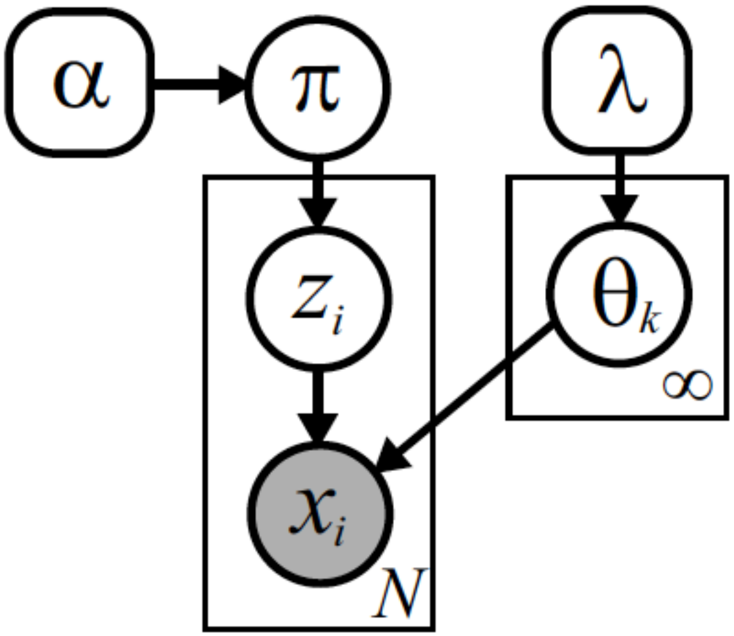
Applied Bayesian Nonparametrics

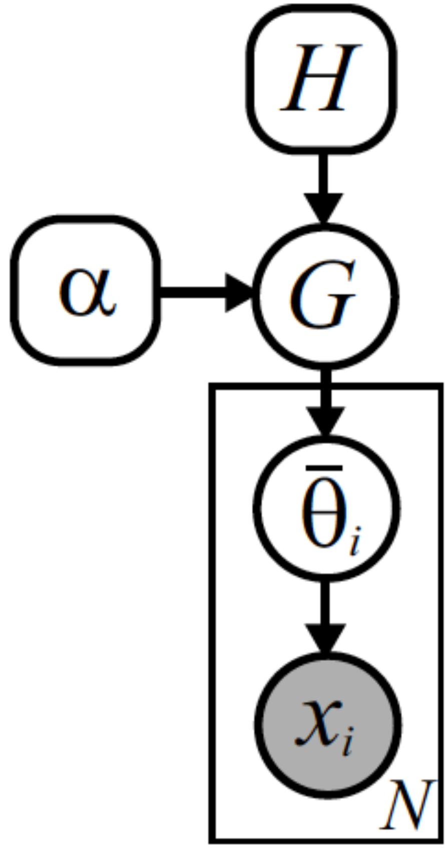
Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

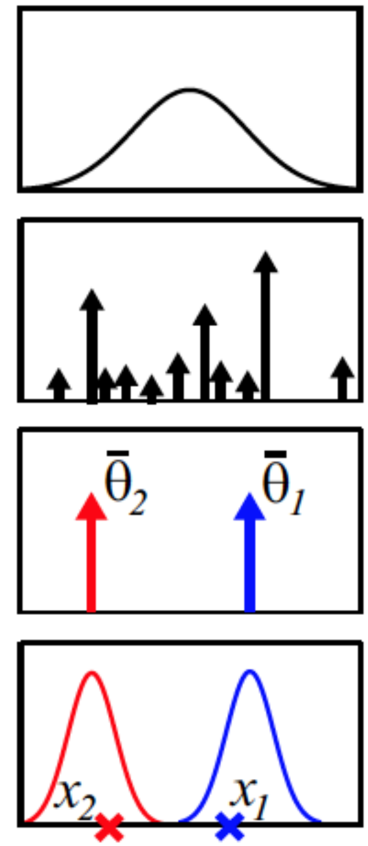
September 29: Dirichlet Process Theory, MCMC for DP Mixture Models

DP Mixture Models

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







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

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

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

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

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

$$z_i \sim \pi$$

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

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*

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

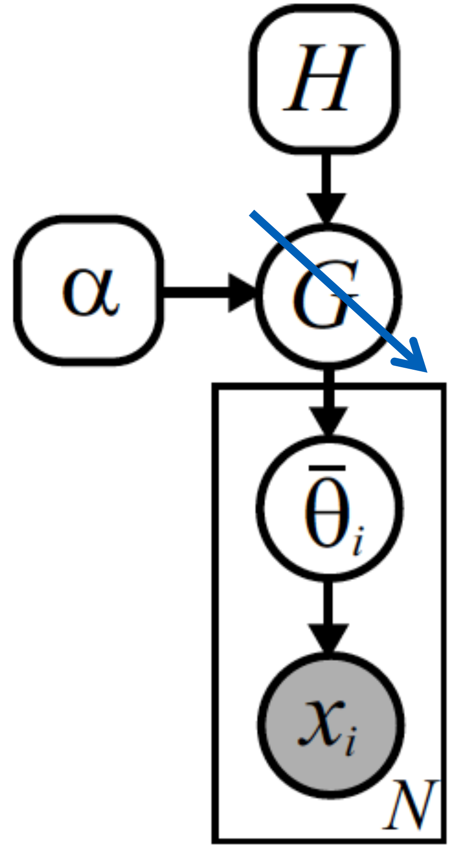
where N_k is the number of previous observations of θ_k, as in eq. (2.179).

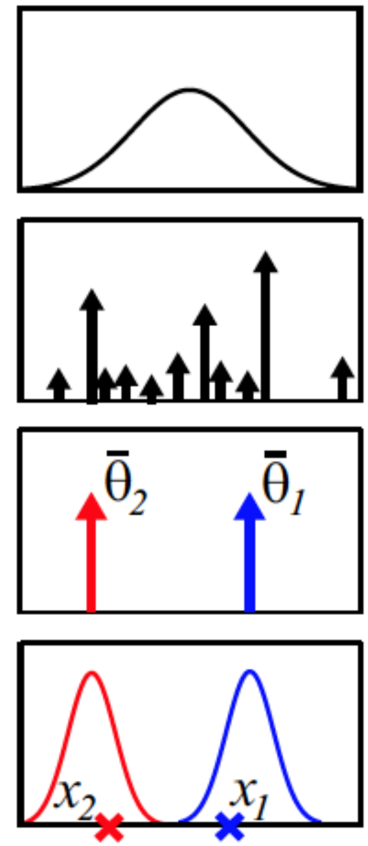
My variation on the classical balls in urns analogy:

- 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…

DP Mixture: Polya Urn Sampler

- Marginalize G to produce Polya urn predictive rule
- Escobar & West (1995)
- Algorithm 1 of Neal (2000)
- Basic Polya urn sampler of Ishwaran & James (2001)
- ***Slow:*** Can only change cluster centers by destroying and recreating that cluster

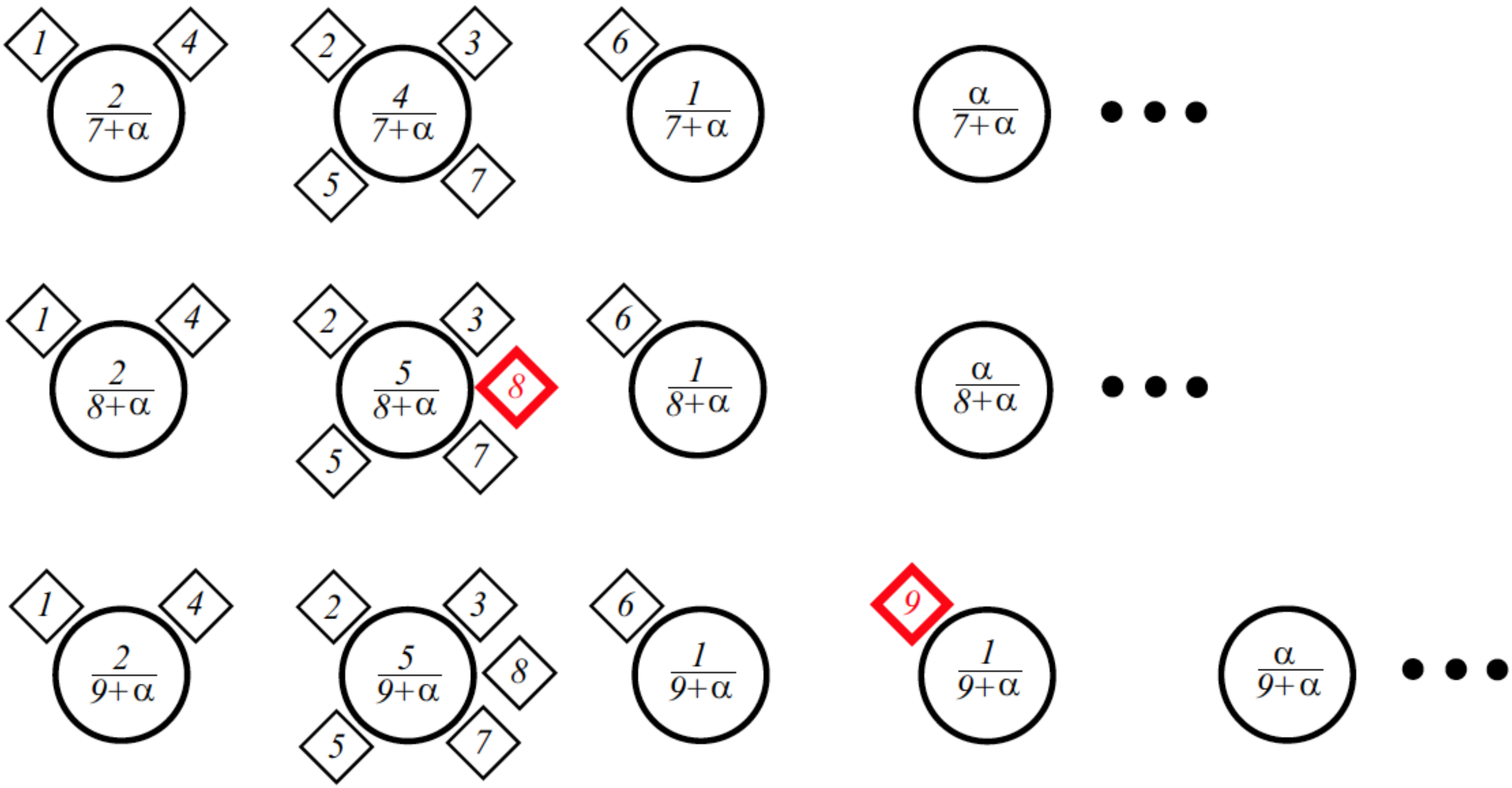




$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k)$$

$$\begin{aligned} \pi &\sim \mathrm{GEM}(\alpha) \\ \theta_k &\sim H(\lambda) \qquad k = 1, 2, \ldots \\ \bar{\theta}_i &\sim G \\ x_i &\sim F(\bar{\theta}_i) \end{aligned}$$

Chinese Restaurant Process

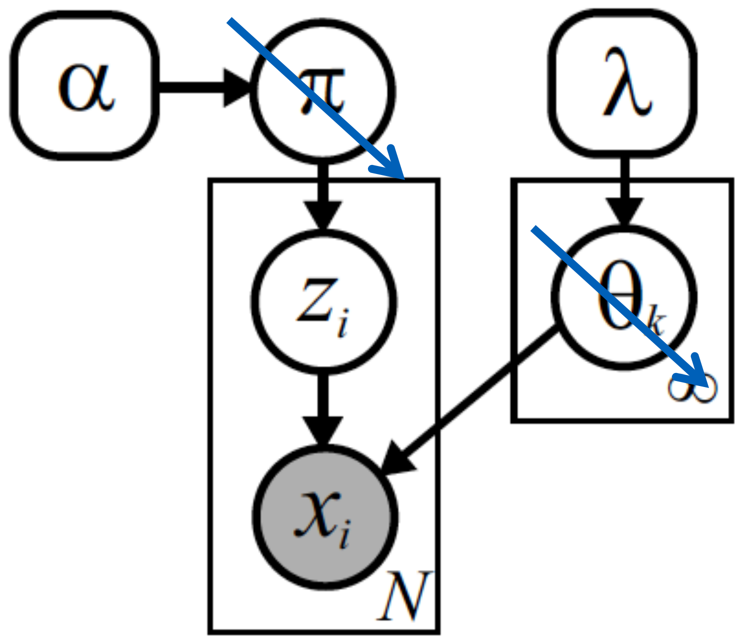


$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \alpha) = \frac{1}{\alpha + N} \left(\sum_{k=1}^{K} N_k \delta(z, k) + \alpha \delta(z, \bar{k}) \right)$$

DP Mixture: CRP Sampler

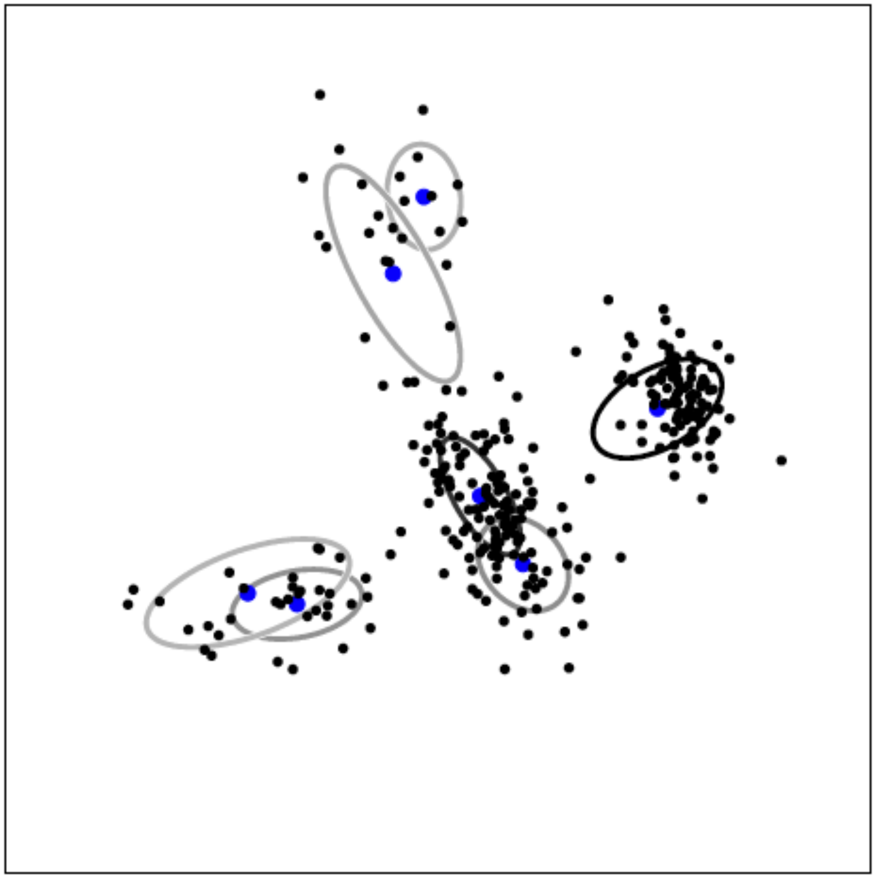
- Conceptually separates cluster allocations and parameters
- Marginalize cluster sizes to give Chinese restaurant process prior on data partitions
- Accelerated Polya urn sampler of Ishwaran & James (2001)
- Algorithm 2 of Neal (2000)

- Algorithm 3 of Neal (2000) also marginalizes (collapses) cluster parameters (needs conjugacy)
- Rasmussen (2001) elaborates
- Effective for limited range of models it applies to…

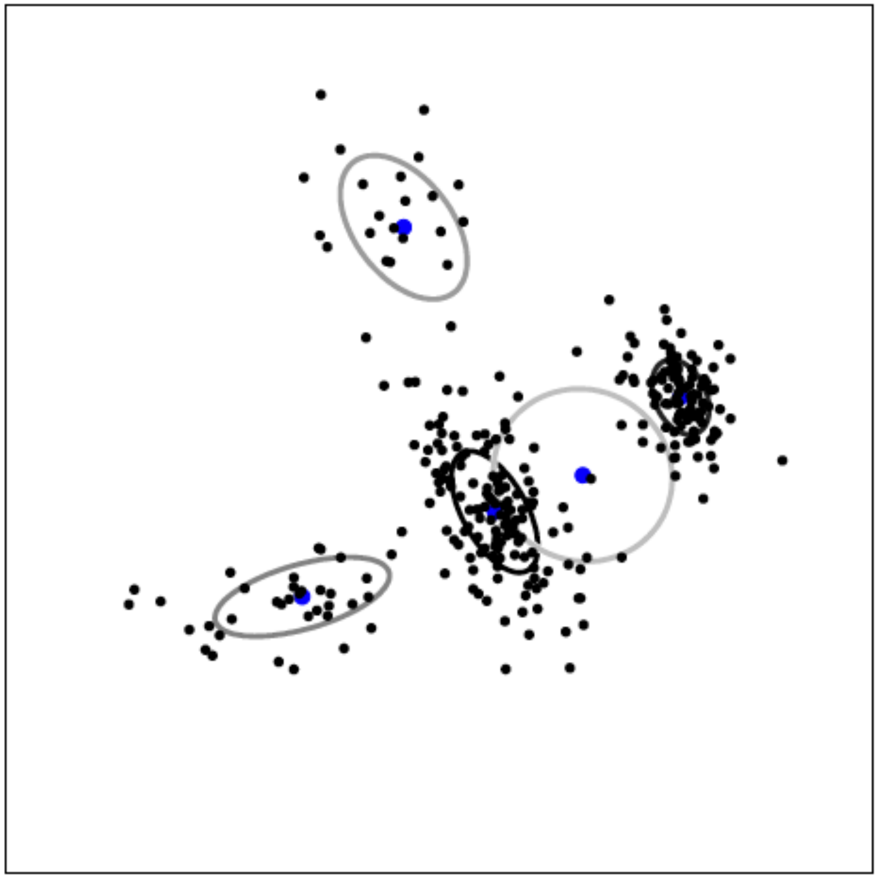


$$\pi \sim \mathrm{GEM}(\alpha)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$
$$z_i \sim \pi$$
$$x_i \sim F(\theta_{z_i})$$

Collapsed DP Sampler: 2 Iterations

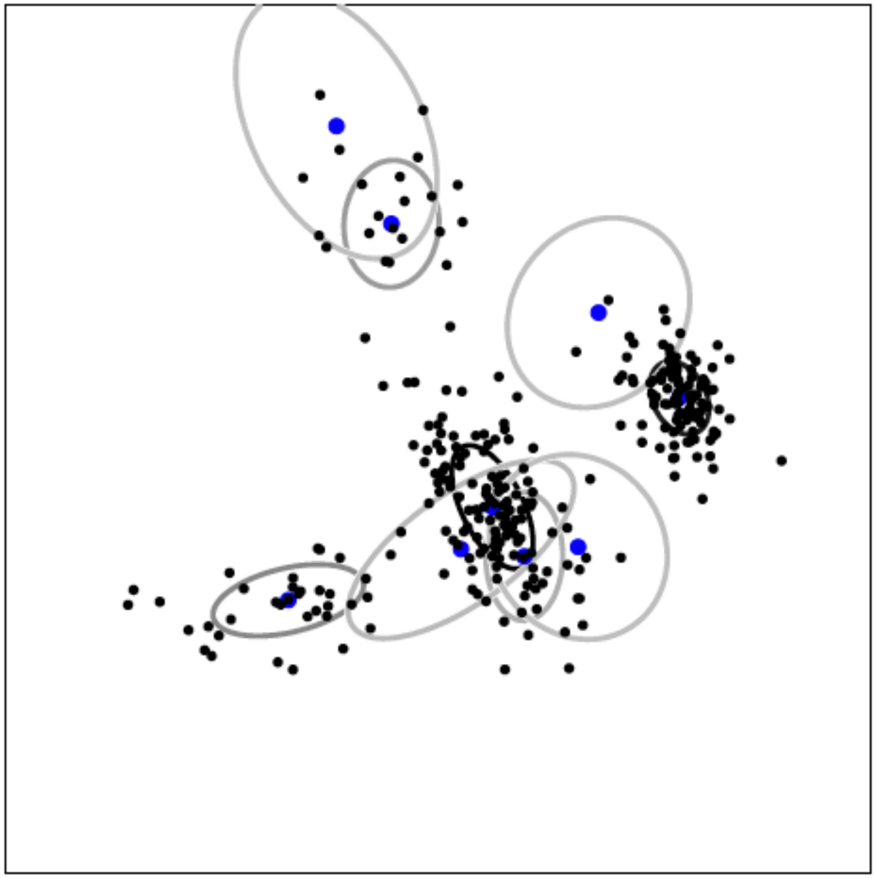


log p(x | π, θ) = –462.25

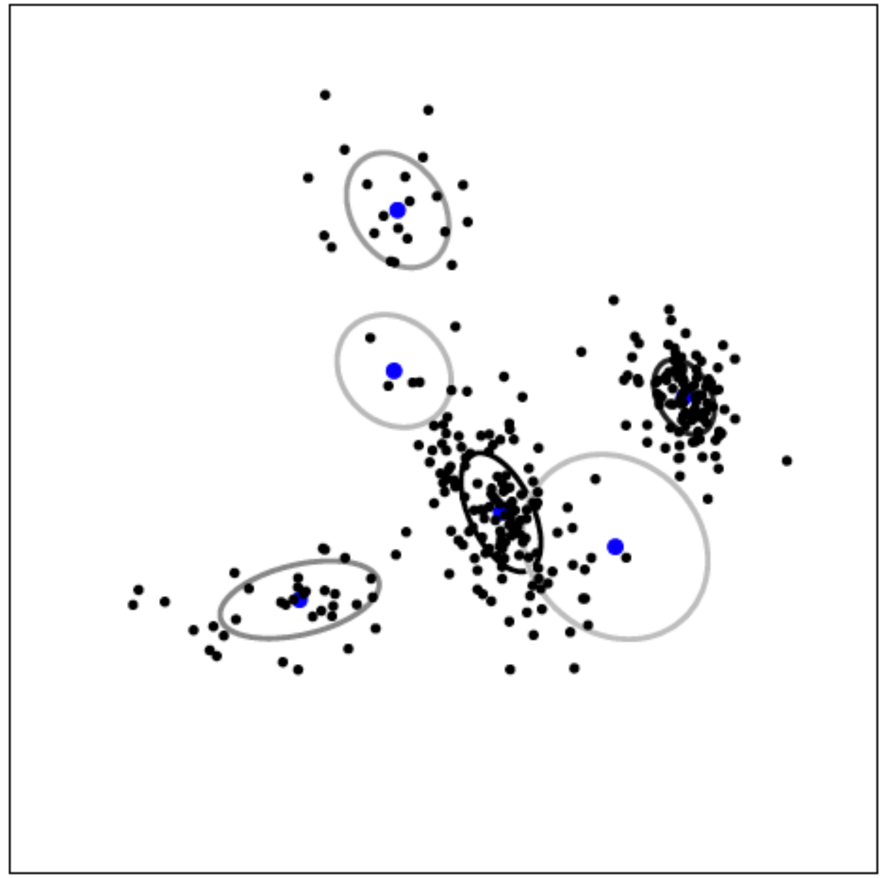


log p(x | π, θ) = –399.82

Collapsed DP Sampler: 10 Iterations

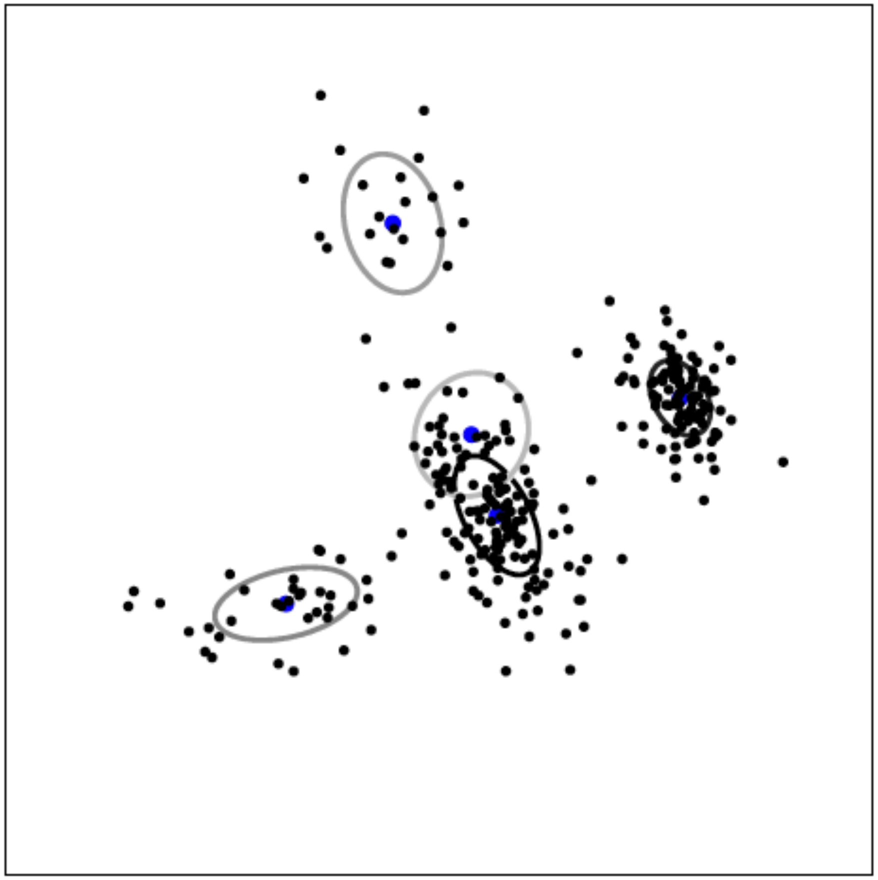


log p(x | π, θ) = –398.32

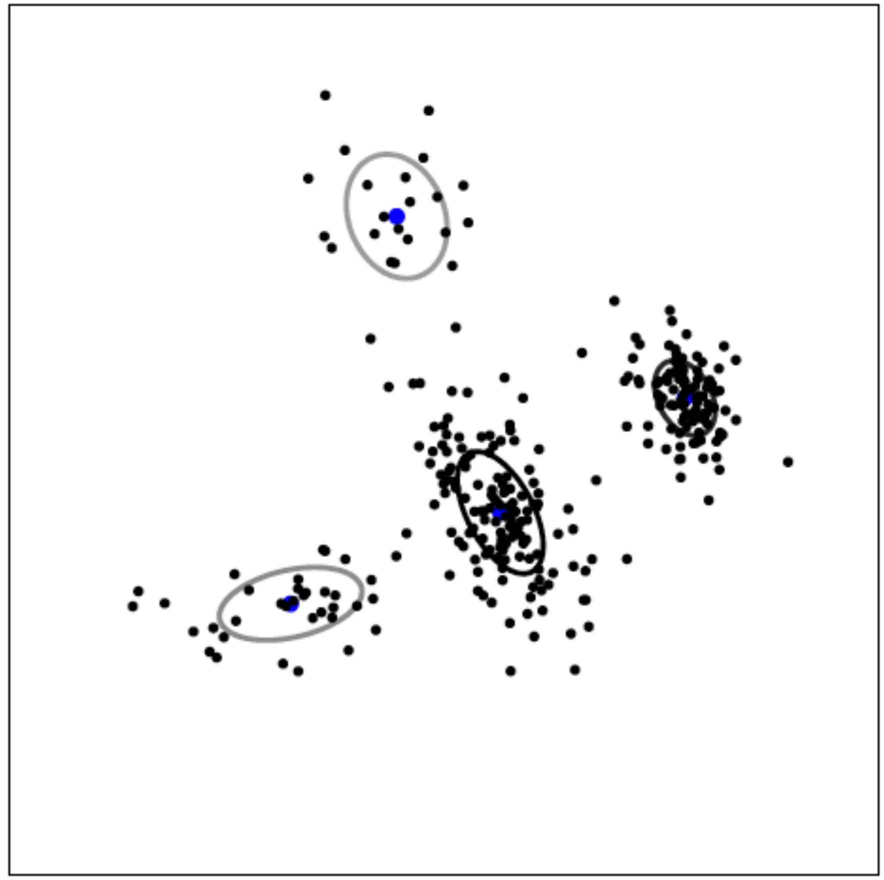


log p(x | π, θ) = –399.08

Collapsed DP Sampler: 50 Iterations

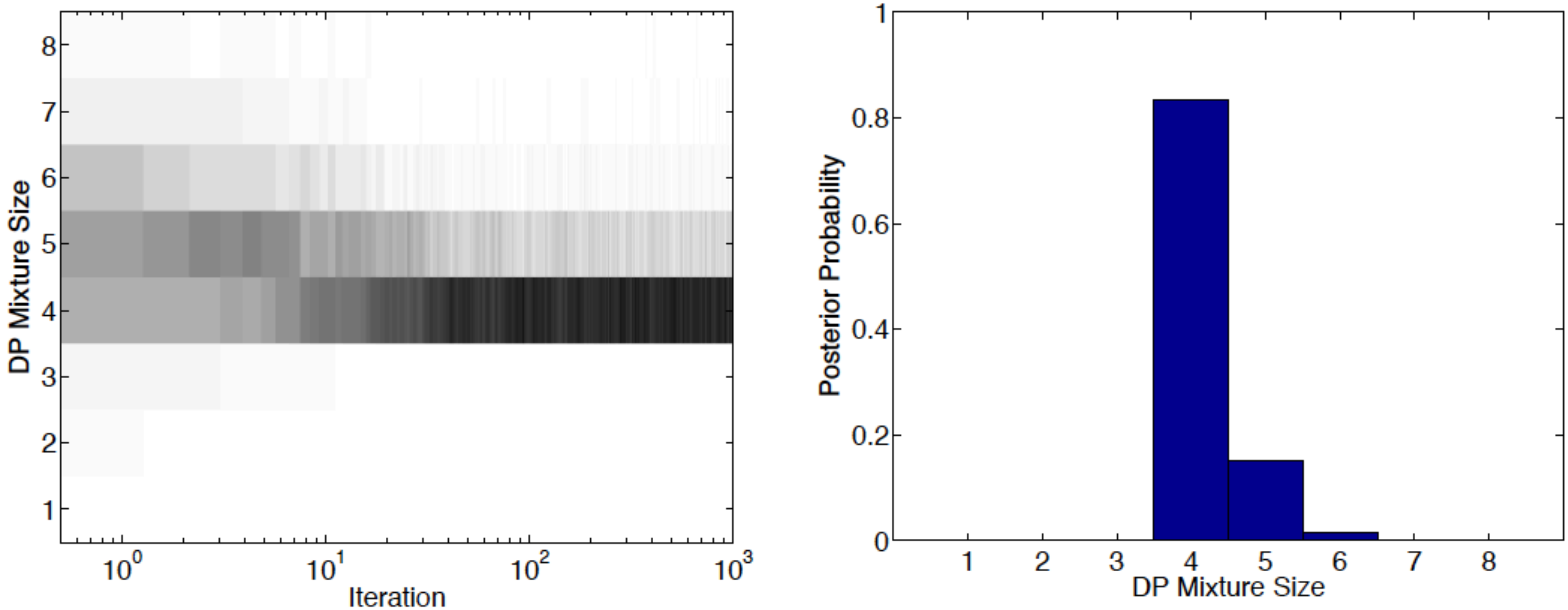


log p(x | π, θ) = –397.67



log p(x | π, θ) = –396.71

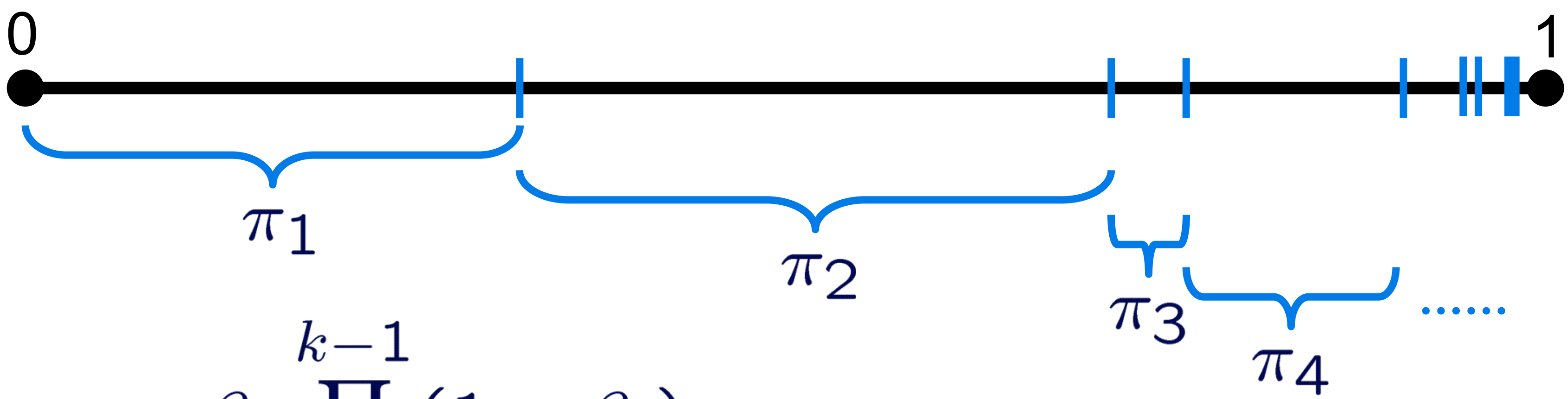
DP Posterior Number of Clusters



These results also place a prior distribution on the DP concentration parameter α, and resample it as part of the MCMC inference (Escobar & West, 1995)

DP Stick-Breaking Construction

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



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

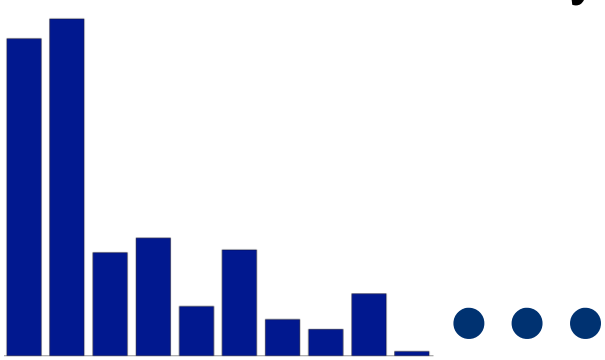
$$\beta_k \sim \text{Beta}(1, \alpha)$$

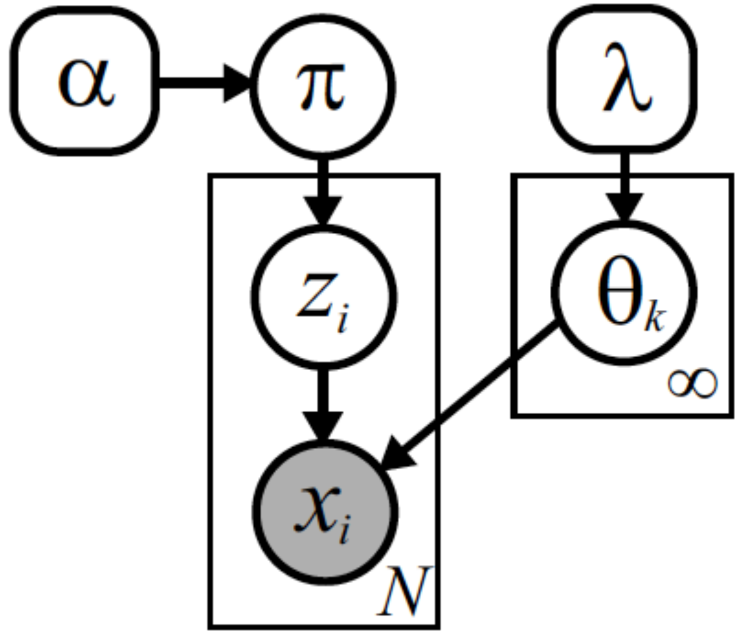
$\alpha \longrightarrow$ concentration parameter

Stick-Breaking Construction: *Sethuraman, 1994*

DP Mixture: Stick-Breaking Sampler

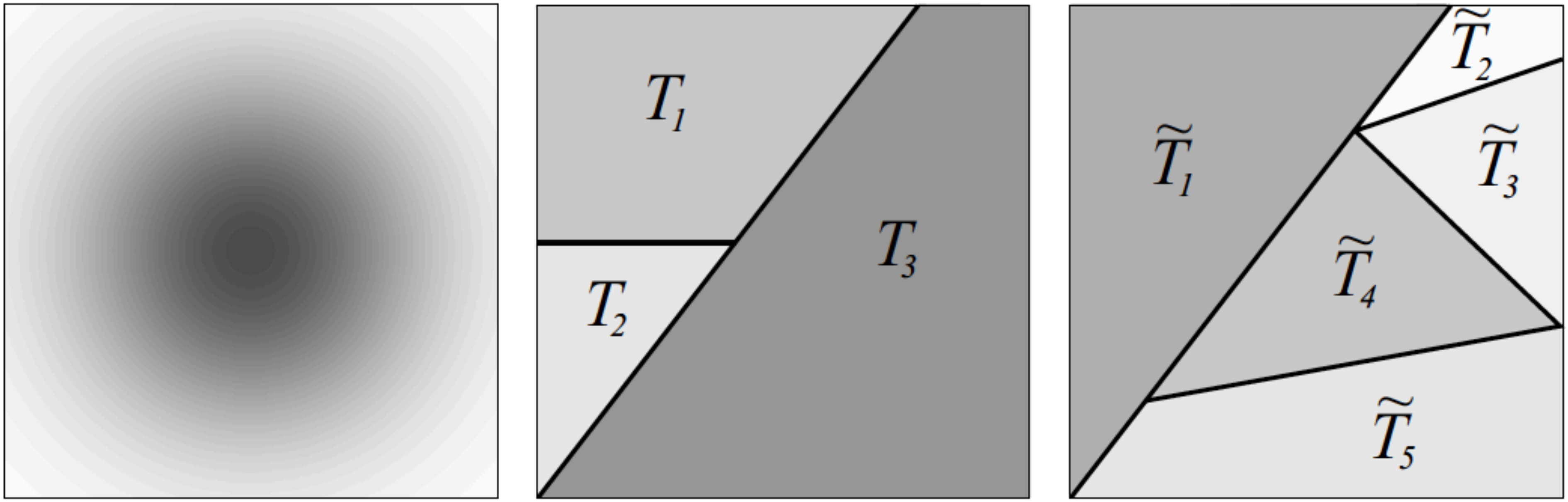
- Explicitly instantiate and resample cluster sizes (stick-breaking prior)
- Without marginalization there are infinitely many cluster size parameters
- Blocked Gibbs sampler of Ishwaran & James (2001) uses analytic bounds to build a finite truncation
- Main benefit: Flexibility





$$\pi \sim \mathrm{GEM}(\alpha)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$
$$z_i \sim \pi$$
$$x_i \sim F(\theta_{z_i})$$

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:

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

Properties of the Dirichlet Process

$(\mathcal{X}, \mathcal{B})$ is some measurable space (the sigma-algebra $\mathcal{B}$ is a collection of sets, and defines the events to be assigned probabilities)

$\mathcal{P}$ is the collection of all probability measures P on $(\mathcal{X}, \mathcal{B})$

ν^X is the posterior distribution of a random probability measure P, with prior distribution ν, given observed data $X \sim P$

P1 $\mathcal{D}_\alpha$ is a probability measure on $(\mathcal{P}, \mathcal{C})$,

P2 $\mathcal{D}_\alpha$ gives probability one to the subset of all discrete probability measures on $(\mathcal{X}, \mathcal{B})$, and

P3 the posterior distribution $\mathcal{D}_\alpha^X$ is the Dirichlet measure $\mathcal{D}_{\alpha+\delta_X}$ where δ_X is the probability measure degenerate at X.

The approach of Sethuraman (1994, 1980):

1. Explicitly construct a process which trivially satisfies P1-P2
2. Show that this process has Dirichlet marginals, and thus is in fact the Dirichlet process
3. Use this construction to establish P3

The Stick-Breaking Construction: Trivially A Discrete Probability Measure

In my notation from earlier this lecture, and past lectures:

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)*.*

From Stick-Breaking to Dirichlet: Setup

In Sethuraman's notation:

$$P(\boldsymbol{\theta}, \mathbf{Y}; B) = P(B) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(B)$$

$$p_n = \theta_n \prod_{1 \le m \le n-1} (1 - \theta_m)$$

$(\theta_1, \theta_2, \ldots)$ are i.i.d. with distribution $B(1, \alpha(\mathcal{X}))$

$(Y_1, Y_2, \ldots)$ are i.i.d. with distribution $\beta(B) = \alpha(B)/\alpha(\mathcal{X})$

A key consequence of the stick-breaking recursion:

$$P(\boldsymbol{\theta}, \mathbf{Y}; B) = \theta_1 \delta_{Y_1}(B) + (1 - \theta_1) P(\boldsymbol{\theta}^*, \mathbf{Y}^*; B)$$

where $\theta_n^* = \theta_{n+1}$ $\quad Y_n^* = Y_{n+1}$

Equality in distribution: $P \stackrel{\text{st}}{=} \theta_1 \delta_{Y_1} + (1 - \theta_1) P$

From Stick-Breaking to Dirichlet: Step 1

Theorem 3.4. *Let $\{B_1, B_2, \ldots, B_k\}$ be a measurable partition of $\mathcal{X}$ and let $\mathbf{P} = (P(B_1), P(B_2), \ldots, P(B_k))$. Then the distribution of $\mathbf{P}$ is the k-dimensional Dirichlet measure $\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\ldots,\alpha(B_k))}$.*

Stick-breaking measure: $$P \stackrel{\text{st}}{=} \theta_1 \delta_{Y_1} + (1-\theta_1)P$$

Evaluating on finite partition: $$\mathbf{P} \stackrel{\text{st}}{=} \theta_1 \mathbf{D} + (1-\theta_1)\mathbf{P}$$

$\mathbf{D}$ takes the value $\mathbf{e}_j$ with probability $\beta(B_j)$

The plan:

We first verify that the k-dimensional Dirichlet measure for $\mathbf{P}$ satisfies the distributional equation (3.4) and then show that this solution is the unique solution.

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:

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

From Stick-Breaking to Dirichlet: Step 2

Evaluating on finite partition: $\mathbf{P} \stackrel{\text{st}}{=} \theta_1 \mathbf{D} + (1-\theta_1)\mathbf{P}$

$\mathbf{D}$ takes the value $\mathbf{e}_j$ with probability $\beta(B_j)$

- Assume that P has distribution $\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\ldots,\alpha(B_k))}$
- Suppose first that $\mathbf{D} = \mathbf{e}_j$, we are interested in

  $$\theta_1 \mathcal{D}_{\mathbf{e}_j} + (1-\theta_1)\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\ldots,\alpha(B_k))}$$

  where samples from $\mathcal{D}_{\mathbf{e}_j}$ equal $\mathbf{e}_j$ with probability one
- This has distribution $\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\ldots,\alpha(B_k))+\mathbf{e}_j}$

Lemma 3.1. *Let $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ and $\delta = (\delta_1, \delta_2, \ldots, \delta_k)$ be k-dimensional vectors. Let U, V be independent k-dimensional random vectors with Dirichlet distributions $\mathcal{D}_\gamma$ and $\mathcal{D}_\delta$, respectively. Let W be independent of (U, V) and have a Beta distribution $B(\gamma, \delta)$, where $\gamma = \sum \gamma_j$ and $\delta = \sum \delta_j$. Then the distribution of $WU + (1-W)V$ is the Dirichlet distribution $\mathcal{D}_{\gamma+\delta}$.*

Intuition via Moments

From Stick-Breaking to Dirichlet: Step 3

Evaluating on finite partition: $\mathbf{P} \stackrel{\text{st}}{=} \theta_1 \mathbf{D} + (1-\theta_1)\mathbf{P}$

$\mathbf{D}$ takes the value $\mathbf{e}_j$ with probability $\beta(B_j)$

- Assume that P has distribution $\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\dots,\alpha(B_k))}$
- Given that $\mathbf{D} = \mathbf{e}_j$, the right-hand-side has distribution

$$\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\dots,\alpha(B_k))+\mathbf{e}_j}$$

- Averaging over $\mathbf{D}$ with weights $\beta(B_j) = \alpha(B_j)/\alpha(\mathcal{X})$ gives

$$\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\dots,\alpha(B_k))}$$

Lemma 3.2. *Let* $\gamma = (\gamma_1, \dots, \gamma_k)$, $\gamma = \sum \gamma_j$ *and let* $\beta_j = \gamma_j/\gamma, j = 1, 2, \dots, k$. *Then*

$$\sum \beta_j \mathcal{D}_{\gamma+\mathbf{e}_j} = \mathcal{D}_\gamma.$$

This conclusion can also be written as $E(\mathcal{D}_{\gamma+\mathbf{Z}}) = \mathcal{D}_\gamma$, *where* $\mathbf{Z}$ *is a random vector that takes the values* $\mathbf{e}_j$ *with probability* $\gamma_j/\gamma, j = 1, \dots, k$.

From Stick-Breaking to Dirichlet: Step 4

Evaluating on finite partition: $\mathbf{P} \stackrel{\text{st}}{=} \theta_1 \mathbf{D} + (1 - \theta_1)\mathbf{P}$

$\mathbf{D}$ takes the value $\mathbf{e}_j$ with probability $\beta(B_j)$

- We have shown that $\mathcal{D}_{(\alpha(B_1),\alpha(B_2),\ldots,\alpha(B_k))}$ is a solution of this recurrence
- In fact, it is the unique solution (proof by contradiction)
- Intuition for Lemma 3.2: Prior distribution can always be written as a weighted combination of posteriors

DP Posteriors and Conjugacy

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:*

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

Proof Hint: For any finite partition, we have

$$p\big((G(T_1), \ldots, G(T_K)) \mid \bar{\theta} \in T_k\big) = \mathrm{Dir}(\alpha H(T_1), \ldots, \alpha H(T_k) + 1, \ldots, \alpha H(T_K))$$

An observation must be of one of the countably infinite atoms which compose the random Dirichlet measure

DPs are Neutral: "Almost" independent

The distribution of a random probability measure G is neutral *with respect to a finite partition* $(T_1, \ldots, T_K)$ *iff*

$$G(T_k) \quad \text{is independent of} \quad \left\{ \frac{G(T_\ell)}{1 - G(T_k)} \;\middle|\; \ell \neq k \right\}$$

$$\text{given that } G(T_k) < 1.$$

Theorem 2.5.2. *Consider a distribution $\mathcal{P}$ on probability measures G for some space Θ. Assume that $\mathcal{P}$ assigns positive probability to more than one measure G, and that with probability one samples $G \sim \mathcal{P}$ assign positive measure to at least three distinct points $\theta \in \Theta$. The following conditions are then equivalent:*

(i) $\mathcal{P} = \mathrm{DP}(\alpha, H)$ *is a Dirichlet process for some base measure H on Θ.*

(ii) $\mathcal{P}$ *is neutral with respect to every finite, measurable partition of Θ.*

(iii) For every measurable $T \subset \Theta$, and any N observations $\bar{\theta}_i \sim G$, the posterior distribution $p\big(G(T) \mid \bar{\theta}_1, \ldots, \bar{\theta}_N\big)$ depends only on the number of observations that fall within T (and not their particular locations).