Applied Bayesian Nonparametrics

Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

September 22: Dirichlet Processes Continued, MCMC for DP Mixture Models

Finite Dirichlet Distributions

$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \qquad \alpha_k > 0$$

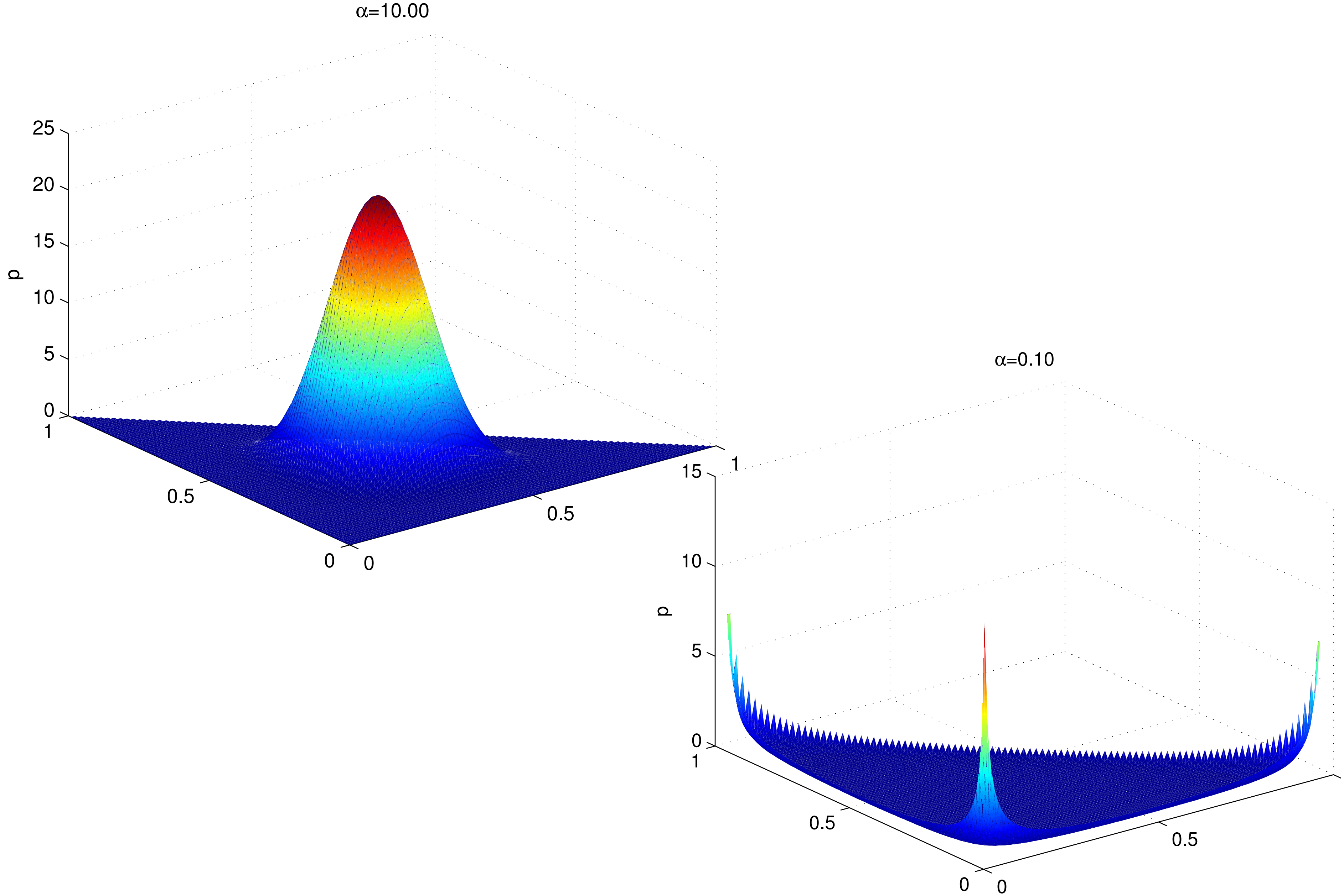
$$\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0} \qquad \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k$$

$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)} \qquad \alpha_k = \frac{\alpha_0}{K}$$

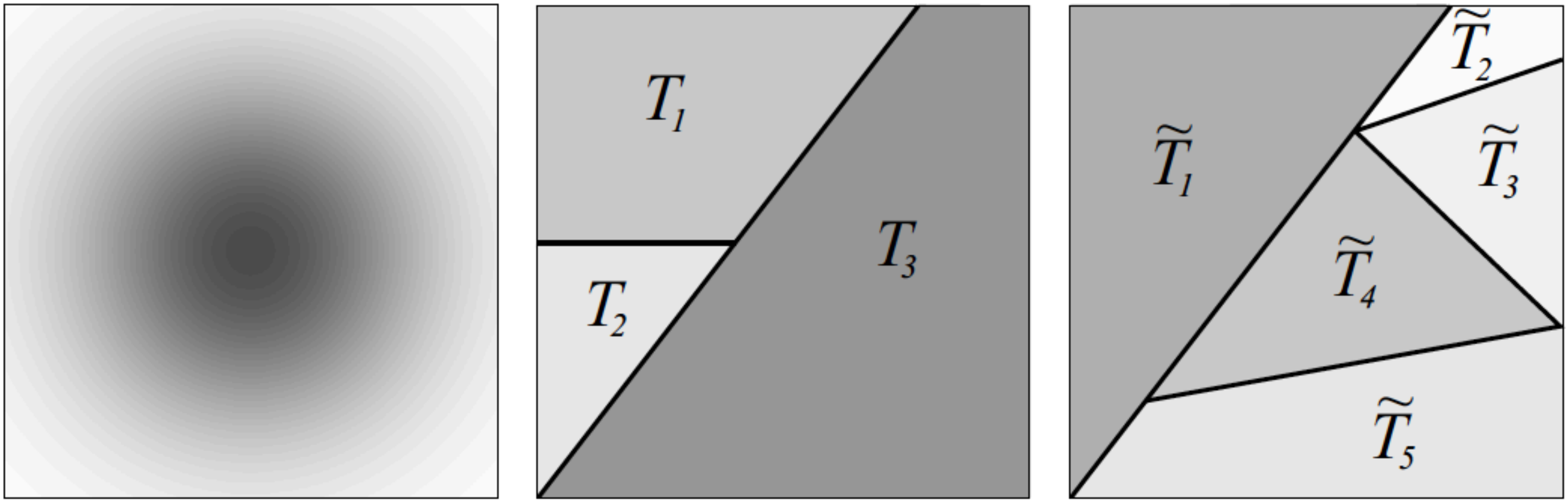
- Beta distribution is special case where K=2:

$$p(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \qquad \alpha, \beta > 0$$

Dirichlet Distributions



Dirichlet Processes



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

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:

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

The Stick-Breaking Construction: DP Realizations are Discrete

Theorem 2.5.3. *Let* $\pi = \{\pi_k\}_{k=1}^{\infty}$ *be an infinite sequence of mixture weights derived from the following stick–breaking process, with parameter* $\alpha > 0$*:*

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad k = 1, 2, \ldots \tag{2.174}$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) = \beta_k \left(1 - \sum_{\ell=1}^{k-1} \pi_\ell\right) \tag{2.175}$$

Given a base measure H *on* Θ*, consider the following discrete random measure:*

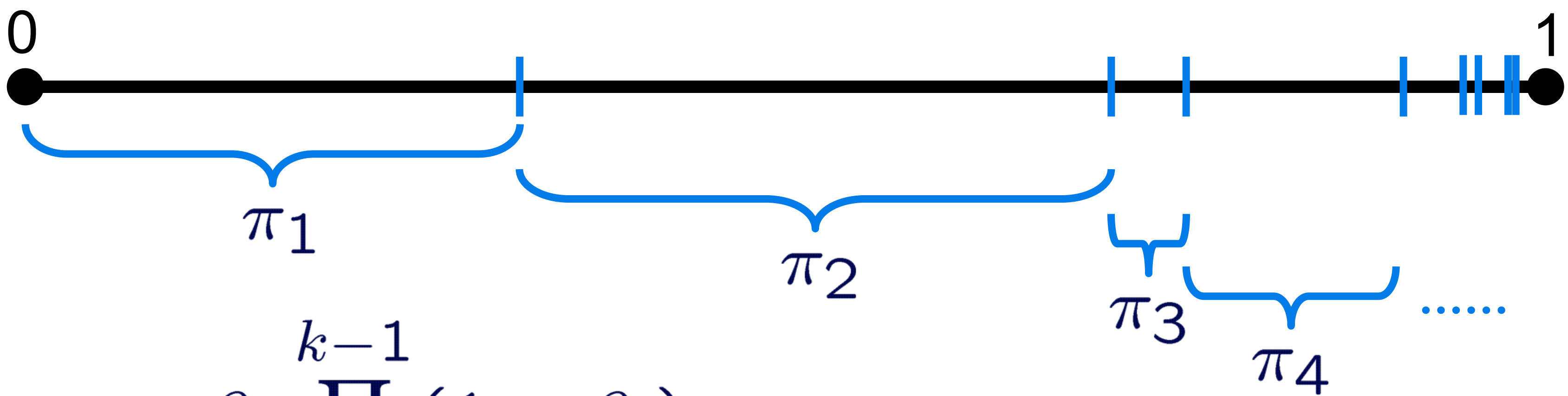
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \theta_k \sim H \tag{2.176}$$

This construction guarantees that $G \sim \mathrm{DP}(\alpha, H)$*. Conversely, samples from a Dirichlet process are discrete with probability one, and have a representation as in eq.* (2.176).

Dirichlet Process Mixtures

$$p(x) = \sum_{k=1}^{\infty} \pi_k f\left(x \mid \theta_k\right)$$

Dirichlet processes define a prior distribution on weights assigned to mixture components:



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

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

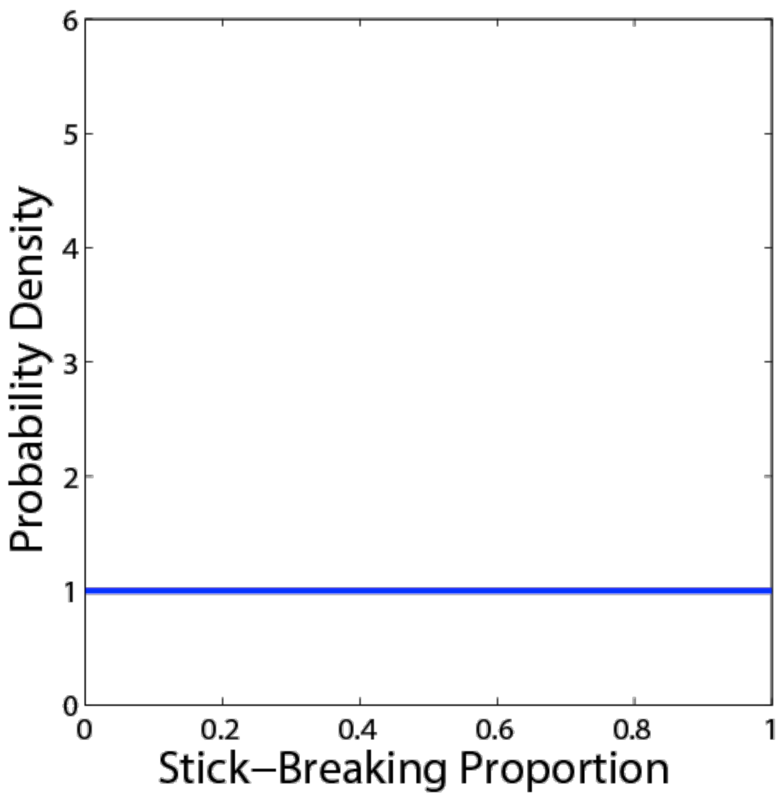
α → concentration parameter

Stick-Breaking Construction: *Sethuraman, 1994*

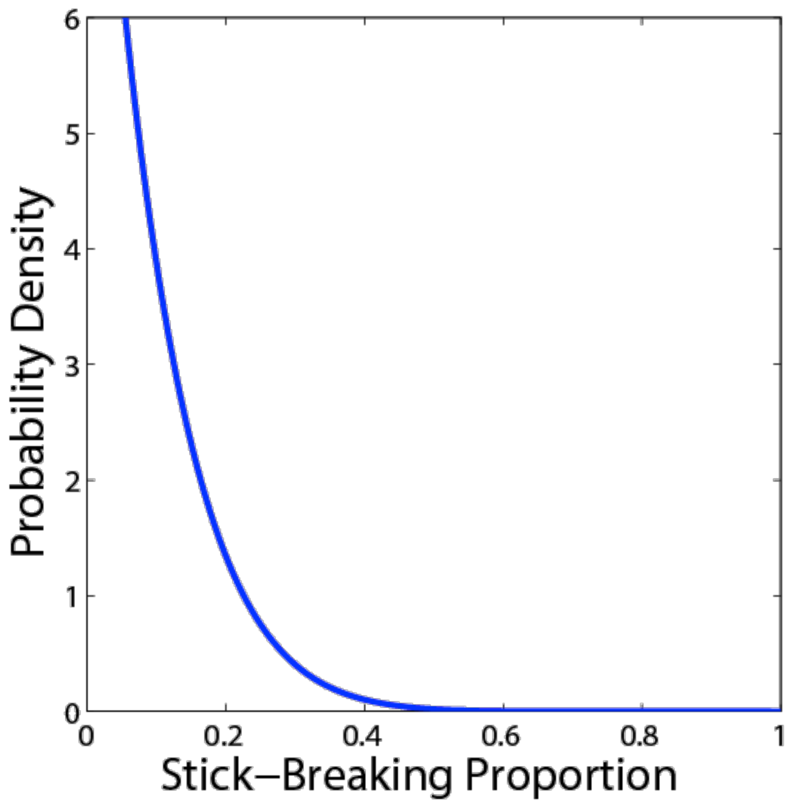
Dirichlet Stick-Breaking

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

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



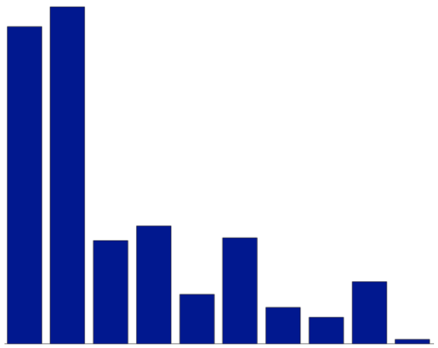
$\alpha = 1$



$\alpha = 10$

Why the Dirichlet Process?

$$p(x) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x \mid 0, \Lambda_k)$$



Nonparametric $\neq$ No Parameters

- Model complexity grows as data observed:
  - Small training sets give *simple, robust* predictions
  - Reduced sensitivity to prior assumptions

Flexible but Tractable

- Literature showing attractive *asymptotic properties*
- Leads to simple, effective *computational methods*
  - Avoids challenging model selection issues

Ferguson 1973; Sethuraman 1994

DPs and Polya Urns

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*

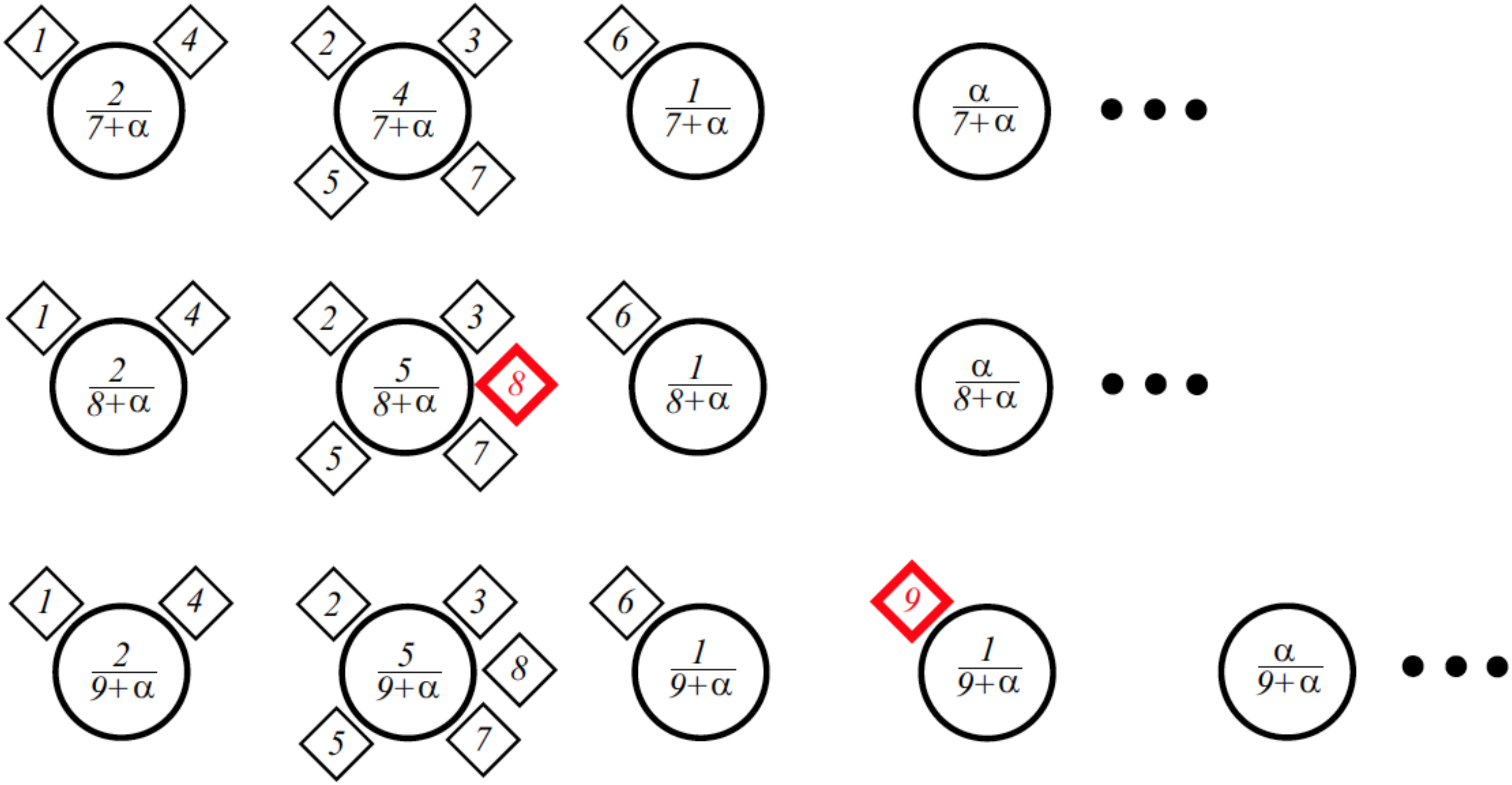
$$p(\bar{\theta}_{N+1} = \theta \mid \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, H) = \frac{1}{\alpha + N}\left(\alpha h(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k)\right) \qquad (2.180)$$

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

My 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 sand
- Repeat this process…

Chinese Restaurant Process



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

Some Informal Intuition

$$(\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

$$z_i \sim \pi$$

$$p(z_i = k \mid z_{\setminus i}, \alpha) = \frac{N_k^{-i} + \alpha/K}{\alpha + N - 1} \qquad N_k^{-i} = \sum_{j \neq i} \delta(z_j, k)$$

$$\lim_{K \to \infty} p(z_i = k \mid z_{\setminus i}, \alpha) = \frac{N_k^{-i}}{\alpha + N - 1}$$

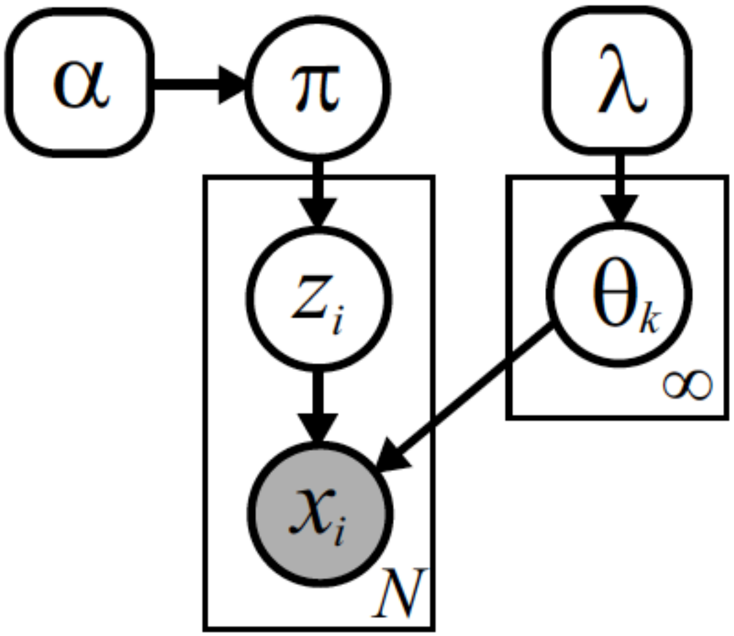
$$p(z_i \neq z_j \text{ for all } j \neq i \mid z_{\setminus i}, \alpha) = 1 - \sum_{k \mid N_k^{-i} > 0} p(z_i = k \mid z_{\setminus i}, \alpha)$$

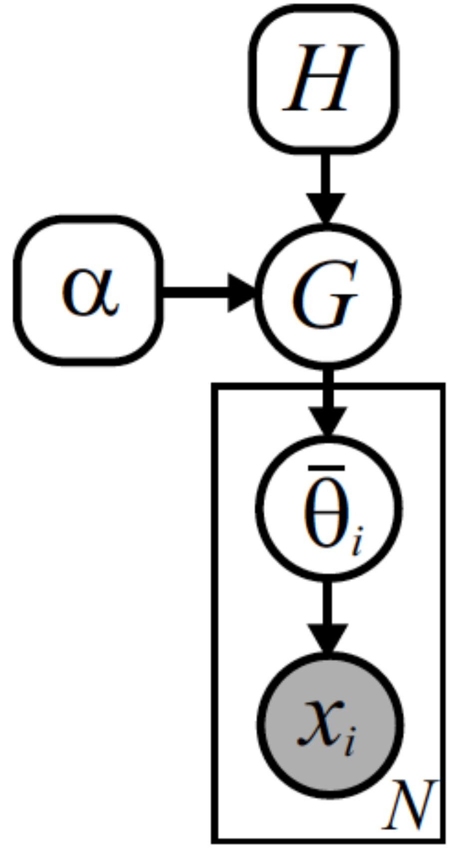
$$\lim_{K \to \infty} p(z_i \neq z_j \text{ for all } j \neq i \mid z_{\setminus i}, \alpha) = 1 - \sum_k \frac{N_k^{-i}}{\alpha + N - 1} = \frac{\alpha}{\alpha + N - 1}$$

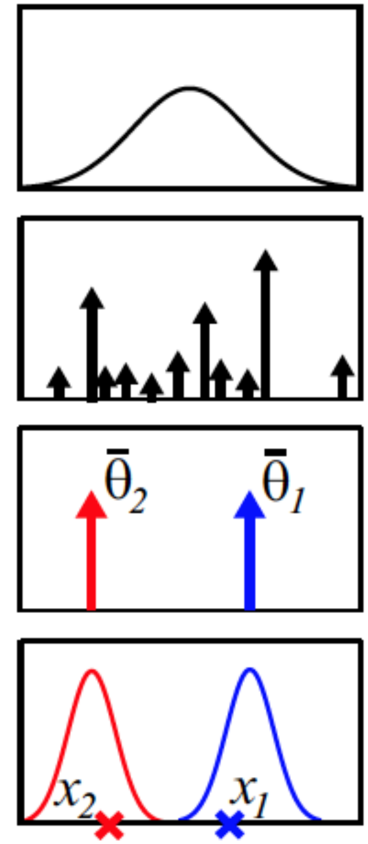
What does this get wrong? Indicators versus partitions…

DP Mixture Models

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







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

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

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

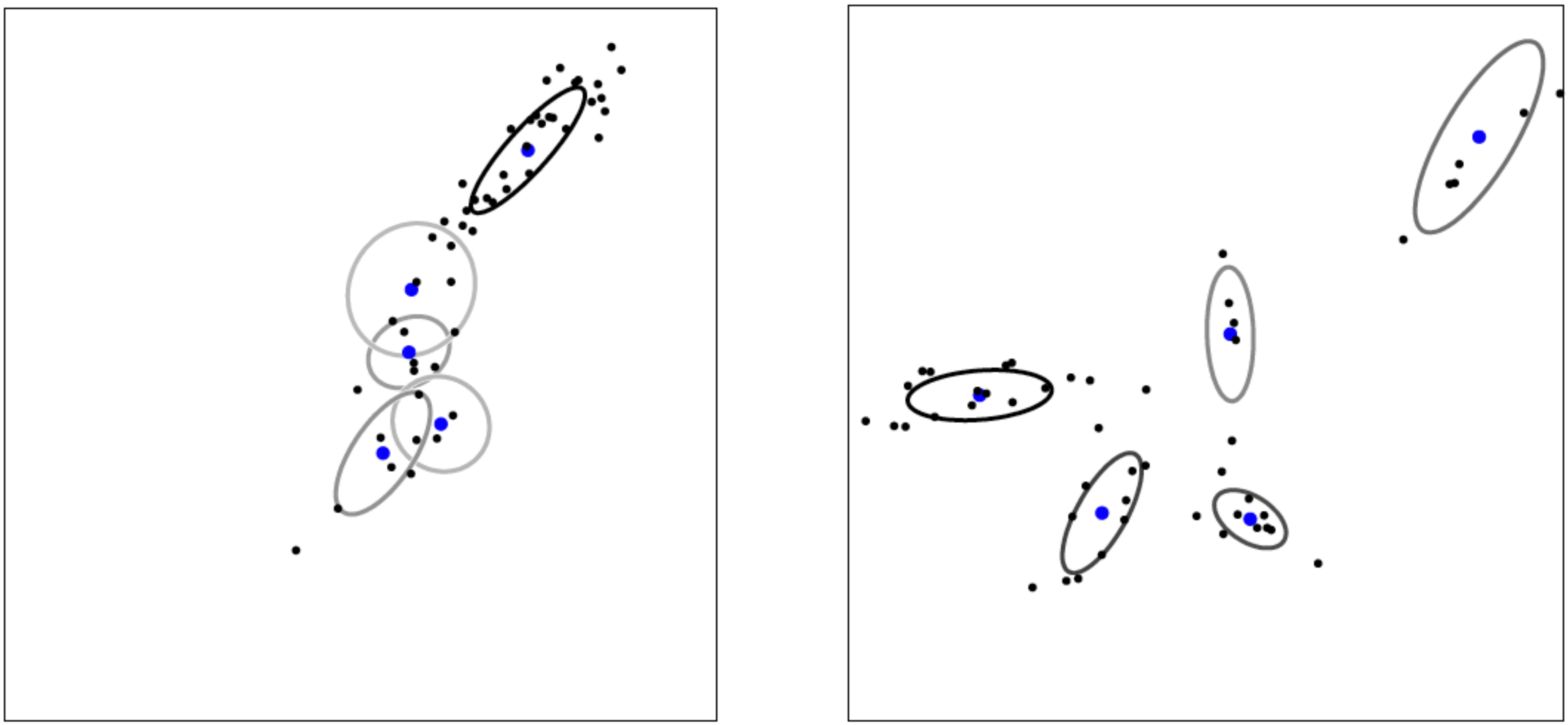
$$\bar{\theta}_i \sim G$$

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

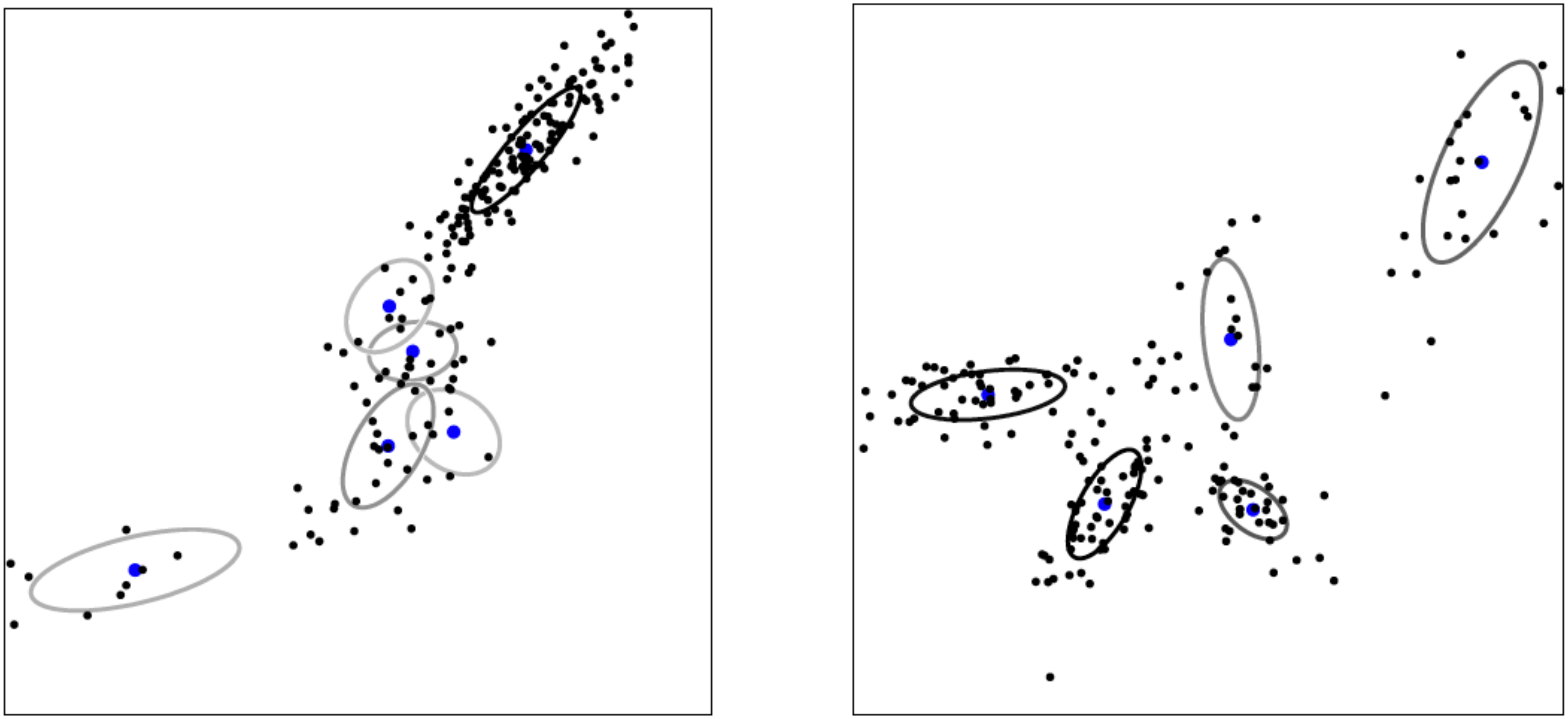
$$z_i \sim \pi$$

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

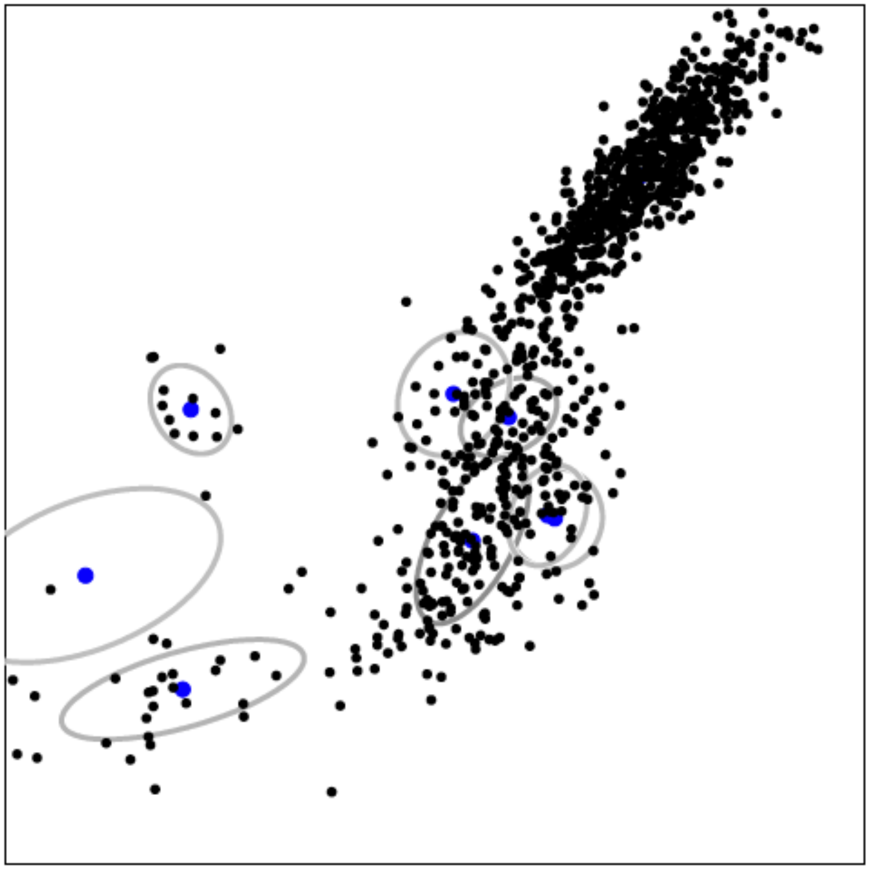
Samples from DP Mixture Priors

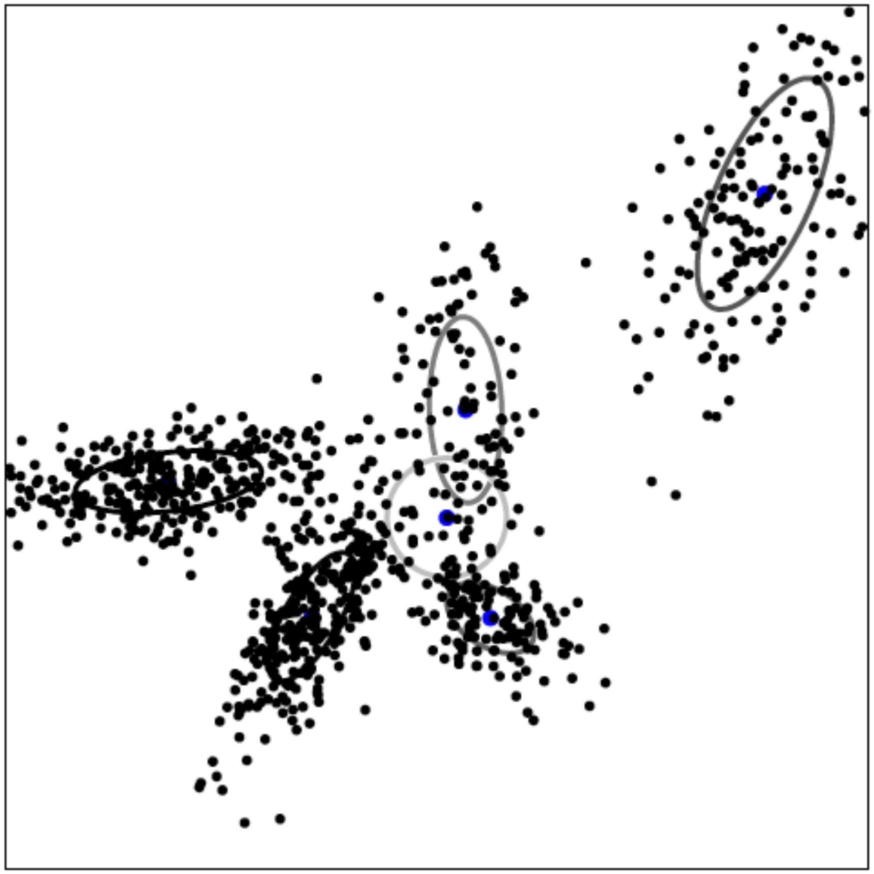


Samples from DP Mixture Priors



Samples from DP Mixture Priors





Views of the Dirichlet Process

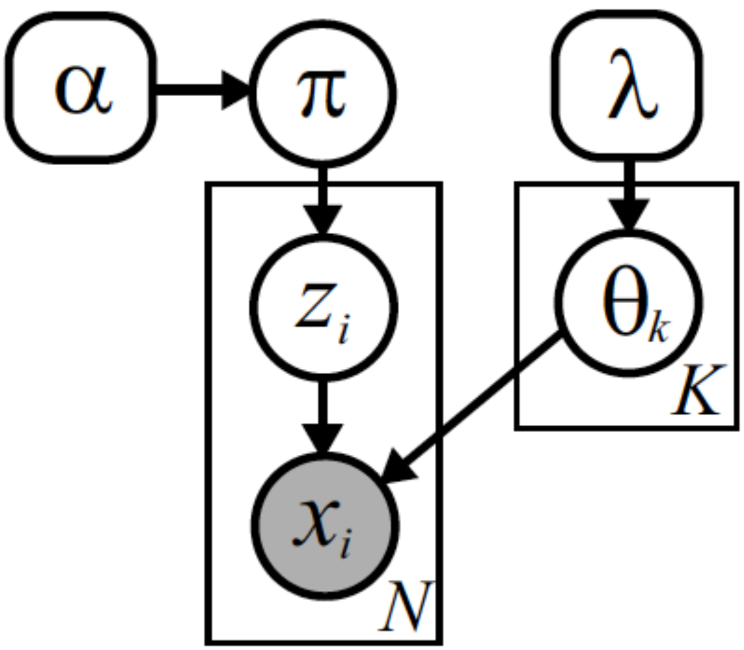
- Implicit stochastic process: Finite Dirichlet marginals
- *Explicit stochastic process: Normalized gamma process*
- Explicit discrete measure: Stick-breaking construction
- Marginalized predictions: Polya urn, or (almost) equivalently the Chinese restaurant process

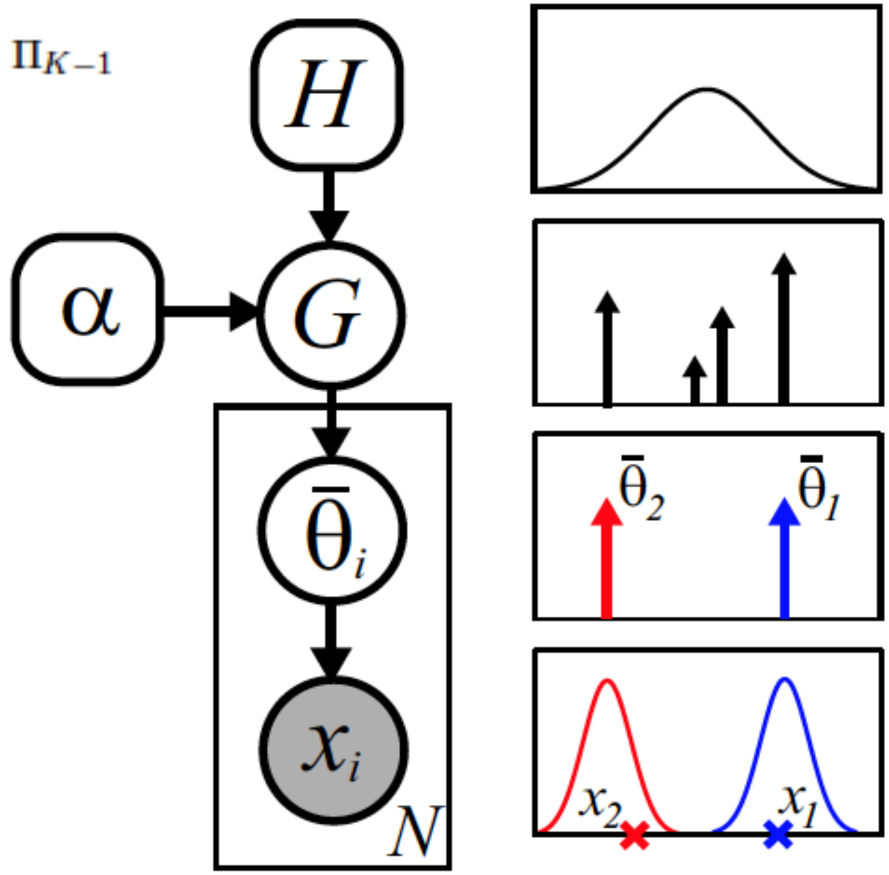
Later in this course:

- Modeling: Generalize one of these representations, to get a fancier (but usually less tractable) process
- Inference: Deal with infinite-dimensional processes by analytic integration, or finite truncation (static or dynamic)

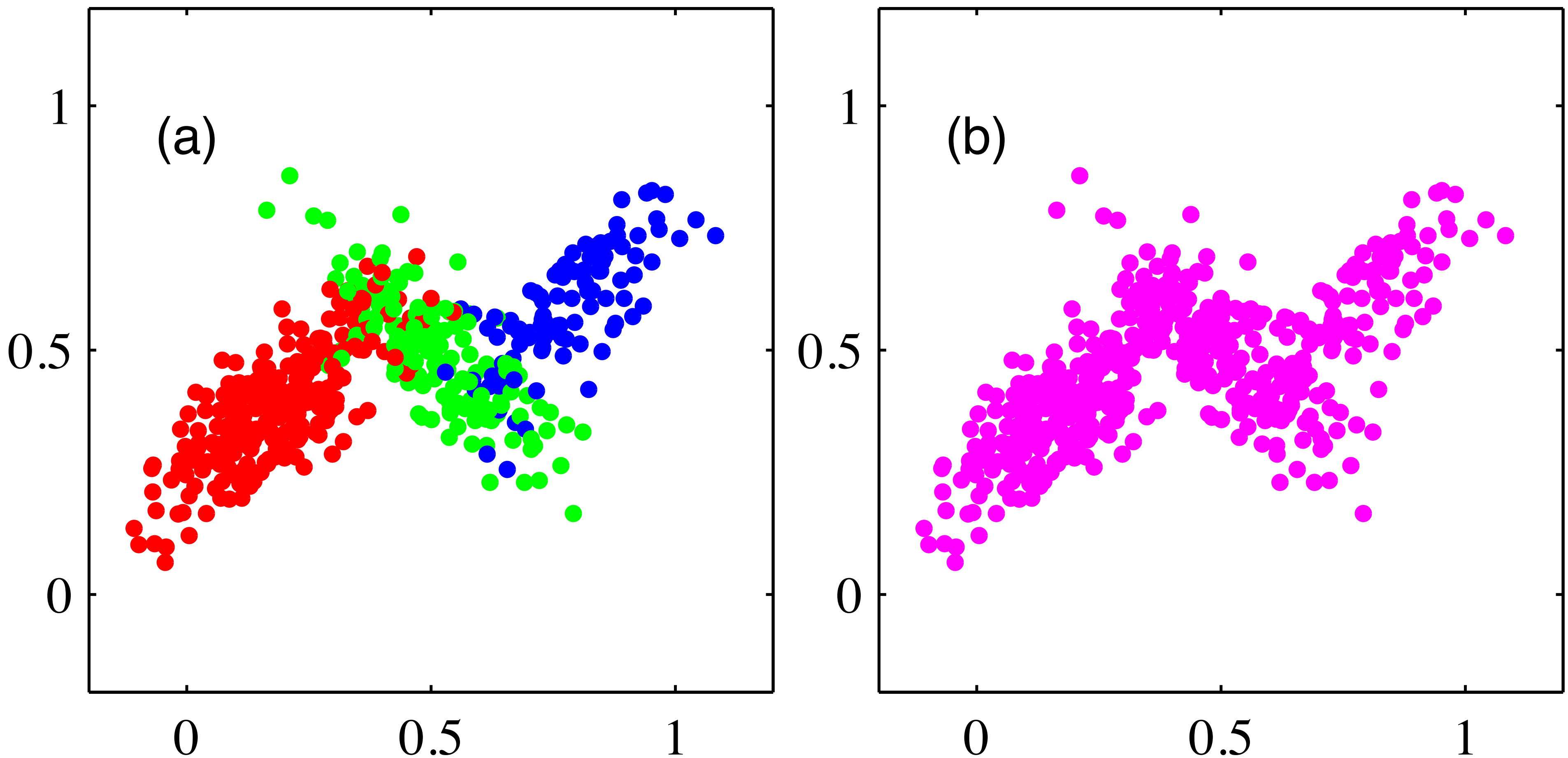
Finite Bayesian Mixture Models

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





Fitting Finite Gaussian Mixtures

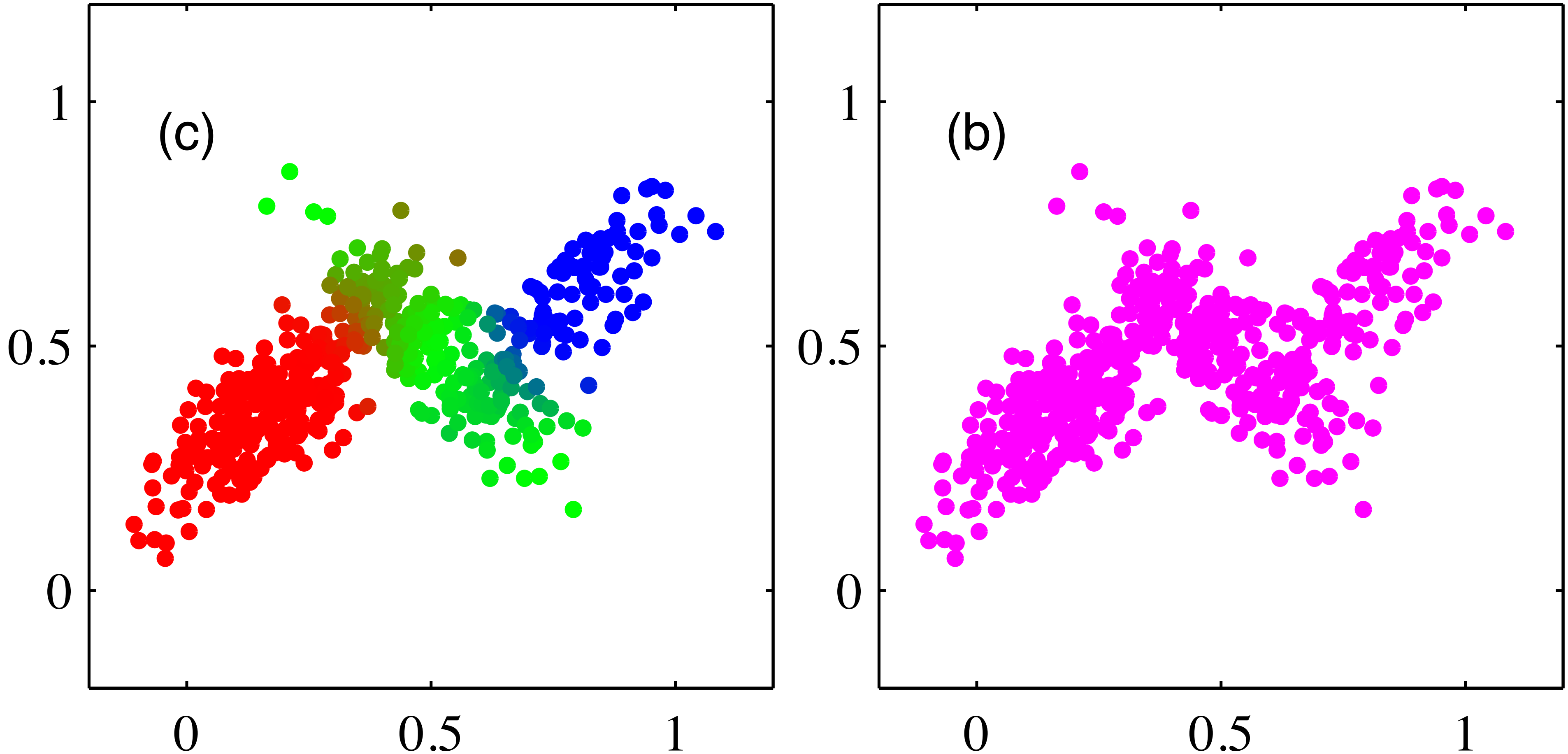


Complete Data Labeled by True Cluster Assignments

Incomplete Data: Points to be Clustered

C. Bishop, Pattern Recognition & Machine Learning

Posterior Assignment Probabilities

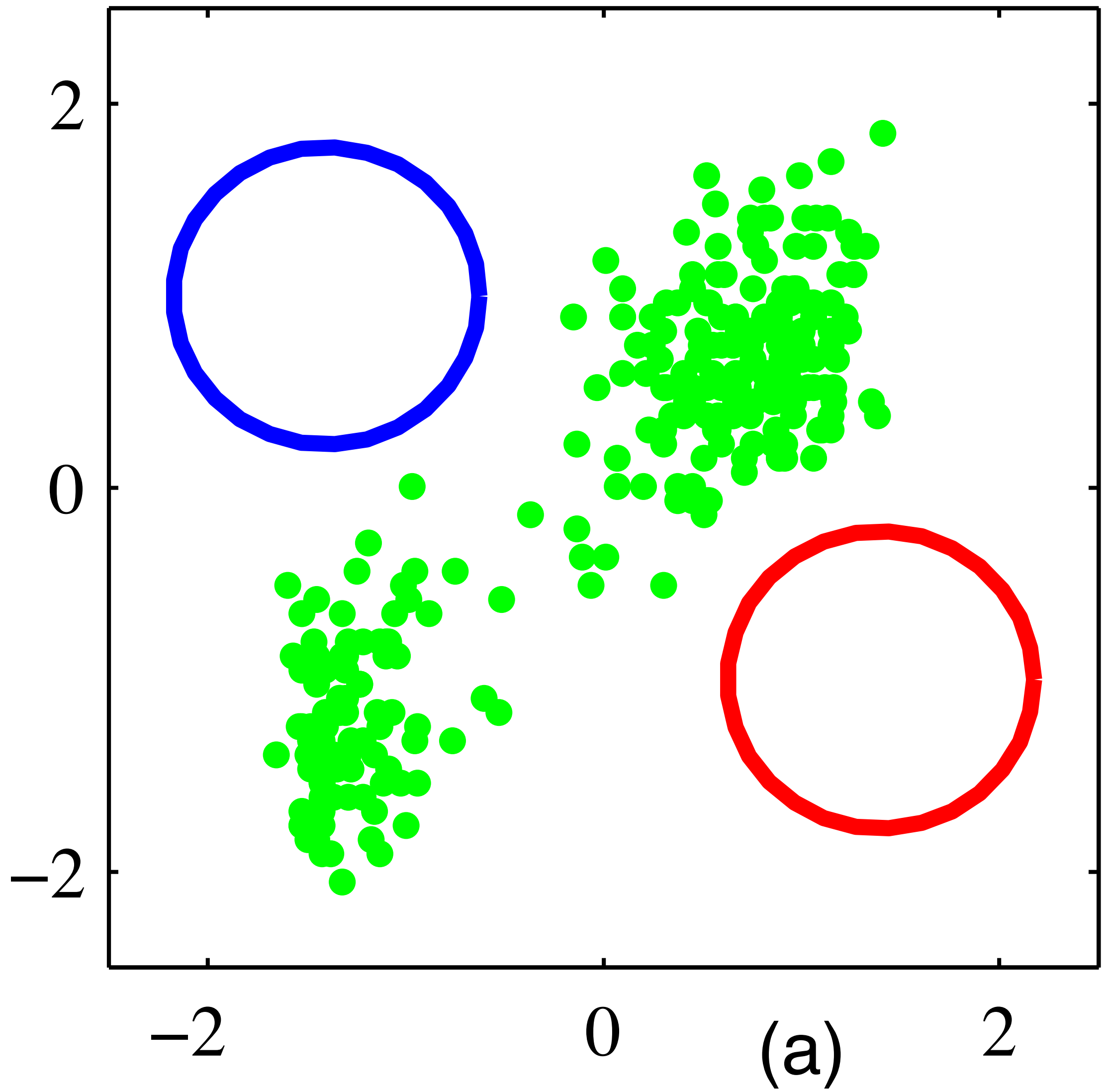


Posterior Probabilities of Assignment to Each Cluster

Incomplete Data: Points to be Clustered

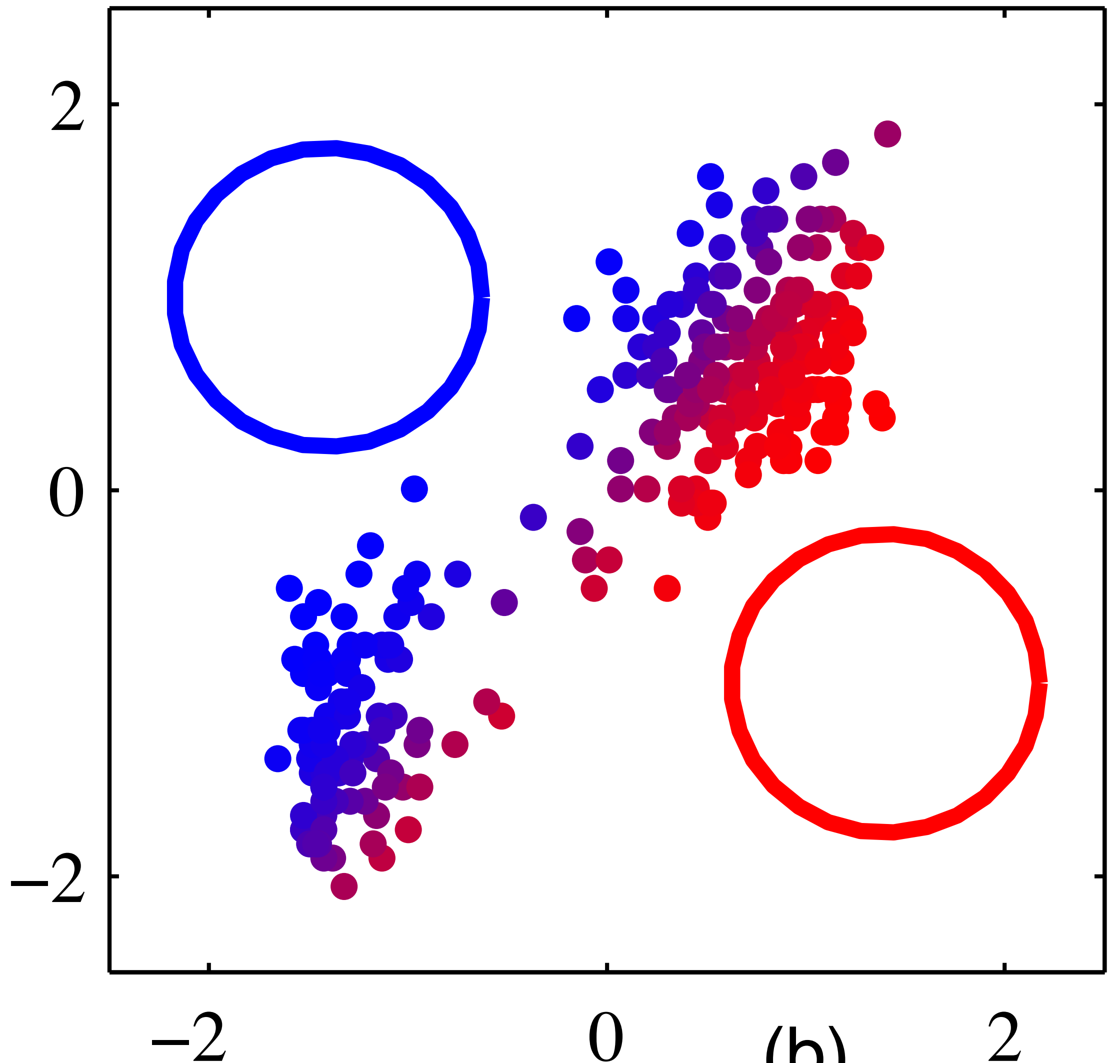
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

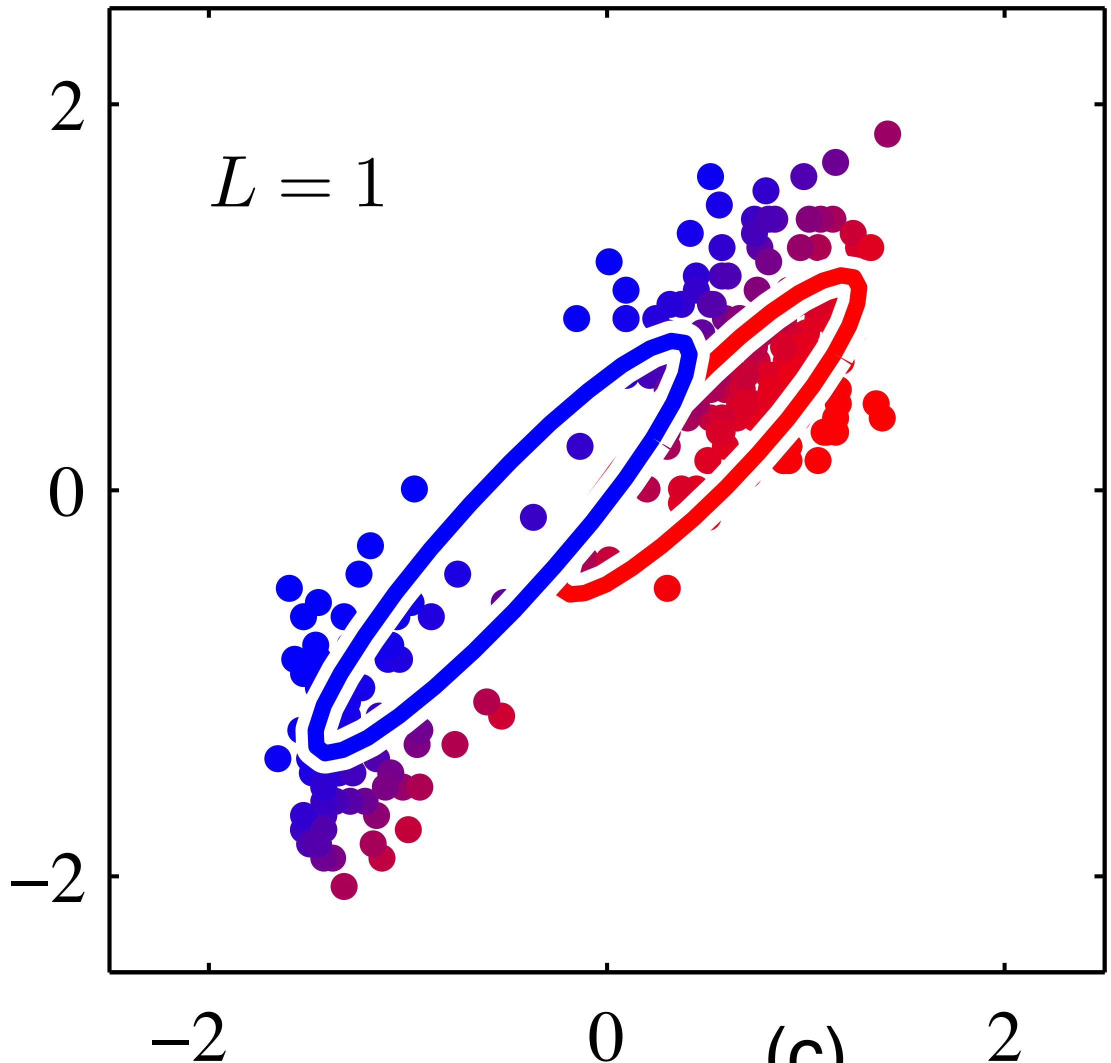
EM Algorithm



(b)

C. Bishop, Pattern Recognition & Machine Learning

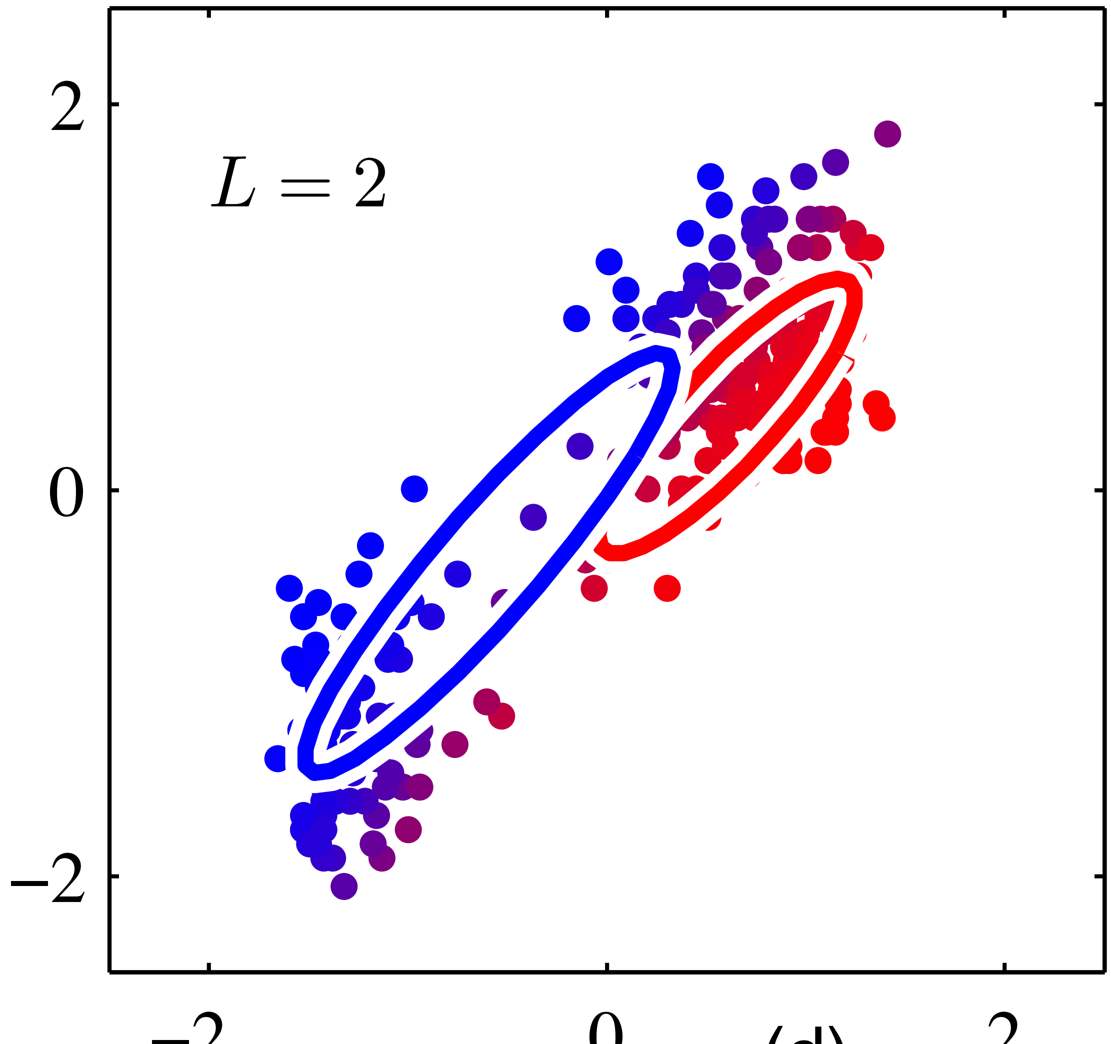
EM Algorithm



(c)

C. Bishop, Pattern Recognition & Machine Learning

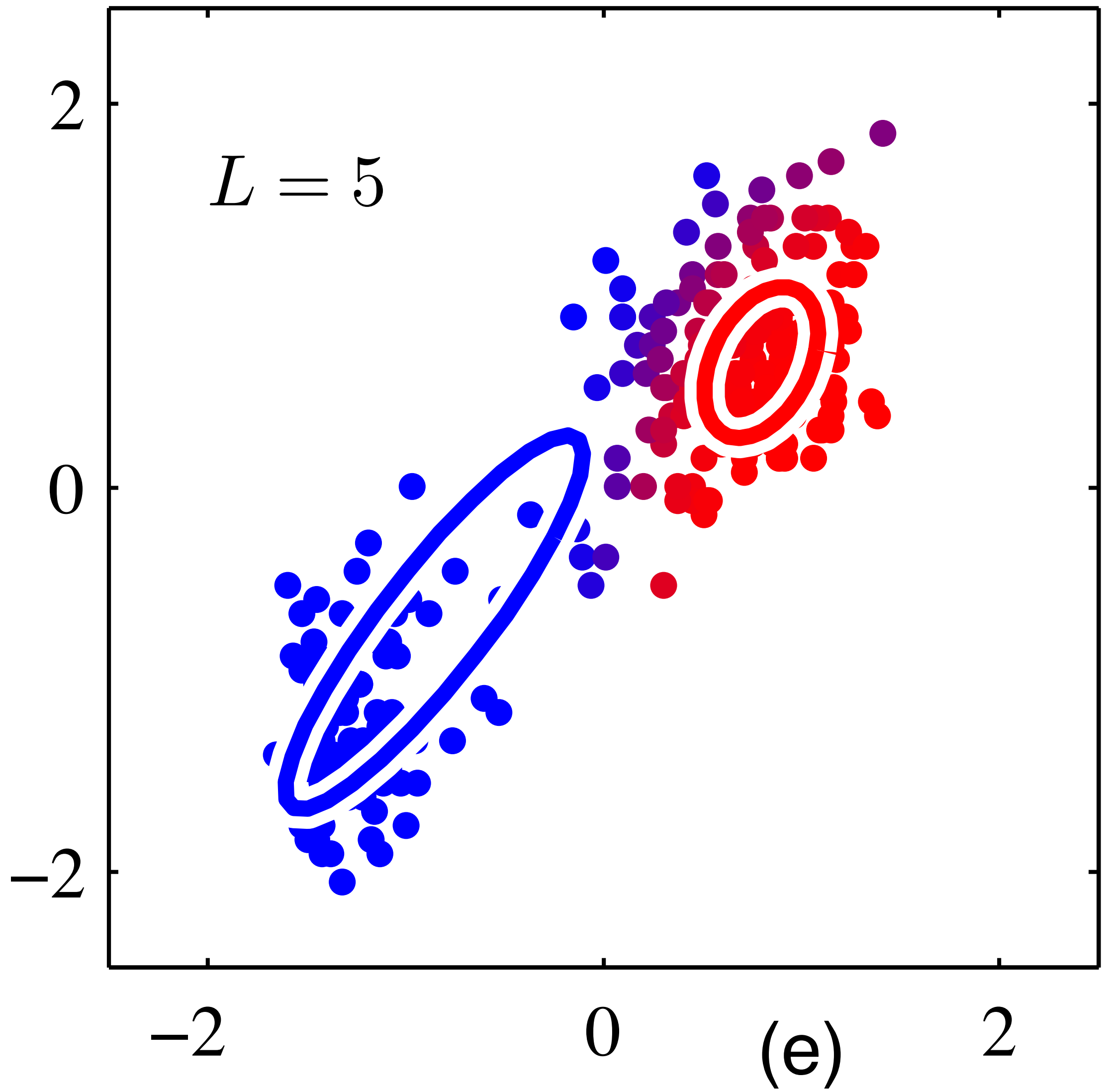
EM Algorithm



(d)

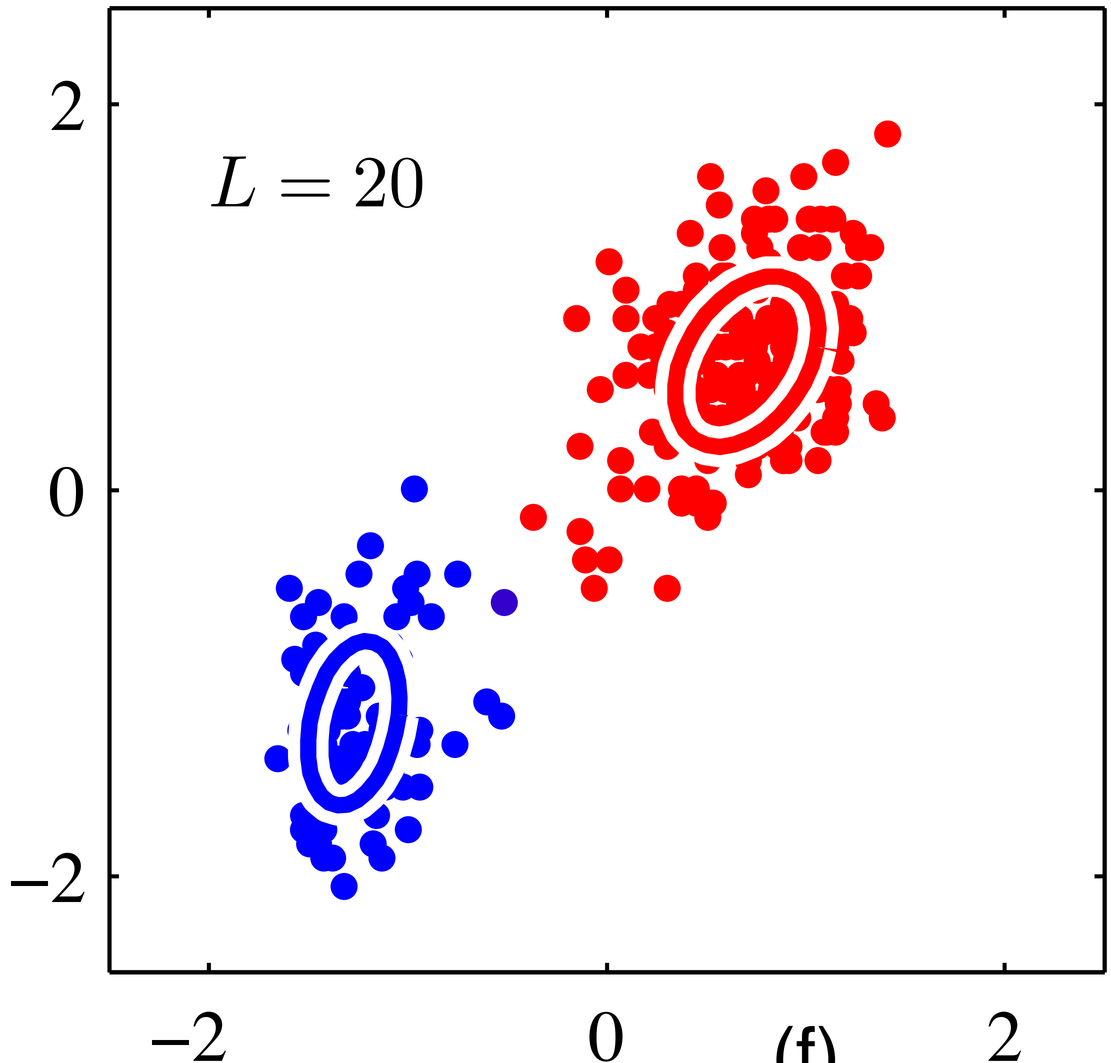
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

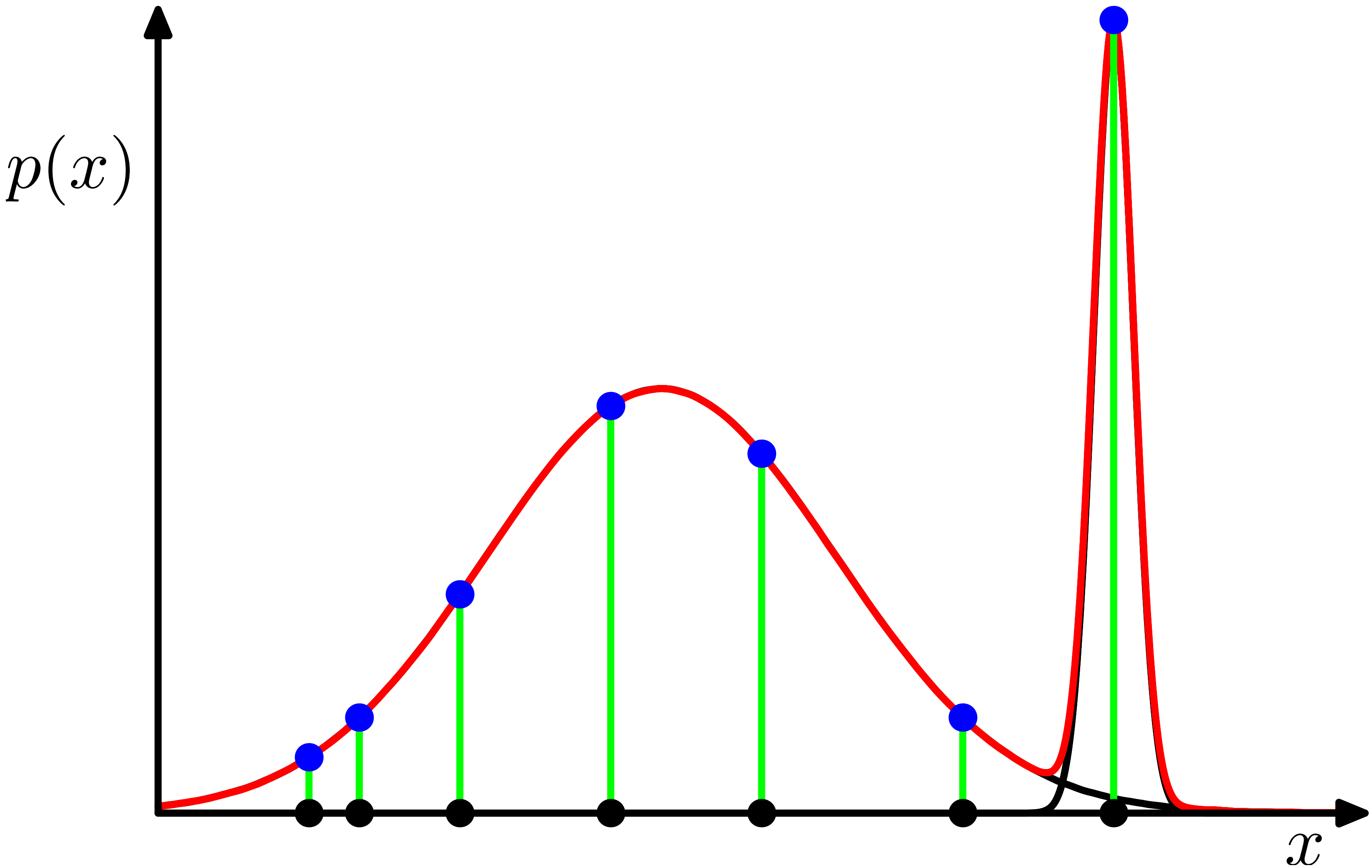
EM Algorithm



(f)

C. Bishop, Pattern Recognition & Machine Learning

Singularities: ML for Gaussian Mixtures

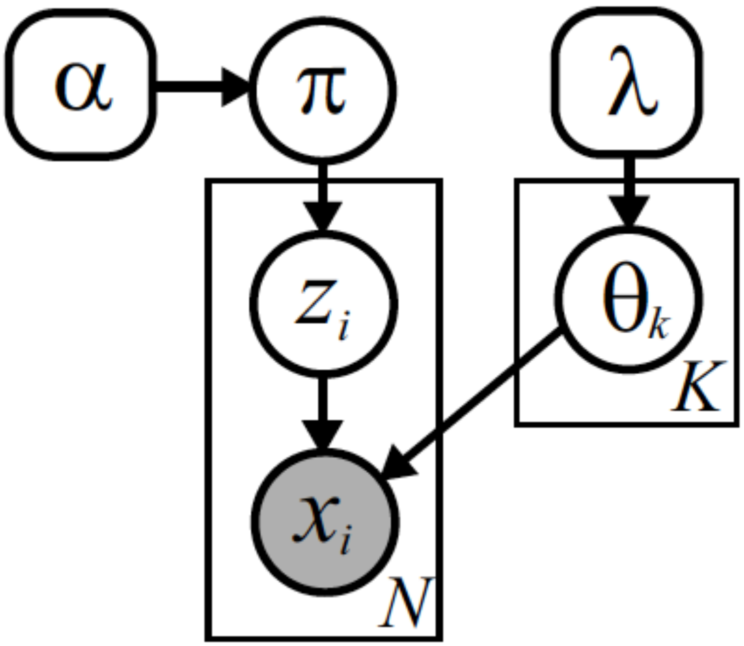


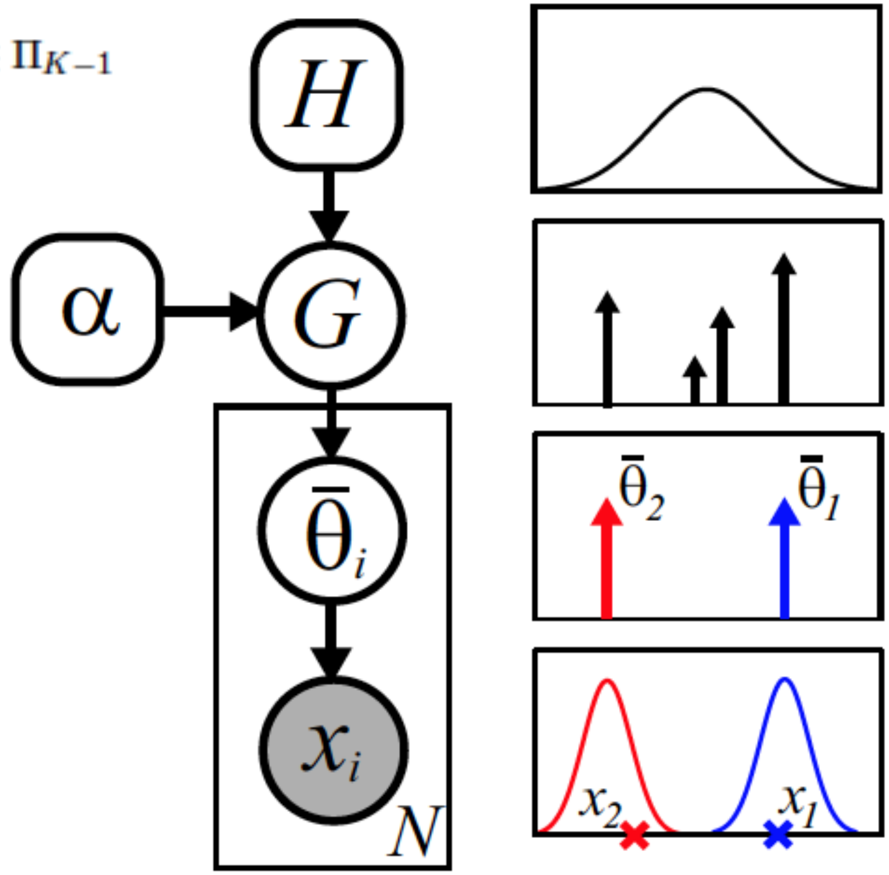
We are hoping EM will find a good local optimum…

C. Bishop, Pattern Recognition & Machine Learning

Finite Bayesian Mixture MCMC

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





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:

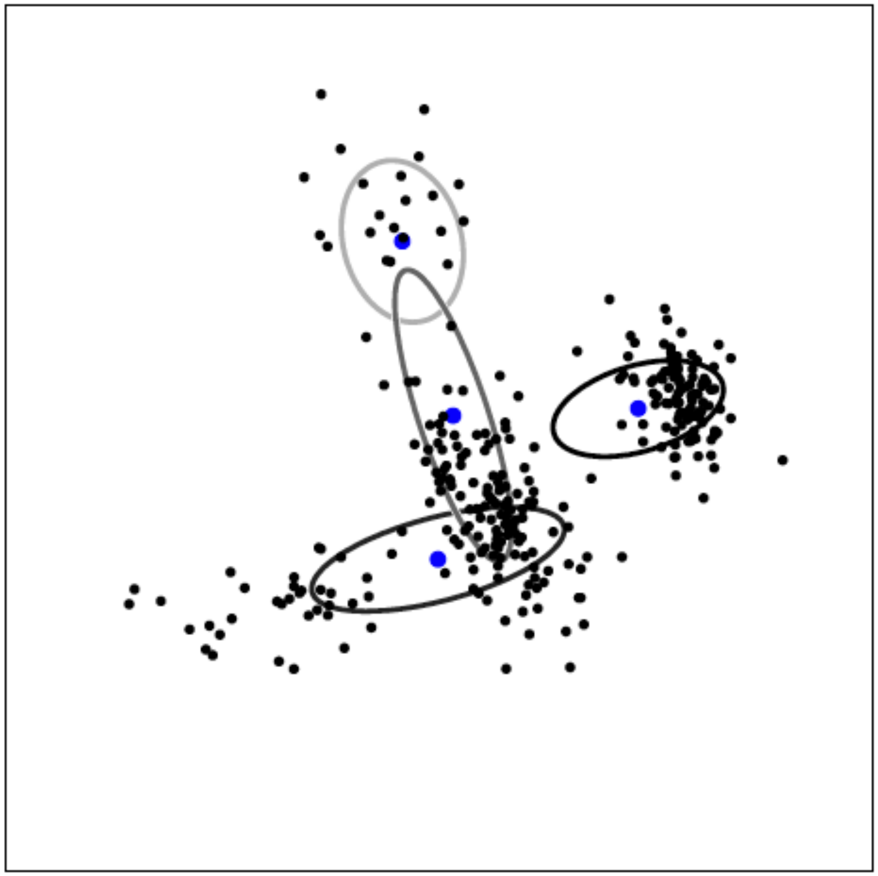
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^N \delta(z_i^{(t)}, k)$$

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

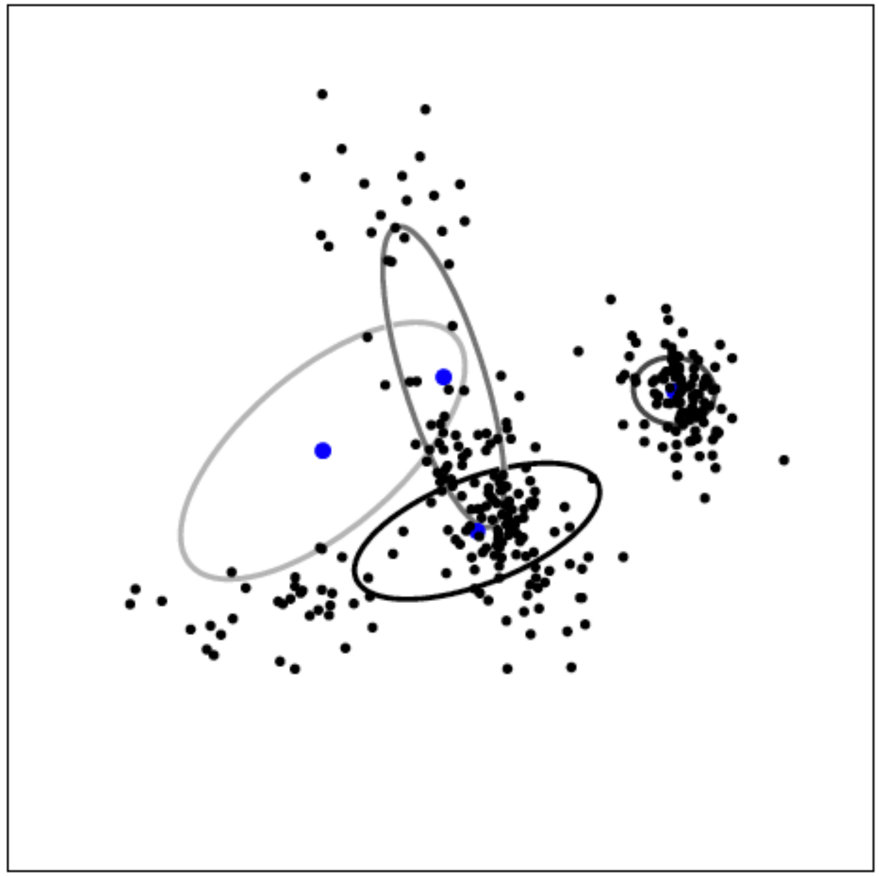
$$\theta_k^{(t)} \sim p(\theta_k \mid \{x_i \mid z_i^{(t)} = k\}, \lambda)$$

When λ defines a conjugate prior, this posterior distribution is given by Prop. 2.1.4.

Standard Sampler: 2 Iterations

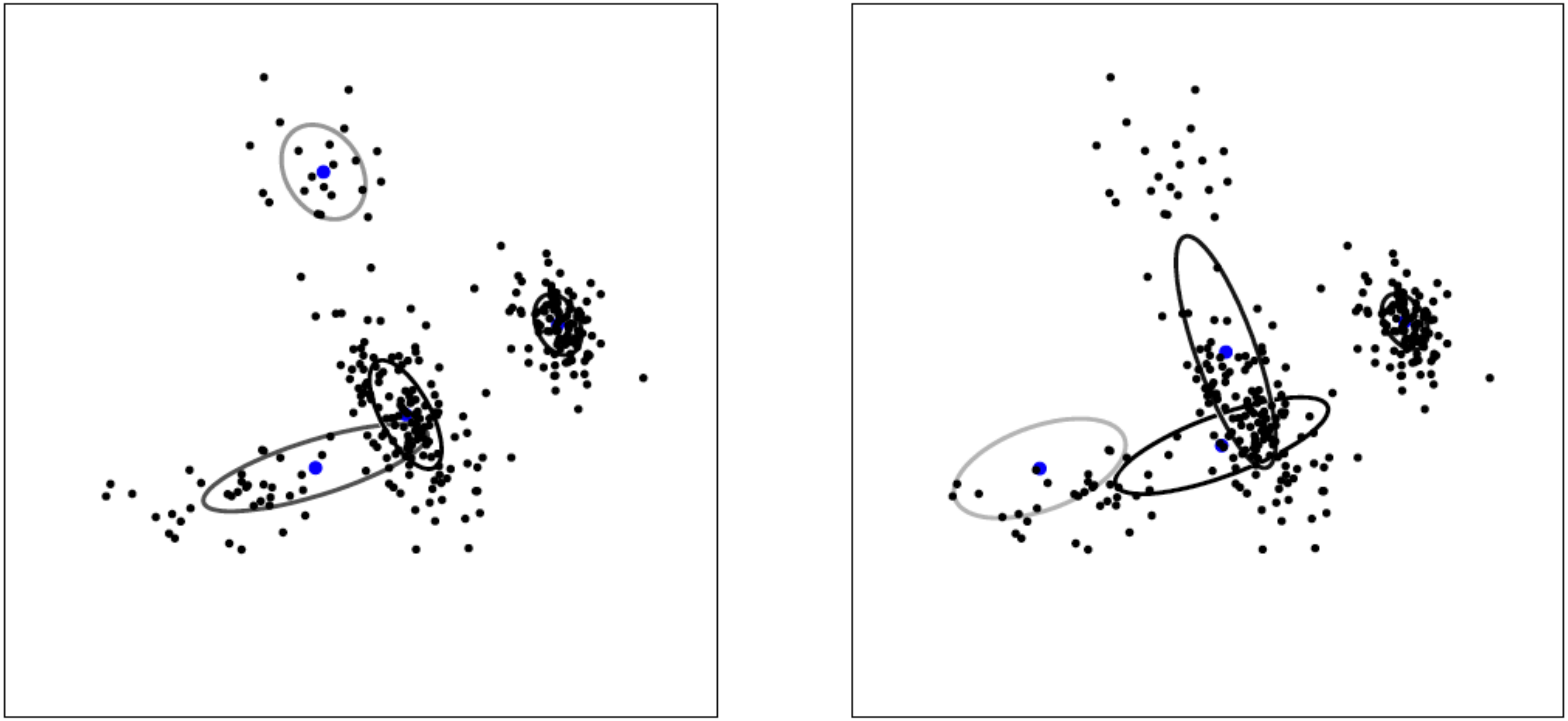


log p(x | π, θ) = –539.17



log p(x | π, θ) = –497.77

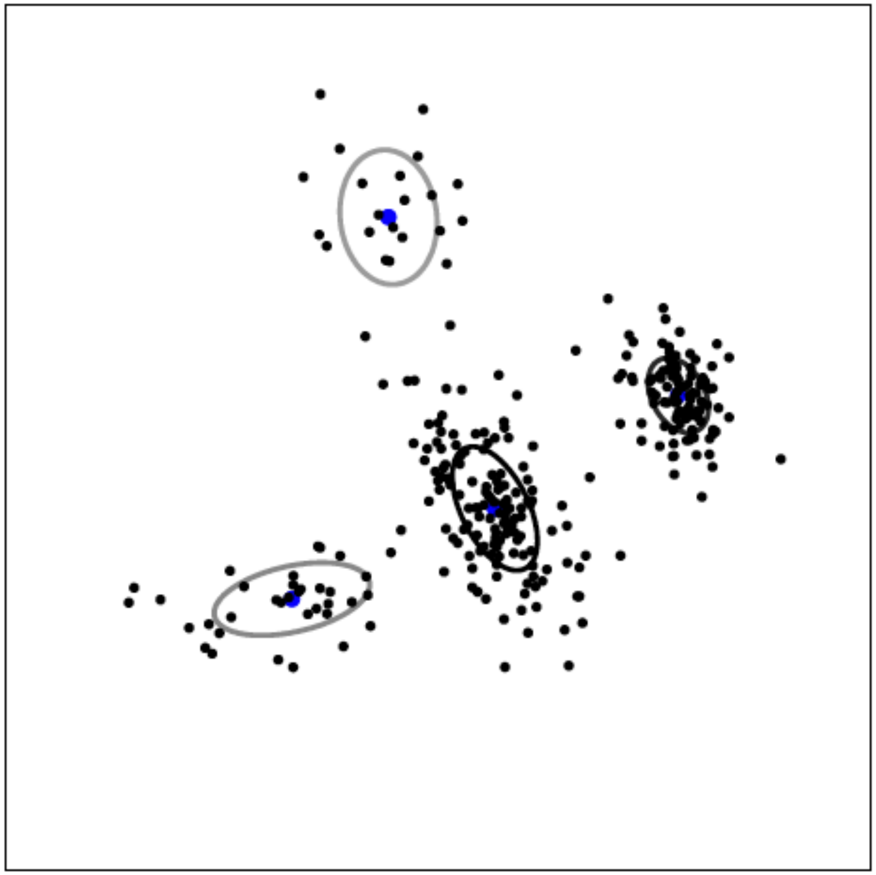
Standard Sampler: 10 Iterations



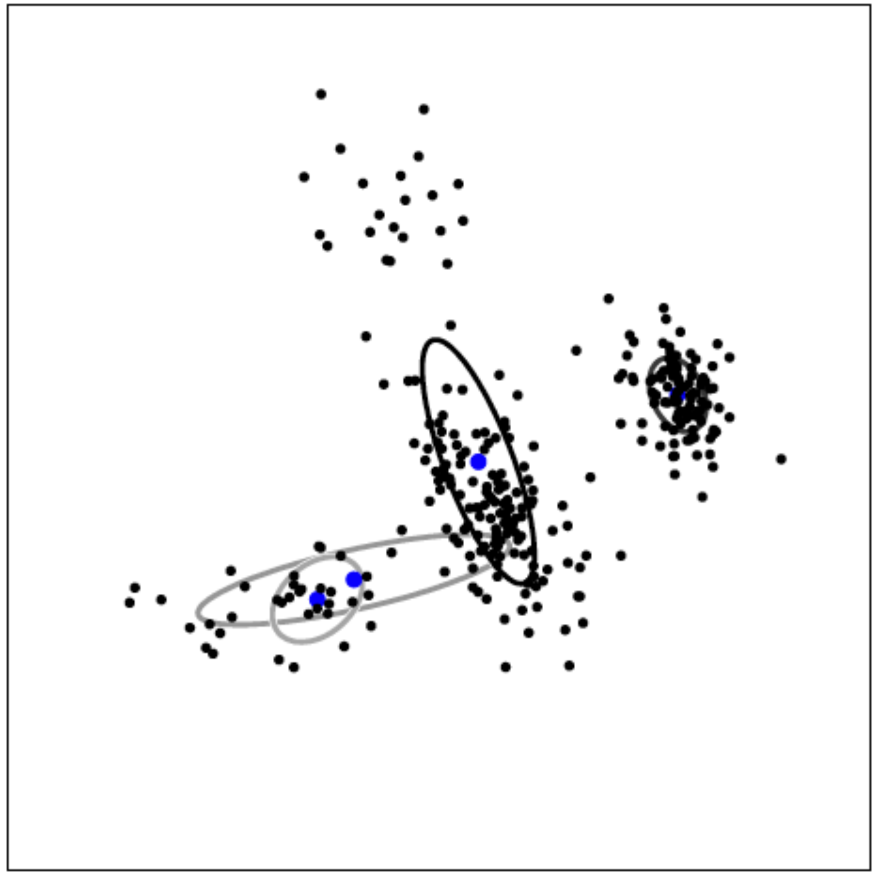
log p(x l π, θ) = –404.18

log p(x l π, θ) = –454.15

Standard Sampler: 50 Iterations



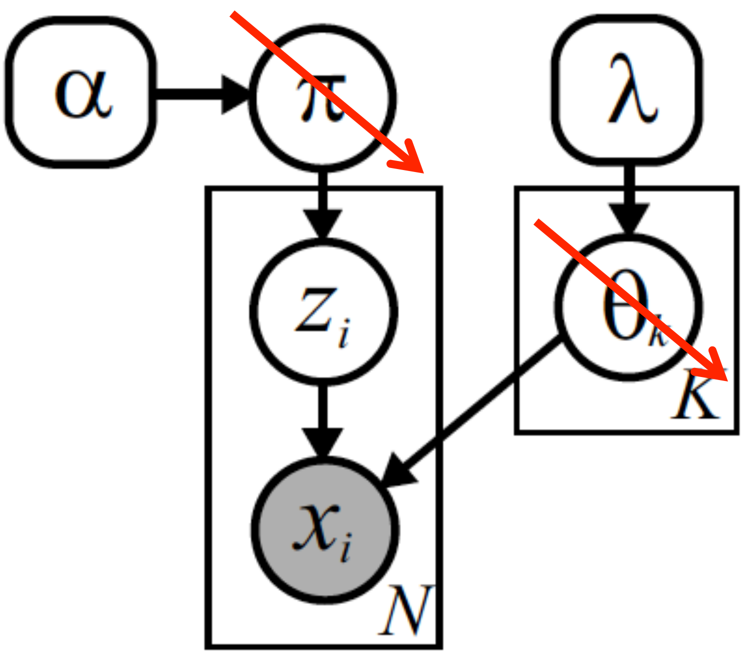
log p(x | π, θ) = –397.40

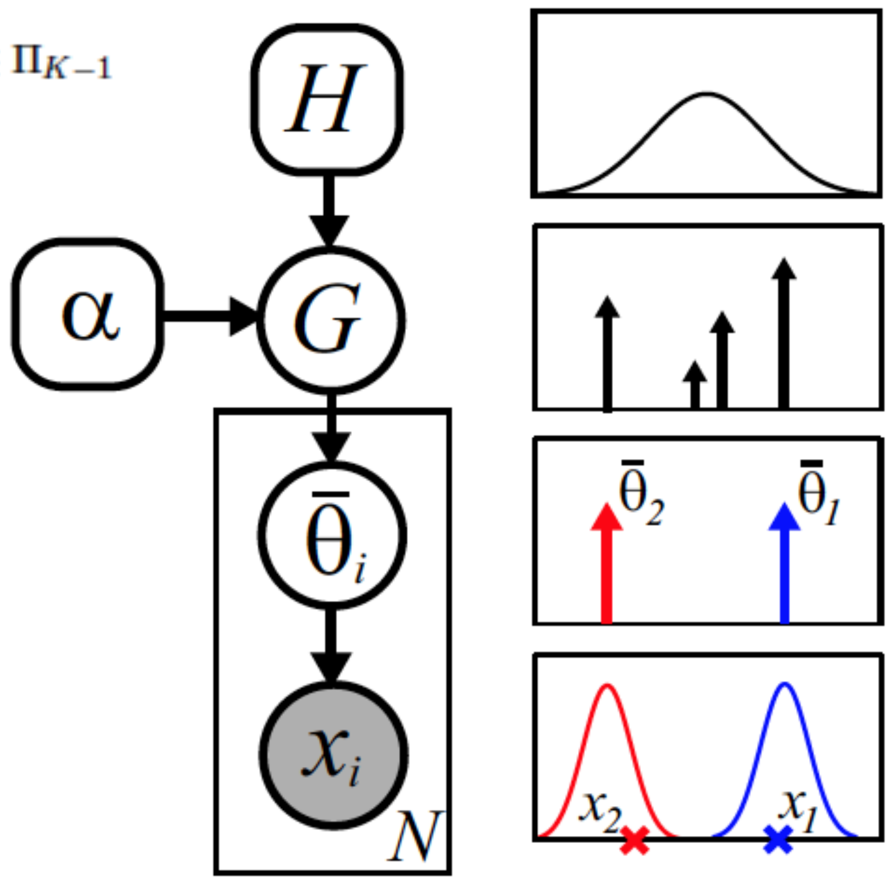


log p(x | π, θ) = –442.89

Collapsed Finite Bayesian Mixture

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





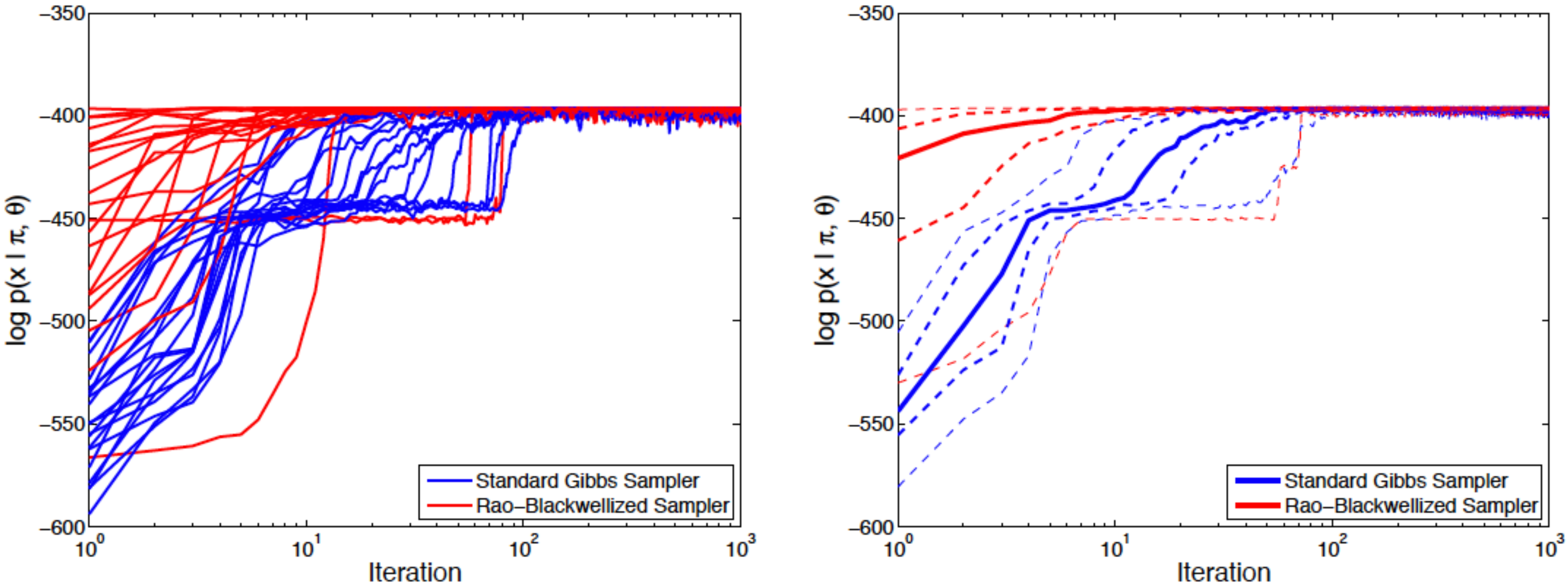
- Conjugate priors allow analytic integration of some parameters
- Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)

Collapsed Finite Mixture Sampler

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

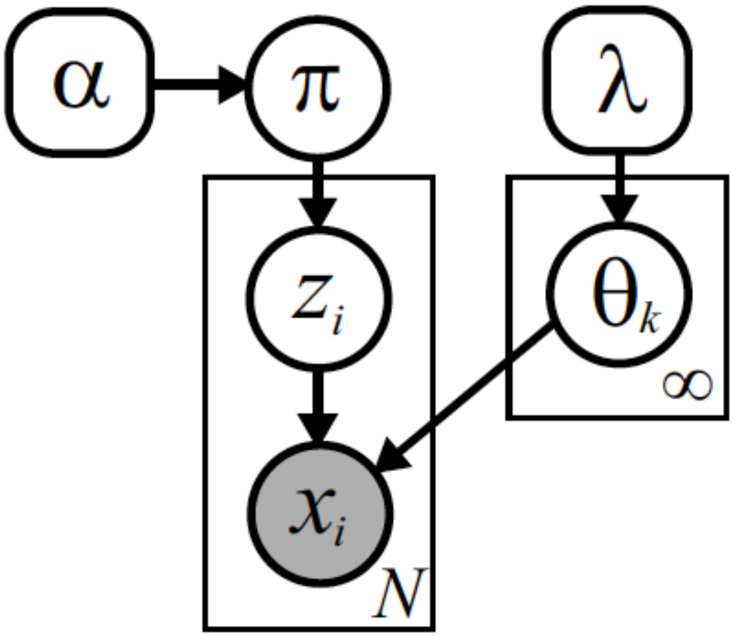
1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1,\ldots,N\}$.
2. Set $z = z^{(t-1)}$. For each $i \in \{\tau(1),\ldots,\tau(N)\}$, sequentially resample z_i as follows:
   (a) For each of the K clusters, determine the predictive likelihood
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   This likelihood can be computed from cached sufficient statistics via Prop. 2.1.4.
   (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.

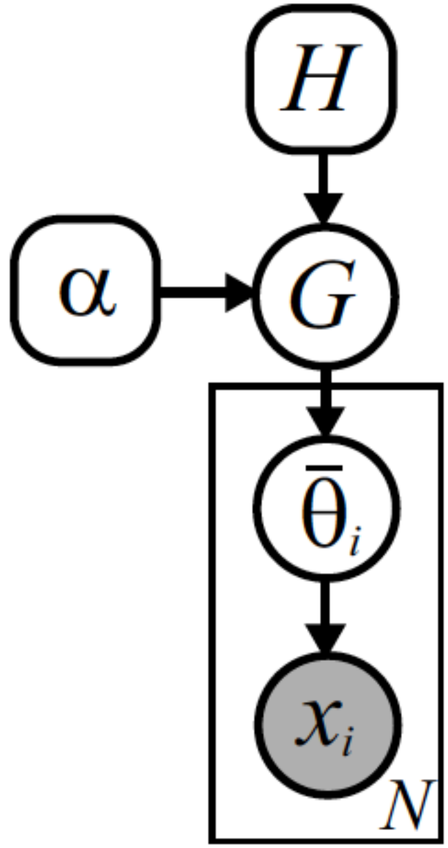
Standard versus Collapsed Samplers

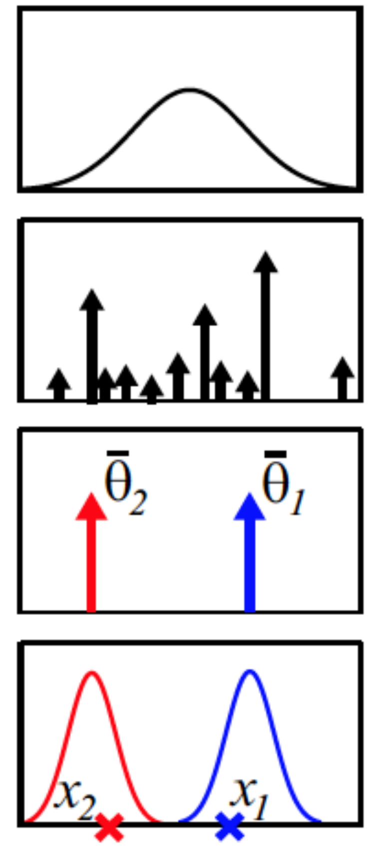


DP Mixture Models

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







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

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

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

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

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

$$z_i \sim \pi$$

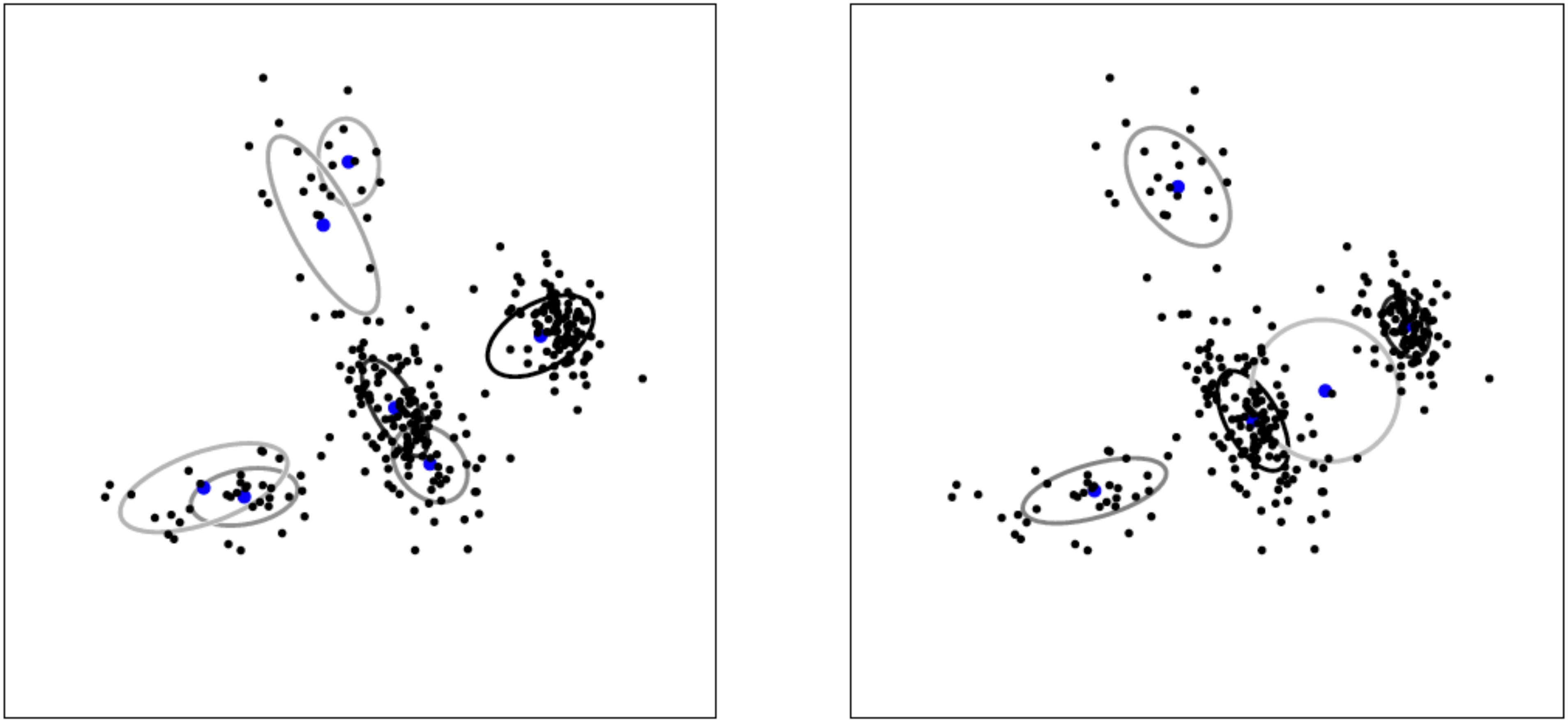
$$x_i \sim F(\theta_{z_i})$$

Collapsed DP Mixture Sampler

Given the previous concentration parameter $\alpha^{(t-1)}$, cluster assignments $z^{(t-1)}$, and cached statistics for the K current clusters, sequentially sample new assignments as follows:

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
   $$f_k(x_i) = p(x_i \mid \{x_j \mid z_j = k, j \neq i\}, \lambda)$$
   This likelihood can be computed from cached sufficient statistics via Prop. 2.1.4. Also determine the likelihood $f_{\bar{k}}(x_i)$ of a potential new cluster $\bar{k}$ via eq. (2.189).
   (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.
5. If $\alpha \sim \text{Gamma}(a, b)$, sample $\alpha^{(t)} \sim p(\alpha \mid K, N, a, b)$ via auxiliary variable methods [76].

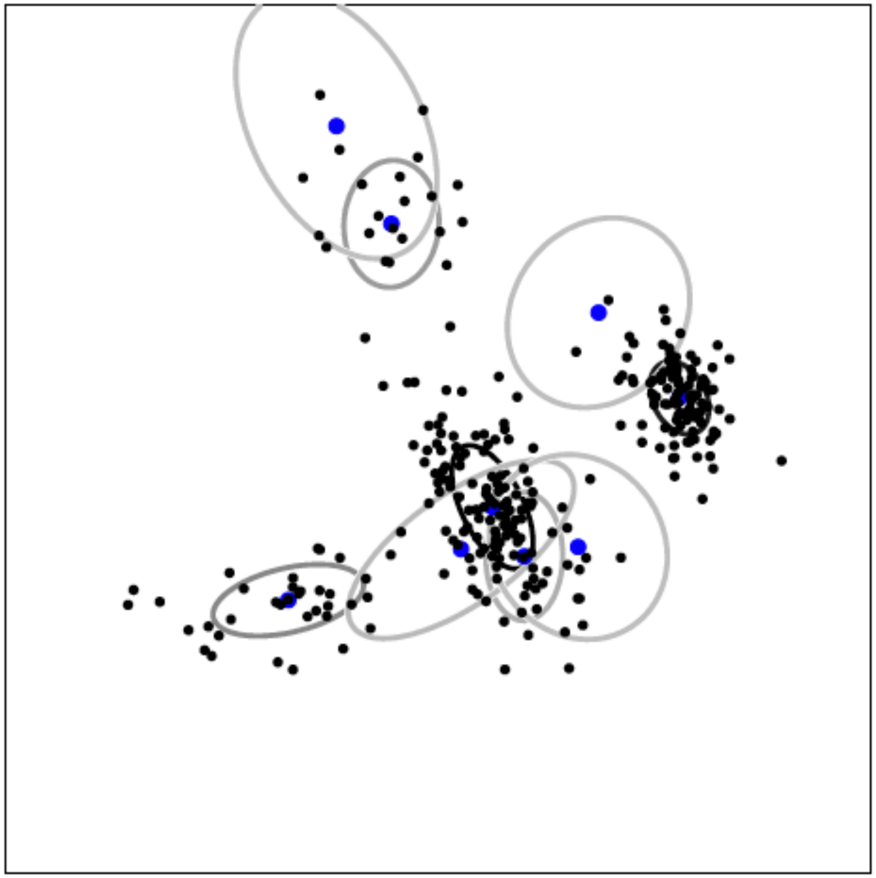
Collapsed DP Sampler: 2 Iterations



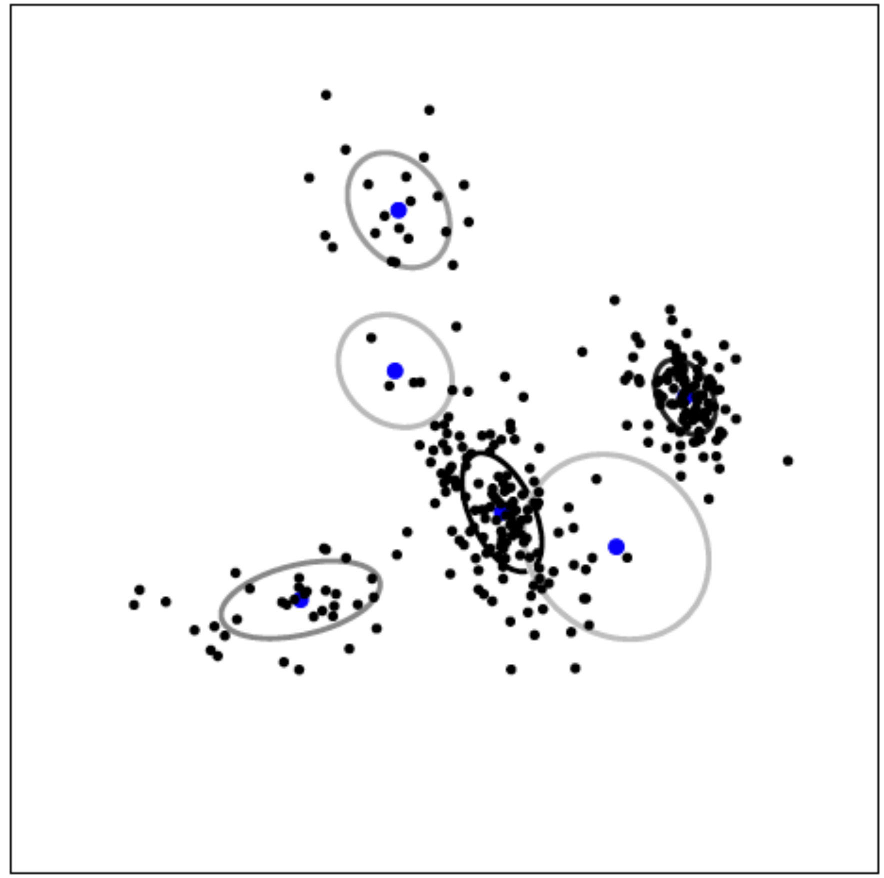
log p(x | π, θ) = –462.25

log p(x | π, θ) = –399.82

Standard Sampler: 10 Iterations

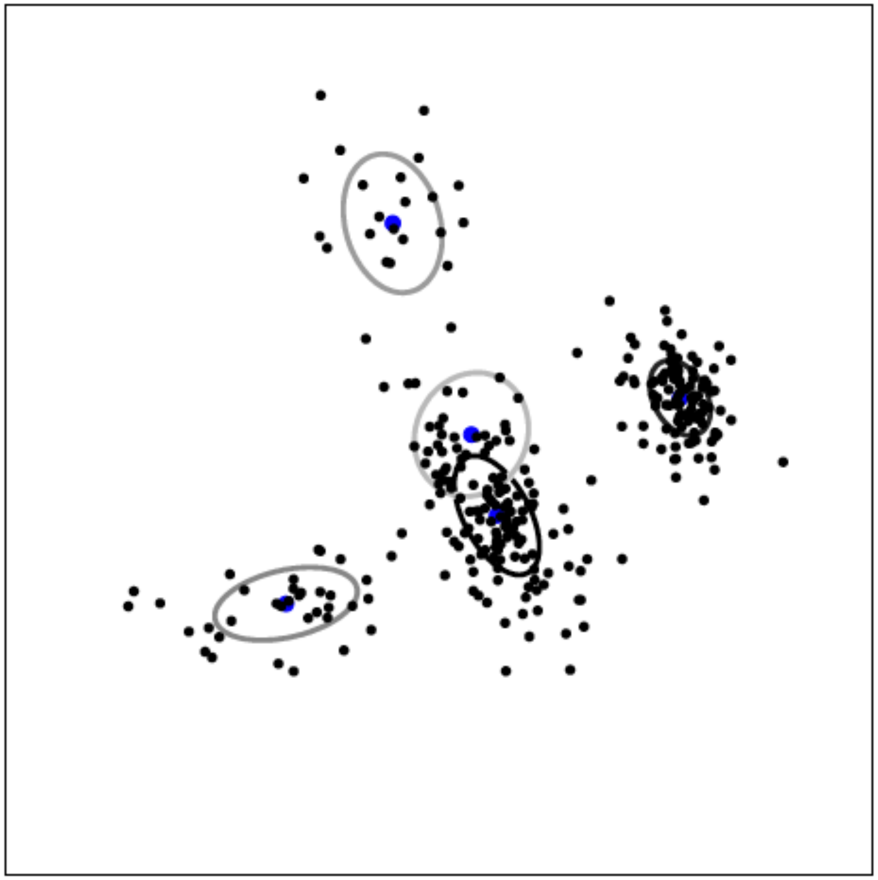


$\log p(x \mid \pi, \theta) = -398.32$

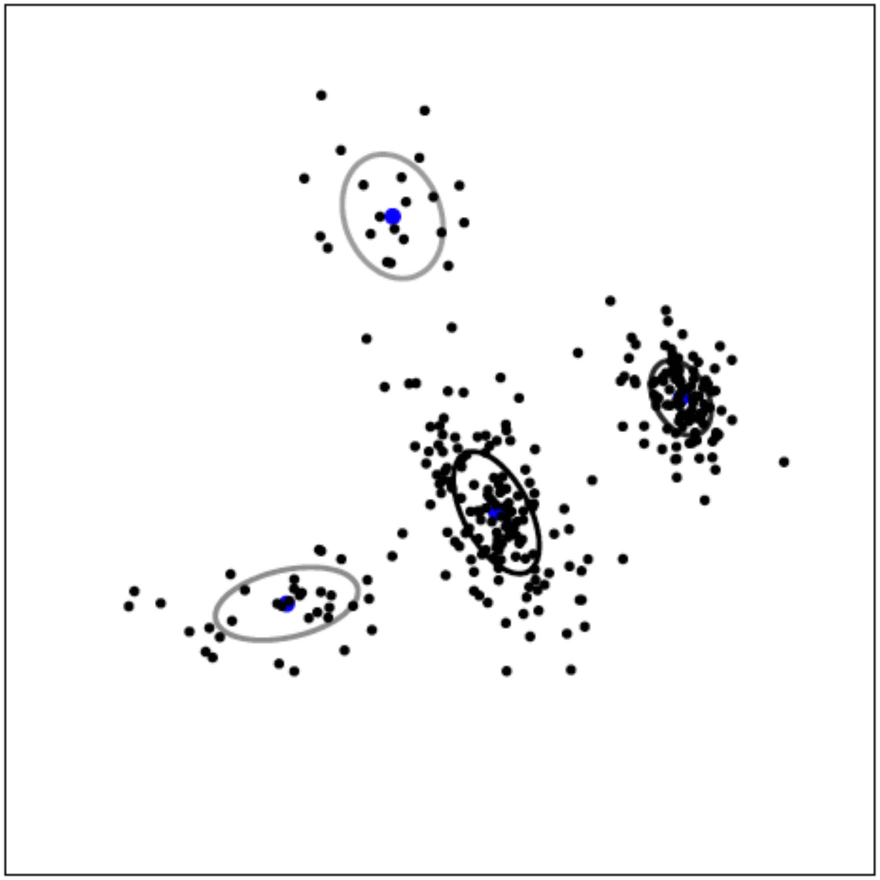


$\log p(x \mid \pi, \theta) = -399.08$

Standard Sampler: 50 Iterations

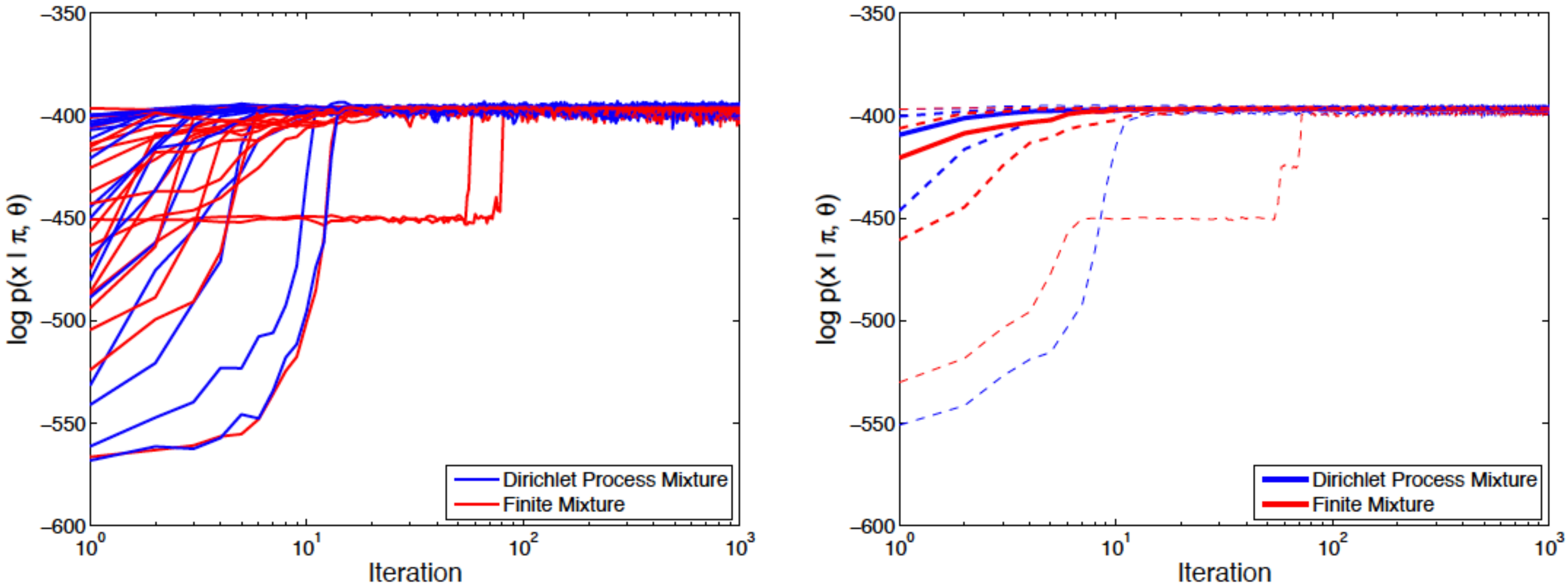


$\log p(x \mid \pi, \theta) = -397.67$

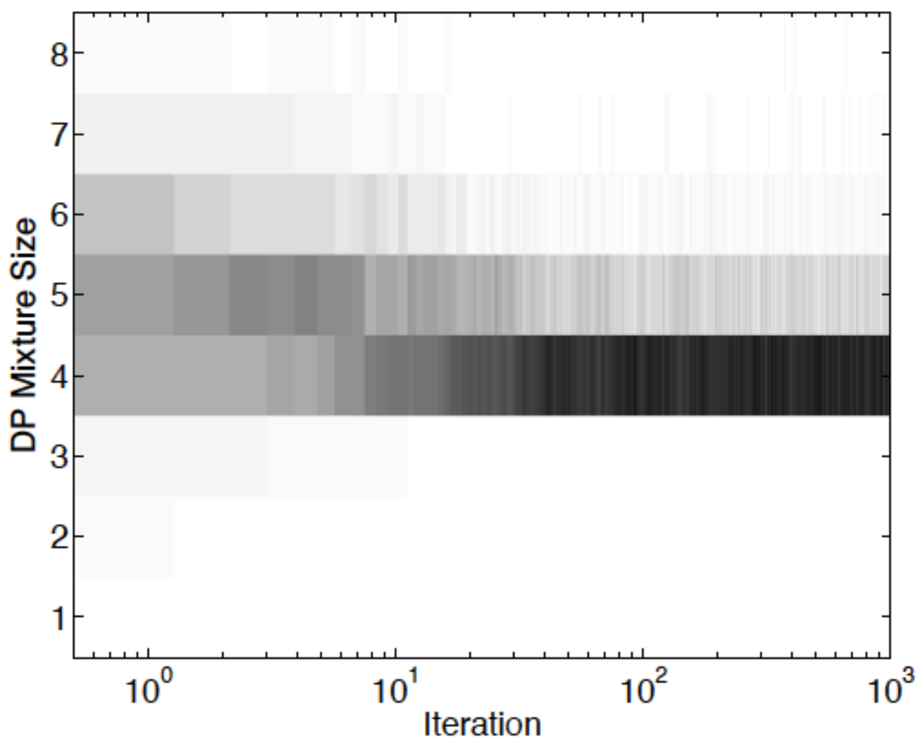


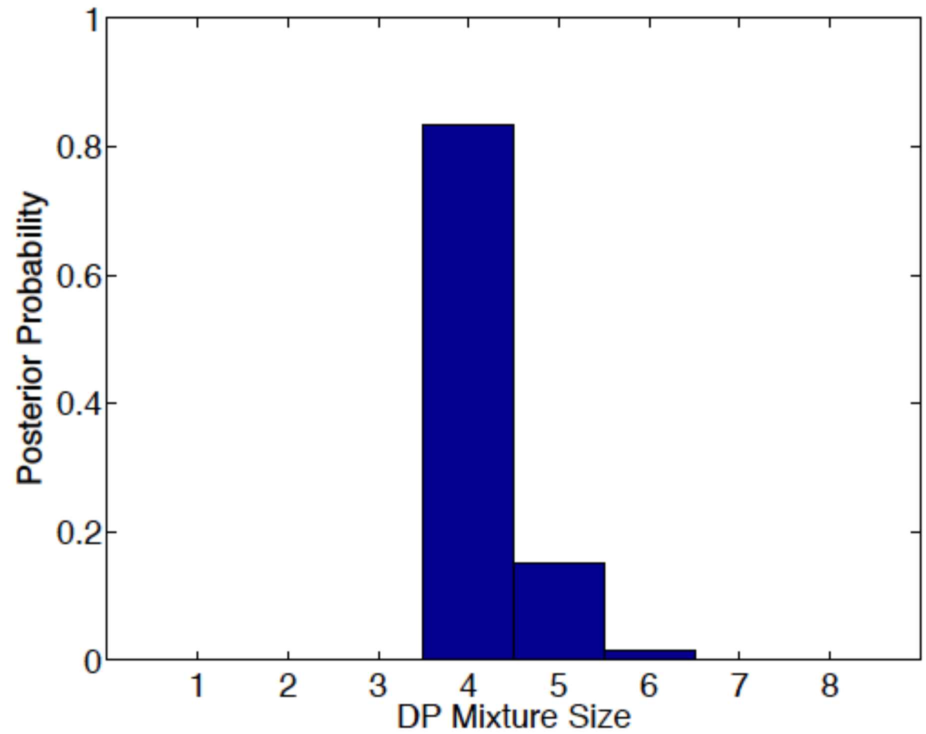
$\log p(x \mid \pi, \theta) = -396.71$

DP versus Finite Mixture Samplers



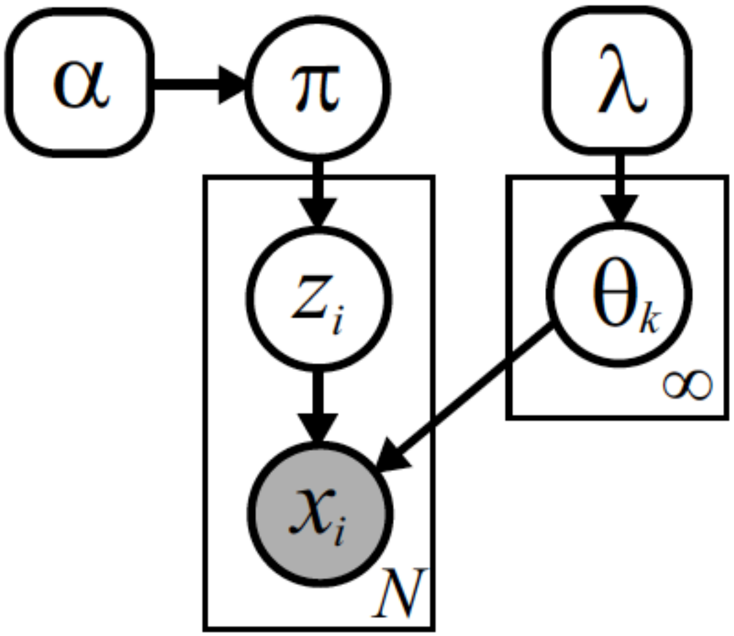
DP Posterior Number of Clusters

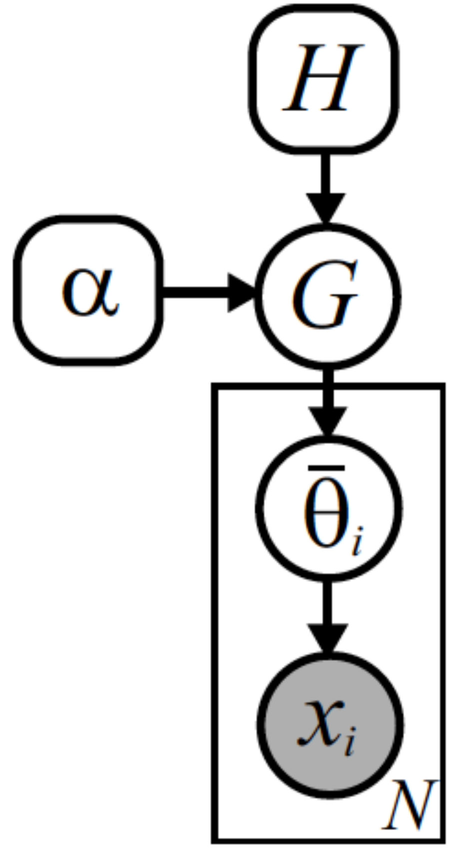


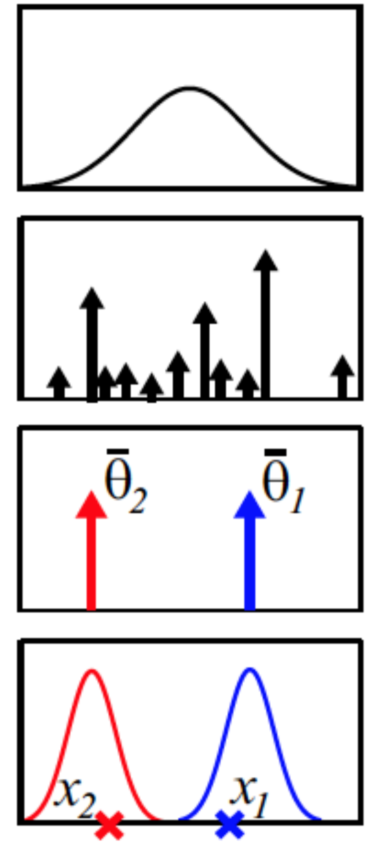


DP Mixture Models

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







- Neal's Alg. 1: Sample $\bar{\theta}_i$
- Neal's Alg. 2: Sample z and θ_k
- Neal's Alg. 3: Sample z (preceding slides)
- Neal's Alg. 4+: If can't integrate θ_k

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

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

$z_i \sim \pi$

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