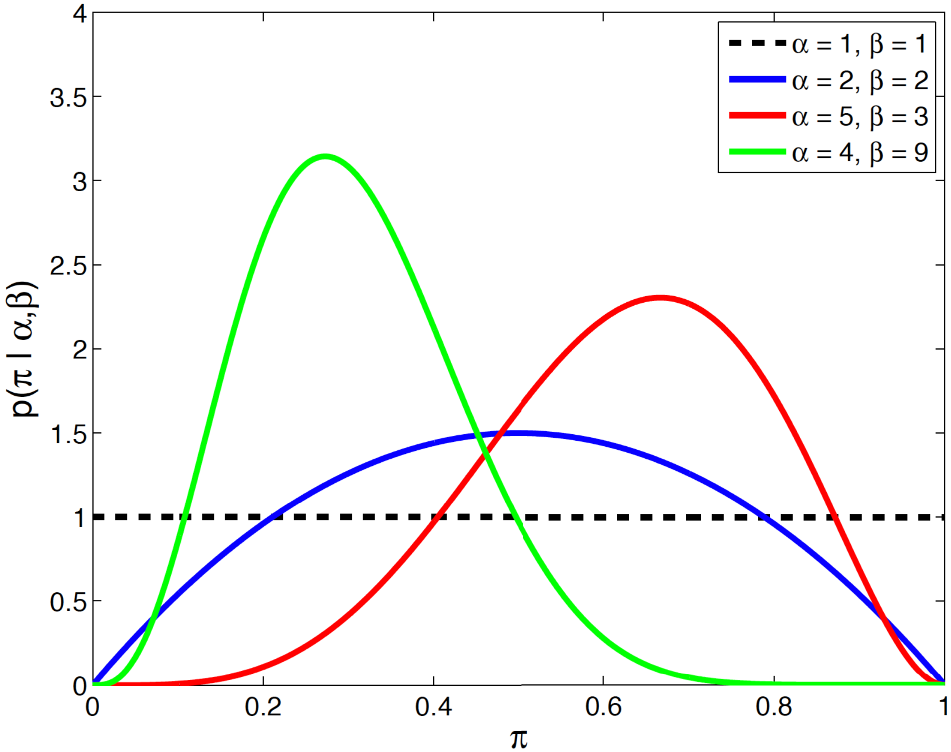
$$p(x_1, \ldots, x_N \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

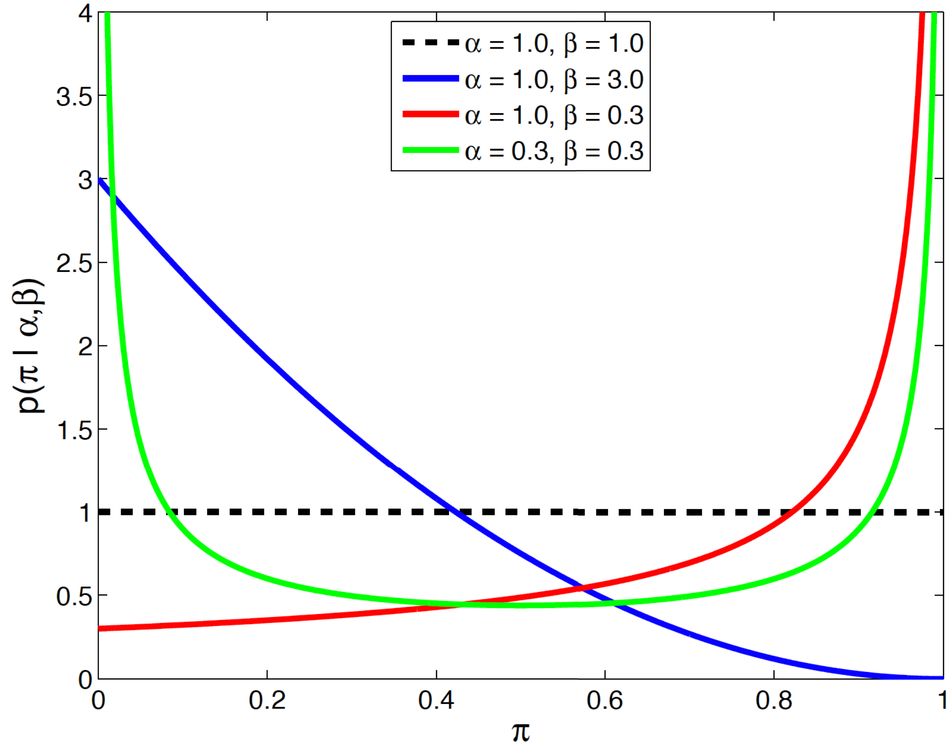
$$N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1) \qquad N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$$

- What is the *maximum likelihood* parameter estimate?

$$\hat{\theta} = \arg\max_{\theta} \log p(x \mid \theta) = \frac{N_1}{N}$$

Beta Distributions

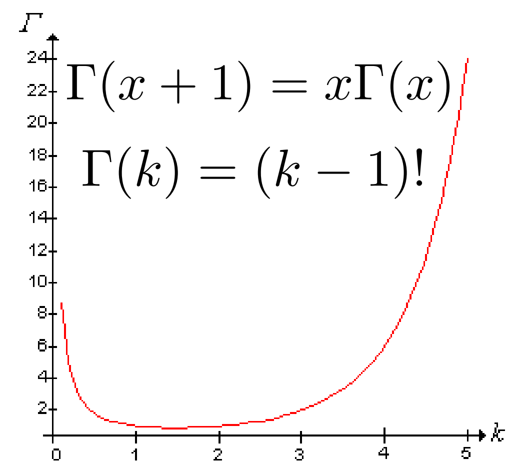




Probability density function: $x \in [0, 1]$

$$\mathrm{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$$

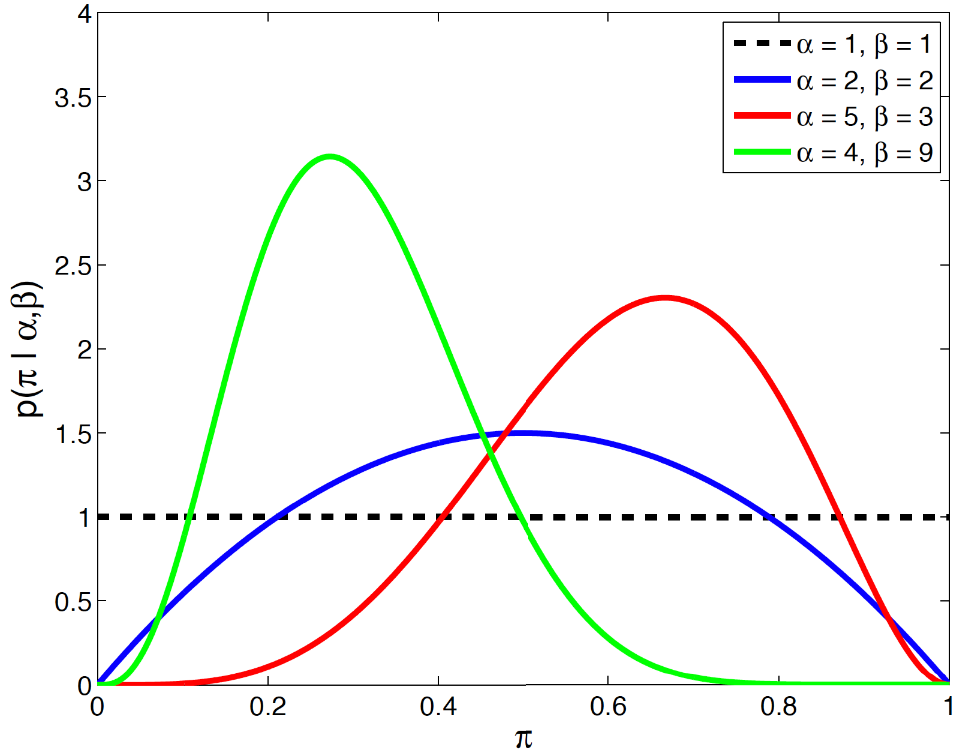
$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad a, b > 0$$

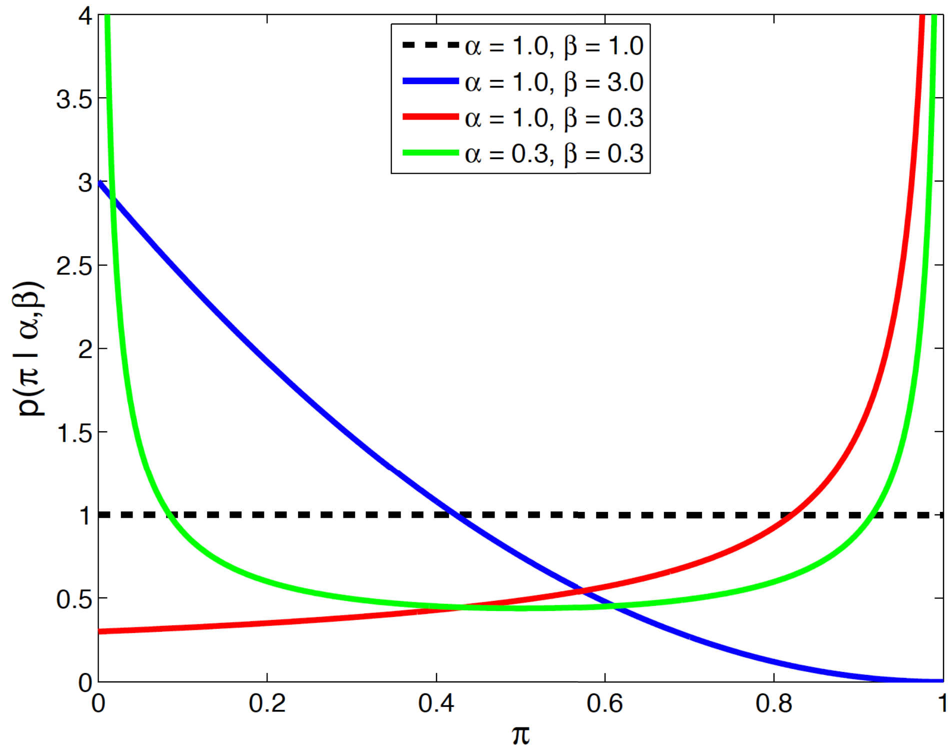


$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(k) = (k-1)!$$

Beta Distributions





$$\mathbb{E}[x] = \frac{a}{a+b} \qquad \mathbb{V}[x] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\mathrm{Mode}[x] = \arg\max_{x\in[0,1]} \mathrm{Beta}(x \mid a, b) = \frac{a-1}{(a-1)+(b-1)}$$

Bayesian Learning of Probabilities

Bernoulli Likelihood: Single toss of a (possibly biased) coin

$$\mathrm{Ber}(x \mid \theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \qquad 0 \le \theta \le 1$$

$$p(x_1, \ldots, x_N \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Beta Prior Distribution:

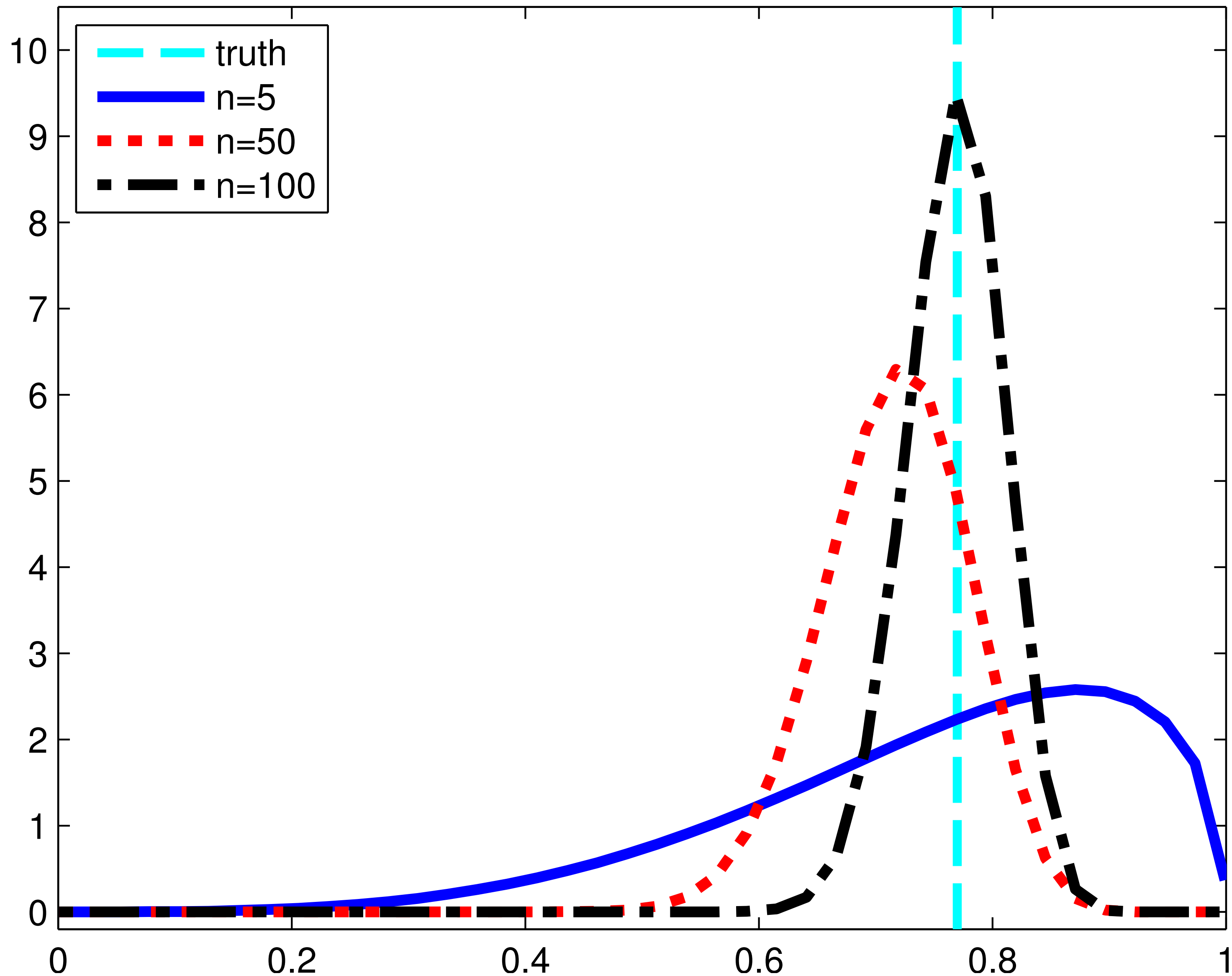
$$p(\theta) = \mathrm{Beta}(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Posterior Distribution:

$$p(\theta \mid x) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \propto \mathrm{Beta}(\theta \mid N_1+a, N_0+b)$$

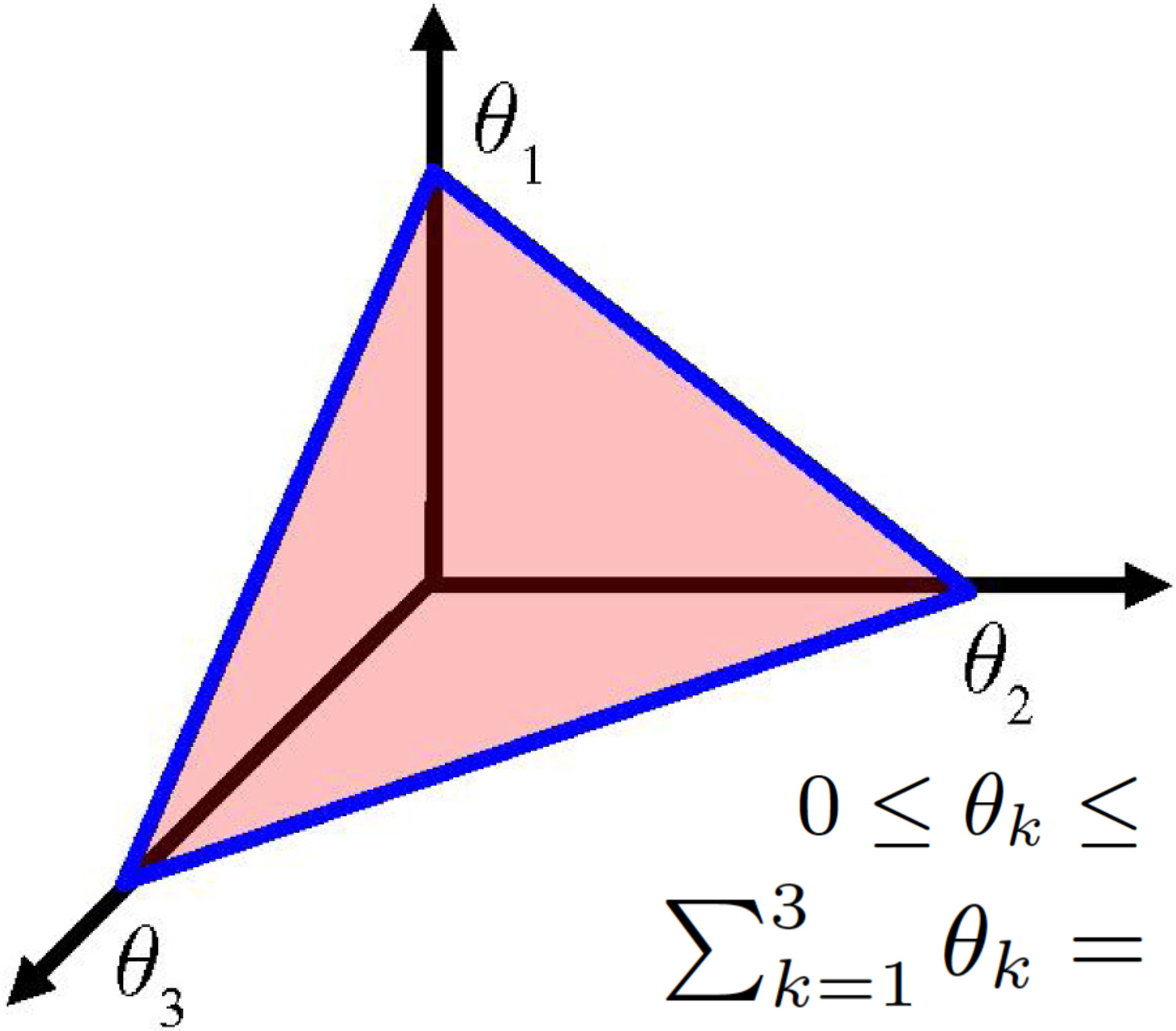
- *This is a conjugate prior, because posterior is in same family*
- *Estimate by posterior mode (MAP) or mean (preferred)*

Sequence of Beta Posteriors



Murphy 2012

Multinomial Simplex



$$0 \leq \theta_k \leq 1$$

$$\sum_{k=1}^{3} \theta_k = 1$$

Learning Categorical Probabilities

Categorical Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad \mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$

- If we have N_k observations of outcome *k* in *N* trials:

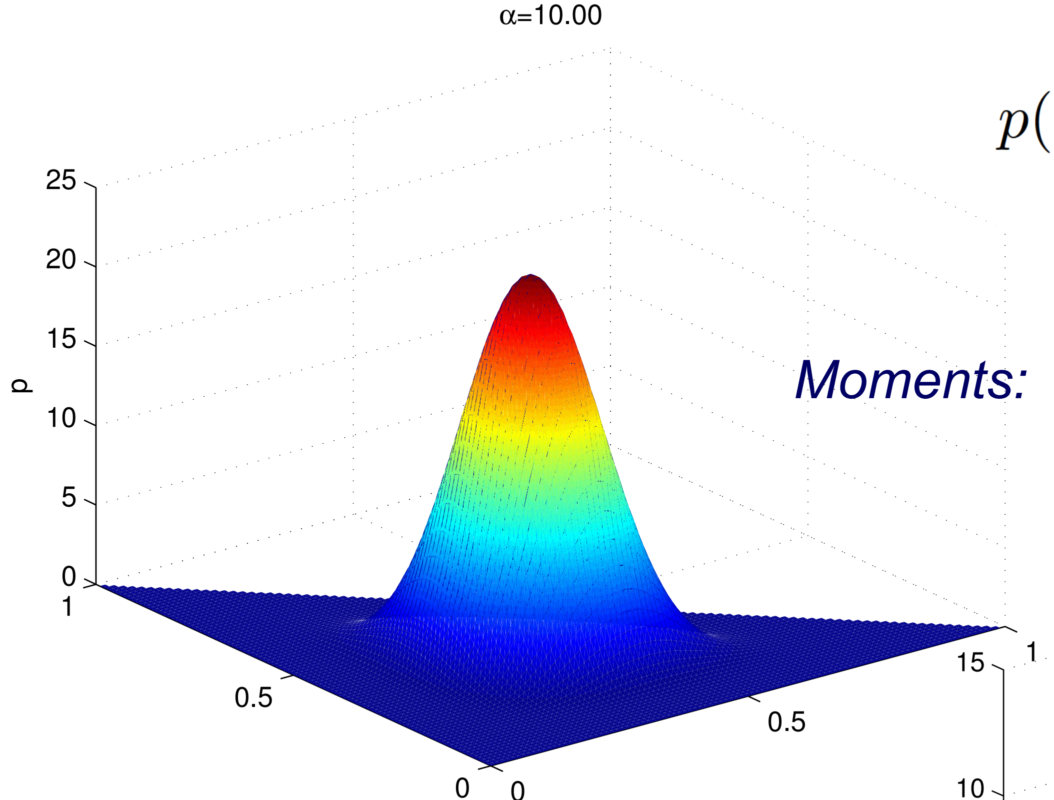
$$p(x_1, \ldots, x_N \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

- The *maximum likelihood* parameter estimates are then:

$$\hat{\theta} = \arg\max_{\theta} \log p(x \mid \theta) \qquad \hat{\theta}_k = \frac{N_k}{N}$$

- Will this produce sensible predictions when *K* is large?
  For nonparametric models we let *K* approach infinity…

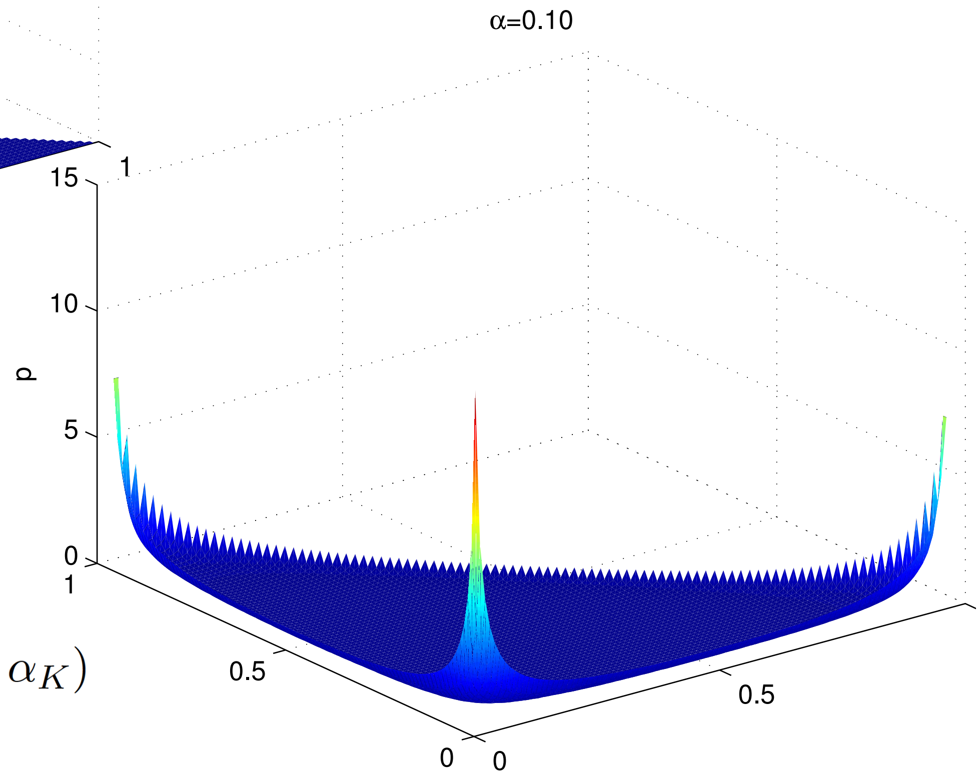
Dirichlet Distributions



$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$\alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$$

Moments: $\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$ $\quad \mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0+1)}$

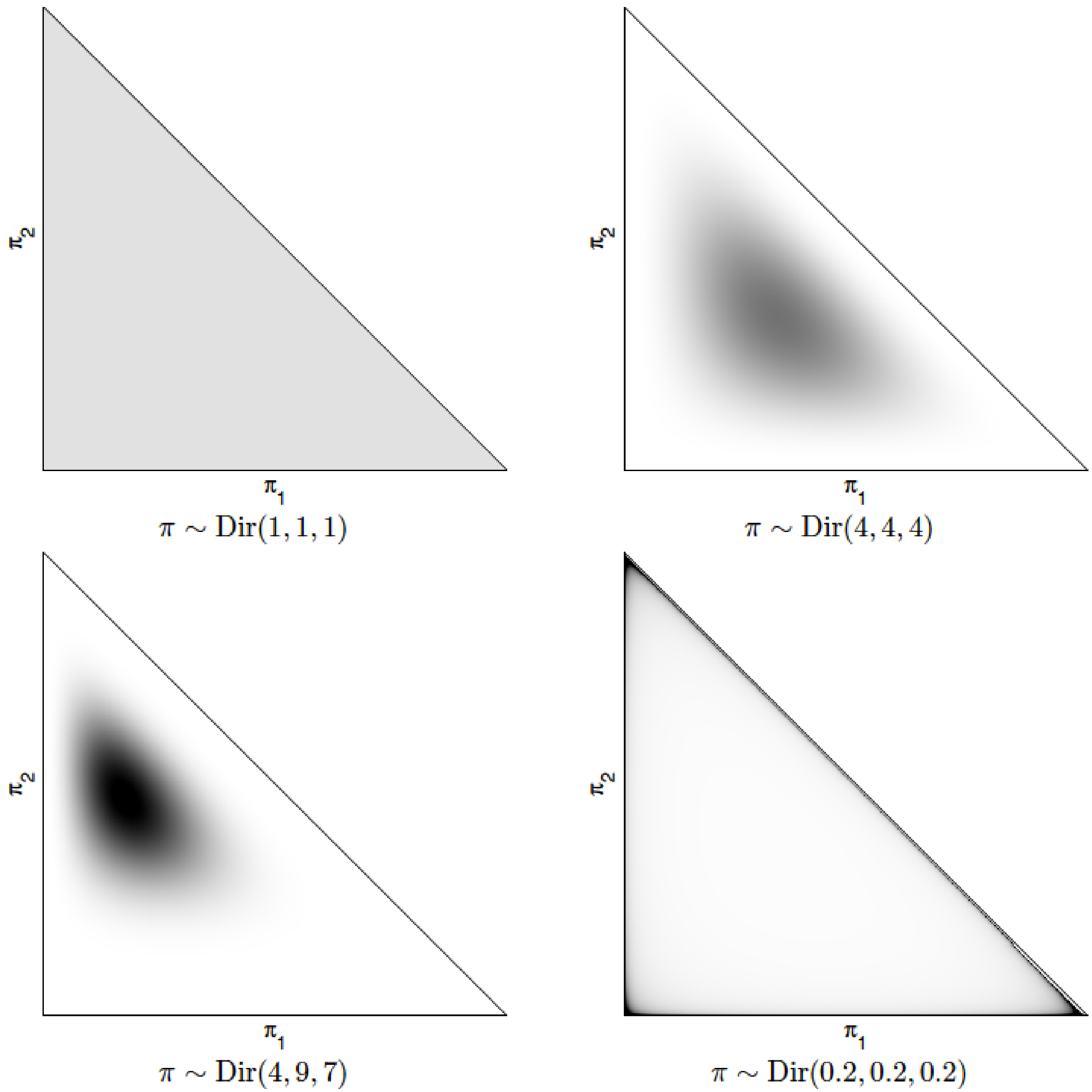


Marginal Distributions:

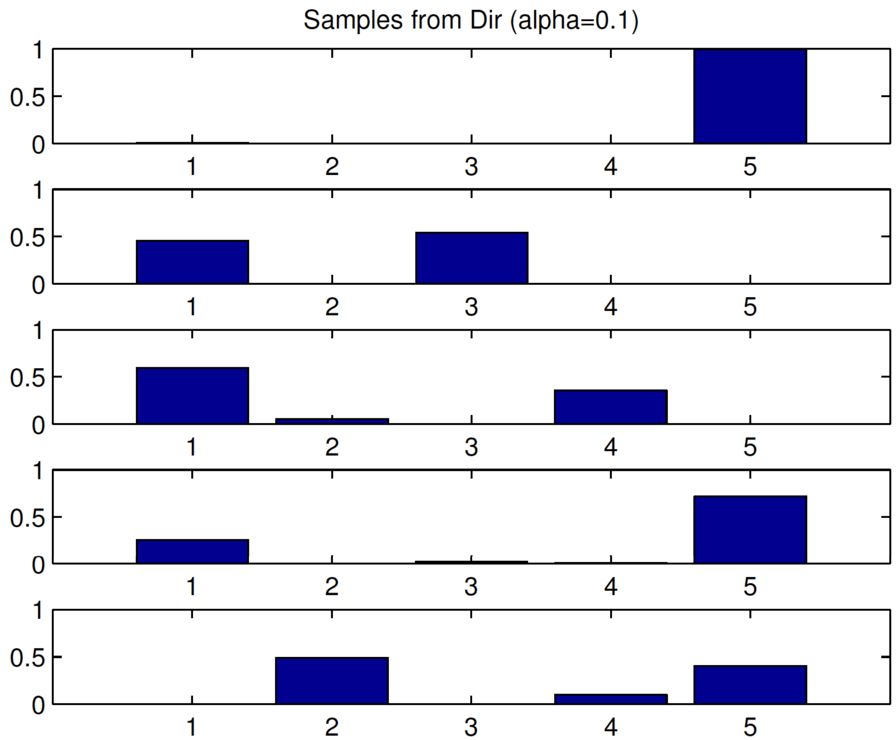
$$\pi_k \sim \mathrm{Beta}(\alpha_k, \alpha_0 - \alpha_k)$$

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

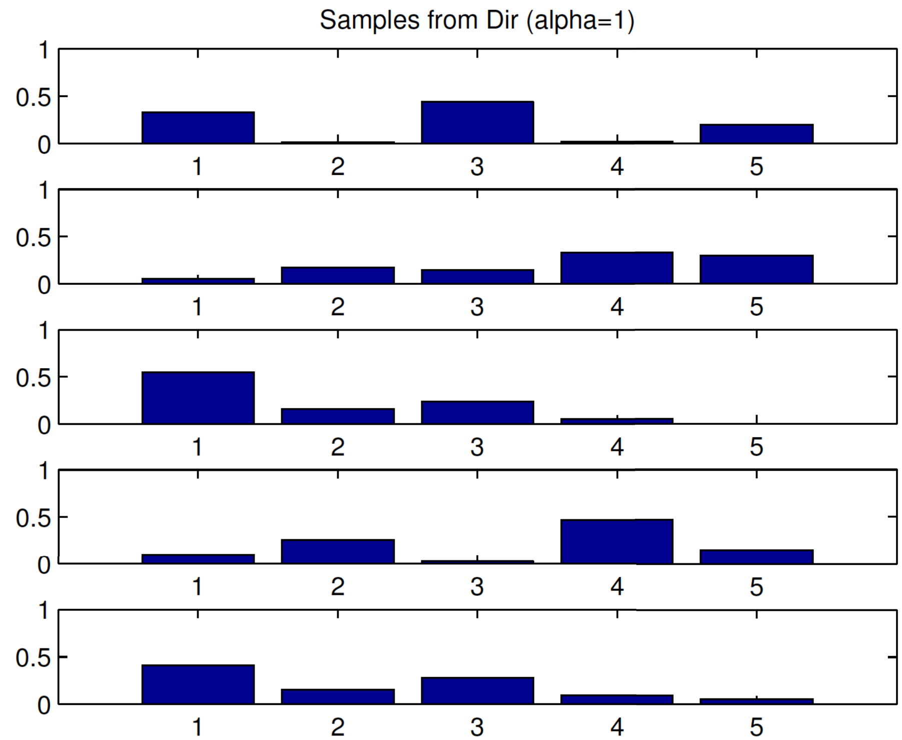
Dirichlet Probability Densities



Dirichlet Samples



$$\mathrm{Dir}(\theta \mid 0.1, 0.1, 0.1, 0.1, 0.1)$$



$$\mathrm{Dir}(\theta \mid 1.0, 1.0, 1.0, 1.0, 1.0)$$

Bayesian Learning of Probabilities

Categorical Distribution: Single roll of a (possibly biased) die

$$\mathrm{Cat}(x \mid \theta) = \prod_{k=1}^{K} \theta_k^{x_k} \qquad \mathcal{X} = \{0,1\}^K, \sum_{k=1}^{K} x_k = 1$$

$$p(x_1, \ldots, x_N \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

Dirichlet Prior Distribution:

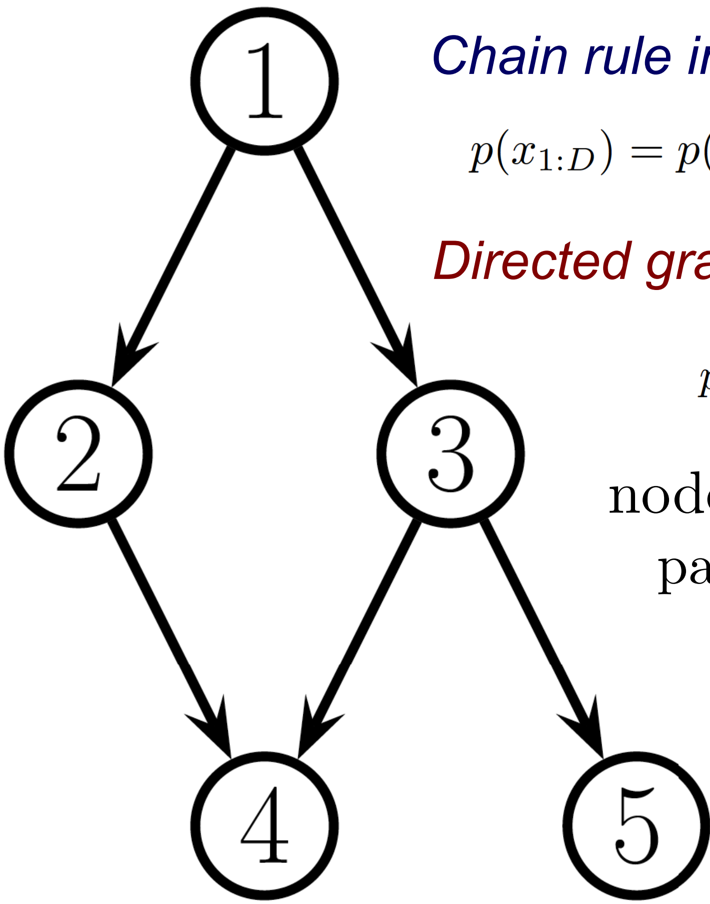
$$p(\theta) = \mathrm{Dir}(\theta \mid \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

Posterior Distribution:

$$p(\theta \mid x) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1} \propto \mathrm{Dir}(\theta \mid N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

- *This is a conjugate prior, because posterior is in same family*

Directed Graphical Models



Chain rule implies that any joint distribution equals:

$$p(x_{1:D}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_1,x_2,x_3)\dots p(x_D|x_{1:D-1})$$

Directed graphical model implies a restricted factorization:

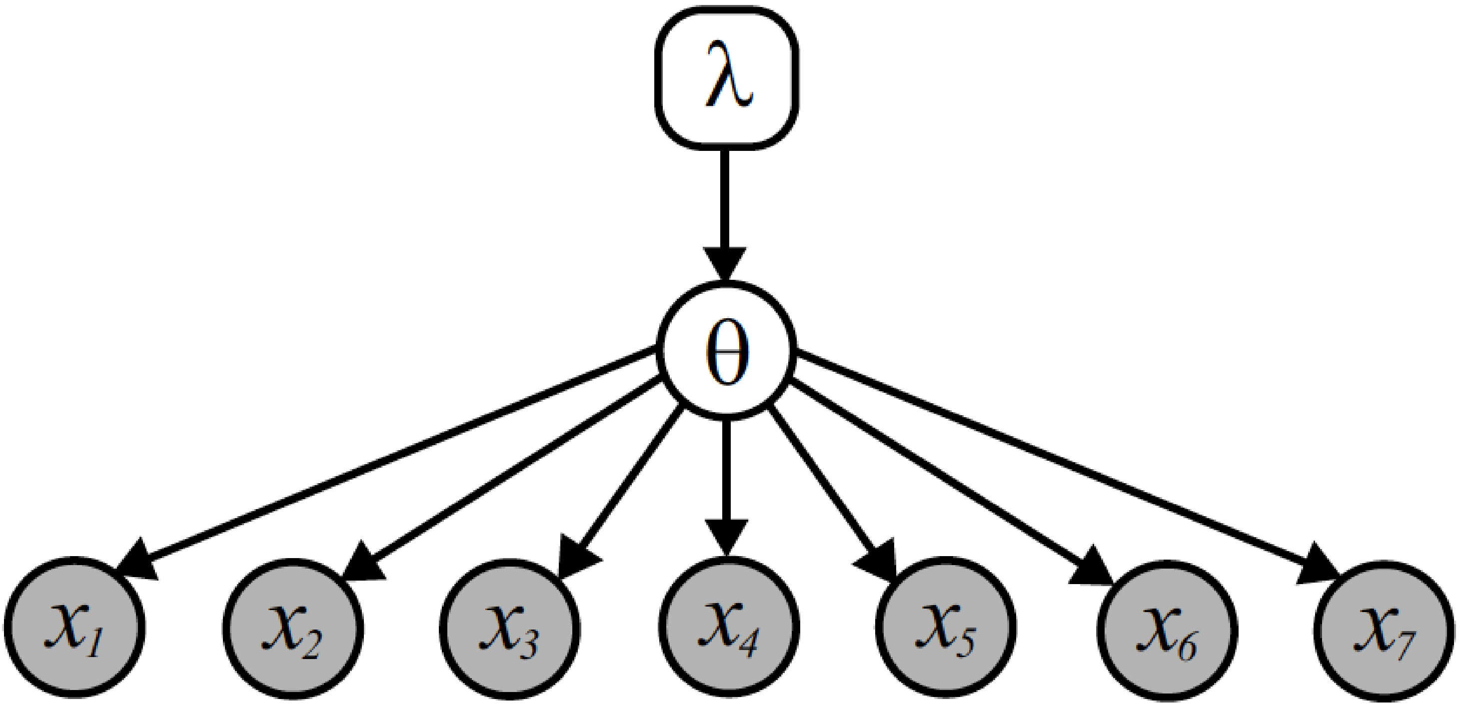
$$p(\mathbf{x}_{1:D}|G) = \prod_{t=1}^{D} p(x_t|\mathbf{x}_{\mathrm{pa}(t)})$$

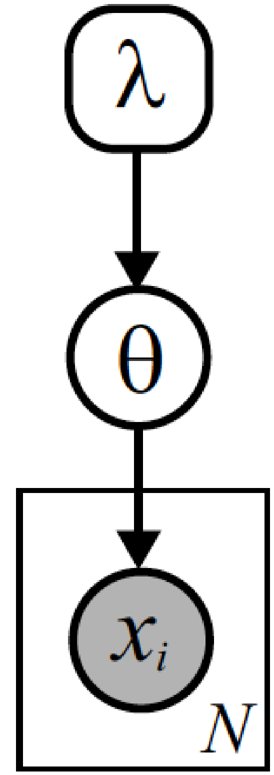
$$\begin{aligned} \text{nodes} &\rightarrow \text{random variables} \\ \mathrm{pa}(t) &\rightarrow \text{parents with edges pointing to node } t \end{aligned}$$

Valid for any directed acyclic graph (DAG): equivalent to dropping conditional dependencies in standard chain rule

$$\begin{aligned} p(\mathbf{x}_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1,\cancel{x_2})p(x_4|\cancel{x_1},x_2,x_3)p(x_5|\cancel{x_1},\cancel{x_2},x_3,\cancel{x_4}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2,x_3)p(x_5|x_3) \end{aligned}$$

Plates: Learning with Priors

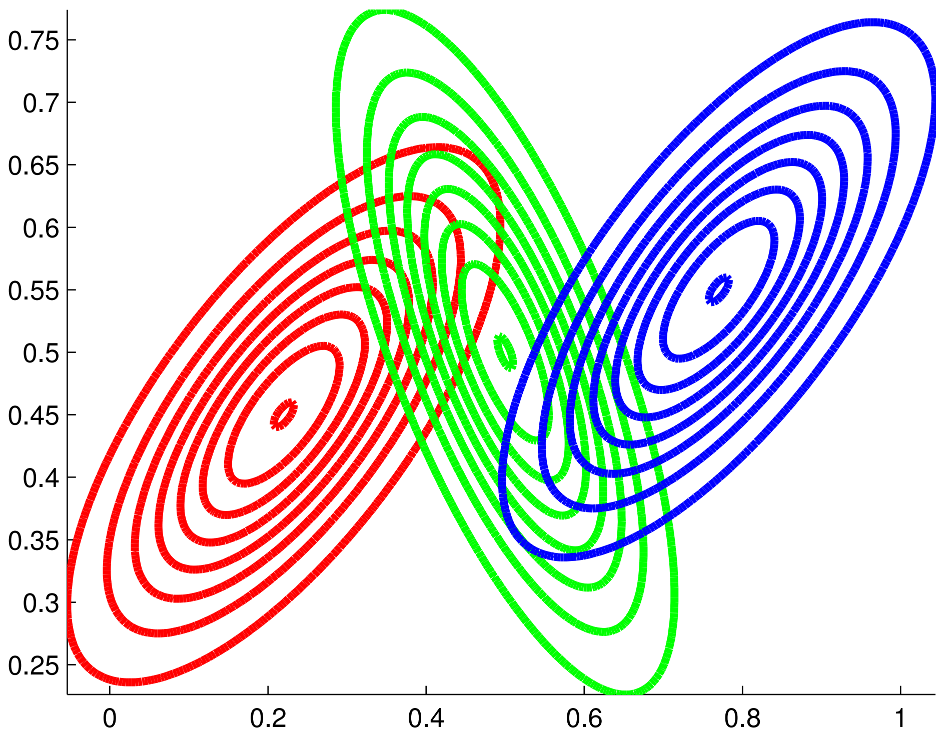


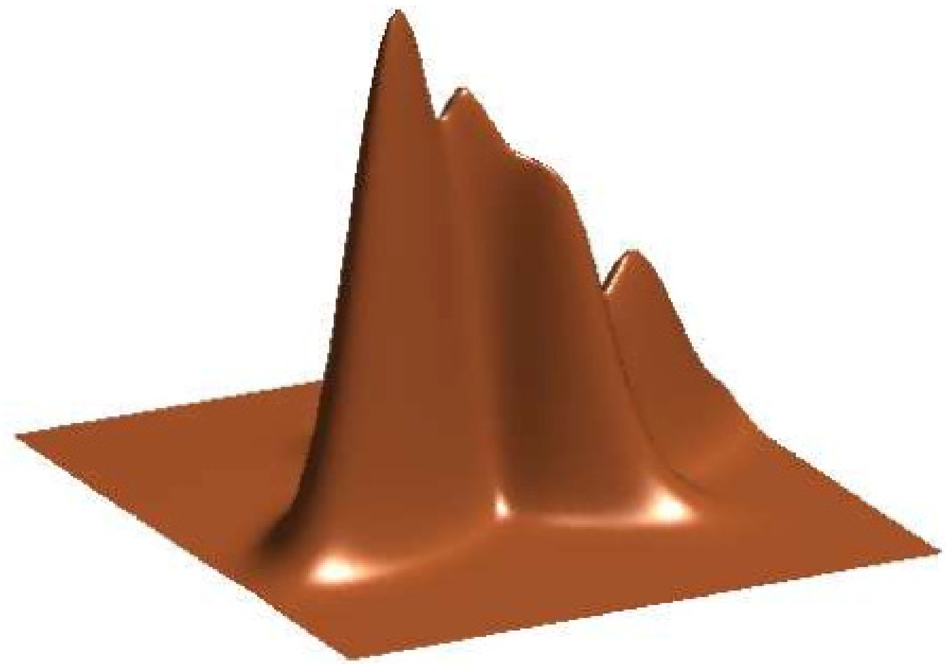


$$p(x_1, \ldots, x_N, \theta \mid \lambda) = p(\theta \mid \lambda) \prod_{i=1}^{N} p(x_i \mid \theta)$$

- Boxes, or *plates*, indicate replication of variables
- Variables which are observed, or fixed, are often *shaded*
- Prior distributions may themselves have *hyperparameters* λ

Gaussian Mixture Models



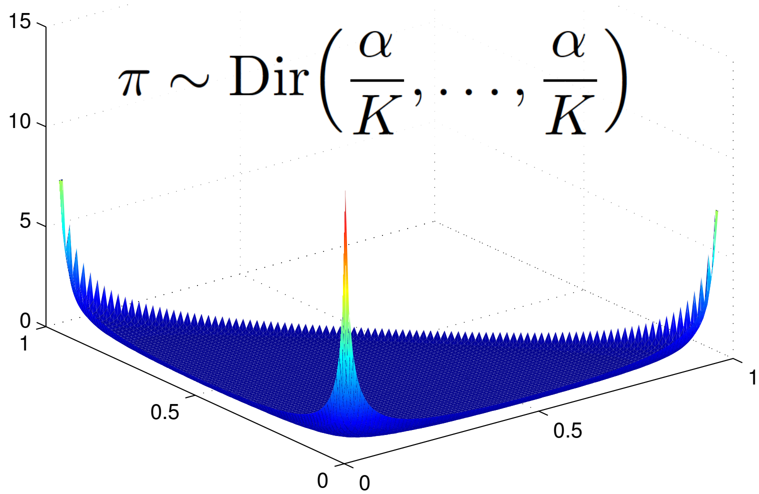


$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

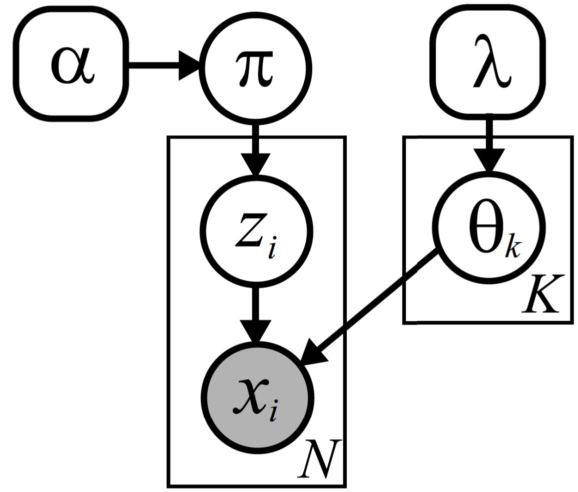
$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Finite Bayesian Mixture Models

- Cluster frequencies: Symmetric Dirichlet



$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$



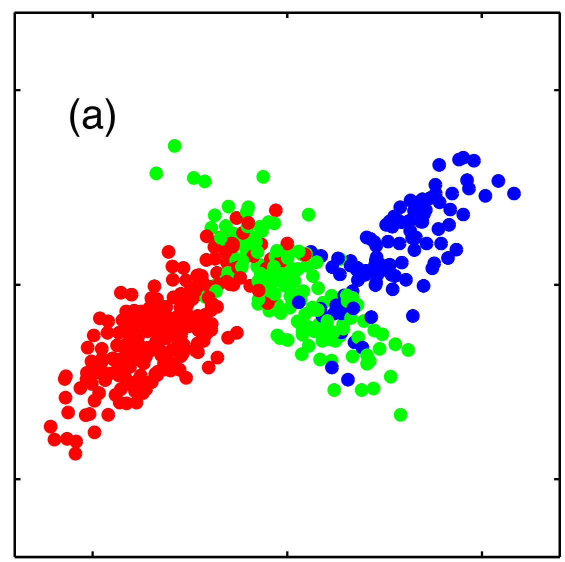
$$p(x \mid \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x \mid \theta_k)$$

- Cluster shapes: Any valid prior on chosen family (e.g., Gaussian mean & covariance)

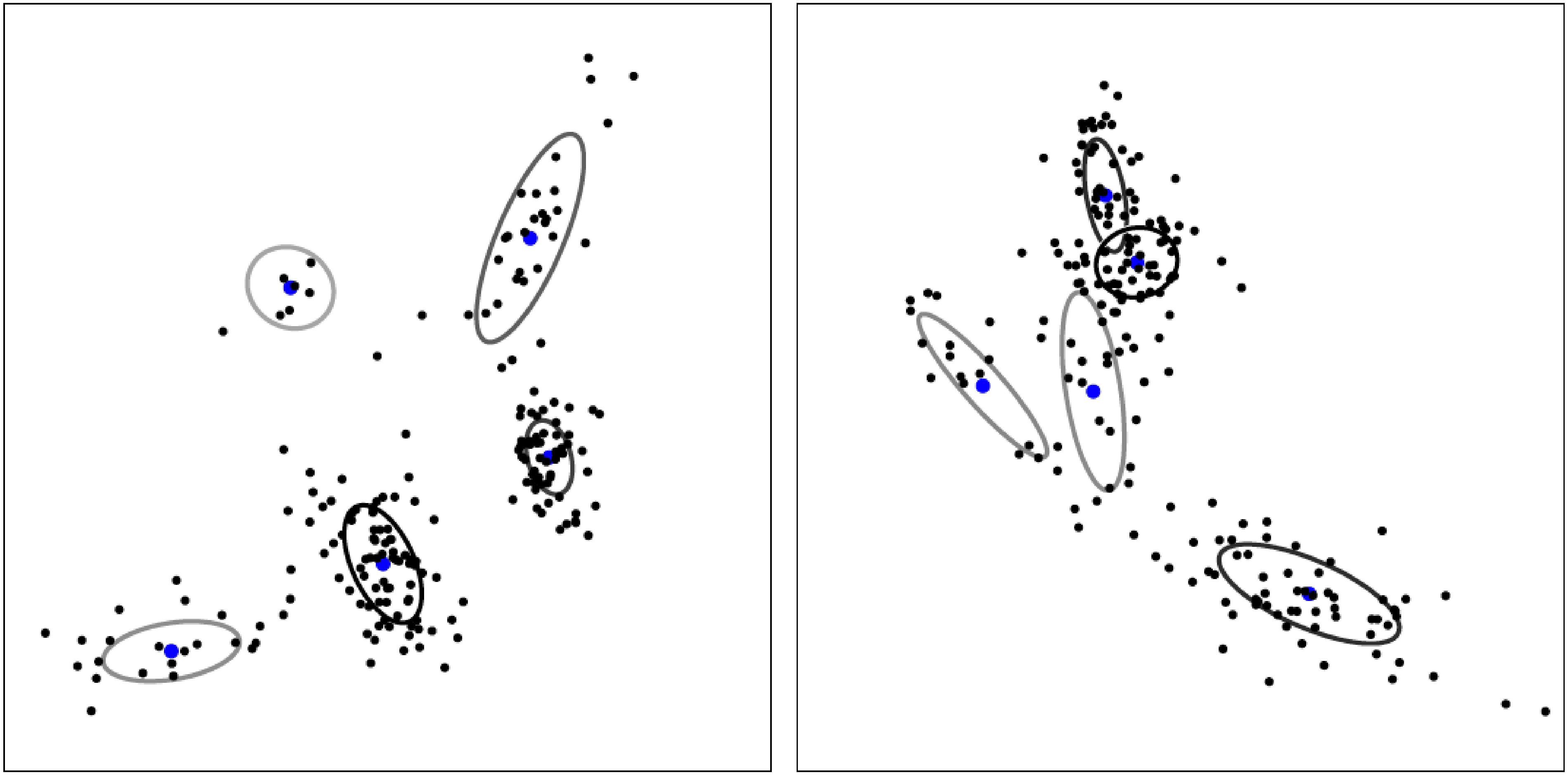
$$\theta_k \sim H(\lambda) \qquad\qquad k = 1, \ldots, K$$

- Data: Assign each data item to a cluster, and sample from that cluster's likelihood

$$z_i \sim \mathrm{Cat}(\pi)$$
$$x_i \sim F(\theta_{z_i})$$



Generative Gaussian Mixture Samples



Learning is simplest with *conjugate* priors on cluster shapes:

- Gaussian with known variance: Gaussian prior on mean
- Gaussian with unknown mean & variance: *normal inverse-Wishart*

Mixtures as Discrete Measures

$$p(x \mid \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x \mid \theta_k)$$

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

$$\theta_k \sim H(\lambda)$$

- Define mixture via a *discrete probability measure* on cluster parameters:

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

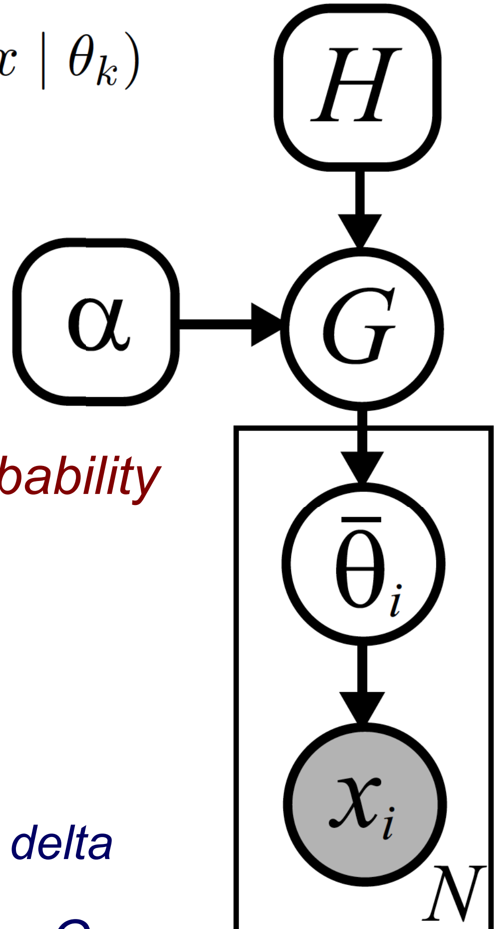
$\delta_{\theta_k} \longrightarrow$ *atom, point mass, Dirac delta*

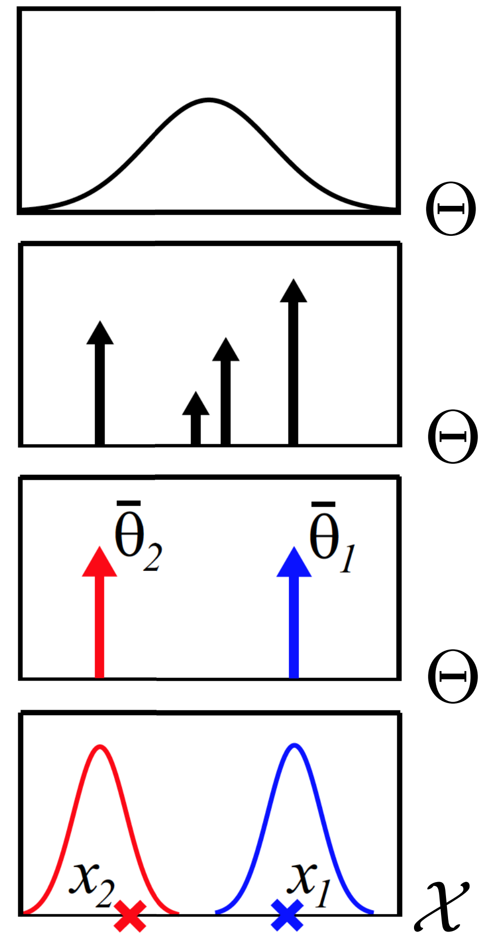
- Generate data via repeated draws *G*:

$$\bar{\theta}_i \sim G$$

$$x_i \sim F(\bar{\theta}_i)$$

$$\bar{\theta}_i = \theta_{z_i}$$





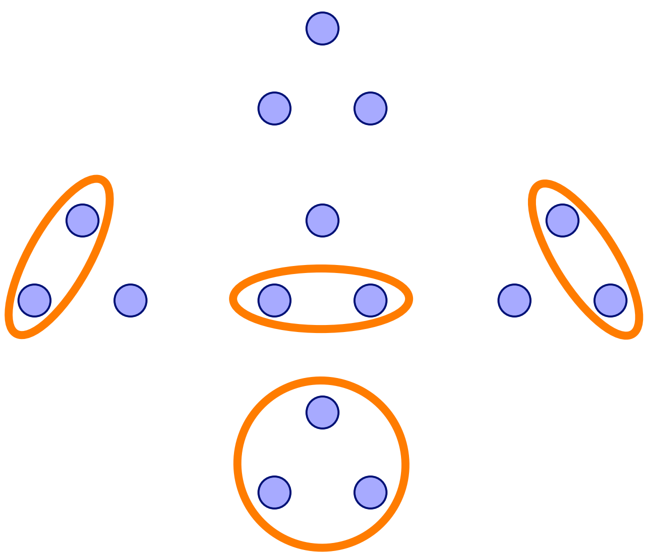
Toy visualization: 1D Gaussian mixture with unknown cluster means and fixed variance

Mixtures Induce Partitions

- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

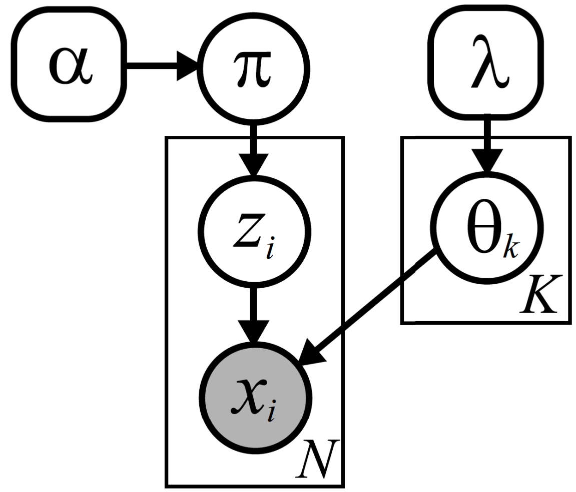
$$z_i \sim \text{Cat}(\pi)$$

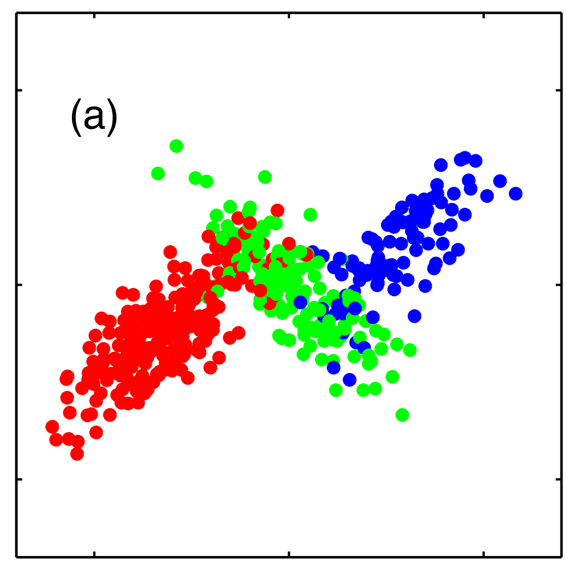
- The number of ways of assigning *N* data points to *K* mixture components is K^N
- If $K \geq N$ this is much larger than the number of ways of partitioning that data:



N=3: 5 partitions versus $3^3 = 27$

$$\pi \sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$





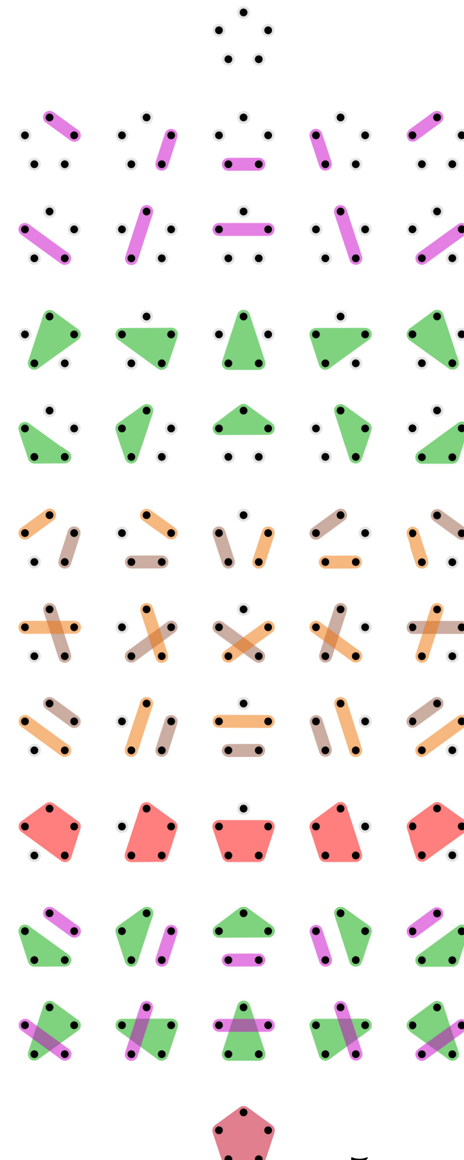
Mixtures Induce Partitions

- If our goal is clustering, the output grouping is defined by assignment *indicator variables:*

$$z_i \sim \mathrm{Cat}(\pi)$$

- The number of ways of assigning *N* data points to *K* mixture components is K^N
- If $K \geq N$ this is much larger than the number of ways of partitioning that data:

For any clustering, there is a unique partition, but many ways to label that partition's blocks.



Courtesy Wikipedia

N=5: 52 partitions versus $5^5 = 3125$

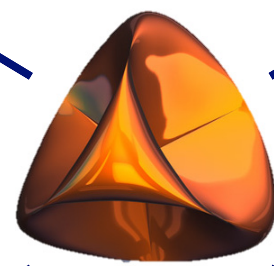
Dirichlet Process Mixtures

The Dirichlet Process (DP)

*A distribution on countably infinite discrete probability measures. Sampling yields a **Polya urn**.*

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Stick-Breaking

An explicit construction for the weights in DP realizations

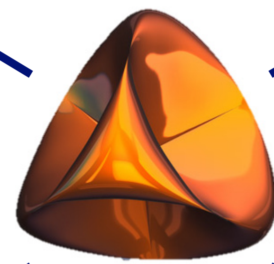
Infinite Mixture Models

As an infinite limit of finite mixtures with Dirichlet weight priors

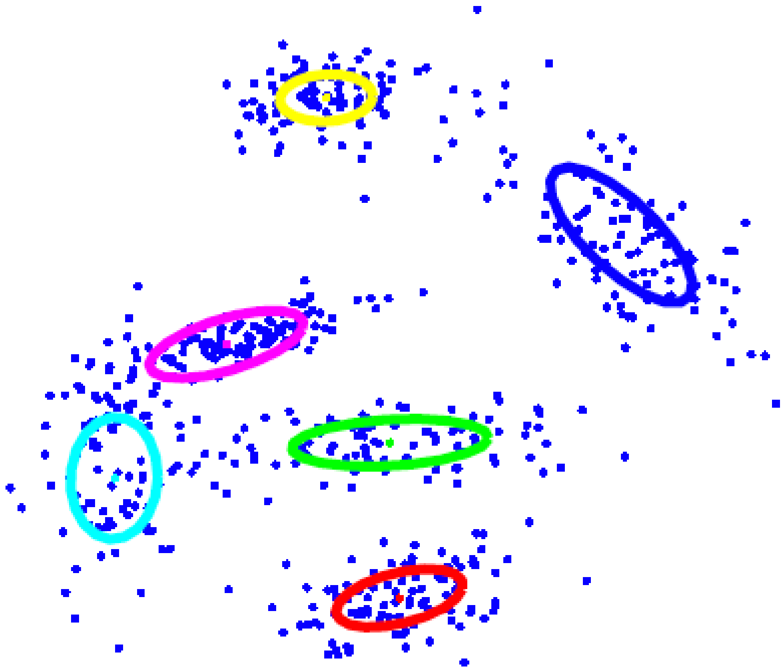
Dirichlet Process Mixtures

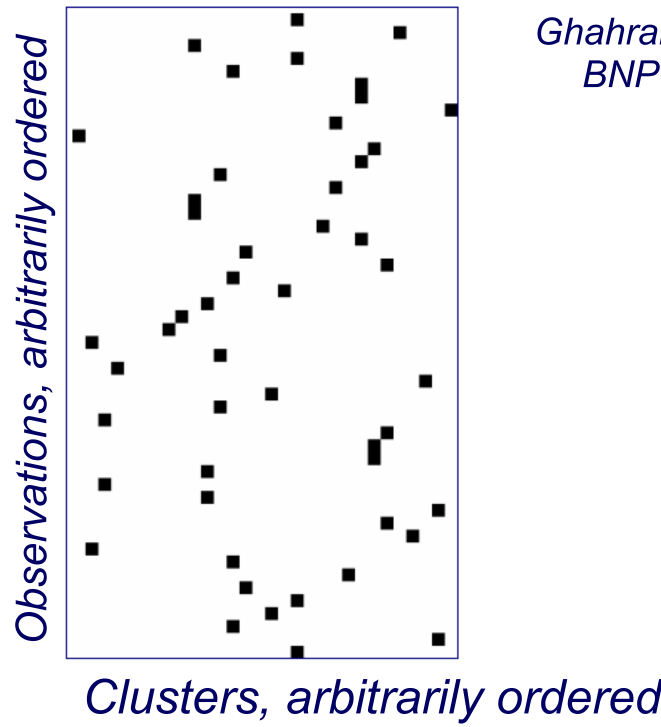
Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Nonparametric Clustering





Ghahramani, BNP 2009

- *Large Support:* All partitions of the data, from one giant cluster to N singletons, have positive probability under prior
- *Exchangeable:* Partition probabilities are invariant to permutations of the data
- *Desirable:* Good asymptotics, computational tractability, flexibility and ease of generalization…

Chinese Restaurant Process (CRP)

- 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*

- The first customer sits at a table. Subsequent customers randomly select a table according to:

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

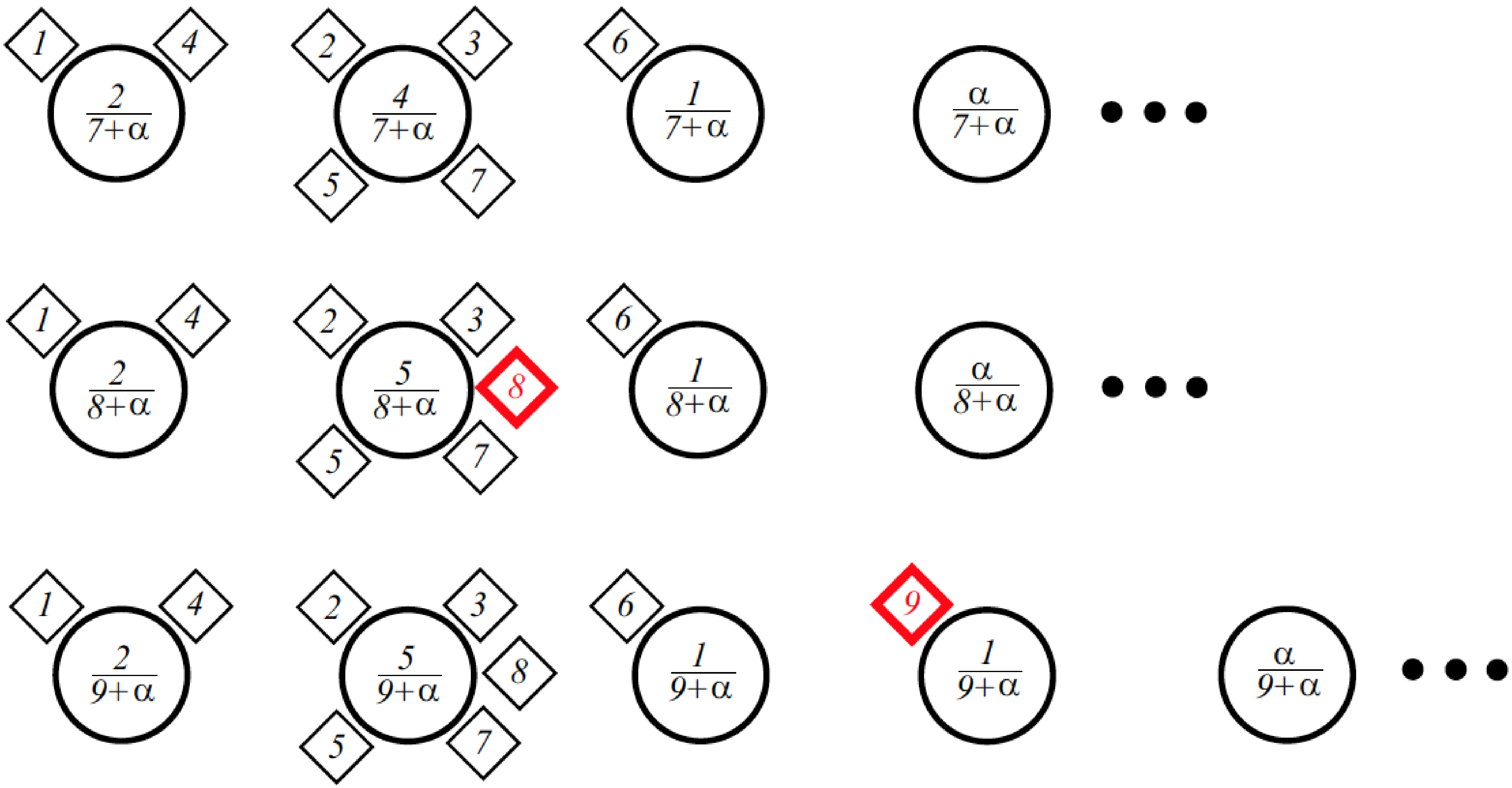
$K \longrightarrow$ number of tables occupied by the first *N* customers

$N_k \longrightarrow$ number of customers seated at table *k*

$\bar{k} \longrightarrow$ a new, previously unoccupied table

$\alpha \longrightarrow$ positive concentration parameter

Chinese Restaurant Process (CRP)



$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

CRPs & Exchangeable Partitions

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

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

$$p(z_1, \ldots, z_N \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k)$$

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

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

- The CRP is thus a prior on *infinitely exchangeable* partitions

De Finetti's Theorem

- Finitely exchangeable random variables satisfy:

$$p(x_1, \ldots, x_N) = p\big(x_{\tau(1)}, \ldots, x_{\tau(N)}\big) \qquad \text{for any permutation } \tau(\cdot)$$

- A sequence is infinitely exchangeable if every finite subsequence is exchangeable
- Exchangeable variables need not be independent, but always have a representation with conditional independencies:

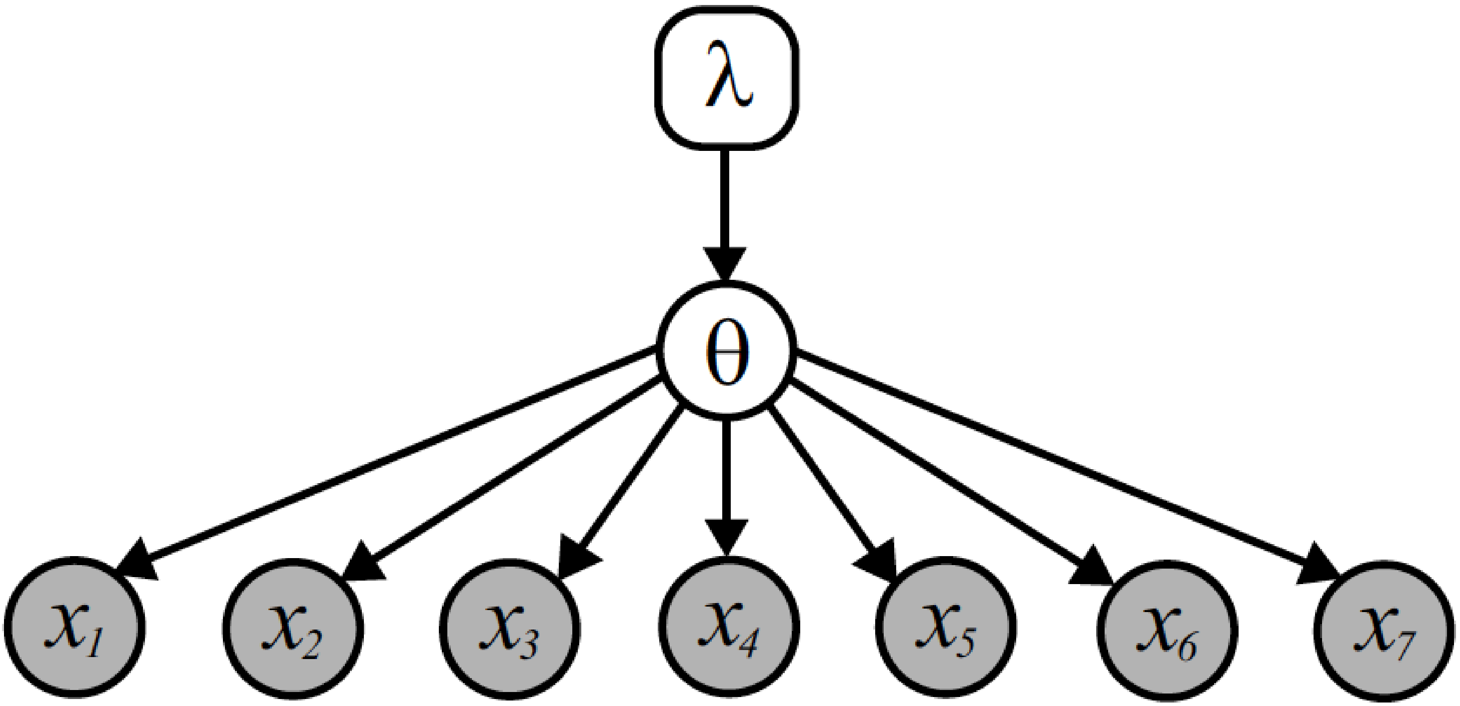
Theorem 2.2.2 (De Finetti). *For any infinitely exchangeable sequence of random variables* $\{x_i\}_{i=1}^{\infty}$, $x_i \in \mathcal{X}$, *there exists some space* Θ, *and corresponding density* $p(\theta)$, *such that the joint probability of any* N *observations has a mixture representation:*

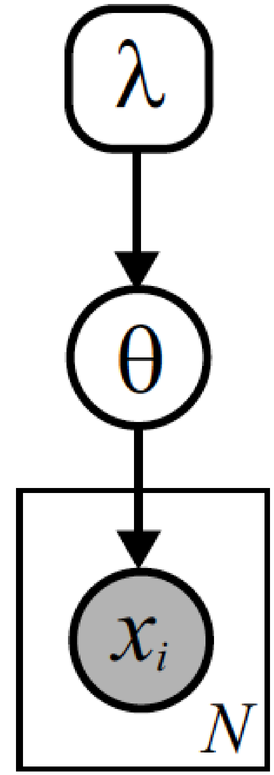
$$p(x_1, x_2, \ldots, x_N) = \int_{\Theta} p(\theta) \prod_{i=1}^{N} p(x_i \mid \theta) \, d\theta \tag{2.77}$$

When $\mathcal{X}$ *is a* K*–dimensional discrete space,* Θ *may be chosen as the* $(K-1)$*–simplex. For Euclidean* $\mathcal{X}$, Θ *is an infinite–dimensional space of probability measures.*

An explicit construction is useful in hierarchical modeling…

De Finetti's Directed Graph





$$p(x_1, \ldots, x_N, \theta \mid \lambda) = p(\theta \mid \lambda) \prod_{i=1}^{N} p(x_i \mid \theta)$$

What distribution underlies the infinitely exchangeable CRP?

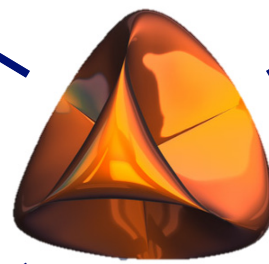
Dirichlet Process Mixtures

The Dirichlet Process (DP)

A distribution on countably infinite discrete probability measures. Sampling yields a ***Polya urn****.*

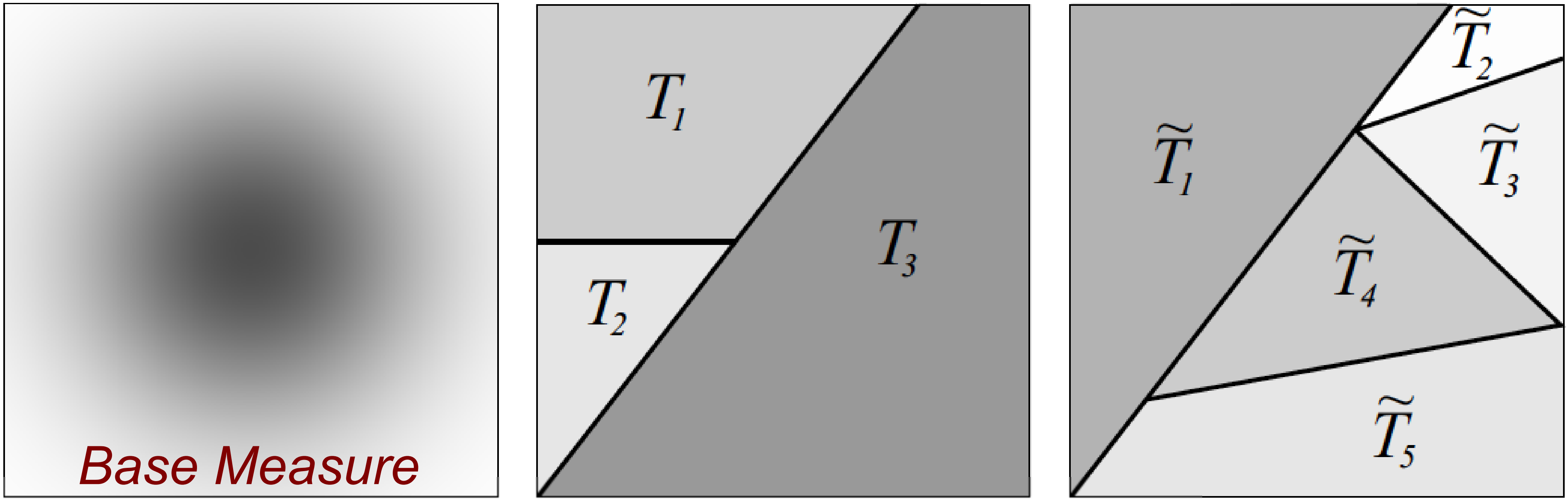
Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Ferguson 1973

Dirichlet Processes



$$\mathbb{E}[G(T)] = H(T) \qquad\qquad G \sim \mathrm{DP}(\alpha, H)$$

- Given a *base measure* (distribution) H & *concentration parameter* $\alpha > 0$
- Then for any finite partition

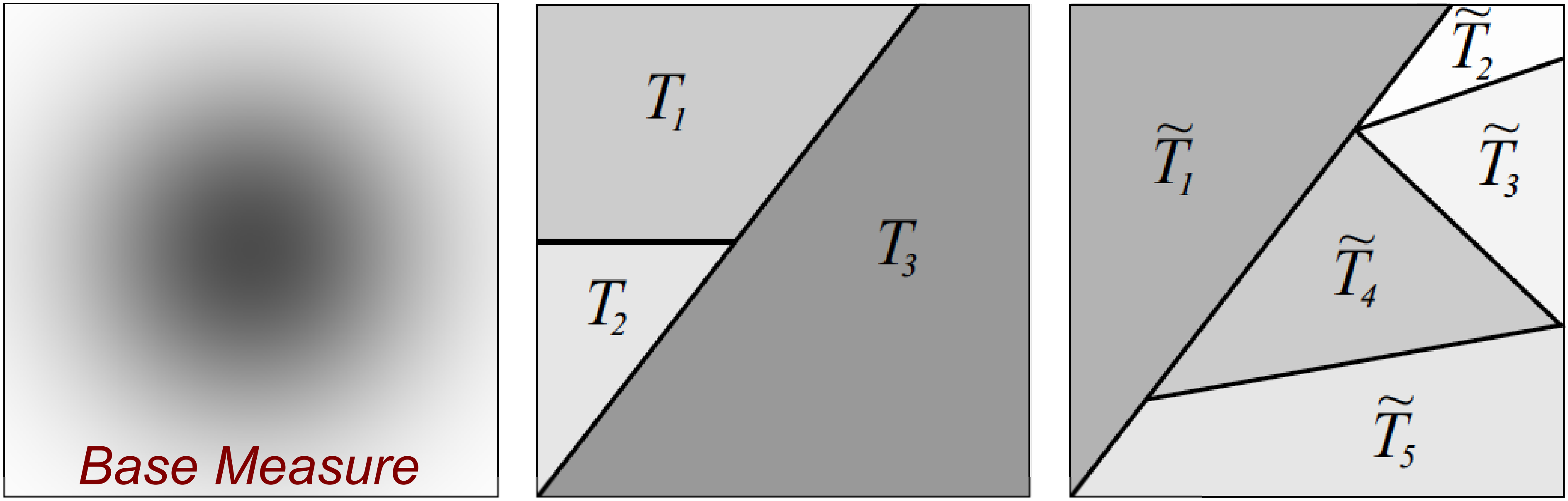
$$\bigcup_{k=1}^{K} T_k = \Theta \qquad T_k \cap T_\ell = \emptyset \qquad k \neq \ell$$

the distribution of the measure of those cells is Dirichlet:

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K))$$

Ferguson 1973

Dirichlet Processes



$$\mathbb{E}[G(T)] = H(T) \qquad\qquad G \sim \mathrm{DP}(\alpha, H)$$

- Marginalization properties of finite Dirichlet distributions satisfy Kolmogorov's extension theorem for stochastic processes:

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)$$

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_K))$$

DP Posteriors and Conjugacy

$$G \sim \mathrm{DP}(\alpha, H) \qquad\qquad \bar{\theta}_i \sim G, i = 1, \dots, N$$

- Does the posterior distribution of *G* have a tractable form?
- For any partition, the posterior mean given N observations is

$$\mathbb{E}\left[G(T) \mid \bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, H\right] = \frac{1}{\alpha + N}\left(\alpha H(T) + \sum_{k=1}^{K} N_k \delta_{\theta_k}(T)\right)$$

$$N_k \triangleq \sum_{i=1}^{N} \delta(\bar{\theta}_i, \theta_k) \qquad\qquad k = 1, \dots, K$$

- In fact, the posterior distribution is another Dirichlet process, with mean that depends on the data's *empirical distribution*:

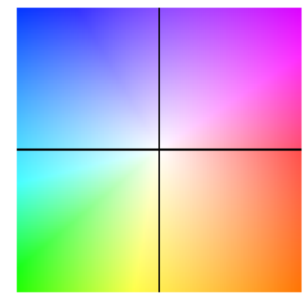
Proposition 2.5.1. *Let $G \sim \mathrm{DP}(\alpha, H)$ be a random measure distributed according to a Dirichlet process. Given N independent observations $\bar{\theta}_i \sim G$, the posterior measure also follows a Dirichlet process:*

$$p\left(G \mid \bar{\theta}_1, \dots, \bar{\theta}_N, \alpha, H\right) = \mathrm{DP}\left(\alpha + N, \frac{1}{\alpha + N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\bar{\theta}_i}\right)\right) \qquad (2.169)$$

DPs and Polya Urns

$$G \sim \mathrm{DP}(\alpha, H) \qquad \bar{\theta}_i \sim G, i = 1, \ldots, N$$

- Can we simulate observations without constructing *G*?
- Yes, by a variation on the classical balls in urns analogy:
  - Consider an urn containing α pounds of very tiny, colored sand (the space of possible colors is Θ)
  - Take out one grain of sand, record its color as $\bar{\theta}_1$
  - Put that grain back, add 1 extra pound of that color
  - Repeat this process…



Theorem 2.5.4. *Let $G \sim \mathrm{DP}(\alpha, H)$ be distributed according to a Dirichlet process, where the base measure H has corresponding density $h(\theta)$. Consider a set of N observations $\bar{\theta}_i \sim G$ taking K distinct values $\{\theta_k\}_{k=1}^K$. The predictive distribution of the next observation then equals*

$$p(\bar{\theta}_{N+1} = \theta \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N}\left(\alpha h(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k)\right) \tag{2.180}$$

where N_k is the number of previous observations of θ_k, as in eq. (2.179).

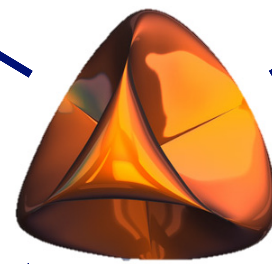
Dirichlet Process Mixtures

The Dirichlet Process (DP)

A distribution on countably infinite discrete probability measures. Sampling yields a ***Polya urn****.*

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Stick-Breaking

An explicit construction for the weights in DP realizations

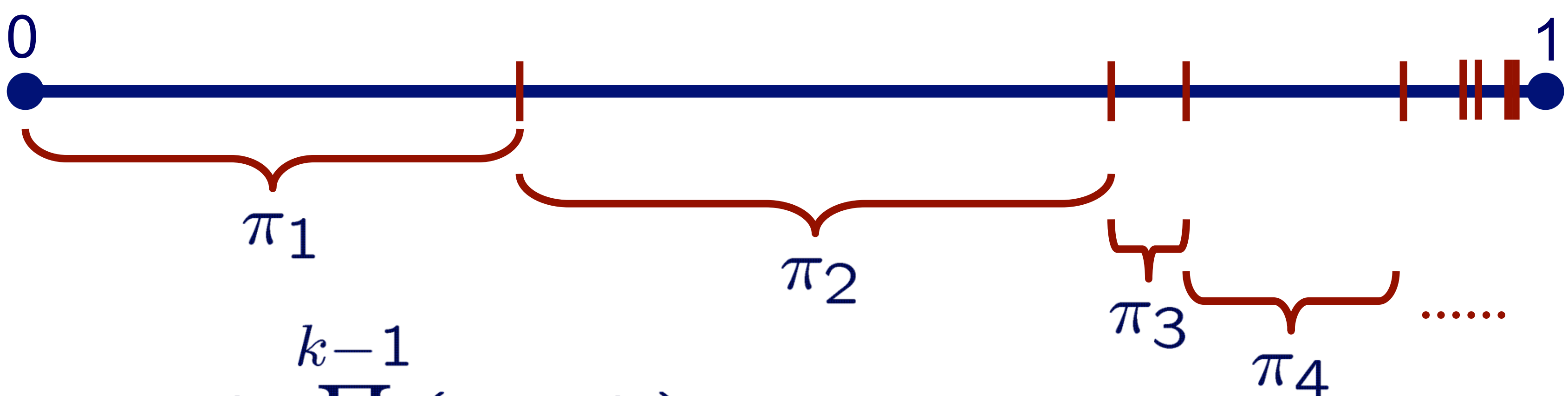
Sethuraman 1994

A Stick-Breaking Construction

- Dirichlet process realizations are discrete with probability one:

$$G \sim \mathrm{DP}(\alpha, H) \qquad\qquad G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

- Cluster shape parameters drawn from base measure: $\theta_k \sim H$
- Cluster weights drawn from a stick-breaking process:



$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

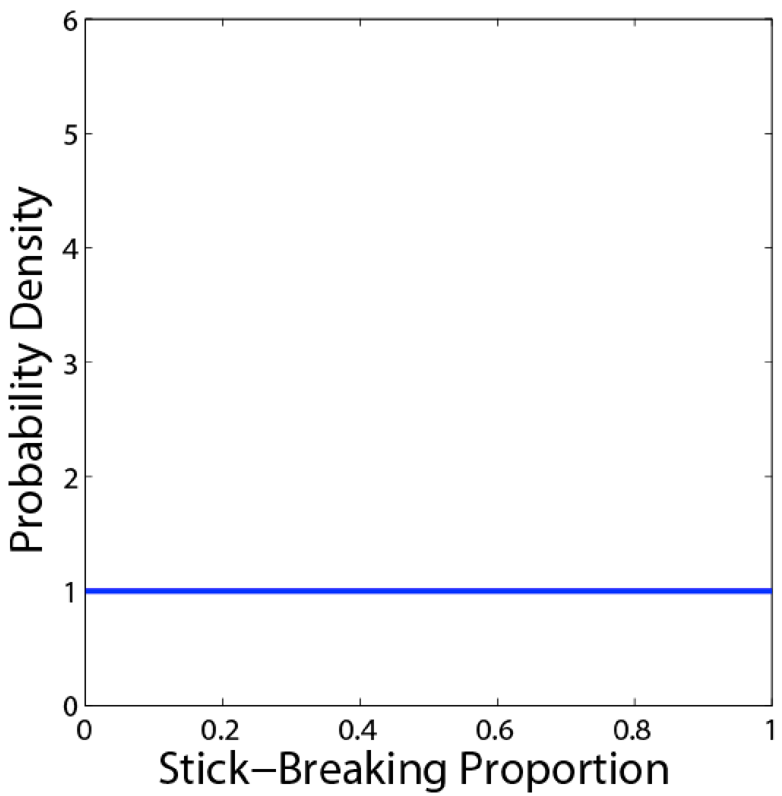
$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

α → concentration parameter

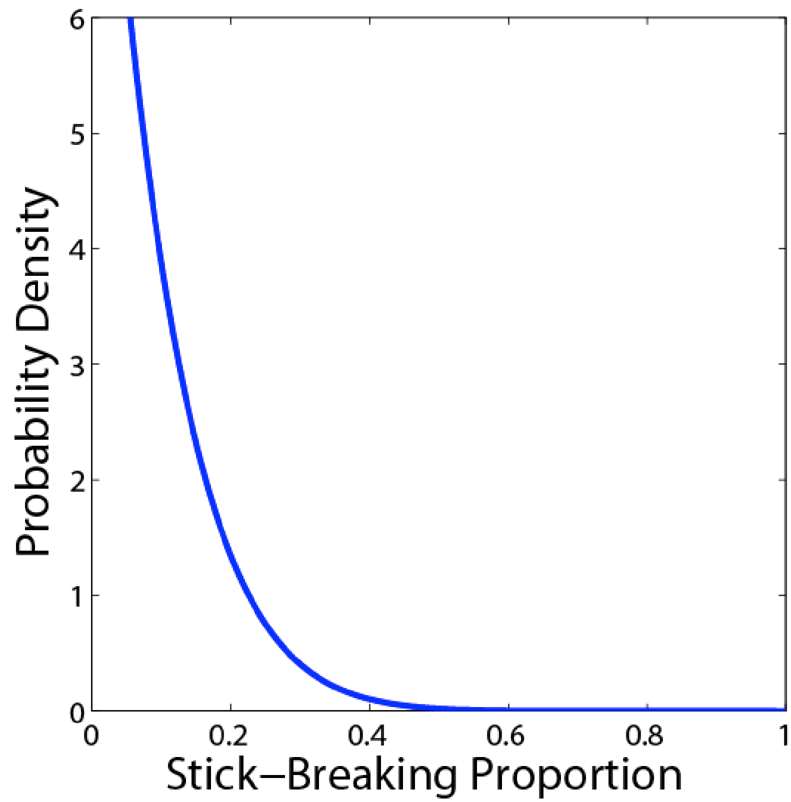
Dirichlet Stick-Breaking

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

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

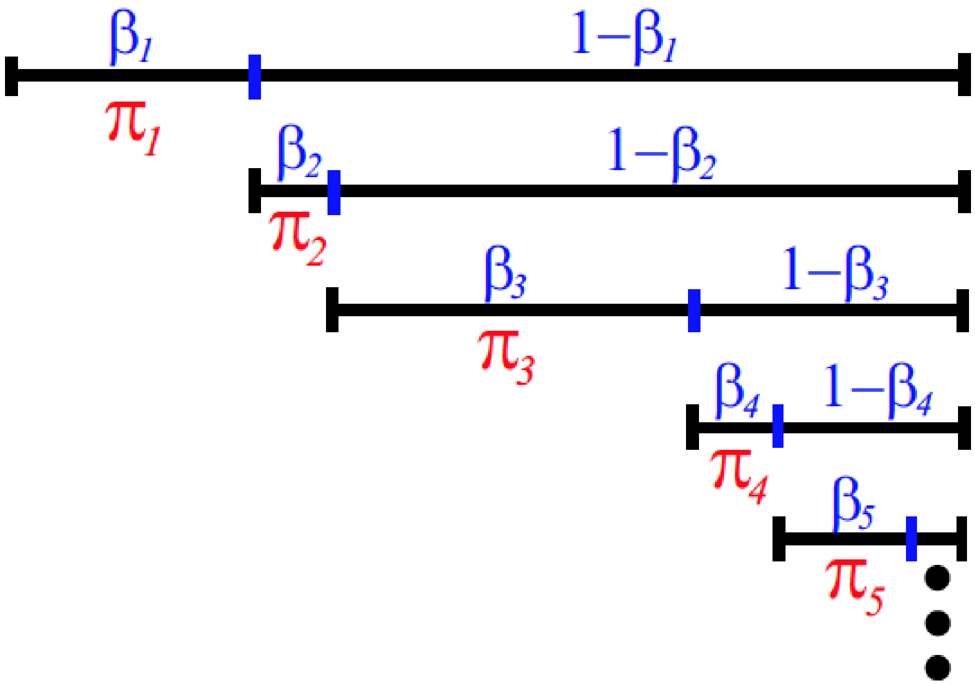


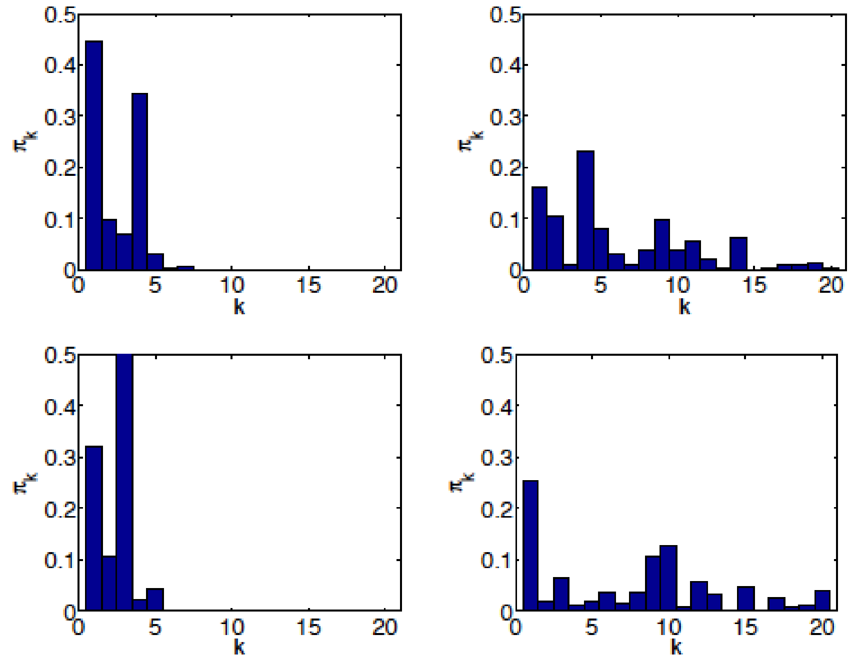
$\alpha = 1$



$\alpha = 10$

DPs and Stick Breaking





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

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right)$$

$$1 - \sum_{k=1}^{K} \pi_k = \prod_{k=1}^{K} (1 - \beta_k) \longrightarrow 0$$

$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

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

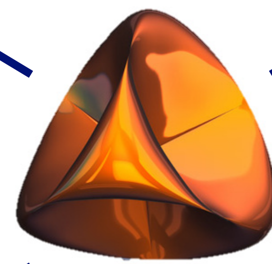
Dirichlet Process Mixtures

The Dirichlet Process (DP)

*A distribution on countably infinite discrete probability measures. Sampling yields a **Polya urn**.*

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Stick-Breaking

An explicit construction for the weights in DP realizations

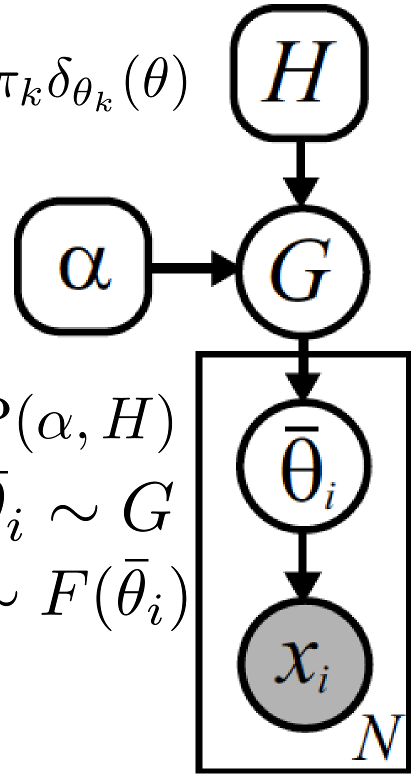
Infinite Mixture Models

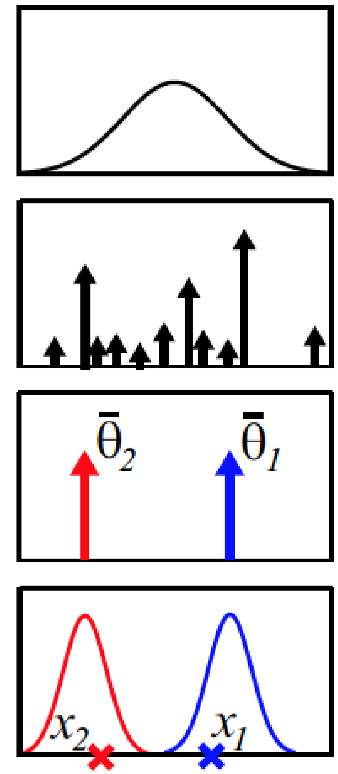
As an infinite limit of finite mixtures with Dirichlet weight priors

DP Mixture Models

$$\theta_k \sim H(\lambda)$$
$$\pi \sim \text{Stick}(\alpha)$$

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





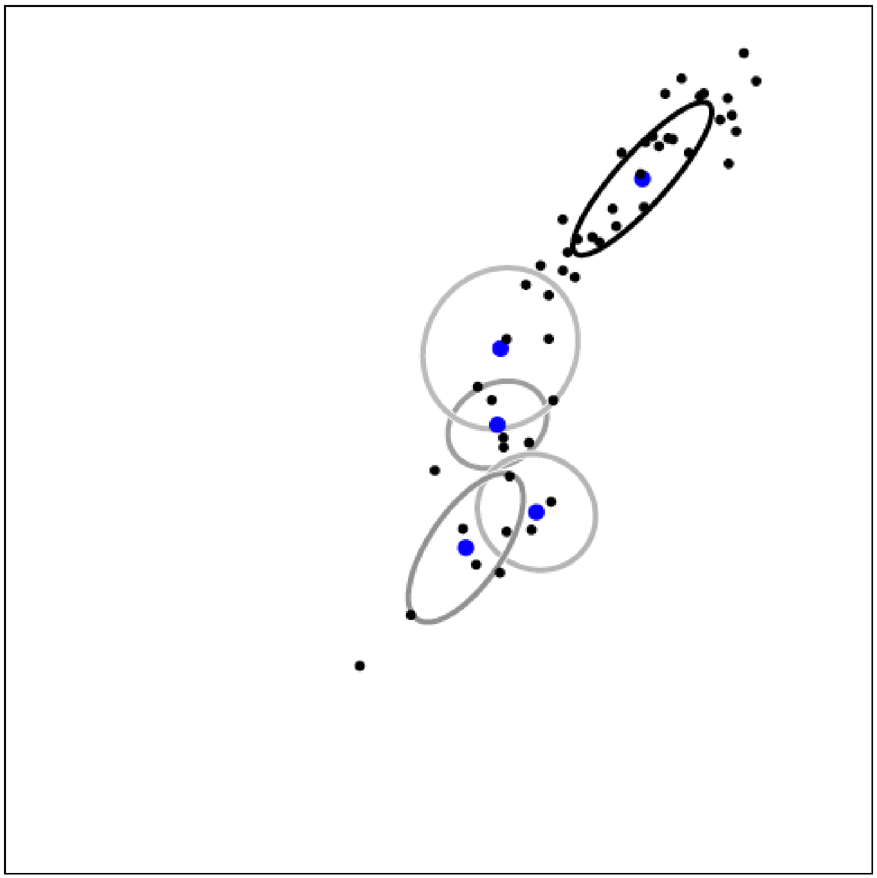
$$z_i \sim \text{Cat}(\pi)$$
$$x_i \sim F(\theta_{z_i})$$

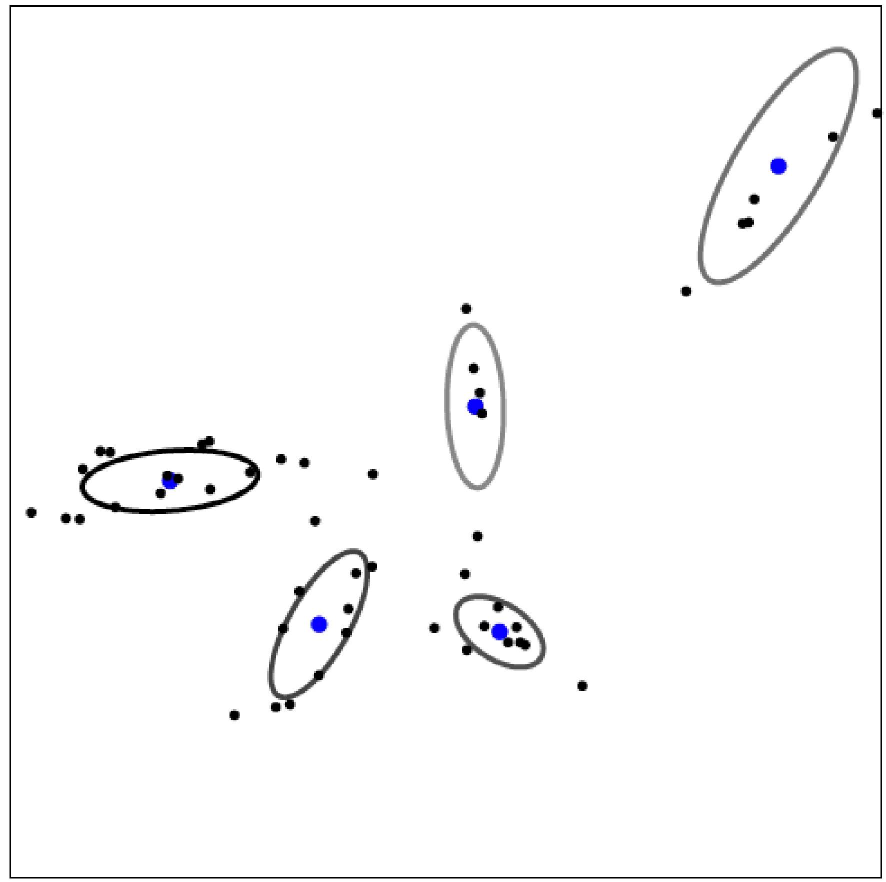
$$G \sim \text{DP}(\alpha, H)$$
$$\bar{\theta}_i \sim G$$
$$x_i \sim F(\bar{\theta}_i)$$

$$p(x \mid \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$

- Stick-breaking: Explicit size-biased sampling of weights π
- Chinese restaurant process: Indicator sequence $z_1, z_2, z_3, \ldots$
- Polya urn: Corresponding parameter sequence $\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3, \ldots$

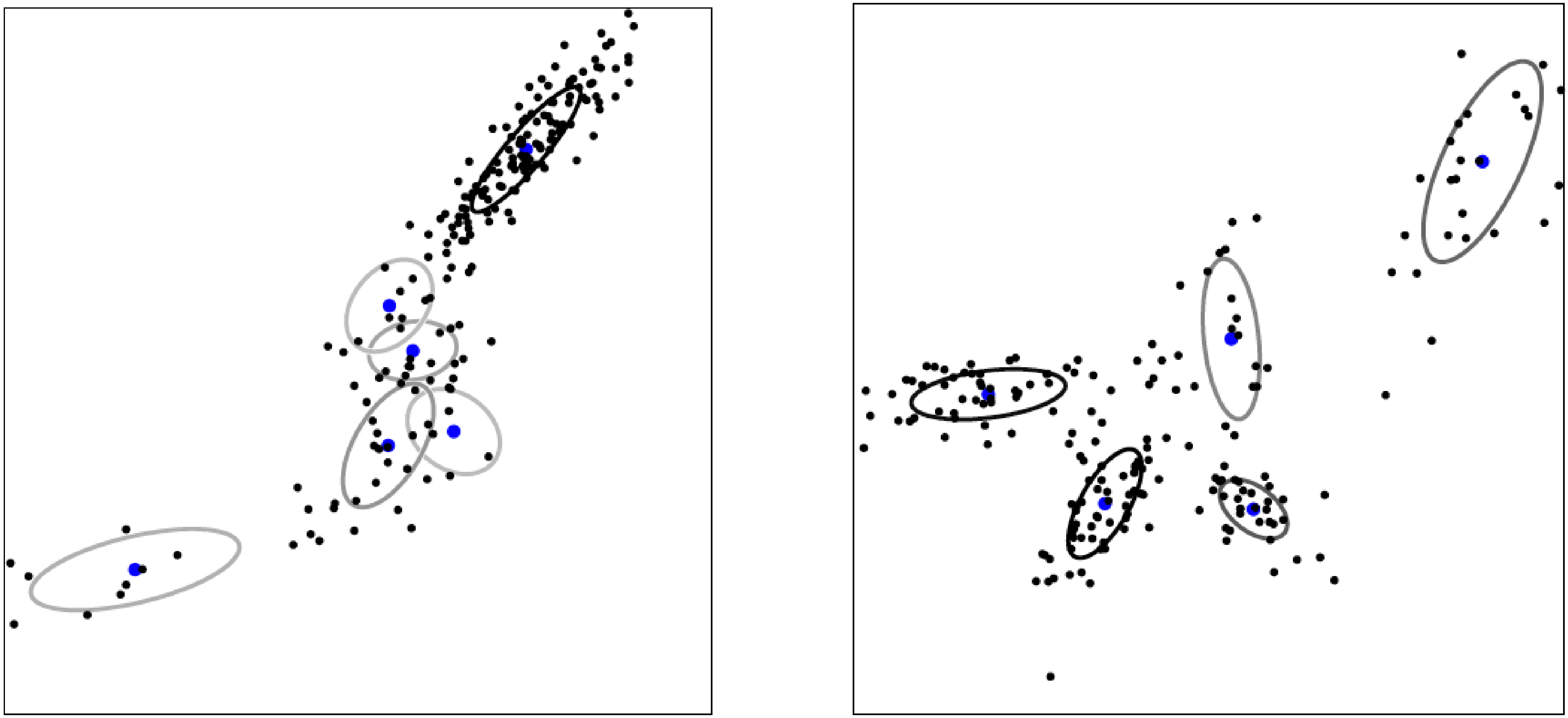
Samples from DP Mixture Priors





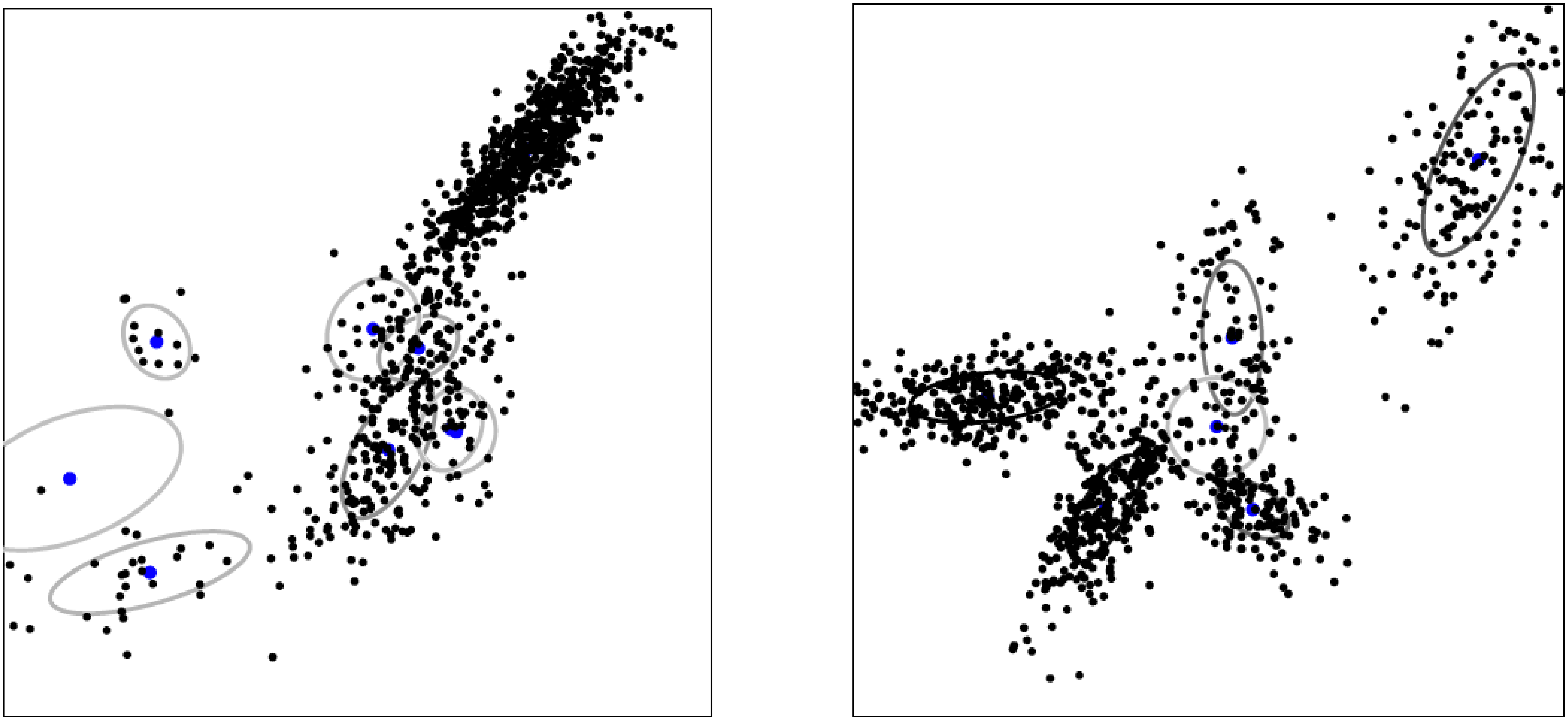
N=50

Samples from DP Mixture Priors



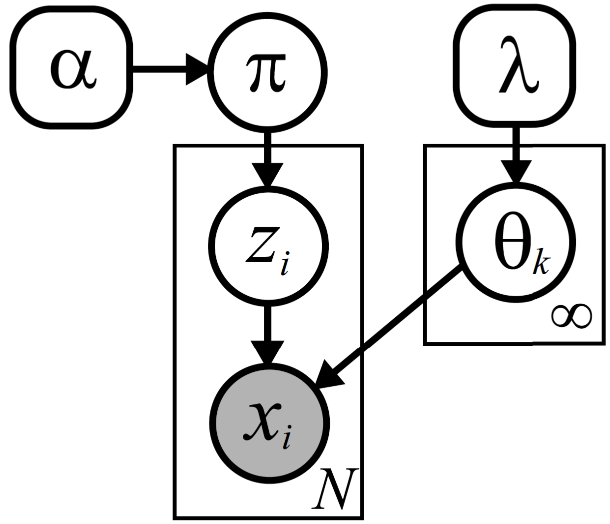
N=200

Samples from DP Mixture Priors



N=1000

Finite versus DP Mixtures



Finite Mixture	*DP Mixture*
$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$	$\pi \sim \mathrm{Stick}(\alpha)$

$$z_i \sim \mathrm{Cat}(\pi)$$

$$x_i \sim F(\theta_{z_i})$$

THEOREM: For any measureable function f, as $K \to \infty$

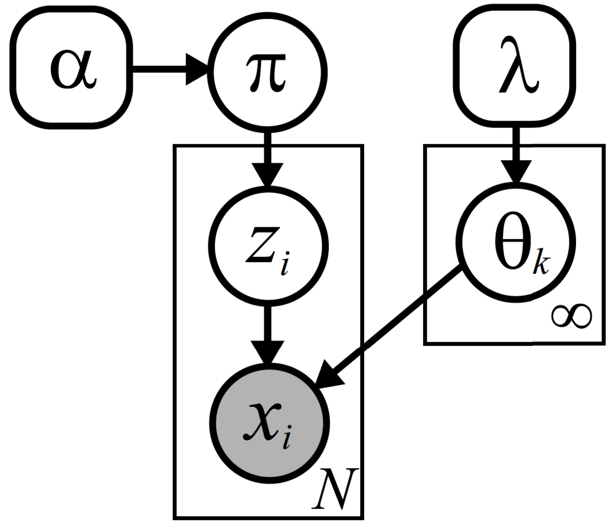
$$\int_\Theta f(\theta)\, dG^K(\theta) \xrightarrow{\mathcal{D}} \int_\Theta f(\theta)\, dG(\theta)$$

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

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

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \qquad \theta_k \sim H$$

Finite versus CRP Partitions



Finite Mixture

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

DP Mixture

$$\pi \sim \mathrm{Stick}(\alpha)$$

$$z_i \sim \mathrm{Cat}(\pi)$$

$K_+ \longrightarrow$ *number of blocks in cluster*

Chinese Restaurant Process:

$$p(z_1, \ldots, z_N \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \alpha^{K_+} \prod_{k=1}^{K_+} (N_k - 1)!$$

Finite Dirichlet:

$$p(z_1, \ldots, z_N \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \left(\frac{\alpha}{K}\right)^{K_+} \prod_{k=1}^{K_+} \prod_{j=1}^{N_k - 1} \left(j + \frac{\alpha}{K}\right)$$

- Probability of Dirichlet *indicators* approach zero as $K \rightarrow \infty$
- Probability of Dirichlet *partition* approaches CRP as $K \rightarrow \infty$

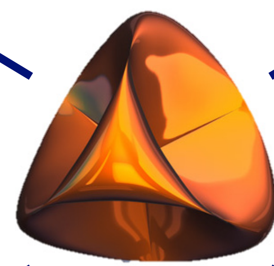
Dirichlet Process Mixtures

The Dirichlet Process (DP)

A distribution on countably infinite discrete probability measures. Sampling yields a ***Polya urn****.*

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a DP prior



Stick-Breaking

An explicit construction for the weights in DP realizations

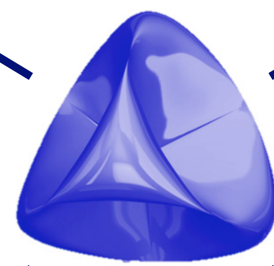
Infinite Mixture Models

As an infinite limit of finite mixtures with Dirichlet weight priors

Pitman-Yor Process Mixtures

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a PY prior



Stick-Breaking

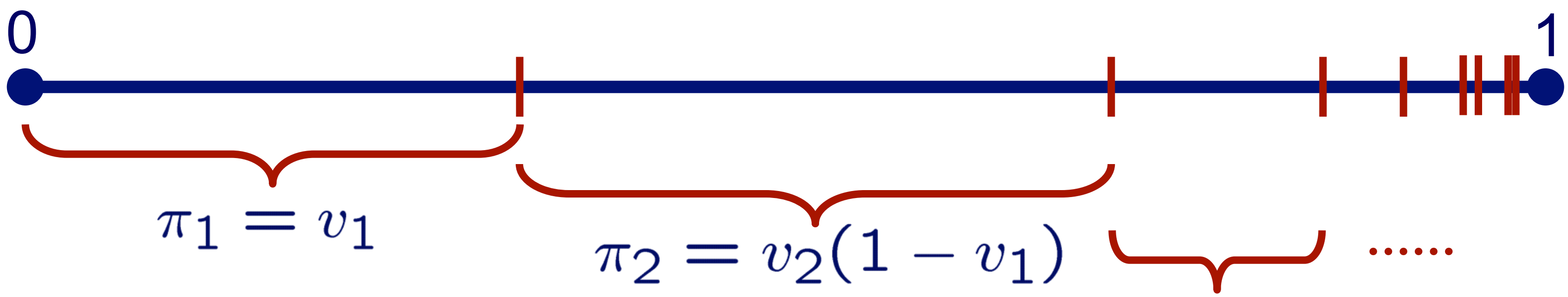
An explicit construction for the weights in PY realizations

Infinite Mixture Models

But not an infinite limit of finite mixtures with symmetric weight priors

Pitman-Yor Processes

The *Pitman-Yor process* defines a distribution on infinite discrete measures, or *partitions*



$$\pi_1 = v_1$$

$$\pi_2 = v_2(1 - v_1)$$

$$\pi_3 = v_3(1 - v_2)(1 - v_1)$$

$$\pi_k = v_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell \right) = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$$

$$v_k \sim \text{Beta}(1 - a, b + ka)$$

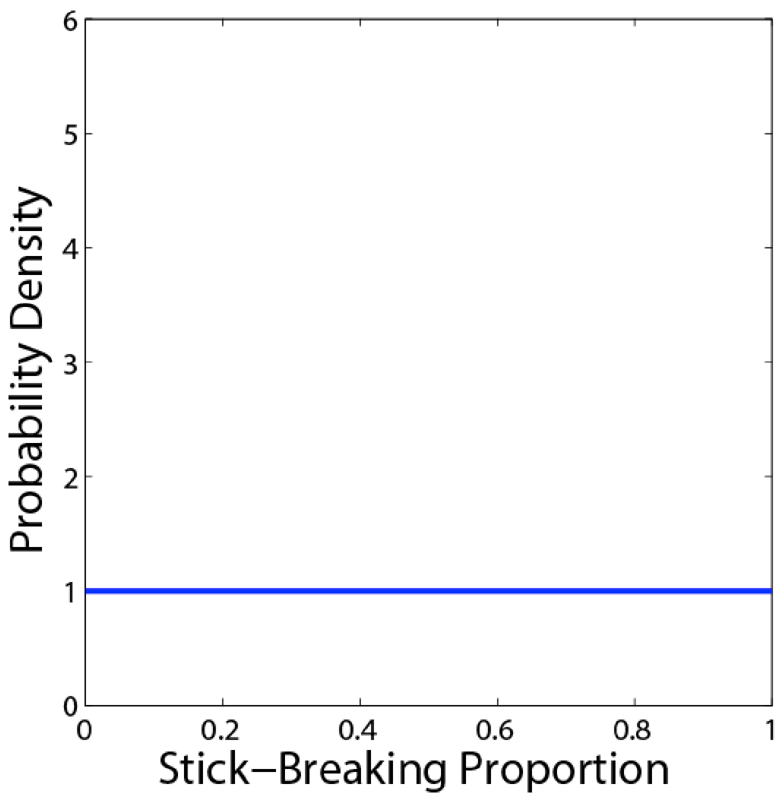
Dirichlet process:

$a = 0$

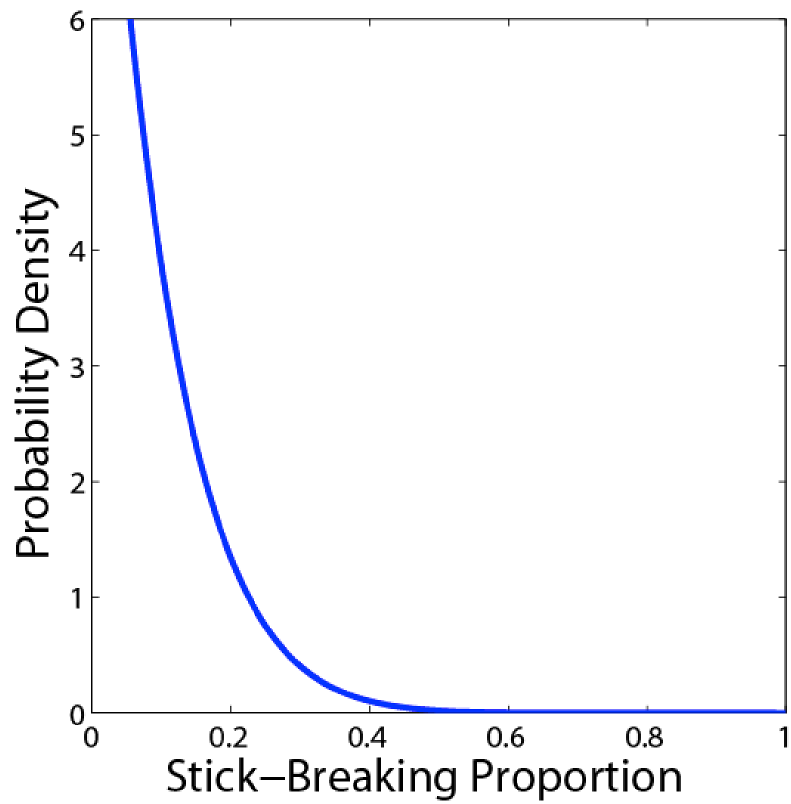
Dirichlet Stick-Breaking

$$v_k \sim \text{Beta}(1, \alpha)$$

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



$\alpha = 1$

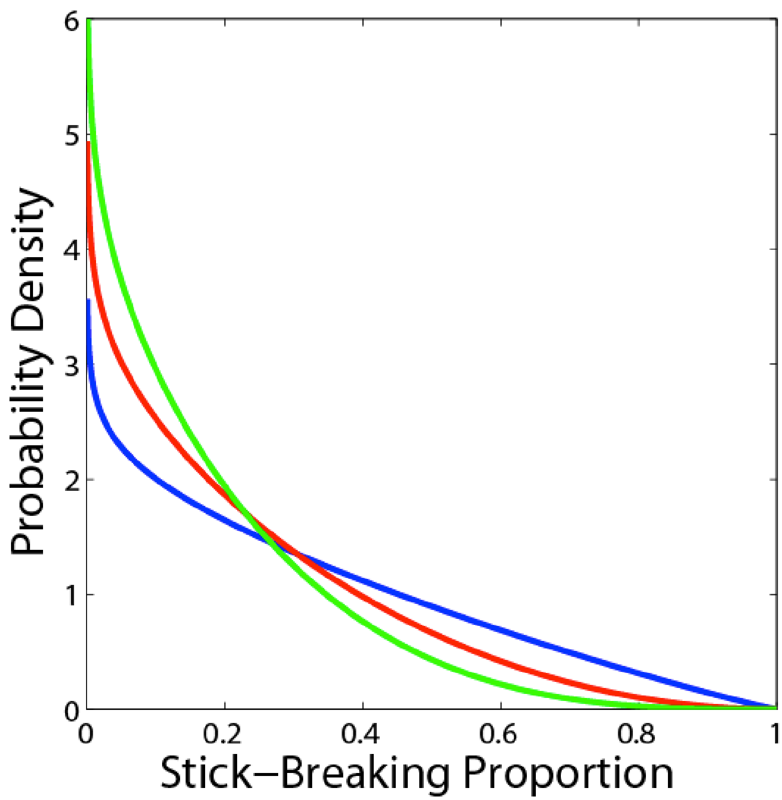


$\alpha = 10$

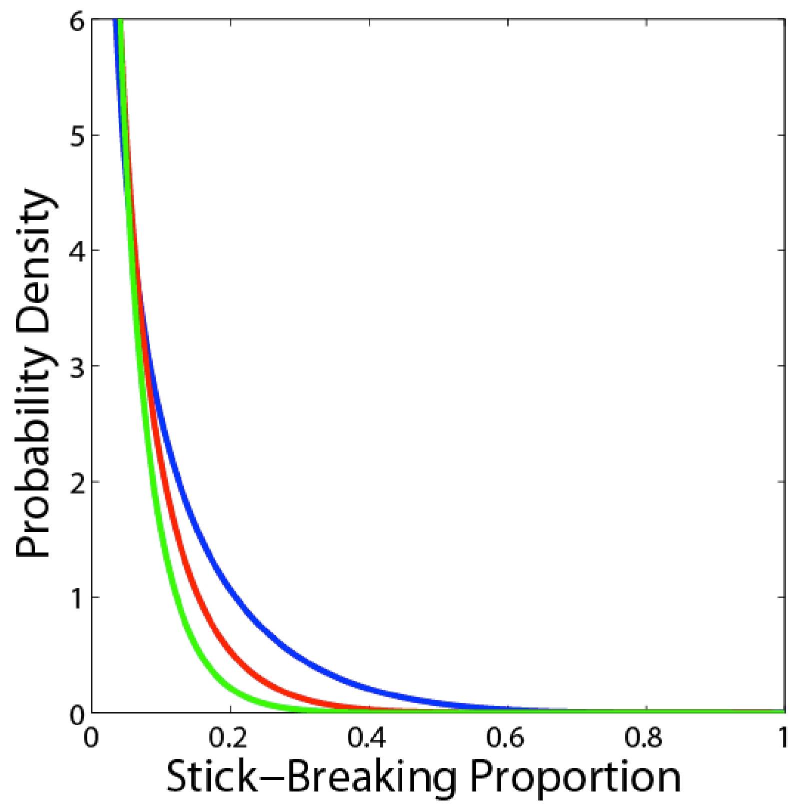
All stick indices k

Pitman-Yor Stick-Breaking

$$v_k \sim \text{Beta}(1-a, b+ka) \qquad E[v_k] = \frac{1-a}{1-a+b+ka}$$



$a = 0.1, b = 3$



$a = 0.5, b = 7$

$k = 1$ $k = 10$ $k = 20$

Chinese Restaurant Process (CRP)

customers ⟷ *observed data to be clustered*

tables ⟷ *distinct blocks of partition, or clusters*

- Partitions sampled from the PY process can be generated via a generalized CRP, which remains *exchangeable*
- The first customer sits at a table. Subsequent customers randomly select a table according to:

$$p(z_{N+1} = z \mid z_1, \ldots, z_N) = \frac{1}{b+N}\left(\sum_{k=1}^{K}(N_k - a)\delta(z,k) + (b + Ka)\delta(z,\bar{k})\right)$$

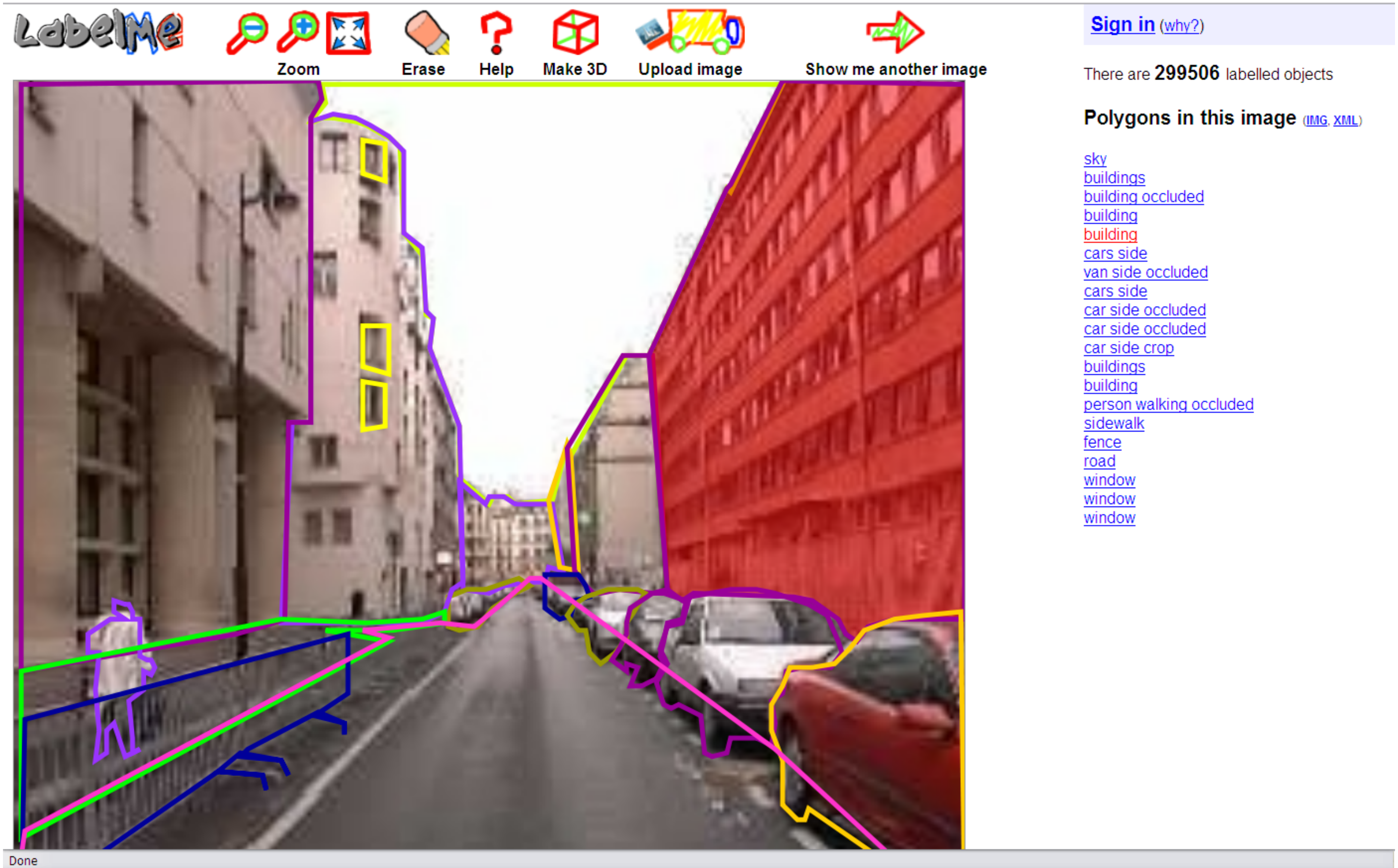
$K \longrightarrow$ number of tables occupied by the first *N* customers

$N_k \longrightarrow$ number of customers seated at table *k*

$\bar{k} \longrightarrow$ a new, previously unoccupied table

$0 \leq a < 1, b > -a \longrightarrow$ discount & concentration parameters

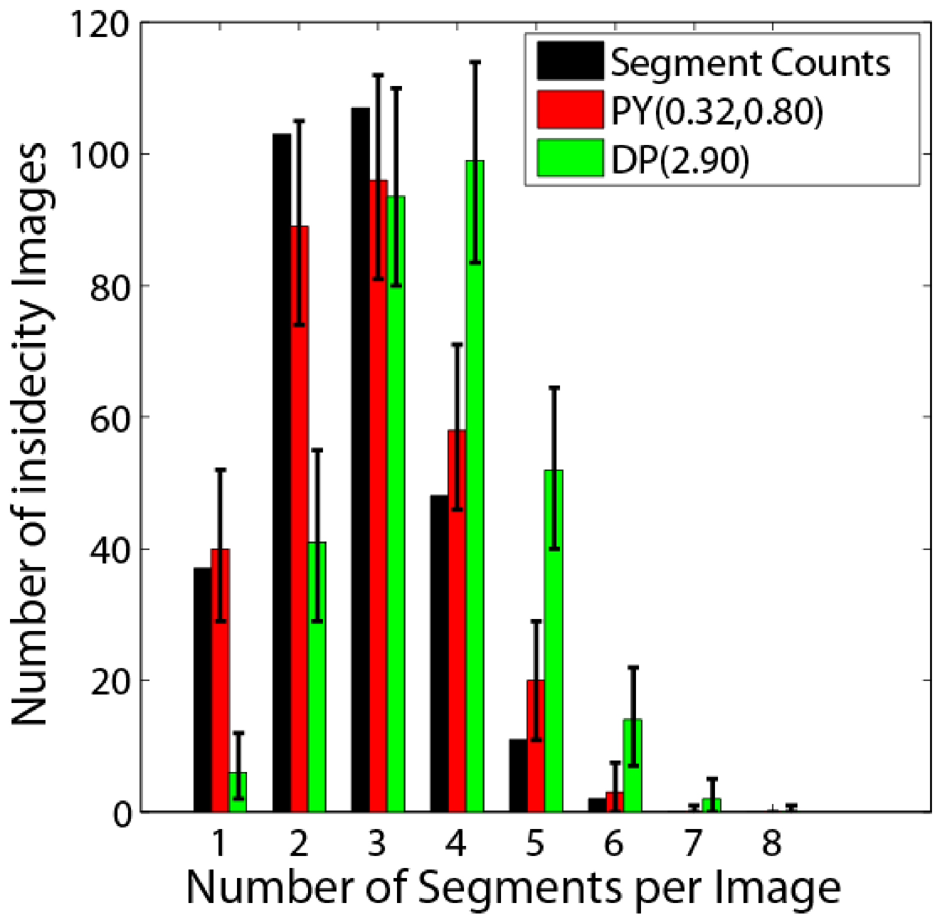
Human Image Segmentations



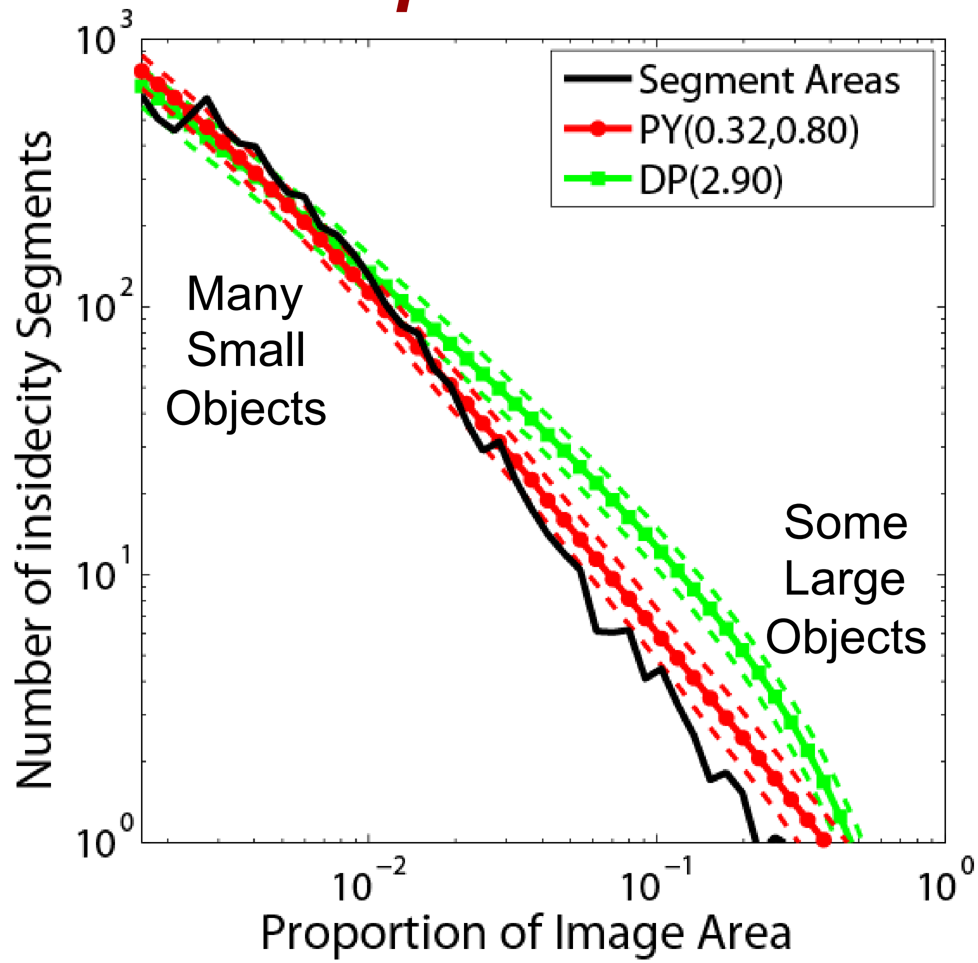
Labels for more than 29,000 segments in 2,688 images of natural scenes

Statistics of Human Segments

How many objects are in this image?



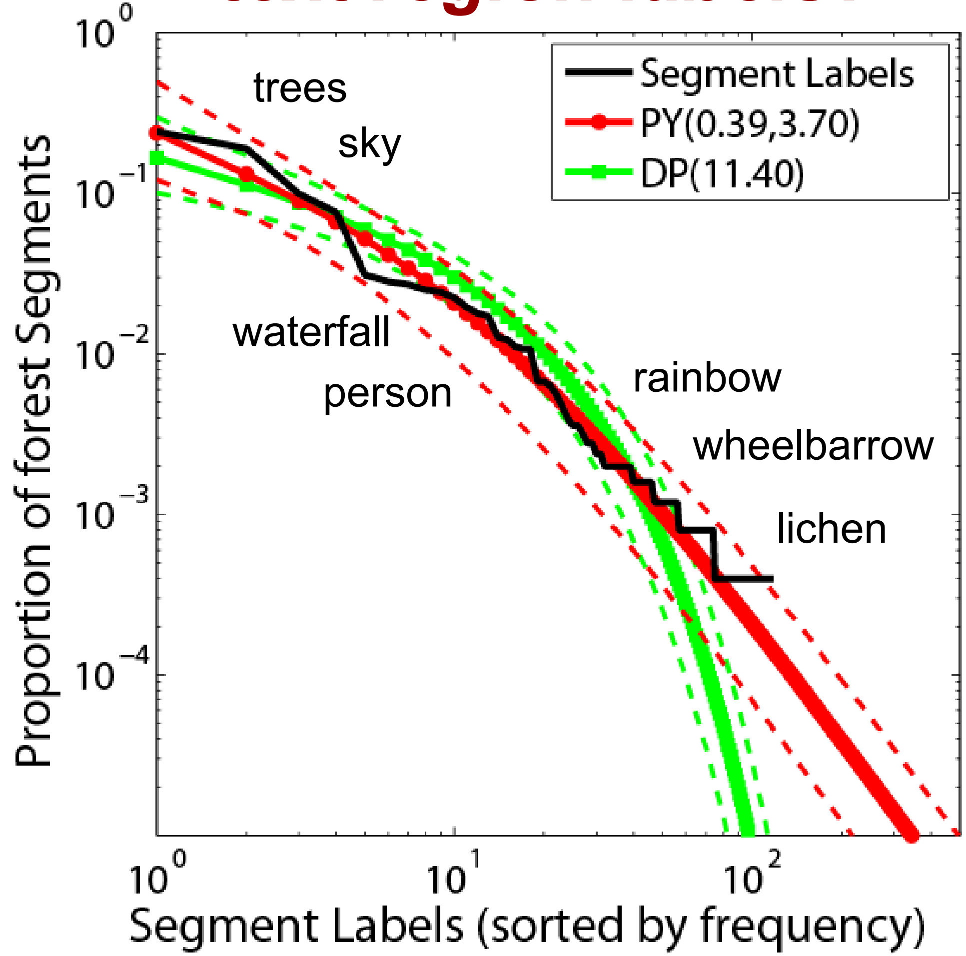
Object sizes follow a power law



Labels for more than 29,000 segments in 2,688 images of natural scenes

Statistics of Semantic Labels

How frequent are text region labels?



Labels for more than 29,000 segments in 2,688 images of natural scenes

Why Pitman-Yor?

Generalizing the Dirichlet Process

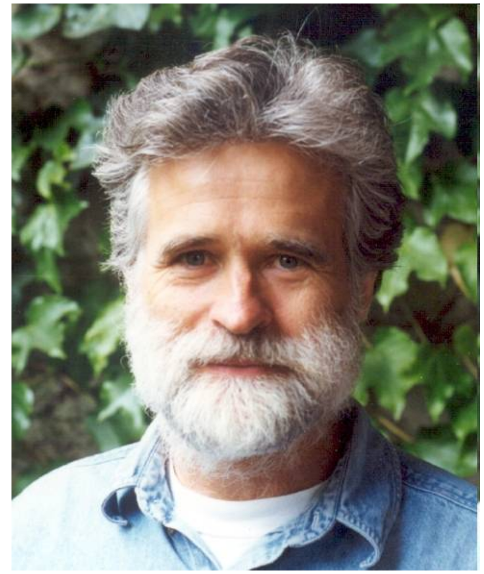
- Distribution on partitions leads to a generalized *Chinese restaurant process*
- Special cases of interest in probability: Markov chains, Brownian motion, …

Power Law Distributions

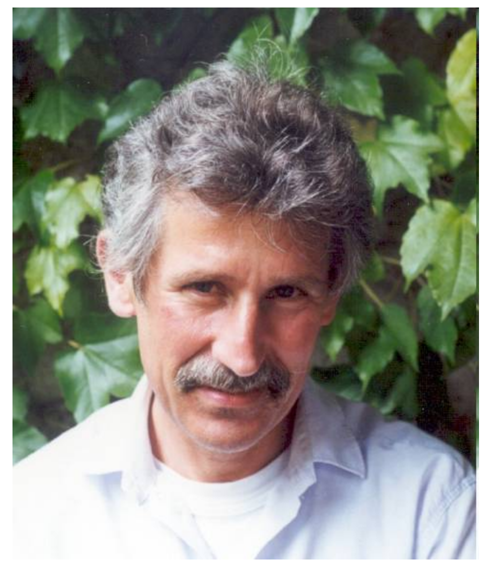
	DP	PY
Number of unique clusters in N observations	$\mathcal{O}(b \log N)$	***Heaps' Law:*** $\mathcal{O}(bN^{a})$
Size of sorted cluster weight k	$\mathcal{O}\left(\alpha_b \left(\frac{1+b}{b}\right)^{-k}\right)$	***Zipf's Law:*** $\mathcal{O}\left(\alpha_{ab} k^{-\frac{1}{a}}\right)$

Natural Language Statistics

Goldwater, Griffiths, & Johnson, 2005
Teh, 2006

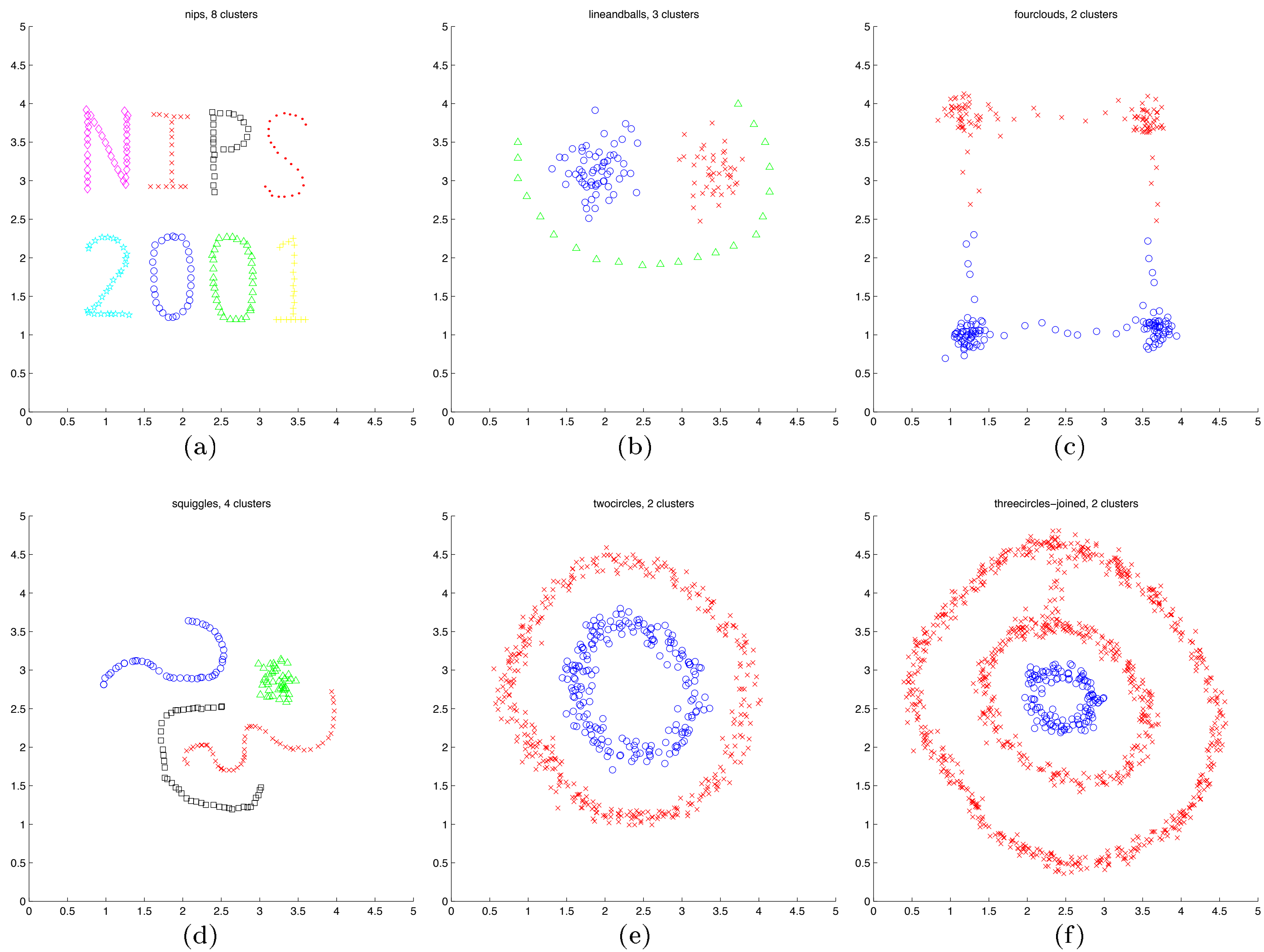


Jim Pitman



Marc Yor

An Aside: Toy Dataset Bias



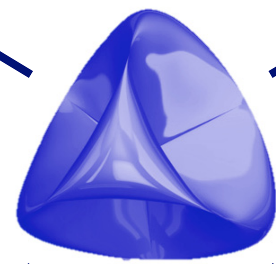
Ng, Jordan, & Weiss, NIPS 2001

Pitman-Yor Process Mixtures

Dirichlet processes and finite Dirichlet distributions do not produce heavy-tailed, power law distributions

Chinese Restaurant Process (CRP)

The distribution on partitions induced by a PY prior



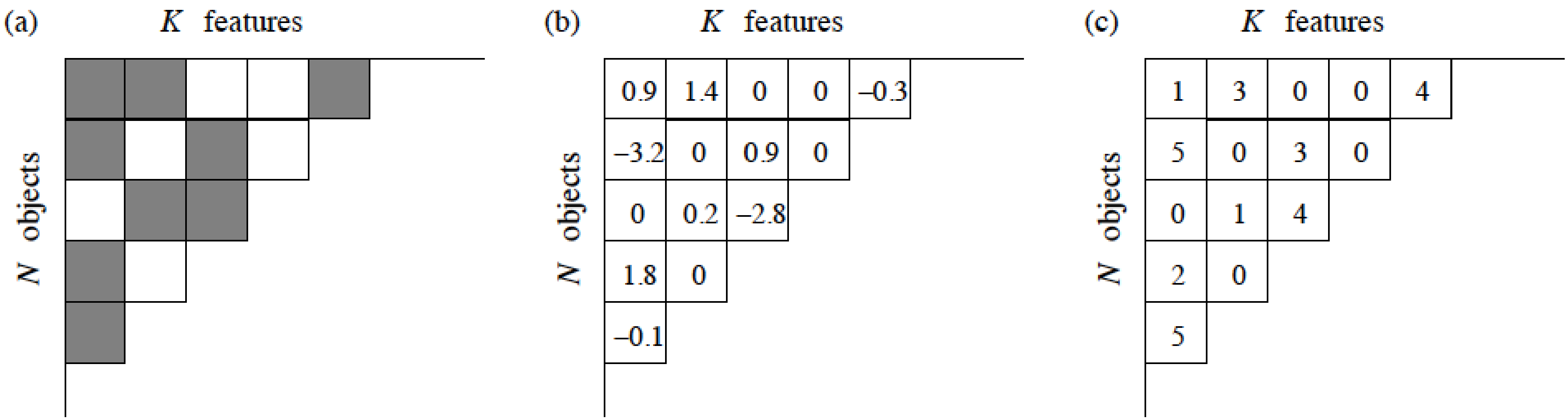
Stick-Breaking

An explicit construction for the weights in PY realizations

Infinite Mixture Models

But not an infinite limit of finite mixtures with symmetric weight priors

Latent Feature Models



Binary matrix indicating feature presence/absence

Depending on application, features can be associated with any parameter value of interest

- *Latent feature modeling*: Each group of observations is associated with a *subset* of the possible latent features
- *Factorial power:* There are 2^K combinations of K features, while accurate mixture modeling may require many more clusters
- *Question:* What is the analog of the DP for feature modeling?

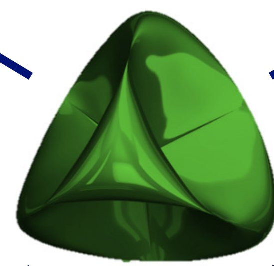
Nonparametric Binary Features

The Beta Process (BP)

A Levy process whose realizations are countably infinite collections of atoms, with mass between 0 and 1.

Indian Buffet Process (IBP)

The distribution on sparse binary matrices induced by a BP



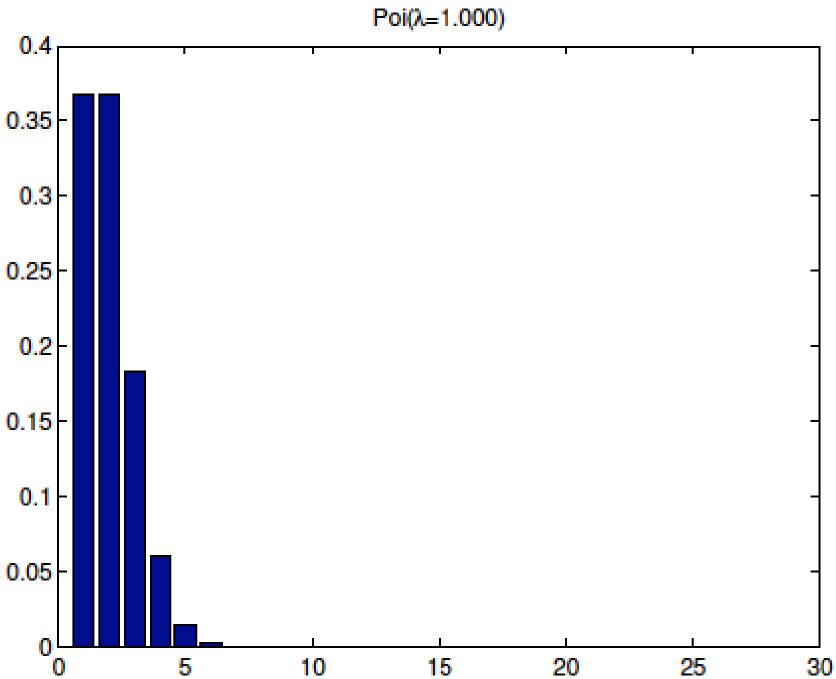
Stick-Breaking

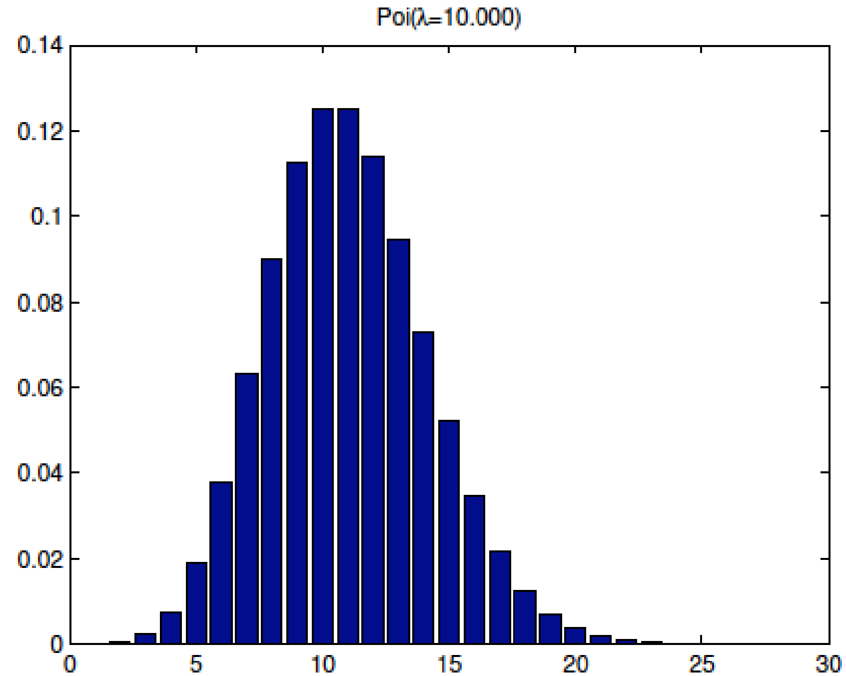
An explicit construction for the feature frequencies in BP realizations

Infinite Feature Models

As an infinite limit of a finite beta-Bernoulli binary feature model

Poisson Distribution for Counts





$$\mathcal{X} = \{0, 1, 2, 3, \ldots\}$$

$$\mathrm{Poi}(x \mid \theta) = e^{-\theta} \frac{\theta^x}{x!} \qquad \theta > 0$$

Indian Buffet Process (IBP)

- Visualize feature assignment as a sequential process of customers sampling dishes from an (infinitely long) buffet:

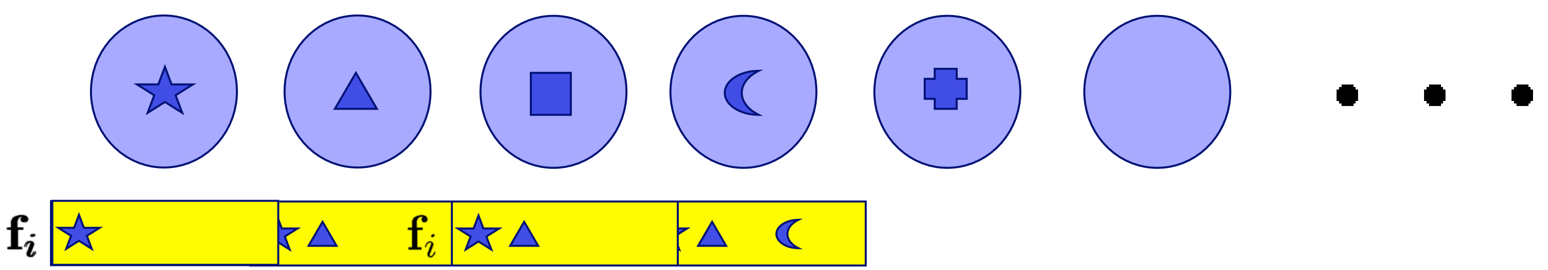
customers ⟷ *observed data to be modeled*

dishes ⟷ *binary features to be selected*

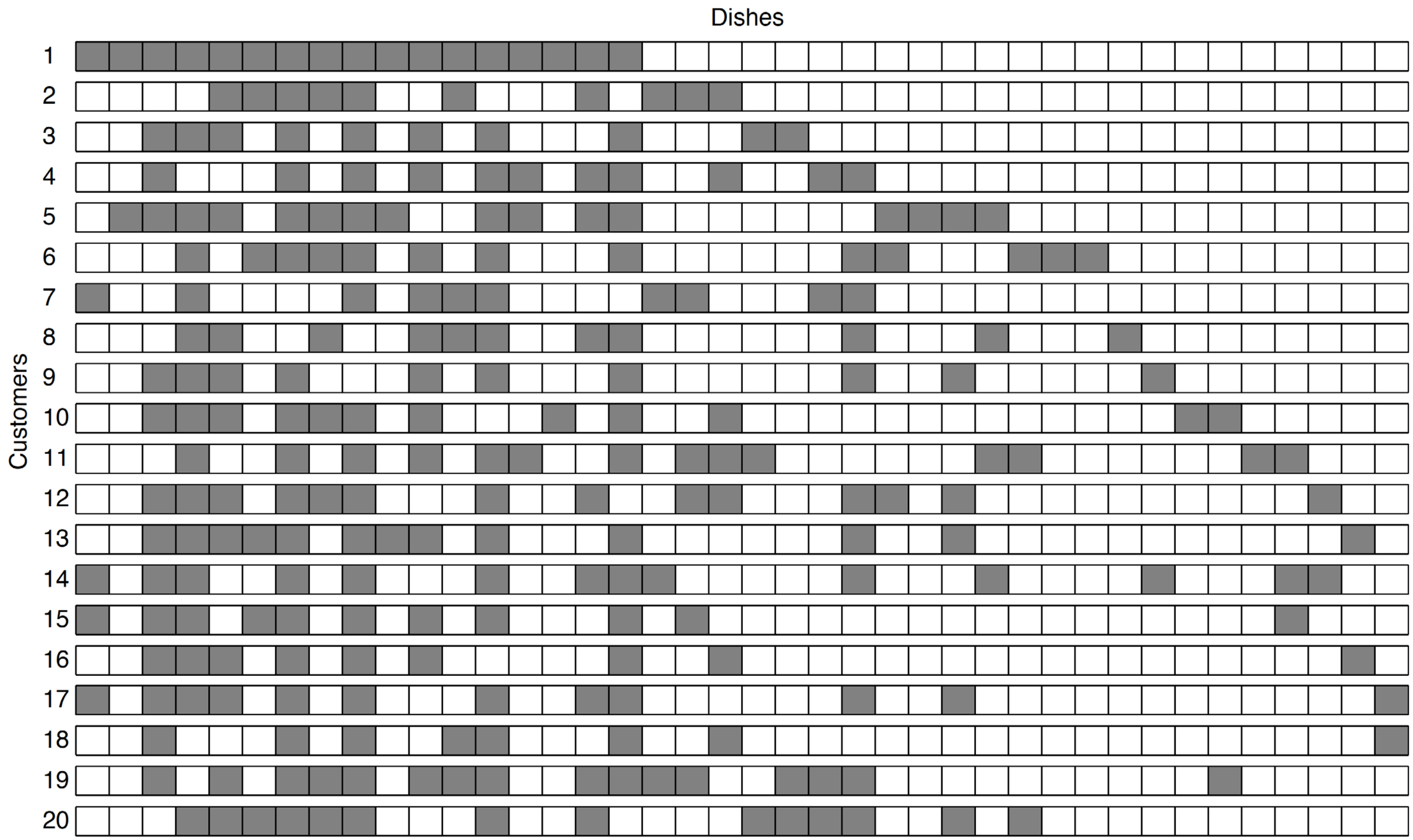
- The first customer chooses $\mathrm{Poisson}(\alpha)$ dishes, $\alpha > 0$
- Subsequent customer *i* randomly samples each previously tasted dish k with probability $f_{ik} \sim \mathrm{Ber}\left(\frac{m_k}{i}\right)$

$m_k \longrightarrow$ number of previous customers to sample dish k

- That customer also samples $\mathrm{Poisson}(\alpha/i)$ new dishes



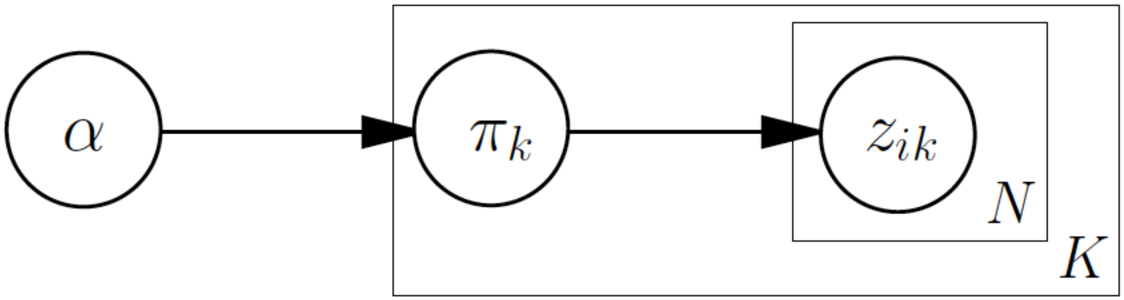
Binary Feature Realizations



Ghahramani, BNP 2009

- IBP is *exchangeable*, up to a permutation of the order with which dishes are listed in the binary feature matrix
- Clustering models like the DP have one “feature” per customer
- The number of features sampled at least once is $\mathcal{O}(\alpha \log N)$

Finite Beta-Bernoulli Features



$$\pi_k \sim \mathrm{Beta}\left(\frac{\alpha}{K}, 1\right) \qquad z_{ik} \sim \mathrm{Ber}(\pi_k)$$

$$P(\mathbf{Z}|\pi) = \prod_{k=1}^{K}\prod_{i=1}^{N} P(z_{ik}|\pi_k) = \prod_{k=1}^{K} \pi_k^{m_k}(1-\pi_k)^{N-m_k} \qquad m_k = \sum_{i=1}^{N} z_{ik}$$

- The expected number of active features in N customers is

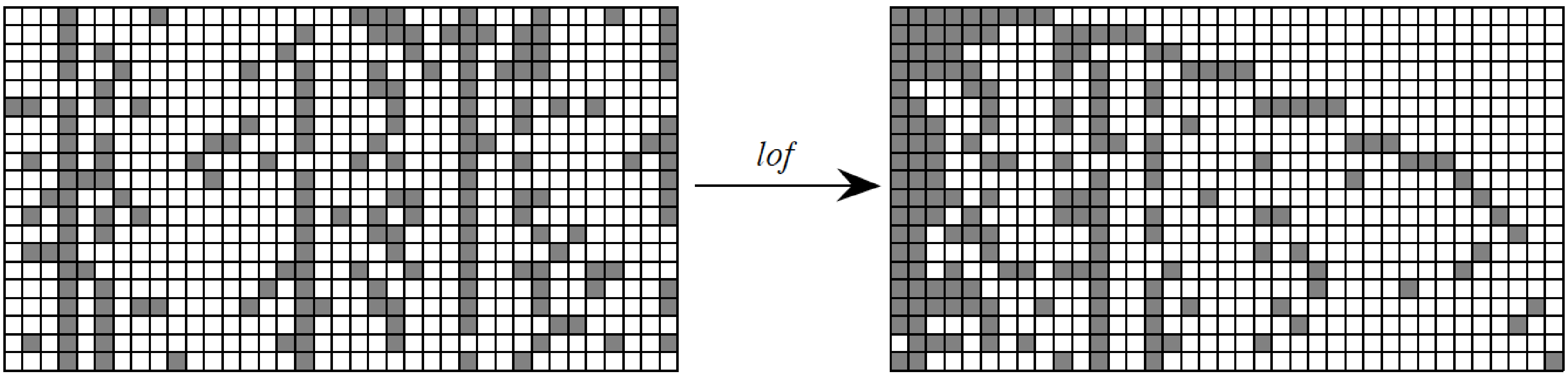
$$\frac{N\alpha}{(1+\alpha/K)} \to N\alpha$$

- The marginal probability of the realized binary matrix equals

$$P(\mathbf{Z}) = \prod_{k=1}^{K}\int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k)\right) p(\pi_k)\, d\pi_k = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(N+1+\frac{\alpha}{K})}$$

Beta-Bernoulli and the IBP

- We can show that the limit of the finite beta-Bernoulli model, and the IBP, produce the same distribution on *left-ordered-form equivalence classes of binary matrices*:



- Poisson distribution in IBP arises from the *law of rare events*:
  - Flip *K* coins with probability of coming up heads α/K
  - As $K \to \infty$ the distribution of the number of total heads approaches $\text{Poisson}(\alpha)$

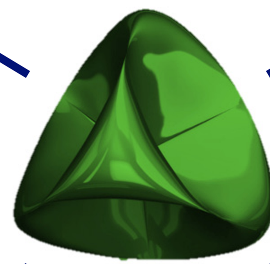
Nonparametric Binary Features

The Beta Process (BP)

A Levy process whose realizations are countably infinite collections of atoms, with mass between 0 and 1.

Indian Buffet Process (IBP)

The distribution on sparse binary matrices induced by a BP



Stick-Breaking

An explicit construction for the feature frequencies in BP realizations

Infinite Feature Models

As an infinite limit of a finite beta-Bernoulli binary feature model

Extensions: Additional control over feature sharing, power laws…

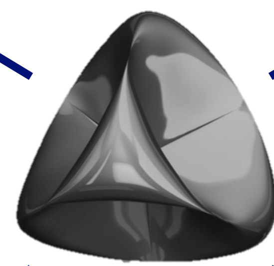
Nonparametric Learning

Infinite Stochastic Processes

Conceptually useful, but usually impractical or impossible for learning algorithms.

CRP & IBP

Tractably learn via finite summaries of true, infinite model.



Stick-Breaking

Truncate stick-breaking to produce provably accurate approximation.

Finite Bayesian Models

Set finite model order to be larger than expected number of clusters or features.

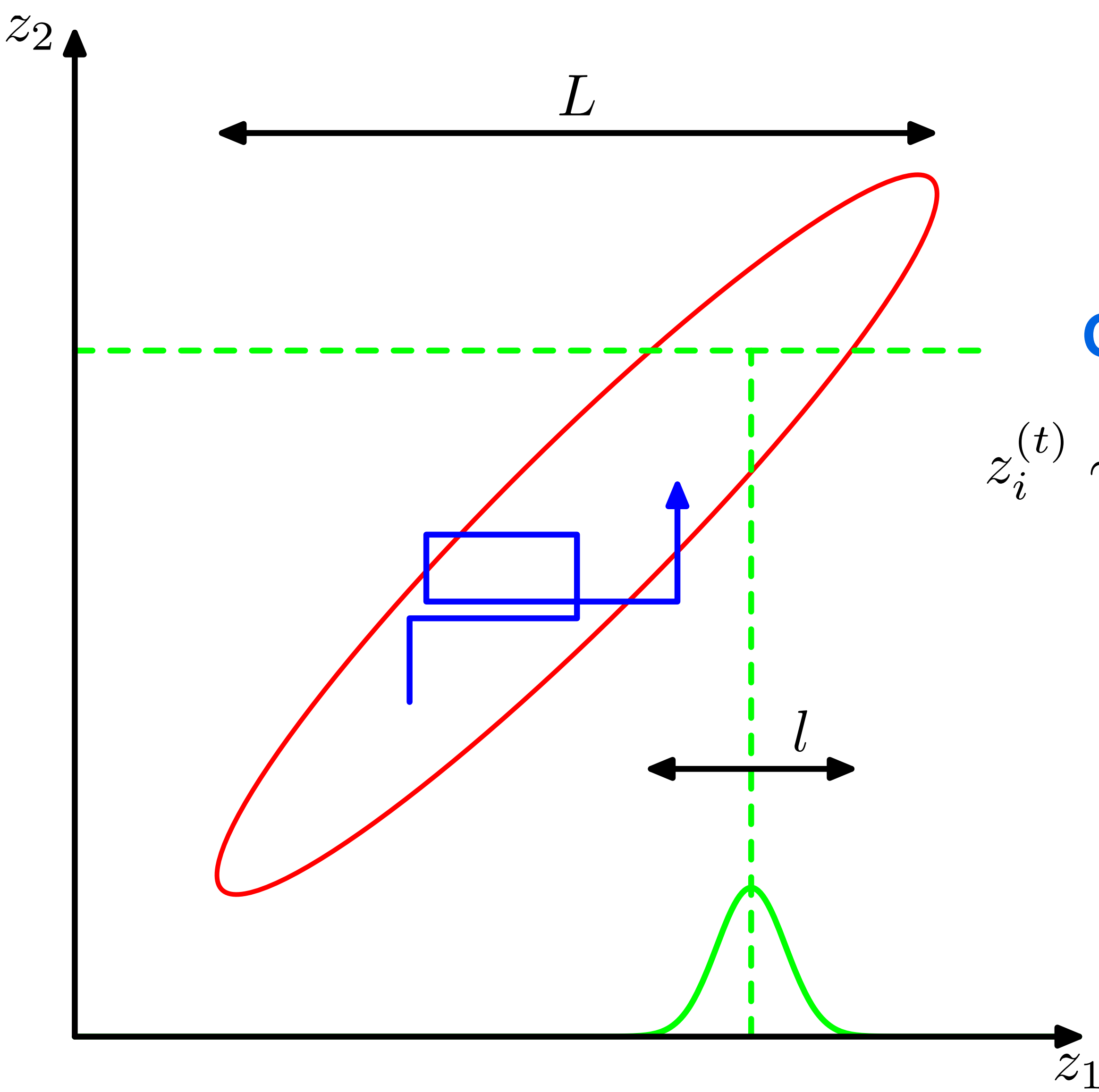
Markov Chain Monte Carlo

- At each time point, state $z^{(t)}$ is a configuration of *all the variables in the model:* parameters, hidden variables, etc.
- We design the transition distribution $q(z \mid z^{(t)})$ so that the chain is *irreducible* and *ergodic*, with a unique stationary distribution $p^*(z)$

$$p^*(z) = \int_{\mathcal{Z}} q(z \mid z')p^*(z')\, dz'$$

- For learning, the target equilibrium distribution is usually the posterior distribution given data *x*: $p^*(z) = p(z \mid x)$
- Popular recipes: *Metropolis-Hastings and Gibbs samplers*

Gibbs Sampler for a 2D Gaussian



General Gibbs Sampler

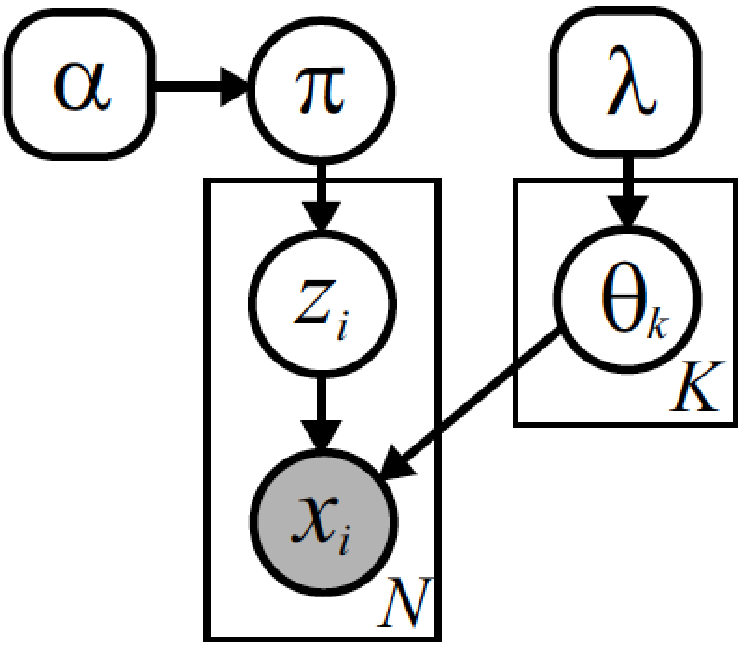
$$z_i^{(t)} \sim p(z_i \mid z_{\setminus i}^{(t-1)}) \qquad i = i(t)$$
$$z_j^{(t)} = z_j^{(t-1)} \qquad j \neq i(t)$$

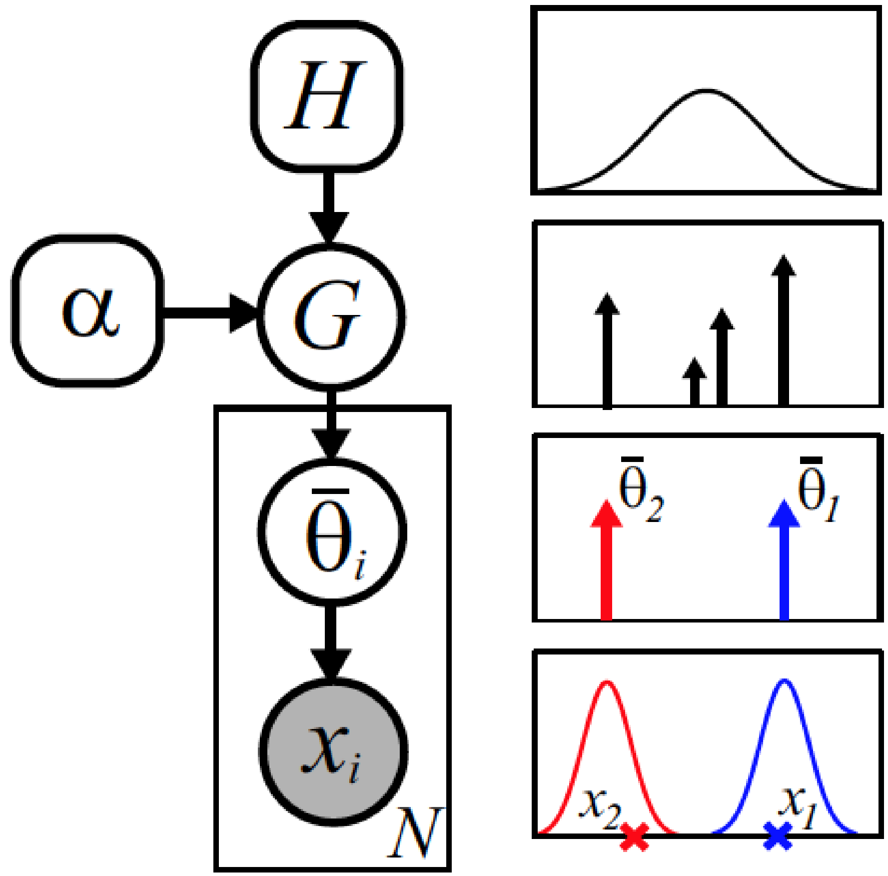
Under mild conditions, converges assuming all variables are resampled infinitely often (order can be fixed or random)

C. Bishop, Pattern Recognition & Machine Learning

Finite Mixture Gibbs Sampler

$$p(x \mid \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x \mid \theta_k)$$





Most basic approach: Sample z, π, θ

Standard Finite Mixture Sampler

Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample a new set of mixture parameters as follows:

1. Independently assign each of the N data points x_i to one of the K clusters by sampling the indicator variables $z = \{z_i\}_{i=1}^N$ from the following multinomial distributions:

$$z_i^{(t)} \sim \frac{1}{Z_i} \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})\, \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})$$

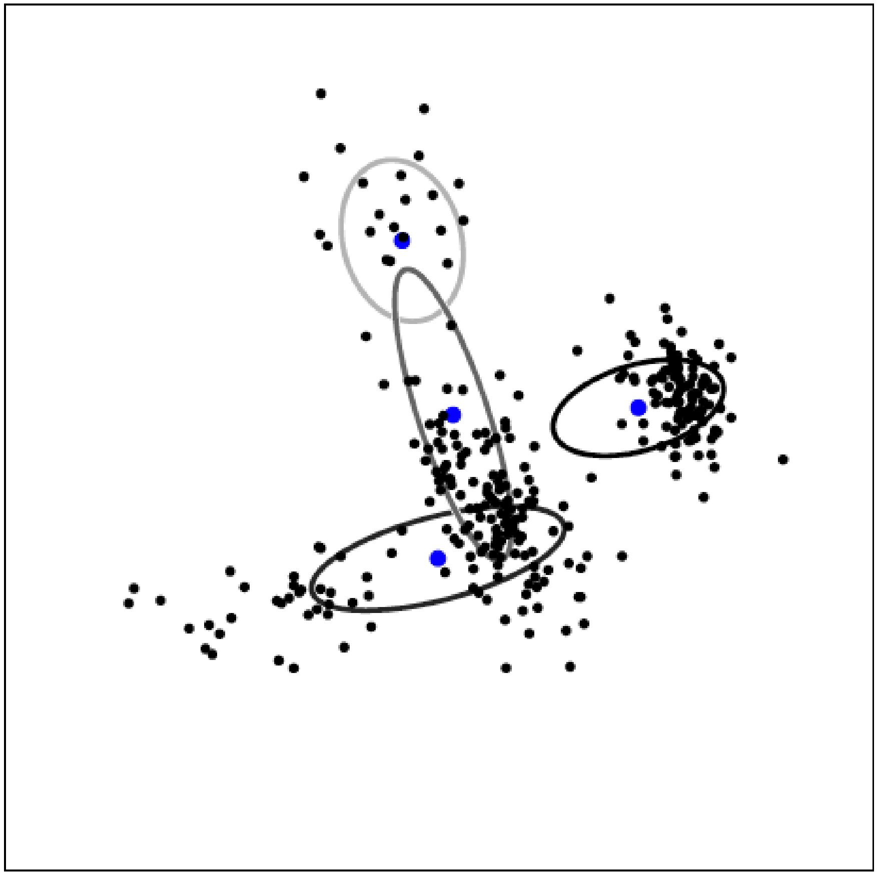
2. Sample new mixture weights according to the following Dirichlet distribution:

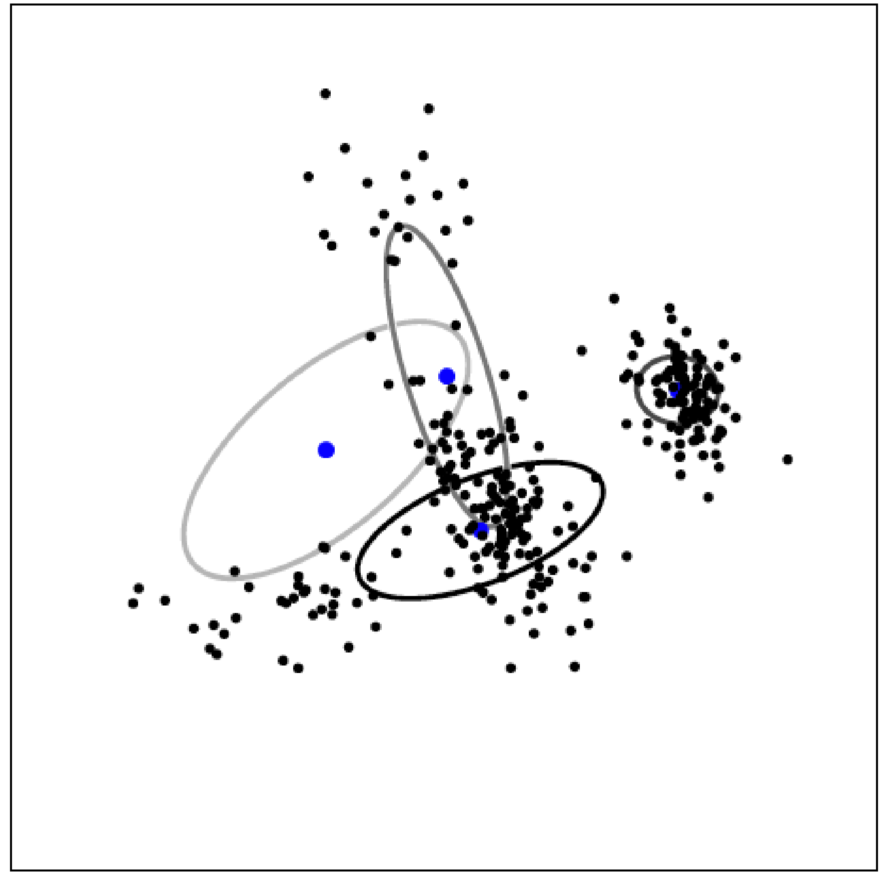
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^{N} \delta(z_i^{(t)}, k)$$

3. For each of the K clusters, independently sample new parameters from the conditional distribution implied by those observations currently assigned to that cluster:

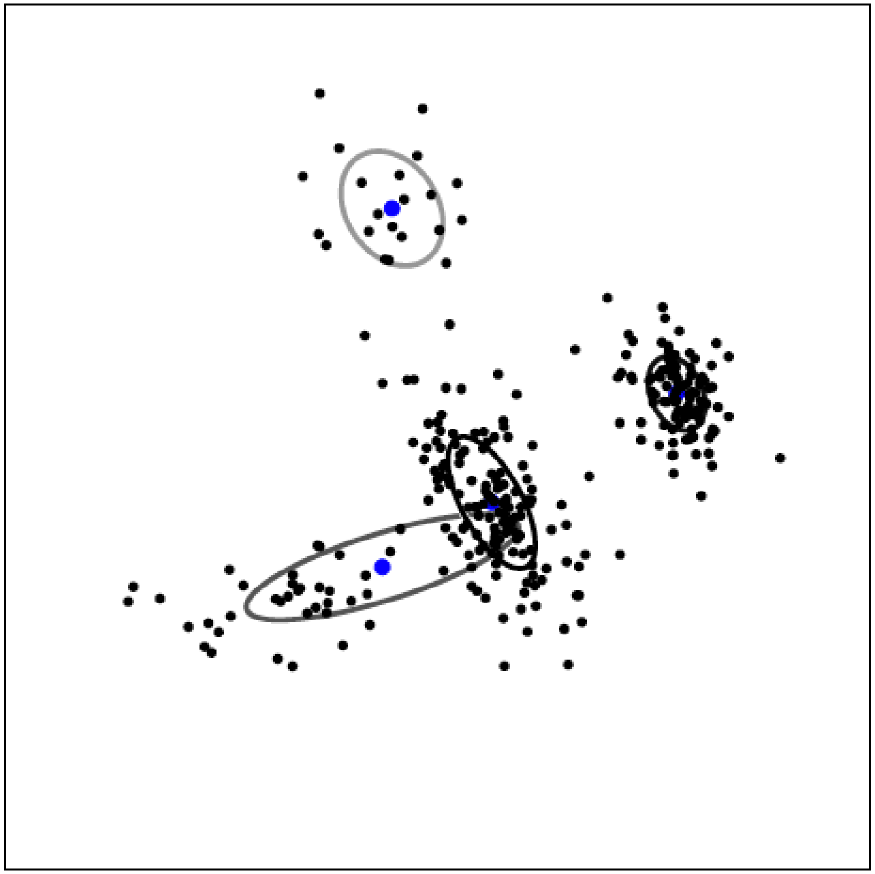
$$\theta_k^{(t)} \sim p(\theta_k \mid \{x_i \mid z_i^{(t)} = k\}, \lambda)$$

Standard Sampler: 2 Iterations

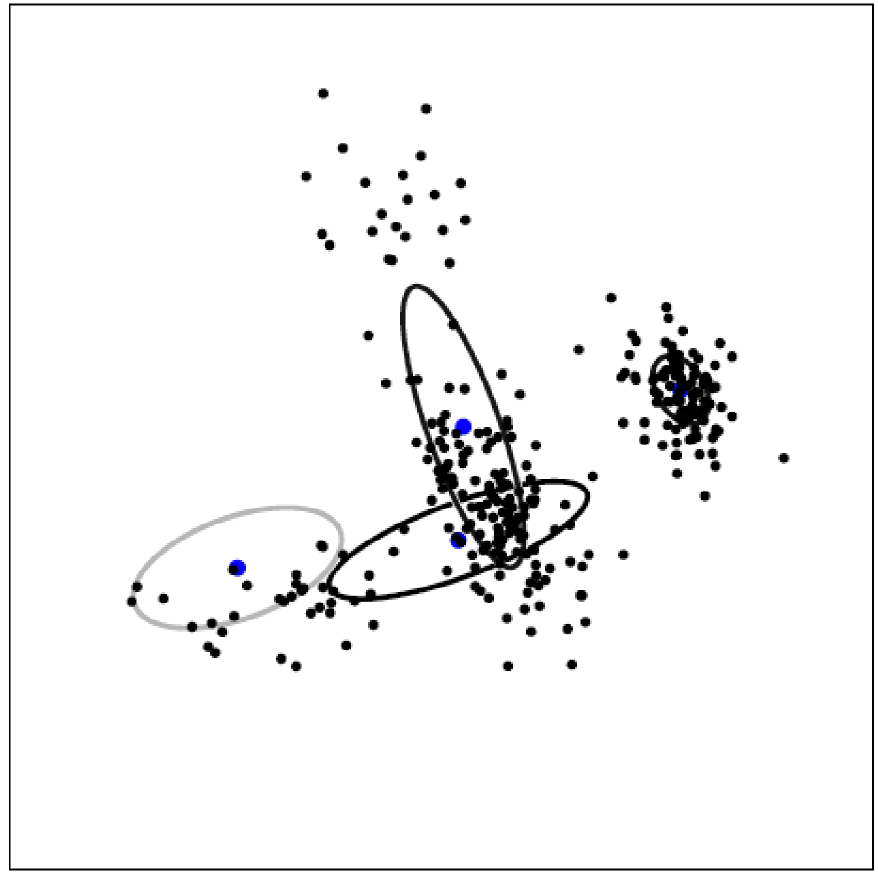

log p(x | π, θ) = –539.17


log p(x | π, θ) = –497.77

Standard Sampler: 10 Iterations

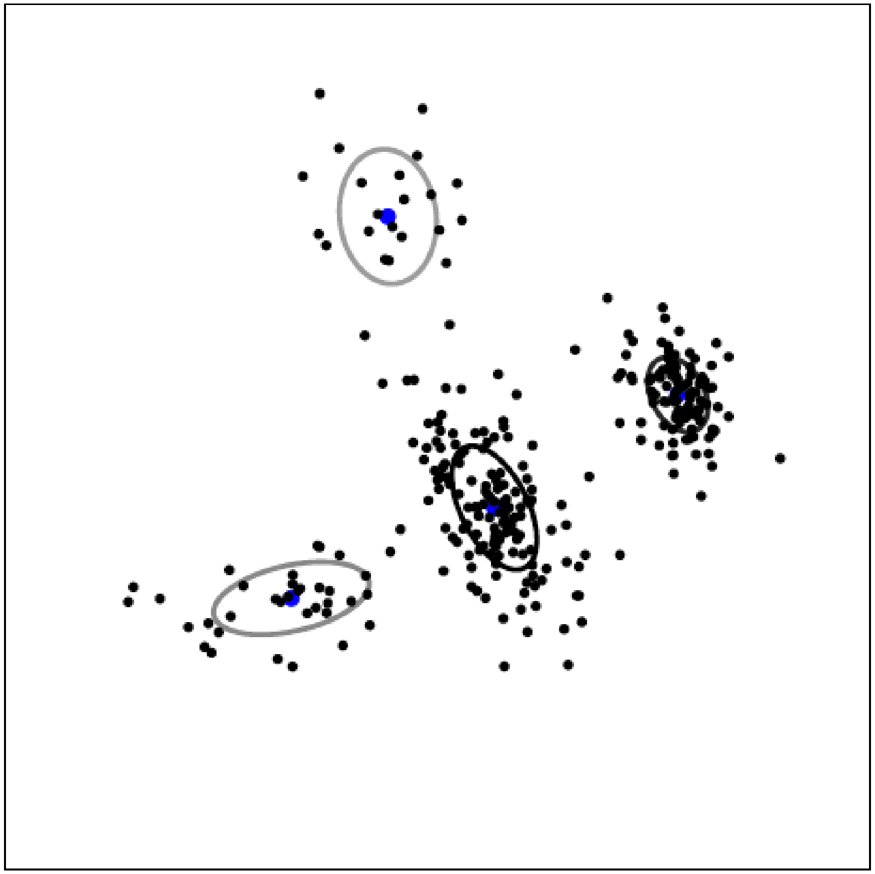


log p(x | π, θ) = –404.18

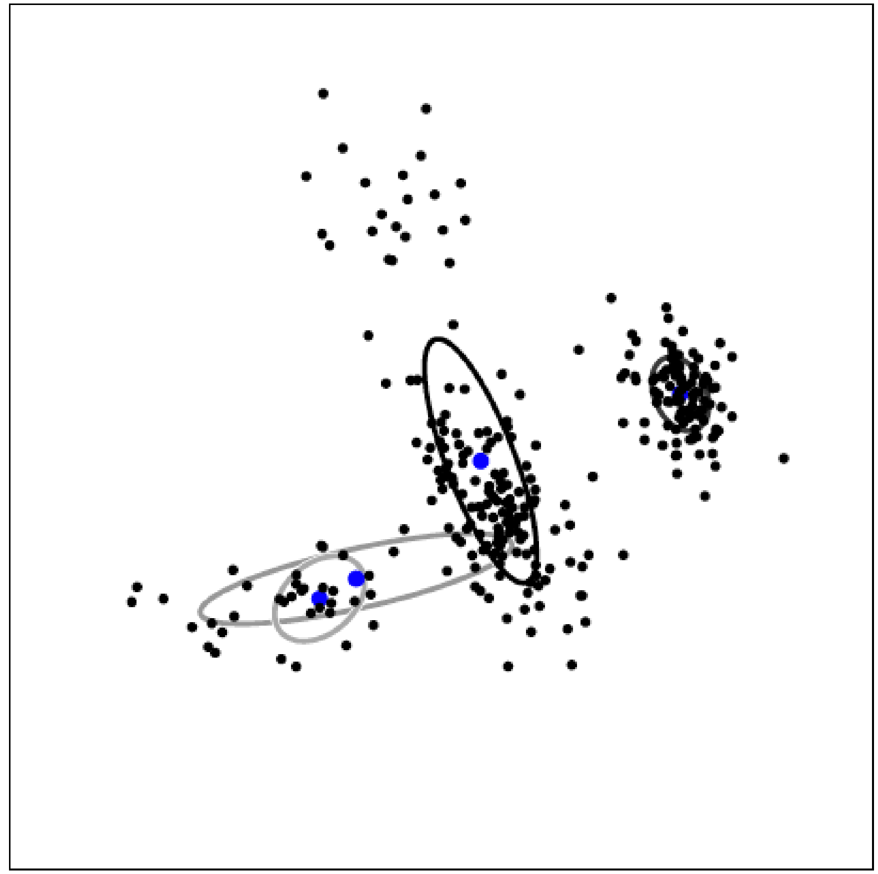


log p(x | π, θ) = –454.15

Standard Sampler: 50 Iterations



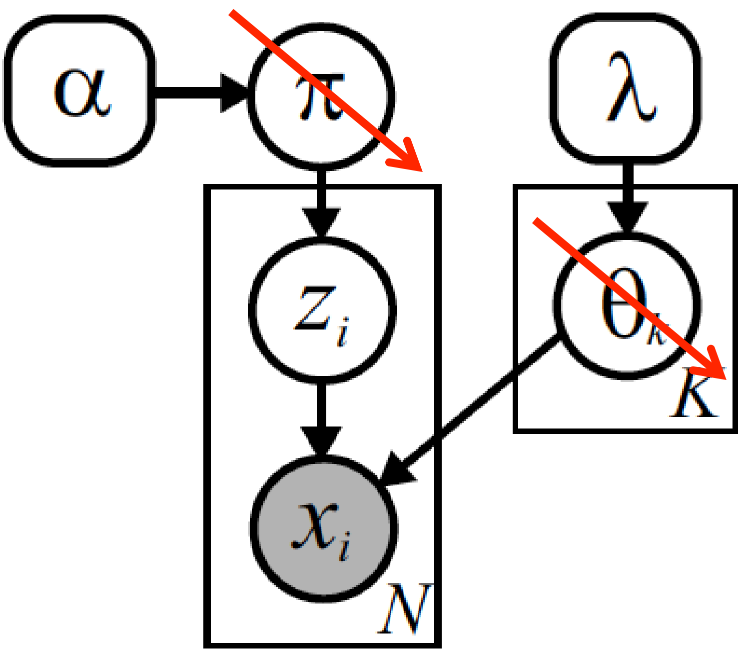
log p(x | π, θ) = –397.40

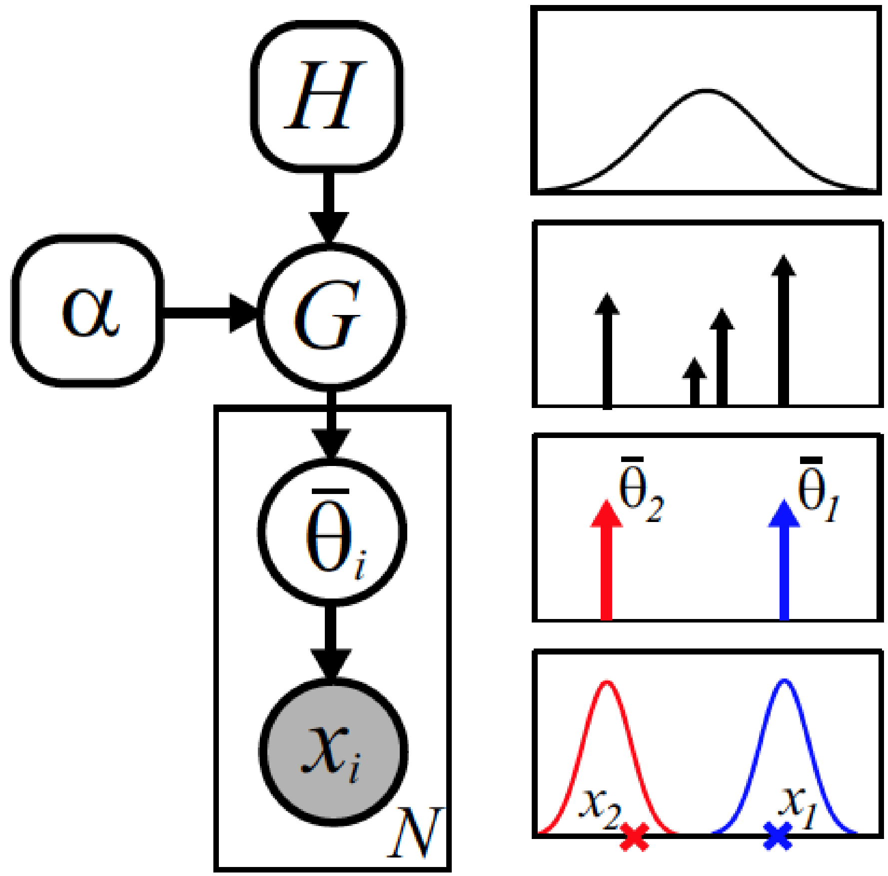


log p(x | π, θ) = –442.89

Collapsed Finite Bayesian Mixture

$$p(x \mid \pi, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k f(x \mid \theta_k)$$





- Conjugate priors allow analytic integration of some parameters
- Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)

Collapsed Finite Mixture Sampler

Given previous cluster assignments $z^{(t-1)}$, sequentially sample new assignments as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, sequentially resample z_i as follows:

   (a) For each of the K clusters, determine the predictive likelihood

   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$

   This likelihood can be computed from cached sufficient statistics

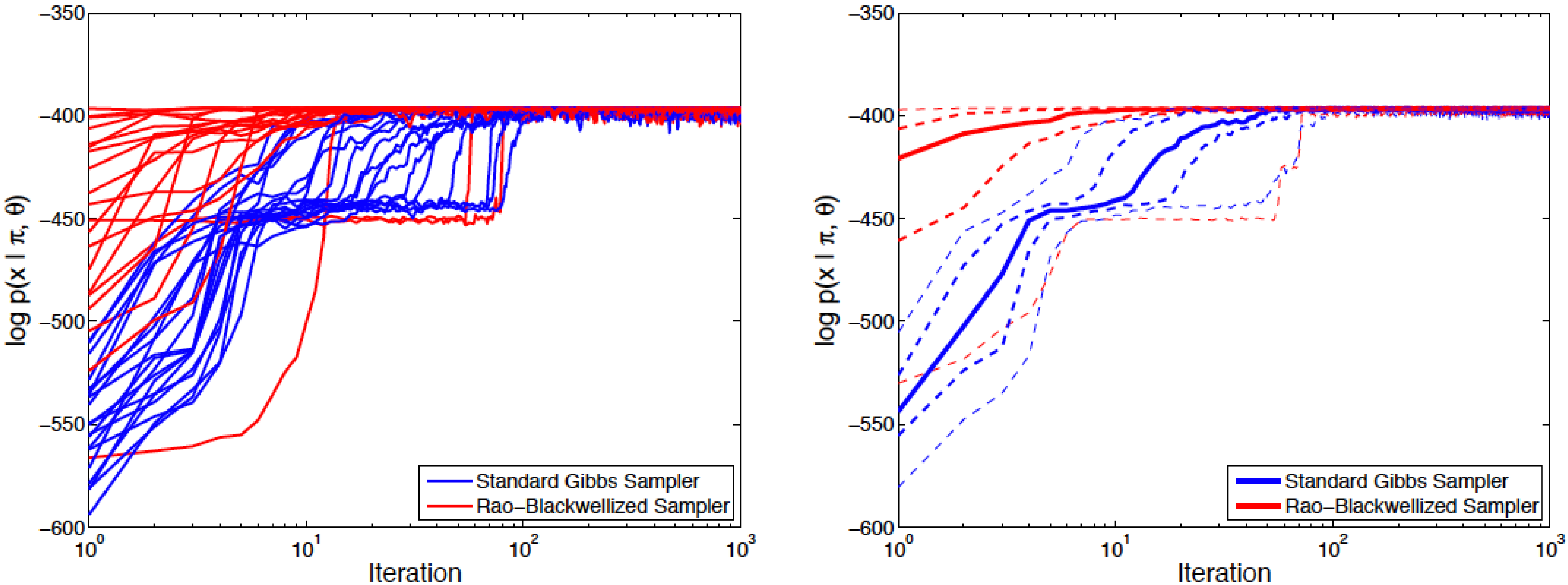
   (b) Sample a new cluster assignment z_i from the following multinomial distribution:

   $$z_i \sim \frac{1}{Z_i} \sum_{k=1}^{K} (N_k^{-i} + \alpha/K) f_k(x_i) \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} (N_k^{-i} + \alpha/K) f_k(x_i)$$

   N_k^{-i} is the number of other observations assigned to cluster k (see eq. (2.162)).

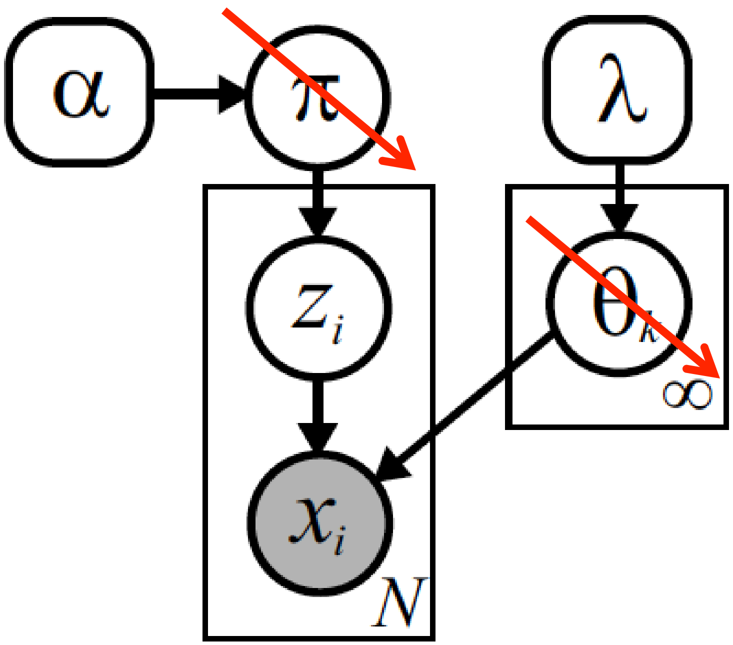
   (c) Update cached sufficient statistics to reflect the assignment of x_i to cluster z_i.
3. Set $z^{(t)} = z$. Optionally, mixture parameters may be sampled via steps 2–3 of Alg. 2.1.

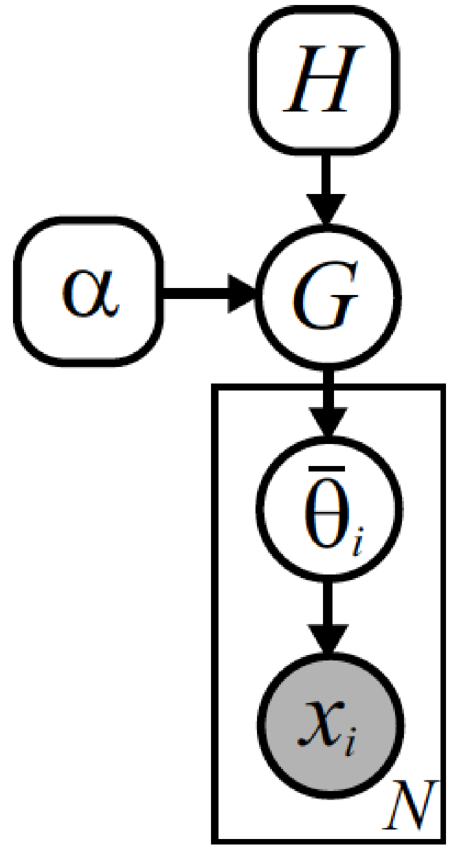
Standard versus Collapsed Samplers

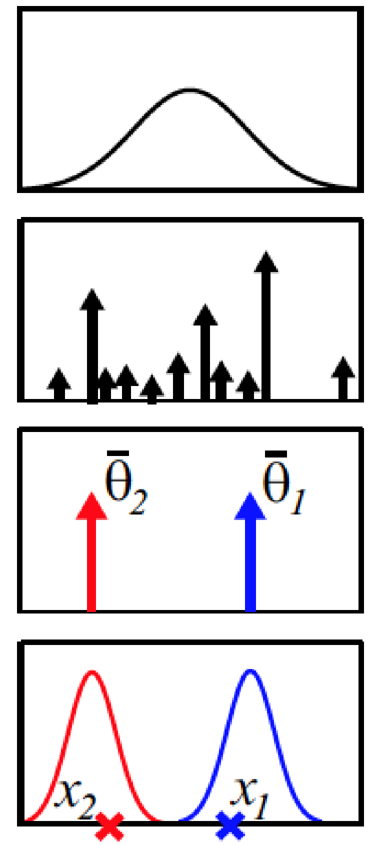


DP Mixture Models

$$p(x \mid \pi, \theta_1, \theta_2, \ldots) = \sum_{k=1}^{\infty} \pi_k f(x \mid \theta_k)$$







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

$$\pi \sim \mathrm{GEM}(\alpha)$$

$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

$$\bar{\theta}_i \sim G$$

$$x_i \sim F(\bar{\theta}_i)$$

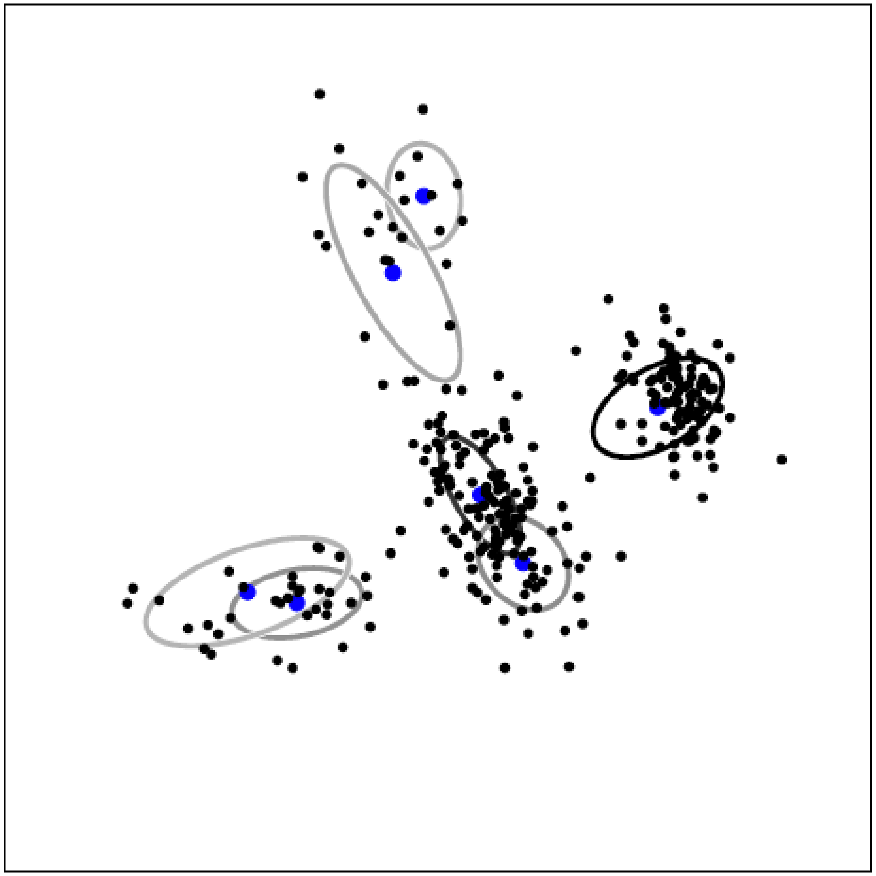
$$z_i \sim \pi$$

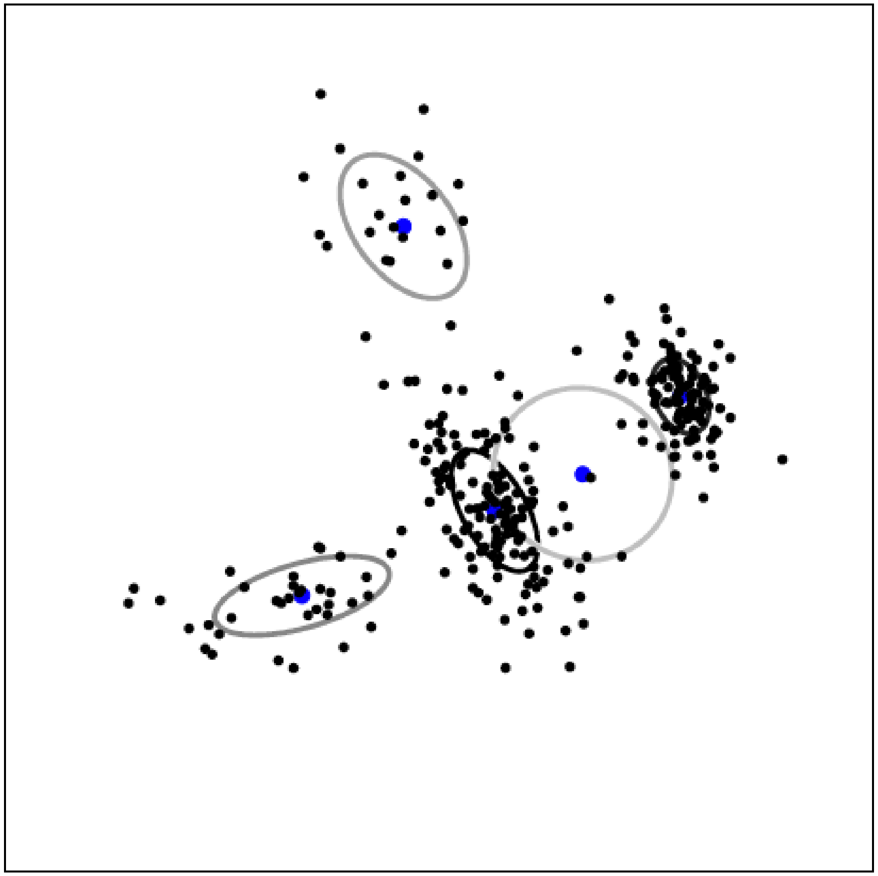
$$x_i \sim F(\theta_{z_i})$$

Collapsed DP Mixture Sampler

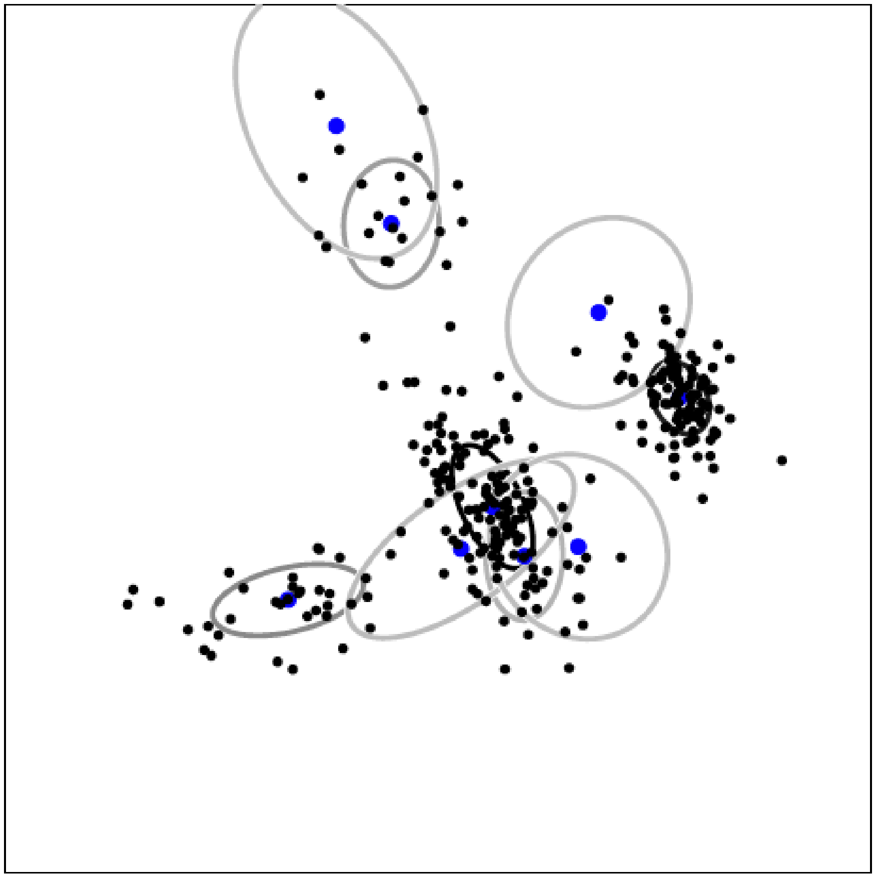
1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $\alpha = \alpha^{(t-1)}$ and $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, resample z_i as follows:
   (a) For each of the K existing clusters, determine the predictive likelihood
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   Also determine the likelihood $f_{\bar{k}}(x_i)$ of a potential new cluster $\bar{k}$
   $$p(x_i \mid \lambda) = \int_\Theta f(x_i \mid \theta)\, h(\theta \mid \lambda)\; d\theta$$
   (b) Sample a new cluster assignment z_i from the following $(K+1)$–dim. multinomial:
   $$z_i \sim \frac{1}{Z_i}\left(\alpha f_{\bar{k}}(x_i)\delta(z_i, \bar{k}) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)\delta(z_i, k)\right) \qquad Z_i = \alpha f_{\bar{k}}(x_i) + \sum_{k=1}^{K} N_k^{-i} f_k(x_i)$$
   N_k^{-i} is the number of other observations currently assigned to cluster k.
   (c) Update cached sufficient statistics to reflect the assignment of x_i to cluster z_i. If $z_i = \bar{k}$, create a new cluster and increment K.
3. Set $z^{(t)} = z$. Optionally, mixture parameters for the K currently instantiated clusters may be sampled as in step 3 of Alg. 2.1.
4. If any current clusters are empty ($N_k = 0$), remove them and decrement K accordingly.

Collapsed DP Sampler: 2 Iterations

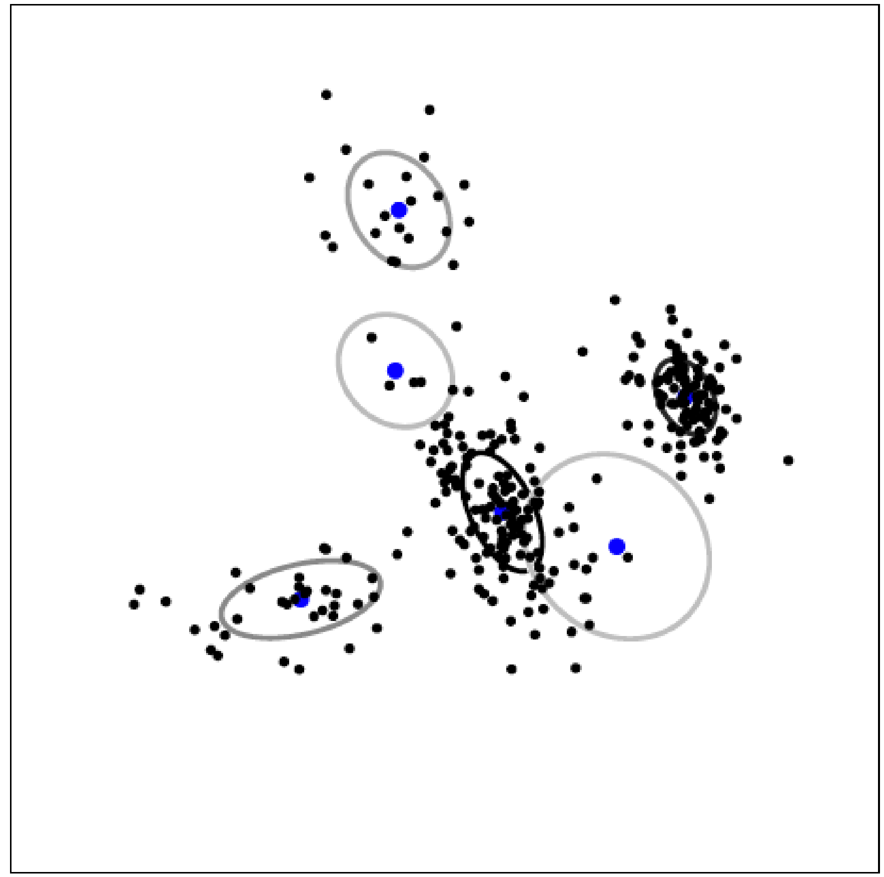

log p(x | π, θ) = –462.25


log p(x | π, θ) = –399.82

Standard Sampler: 10 Iterations

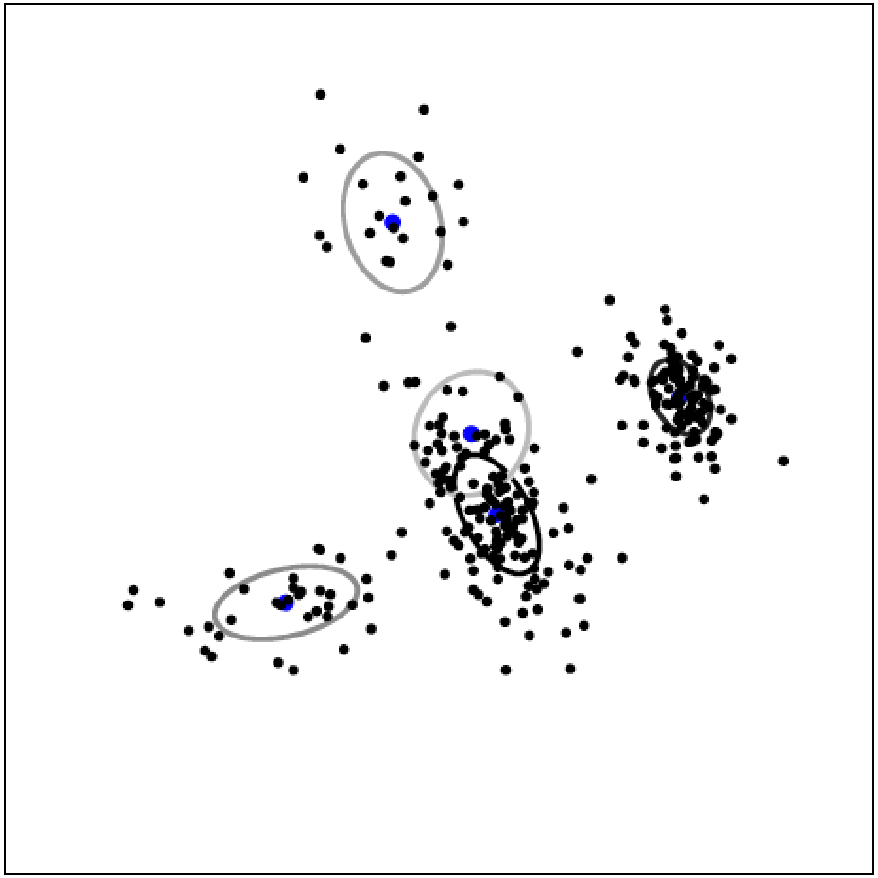


$\log p(x \mid \pi, \theta) = -398.32$

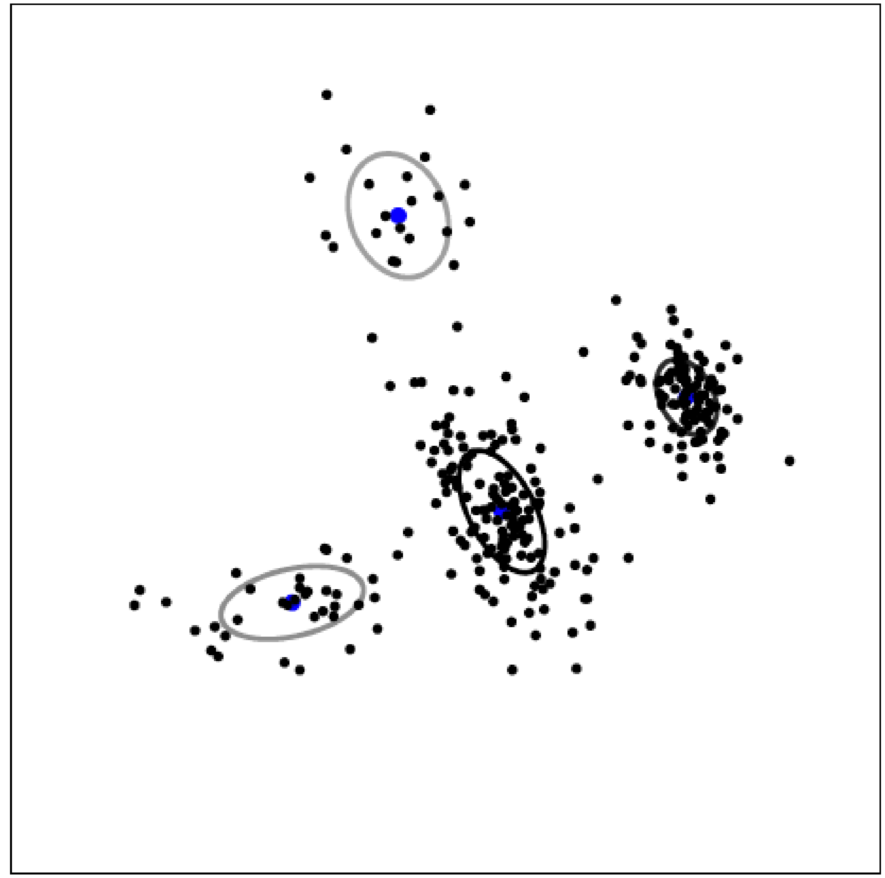


$\log p(x \mid \pi, \theta) = -399.08$

Standard Sampler: 50 Iterations

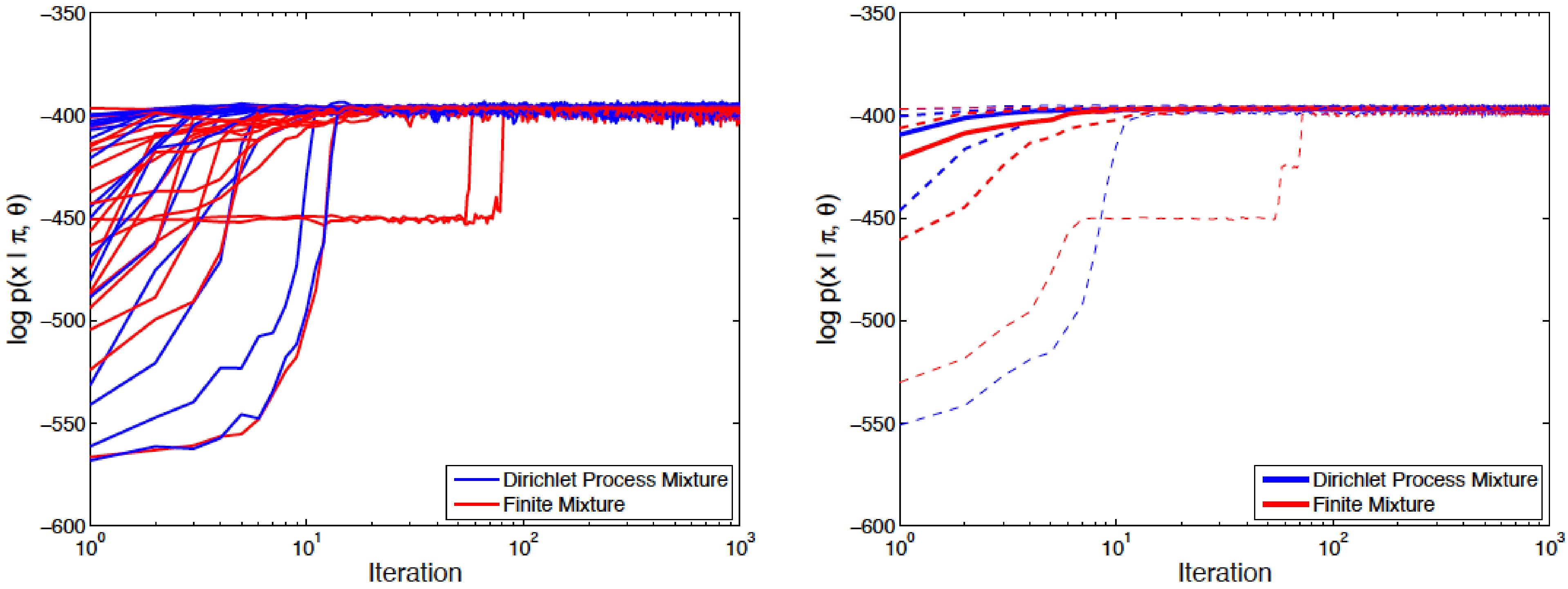


$\log p(x \mid \pi, \theta) = -397.67$

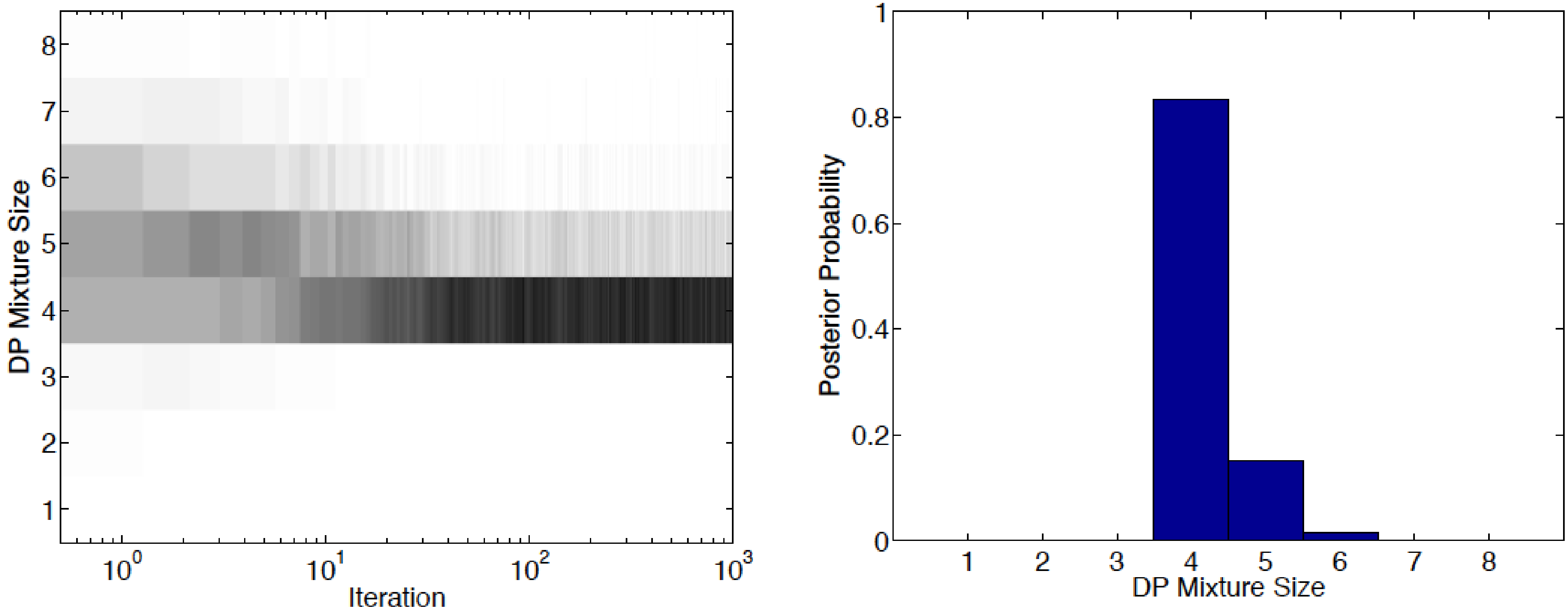


$\log p(x \mid \pi, \theta) = -396.71$

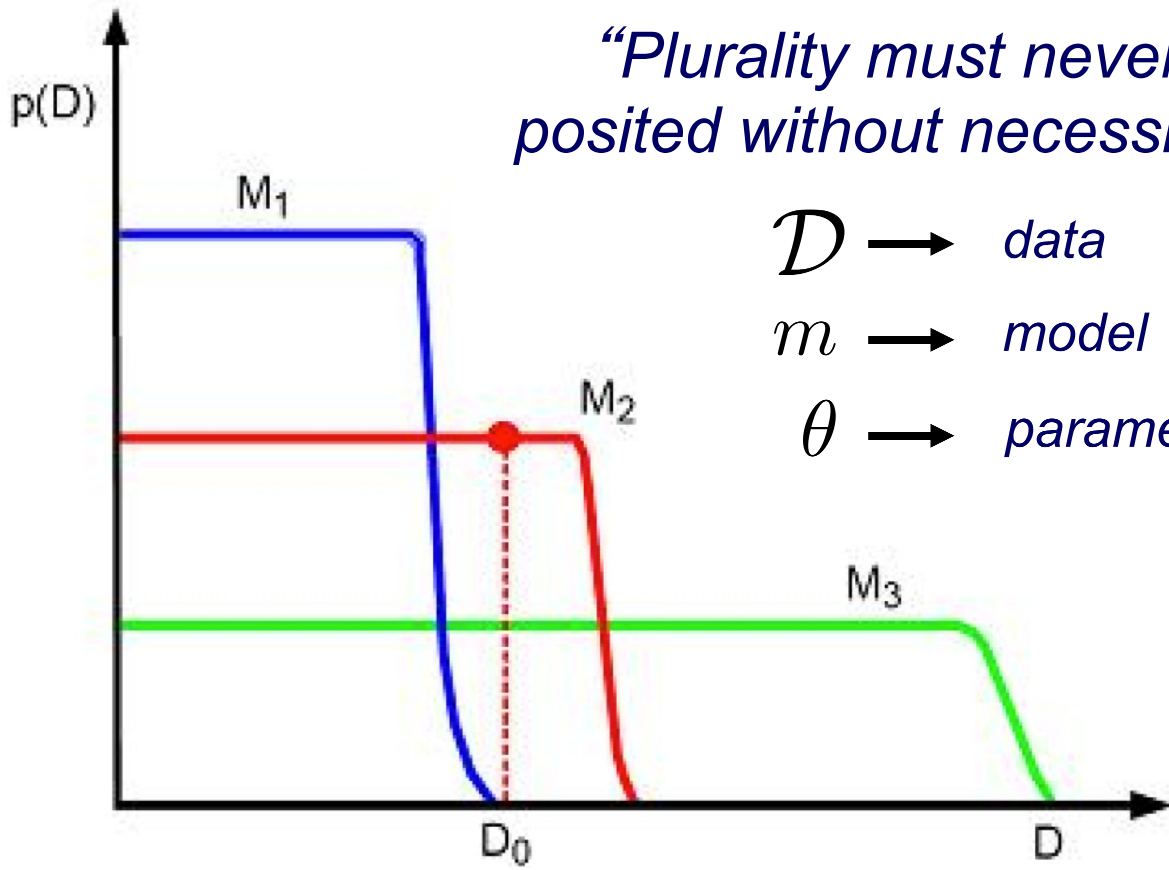
DP versus Finite Mixture Samplers



DP Posterior Number of Clusters



Bayesian Ockham's Razor



"Plurality must never be posited without necessity."

$\mathcal{D} \longrightarrow$ *data*

$m \longrightarrow$ *model*

$\theta \longrightarrow$ *parameters*

William of Ockham

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m, \mathcal{D})} \qquad p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

Even with uniform $p(m)$, marginal likelihood provides a model selection bias

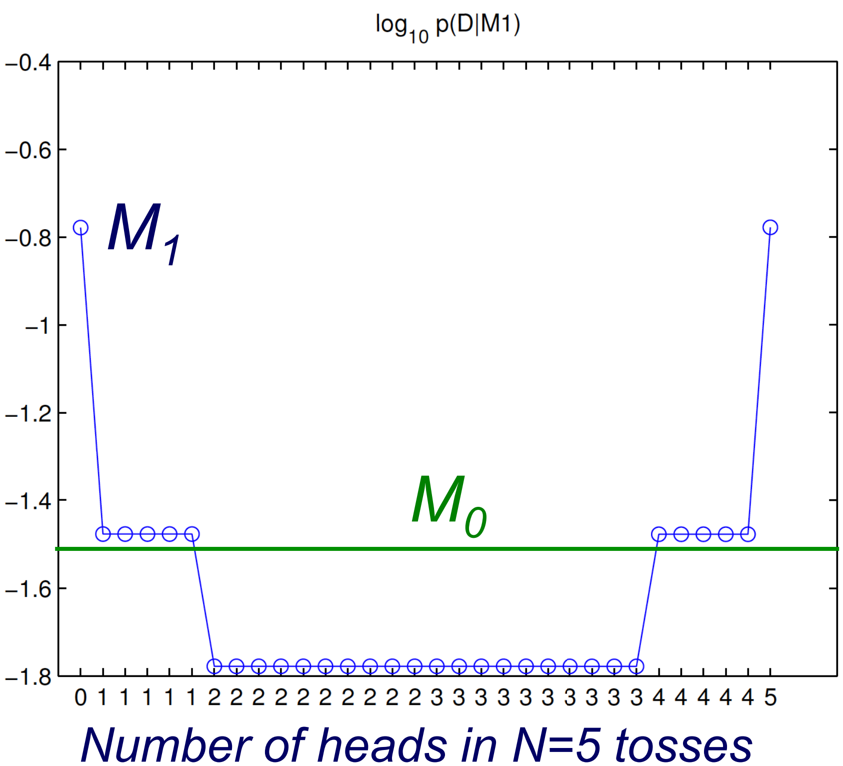
Example: Is this coin fair?

M_0: Tosses are from a fair coin: $\theta = 1/2$

M_1: Tosses are from a coin of unknown bias: $\theta \sim \mathrm{Unif}(0,1)$

Marginal Likelihoods

$$p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N \qquad p(\mathcal{D}|M_1) = \int p(\mathcal{D}|\theta)p(\theta)d\theta = \frac{B(\alpha_1+N_1,\alpha_0+N_0)}{B(\alpha_1,\alpha_0)}$$



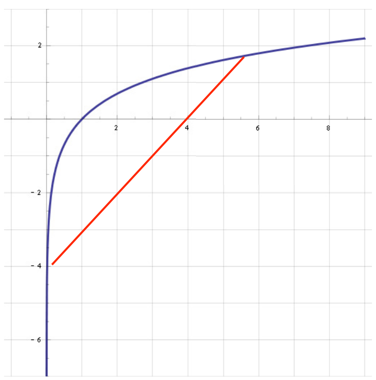
$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- *ML: Always prefer M_1*
- *Bayes: Unbalanced counts are much more likely with a biased coin, so favor M_1*
- *Bayes: Balanced counts only happen with some biased coins, so favor M_0*

Variational Approximations

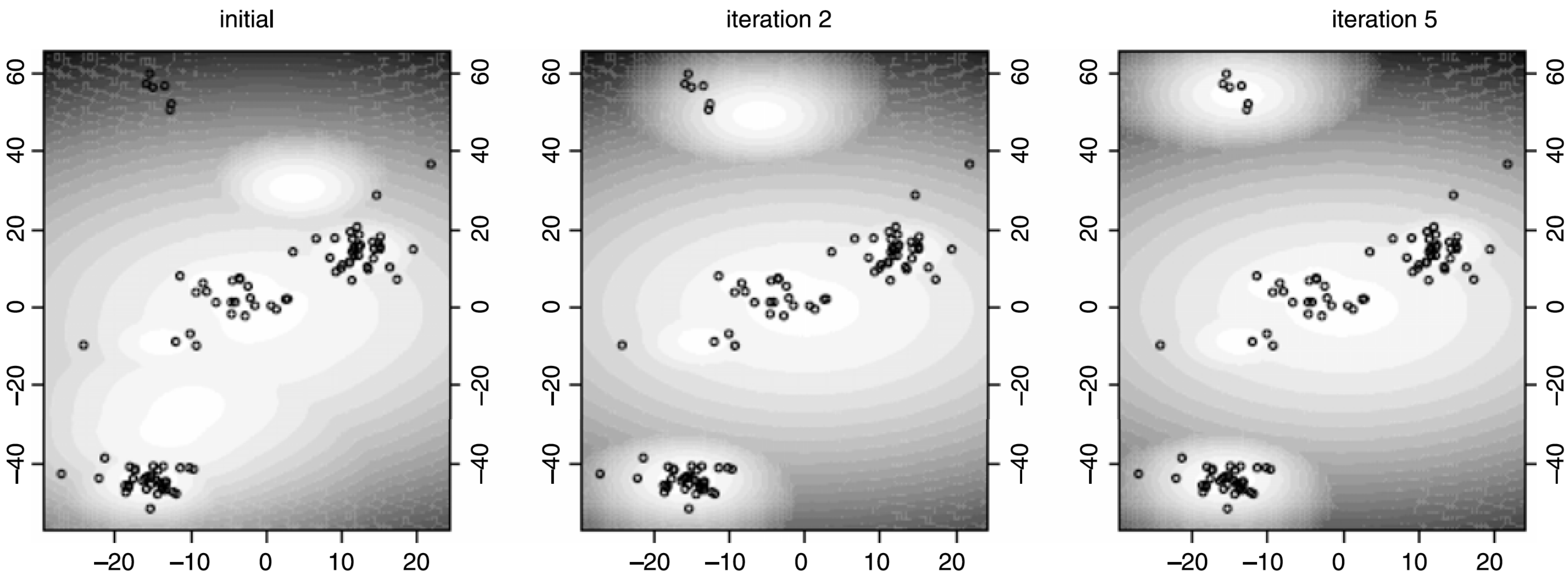
$$D(q(x) \,||\, p(x \mid y)) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

$$\begin{aligned}
\log p(y) &= \log \sum_x p(x, y) \\
&= \log \sum_x q(x) \frac{p(x, y)}{q(x)} && \textit{(Multiply by one)} \\
&\geq \underbrace{\sum_x q(x) \log \frac{p(x, y)}{q(x)}}_{= -D(q(x) \,||\, p(x \mid y)) + \log p(y)} && \textit{(Jensen's inequality)}
\end{aligned}$$



- Minimizing KL divergence maximizes a likelihood bound
- *Variational EM algorithms*, which maximize for *q(x)* within some tractable family, retain BNP model selection behavior

Mean Field for DP Mixtures



Blei & Jordan 2006

- Truncate stick-breaking at some upper bound *K* on the true number of occupied clusters:

$$\beta_k \sim \text{Beta}(1, \alpha) \qquad \pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \qquad k = 1, \ldots, K-1$$

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$$

- Priors encourage assigning data to fewer than *K* clusters

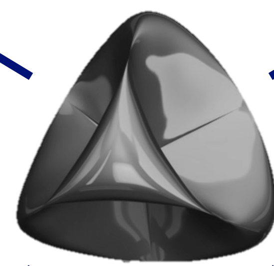
MCMC & Variational Learning

Infinite Stochastic Processes

Conceptually useful, but usually impractical or impossible for learning algorithms.

CRP & IBP

Tractably learn via finite summaries of true, infinite model.



Stick-Breaking

Truncate stick-breaking to produce provably accurate approximation.

Finite Bayesian Models

Set finite model order to be larger than expected number of clusters or features.

Applied BNP: Part II