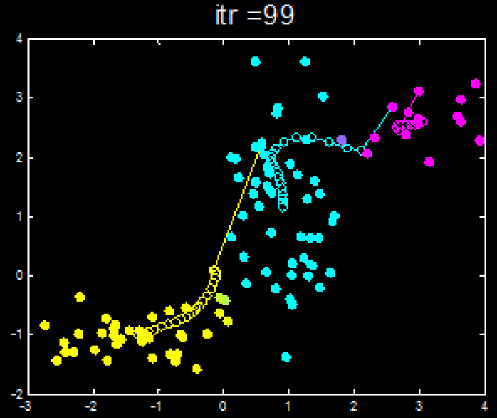
Dirichlet Process

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST

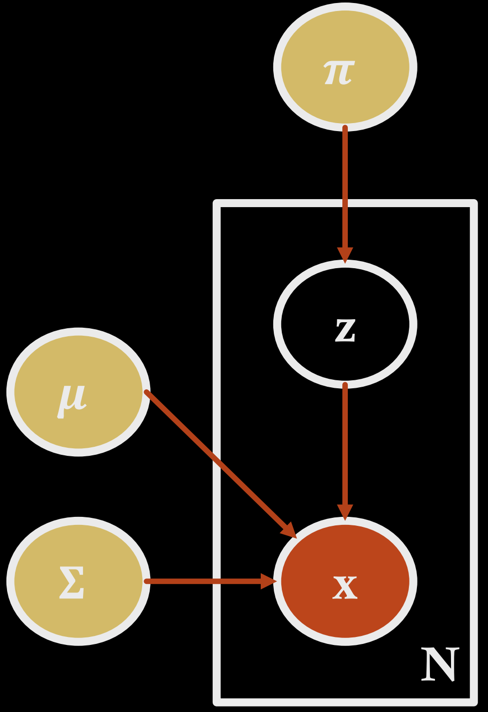
icmoon@kaist.ac.kr

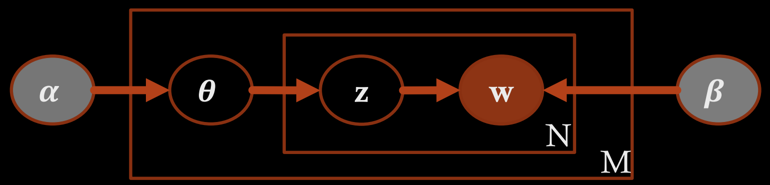
Definition of Dirichlet Process



Detour: Gaussian Mixture Model

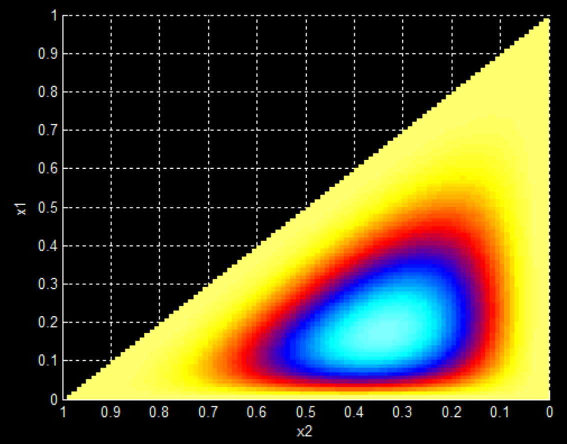
- Let's assume that the data points are drawn from a mixture distribution of multiple multivariate Gaussian distributions
  - $P(x) = \sum_{k=1}^{K} P(z_k)P(x|z) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$
  - How to model such mixture?
    - Mixing coefficient, or Selection variable: z_k
      - The selection is stochastic which follows the multinomial distribution
      - $z_k \in \{0,1\}, \sum_k z_k = 1, P(z_k = 1) = \pi_k, \sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
      - $P(Z) = \prod_{k=1}^{K} \pi_k^{z_k}$
    - Mixture component
      - $P(X|z_k = 1) = N(x|\mu_k, \Sigma_k) \rightarrow P(X|Z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$
  - This is the marginalized probability. How about conditional?
    - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P(x|z_j = 1)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$
- Log likelihood of the entire dataset is
  - $\ln P(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)\}$

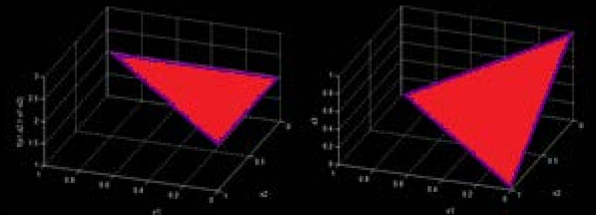




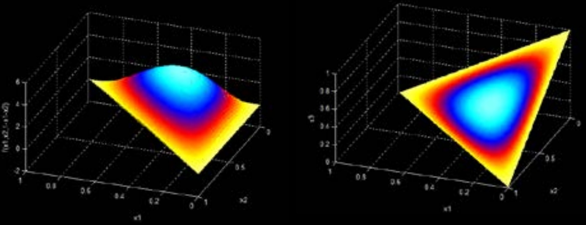
Detour: Dirichlet Distribution

- **Generative Process**
  - $\theta_i \sim Dir(\alpha), i \in \{1, \ldots, M\}$, $\varphi_k \sim Dir(\beta), k \in \{1, \ldots, K\}$
  - $z_{i,l} \sim Mult(\theta_i), i \in \{1, \ldots, M\}, l \in \{1, \ldots, N\}$, $w_{i,l} \sim Mult(\varphi_{z_{i,l}}), i \in \{1, \ldots, M\}, l \in \{1, \ldots, N\}$
- Dirichlet Distribution
  - $P(x_1, \ldots, x_K | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} {x_i}^{\alpha_i - 1}$
    - $x_1, \ldots, x_{K-1} > 0$
    - $x_1 + \cdots + x_{K-1} < 1$
    - $x_K = 1 - x_1 - \cdots - x_{K-1}$
    - $\alpha_i > 0$

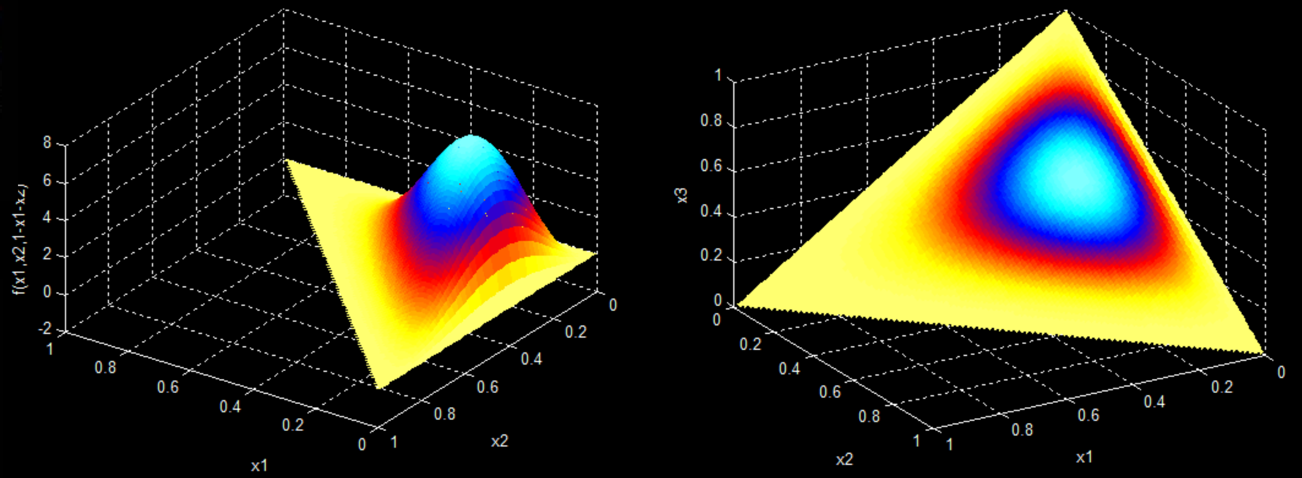




$[\alpha_1, \alpha_2, \alpha_3] = [1,1,1]$



$[\alpha_1, \alpha_2, \alpha_3] = [2,2,2]$

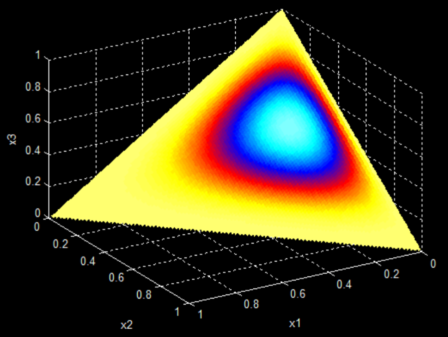


$[\alpha_1, \alpha_2, \alpha_3] = [2,3,4]$

Multinomial-Dirichlet Conjugate Relation

- Multinomial distribution
  - N independently and identically distributed instances, $N = \sum_i c_i$
  - c_i is the number of occurrences of the i-th choice
  - $P(D|\theta) = \frac{N!}{\prod_i c_i!} \prod_i \theta_i^{c_i}$
- Dirichlet distribution
  - $P(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i - 1}$
- Bayesian Posterior
  - $P(\theta|D,\alpha) \propto P(D|\theta)P(\theta|\alpha) = \frac{N!}{\prod_i c_i!} \prod_i \theta_i^{c_i} \frac{1}{B(\alpha)} \prod_i \theta_i^{\alpha_i - 1} = \frac{N!}{B(\alpha)\prod_i c_i!} \prod_i \theta_i^{\alpha_i + c_i - 1} \propto \prod_i \theta_i^{\alpha_i + c_i - 1}$
  - $P(\theta|D,\alpha) = \frac{1}{B(\alpha + c)} \prod_i \theta_i^{\alpha_i + c_i - 1}$
  - Coming back to the Dirichlet distribution : Conjugate Prior
    - The likelihood of the Dirichlet distribution is the conjugate prior of the multinomial distribution
- Dirichlet distribution with D as a single observation with i-th choice
  - $\theta|\alpha \sim Dir(\alpha_1, \dots, \alpha_i, \dots, \alpha_N)$
  - $\theta|\alpha, D \sim Dir(\alpha_1, \dots, \alpha_i + 1, \dots, \alpha_N)$

Dirichlet Process



$Dir(2,3,4)$

- Dirichlet process, $G|\alpha, H \sim DP(\alpha, H)$
  - $\big(G(A_1), \dots, G(A_r)\big)|\alpha, H \sim Dir(\alpha H(A_1), \dots, \alpha H(A_r))$
    - $A_1 \cap \cdots \cap A_r = \emptyset, \quad A_1 \cup \cdots \cup A_r = \Theta$
  - Properties

$$E[G(A)] = H(A)$$

$$V[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

    - H : Base distribution
    - α : Concentration parameter, strength parameter (strength of prior)
- Posterior distribution given a dataset of $\theta_1 \dots \theta_n$
  - $Posterior \propto Likelihood \times Prior$
  - Multinomial-Dirichlet conjugate relationship
    - The posterior becomes the Dirichlet distribution, again, adjusted to reflect the likelihood
  - $\big(G(A_1), \dots, G(A_r)\big)|\theta_1 \dots \theta_n, \alpha, H \sim Dir(\alpha H(A_1) + n_1, \dots, \alpha H(A_r) + n_r)$
    - $n_k = |\{\theta_i | \theta_i \in A_k, 1 \le i \le n\}|$

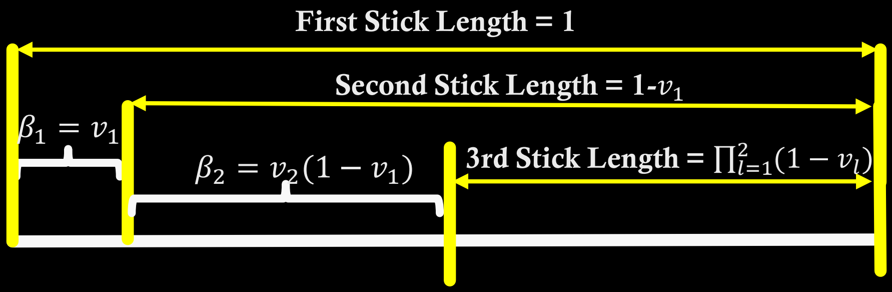
$$G|\theta_1 \dots \theta_n, \alpha, H \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right)$$

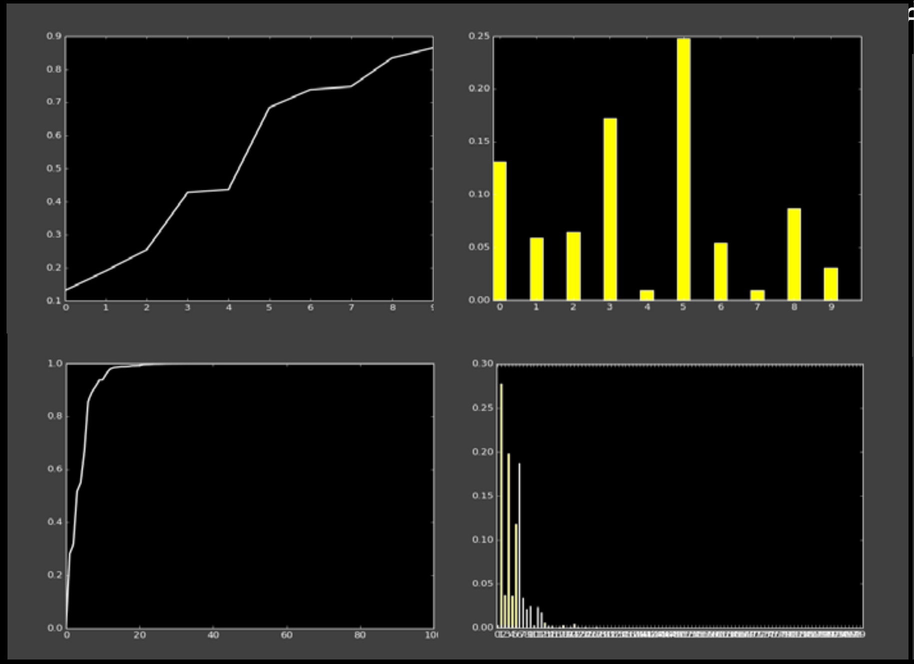
Sampling from Dirichlet Process

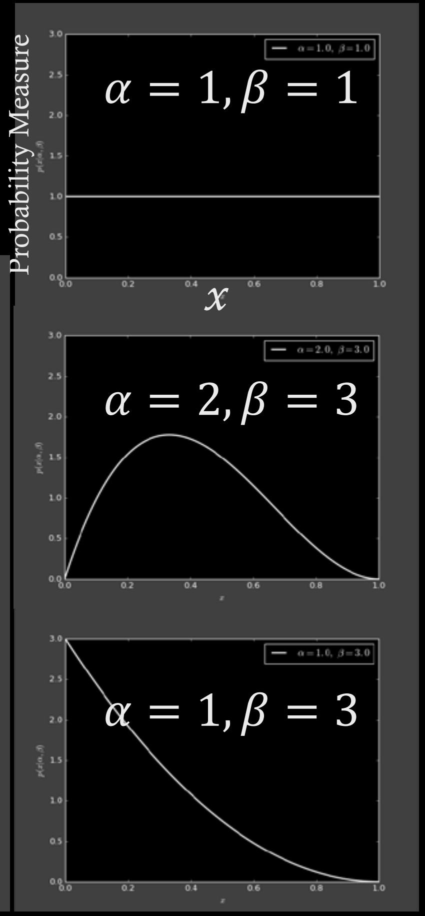
- Dirichlet process
  - $(G(A_1), \ldots, G(A_r))|\alpha, H \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_r))$
  - $G|\theta_1 \ldots \theta_n, \alpha, H \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right)$
- Definition is done, but how to realize the definition?
  - How to draw an instance, or a distribution, *G*, from the Dirichlet process?
  - How to draw an instance, θ_i, from the distribution, *G*?
- Multiple generation *schemes*, or *construction*, exist
  - From the definition of Dirichlet process to the sample from the Dirichlet process
  - Stick Breaking Scheme
  - Polya Urn Scheme
  - Chinese Restaurant Process Scheme

Stick-Breaking Construction

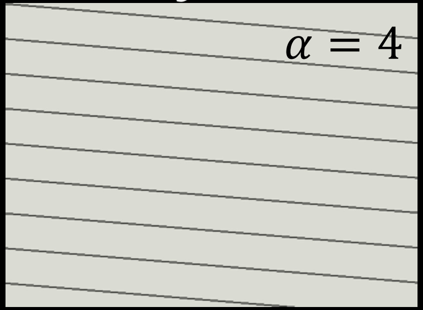
- Imagine that we create a probability mass function on infinite choices
  - $k = 1,2, \ldots, \infty$
  - $v_k|\alpha \sim Beta(1, \alpha)$
  - $\beta_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$
- Common notation is
  - $\beta \sim GEM(\alpha)$
- We were constructing a distribution for the Dirichlet process
  - $G|\alpha, H \sim DP(\alpha, H)$
    - $\beta \sim GEM(\alpha)$
    - $G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$
    - $\theta_k|H \sim H$
  - θ_k chooses a n-th broken stick, and the stick length is the prob.
  - We know the existence of the infinite-th stick length.
- Exponential growth in CDF
- → Discount the growth
- → Pitman-Yor Process
  - Close to Power law dist.
  - Useful for language models…

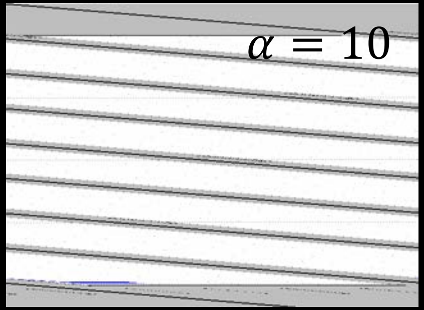






Polya Urn Scheme





- Dirichlet process
  - $G|\theta_1 \dots \theta_n, \alpha, H \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right)$
    - $G|\alpha, H \sim DP(\alpha, H)$
      - $\left(G(A_1), \dots, G(A_r)\right)|\alpha, H \sim Dir(\alpha H(A_1), \dots, \alpha H(A_r))$
      - $E[G(A)] = H(A)$
  - $\theta_n|\theta_1 \dots \theta_{n-1}, \alpha, H \sim DP\left(\alpha + n - 1, \frac{\alpha}{\alpha+n-1} H + \frac{n-1}{\alpha+n-1} \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{n-1}\right)$
  - $E[\theta_n|\theta_1 \dots \theta_{n-1}, \alpha, H] \sim \frac{\alpha}{\alpha+n-1} H + \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1} \sim \frac{\alpha}{\alpha+n-1} H + \frac{\sum_{k=1}^{K} N_k \delta_{\theta_k}}{\alpha+n-1}$, N_k : the number of k-th choice occurrences
  - This enables sampling an observation from the Dirichlet process without constructing $G|\alpha, H \sim DP(\alpha, H)$
  - Stick-breaking (distribution) *construction* vs. Polya Urn *sampling* from distribution
- Polya Urn Scheme
  - Create an empty urn
  - Do
    - toss = Coin toss from $[0, \alpha + n - 1]$
    - If $0 \le$ toss $< \alpha$
      - Add a ball to the urn by paining the ball as a sample from $\theta_n \sim H$
    - If α $\le$ toss $< \alpha + n - 1$
      - Pick a ball from the urn
      - Return the ball and a new ball with the same color to the urn

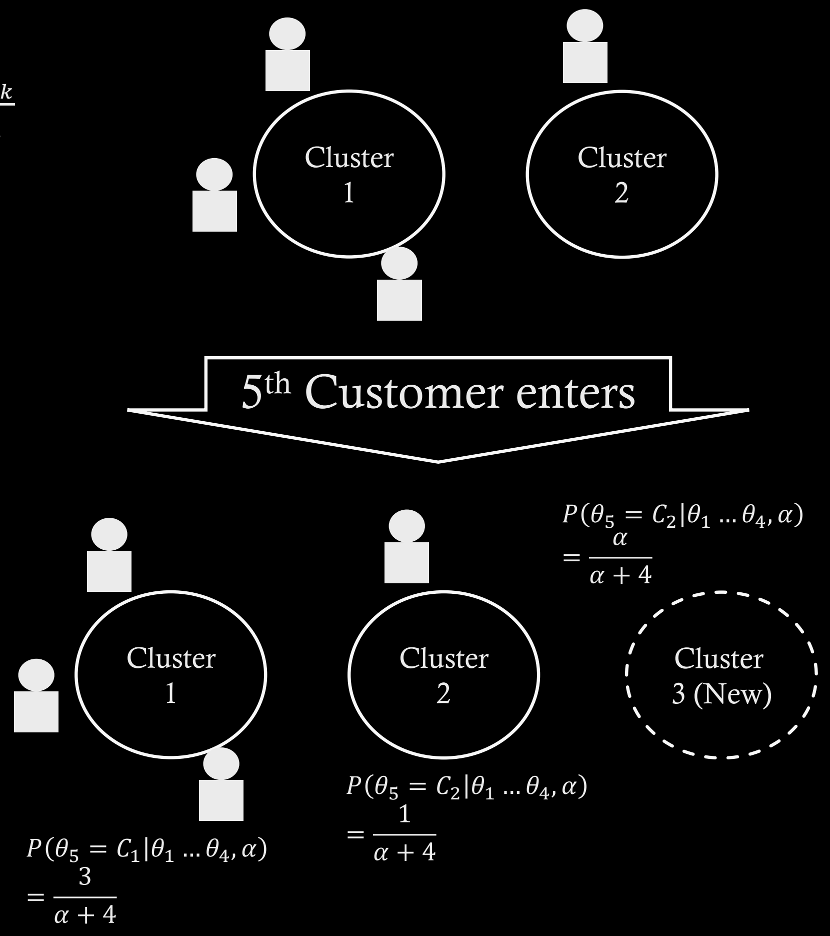
Chinese Restaurant Process

- Dirichlet process
  - $G|\theta_1 \dots \theta_n, \alpha, H \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right)$
  - $E[\theta_n|\theta_1 \dots \theta_{n-1}, \alpha, H]$

$$\sim \frac{\alpha}{\alpha+n-1} H + \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha+n-1}$$
$$\sim \frac{\alpha}{\alpha+n-1} H + \frac{\sum_{k=1}^{K} N_k \delta_{\theta_k}}{\alpha+n-1}$$

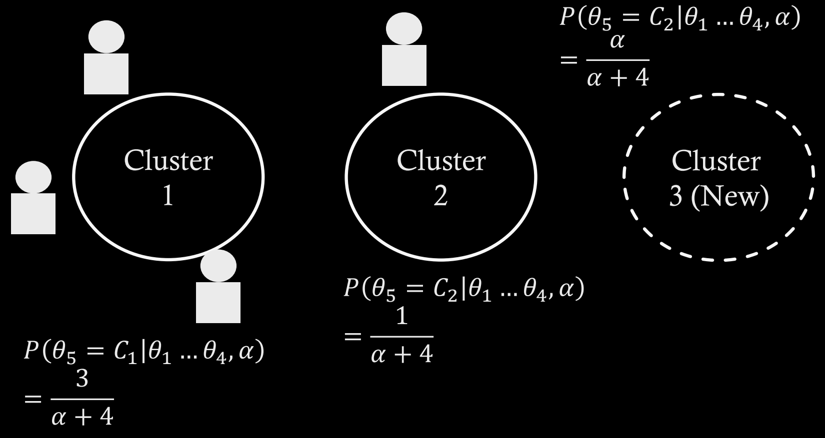
    N_k : the number of k-th choice occurrences

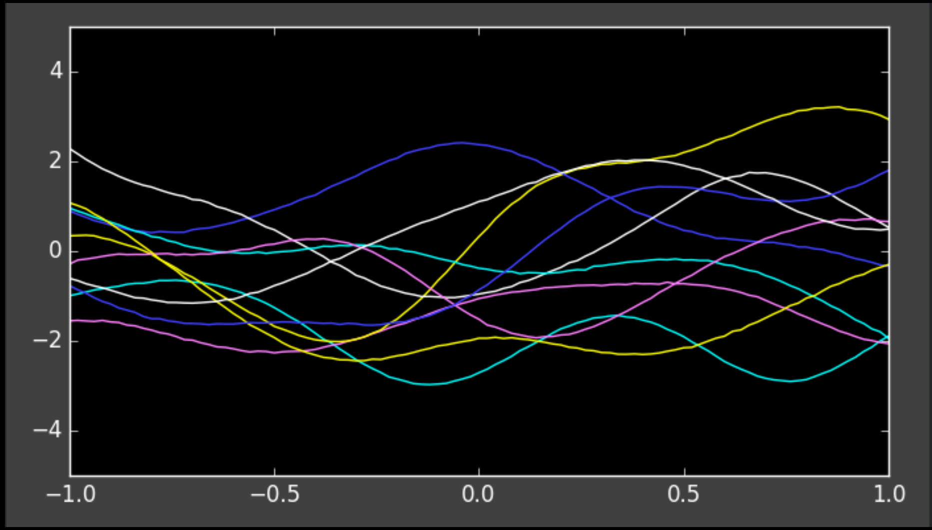
  - $P(\theta_n|\theta_1 \dots \theta_{n-1}, \alpha) = \begin{cases} \frac{N_k}{\alpha+n-1} \\ \frac{\alpha}{\alpha+n-1} \end{cases}$
- Chinese restaurant process
  - Assume Infinite number of tables in a restaurant
  - First customer sits at the first table
  - Loop for Customer N sits at:
    - 1) Table *k* with $P(\theta_n|\theta_1 \dots \theta_{n-1}, \alpha) = \frac{N_k}{\alpha+n-1}$
    - 2) A new table *k+1* with $P(\theta_n|\theta_1 \dots \theta_{n-1}, \alpha) = \frac{\alpha}{\alpha+n-1}$
- Properties of Chinese restaurant process
  - Clustering formation
  - Rich-get-richer property
  - No fixed number of clusters with a fixed number of instances
  - Almost identical to Polya Urn Scheme



Detour: Random Process

- Random process, a.k.a. stochastic process, is
  - An infinite indexed collection of random variables, $\{X(t) | t \in T\}$
    - Index parameter : t
      - Can be time, space….
  - A function, $X(t, \omega)$, where $t \in T$ and $\omega \in \Omega$
    - Outcome of the underlying random experiment : ω
    - Fixed t → $X(t, \omega)$ is a random variable over Ω
    - Fixed ω → $X(t, \omega)$ is a deterministic function of t , a sample function
- Example of random process
  - Gaussian process
    - Fixed t, a random variable following a Gaussian distribution
    - Fixed ω, a deterministic curve of t
  - Dirichlet process
    - Fixed t, a random variable following a Dirichlet distribution
    - Fixed ω, a deterministic placement over clusters



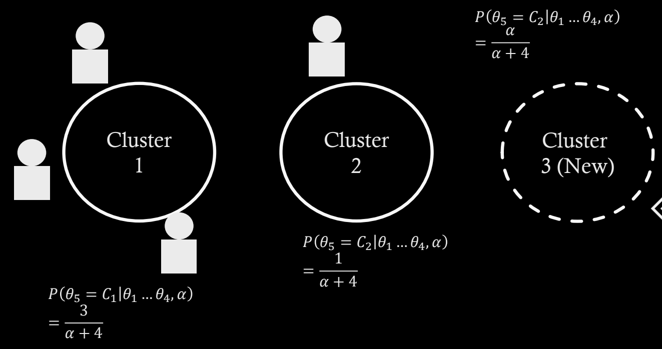


de Finetti's Theorem

- Exchangeability
  - A joint probability distribution is exchangeable if it is invariant to permutation
  - Given a permutation of S
  - $\mathrm{P}(x_1, x_2, \dots, x_N) = \mathrm{P}\left(x_{S(1)}, x_{S(2)}, \dots, x_{S(N)}\right)$
- (De Finetti, 1935) If $(x_1, x_2, \dots)$ are infinitely exchangeable, then the joint probability $\mathrm{P}(x_1, x_2, \dots, x_N)$ has a representation as a mixture

$$\mathrm{P}(x_1, x_2, \dots, x_N) = \int \left(\prod_{i=1}^{N} P(x_i|\theta) \right) dP(\theta) = \int P(\theta) \left(\prod_{i=1}^{N} P(x_i|\theta) \right) d\theta$$

  For some random variable θ
  - Independent and identically distributed → Exchangeable
  - Exchangeable → IID : No. A counter example is the Polya urn sampling
- Chinese restaurant process is an exchangeable process
  - No proof in this scope
  - Why is exchangeability important?
    - Enables a simple derivation of Gibbs sampler for the inference
    - We remove the instance of the next Gibbs sampling from the existing cluster assignment



Detour: Concept of Gibbs Sampling

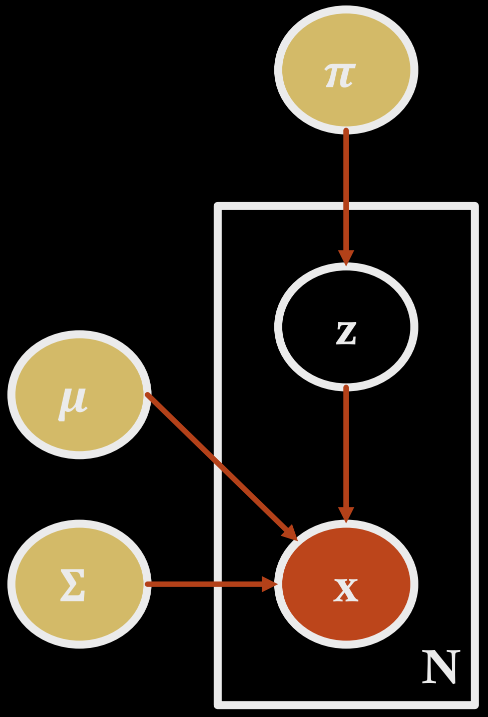
- Each step involves **replacing** the value of one of the variables by a value drawn from the distribution of that variable conditioned on the values of the remaining variables
- Repeated either by cycling through the variables in some particular order or by choosing the variable to be updated at each step at random from some distribution
- Example
  1. Full joint probability : $p(z_1, z_2, z_3)$
  2. Sample $z_1 \sim p\left(z_1 \mid z_2^{(\tau)}, z_3^{(\tau)}\right) \rightarrow$ Obtain a value $z_1^{(\tau+1)}$
  3. Sample $z_2 \sim p\left(z_2 \mid z_1^{(\tau+1)}, z_3^{(\tau)}\right) \rightarrow$ Obtain a value $z_2^{(\tau+1)}$
  4. Sample $z_3 \sim p\left(z_3 \mid z_1^{(\tau+1)}, z_2^{(\tau+1)}\right) \rightarrow$ Obtain a value $z_3^{(\tau+1)}$

$\left\{z_1^{(\tau)}, z_2^{(\tau)}, z_3^{(\tau)}\right\}$ $\left\{z_1^{(\tau+1)}, z_2^{(\tau)}, z_3^{(\tau)}\right\}$ $\left\{z_1^{(\tau+1)}, z_2^{(\tau+1)}, z_3^{(\tau)}\right\}$ $\left\{z_1^{(\tau+1)}, z_2^{(\tau+1)}, z_3^{(\tau+1)}\right\}$

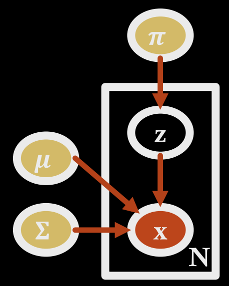
Dirichlet Process Mixture Model

Detour: Gaussian Mixture Model

- Let's assume that the data points are drawn from a mixture distribution of multiple multivariate Gaussian distributions
  - $P(x) = \sum_{k=1}^{K} P(z_k)P(x|z) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$
  - How to model such mixture?
    - Mixing coefficient, or Selection variable: z_k
      - The selection is stochastic which follows the multinomial distribution
      - $z_k \in \{0,1\}, \sum_k z_k = 1, P(z_k = 1) = \pi_k, \sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
      - $P(Z) = \prod_{k=1}^{K} {\pi_k}^{z_k}$
    - Mixture component
      - $P(X|z_k = 1) = N(x|\mu_k, \Sigma_k) \rightarrow P(X|Z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$
  - This is the marginalized probability. How about conditional?
    - $\gamma(z_{nk}) \equiv p(z_k = 1|x_n) = \frac{P(z_k=1)P(x|z_k = 1)}{\sum_{j=1}^{K} P(z_j=1)P(x|z_j = 1)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$
- Log likelihood of the entire dataset is
  - $\ln P(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)\}$



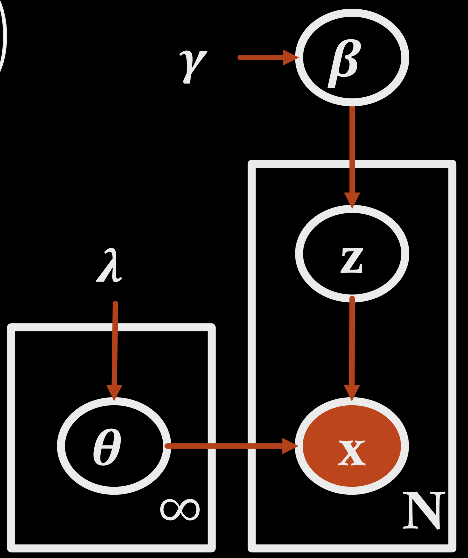
Dirichlet Process Mixture Model



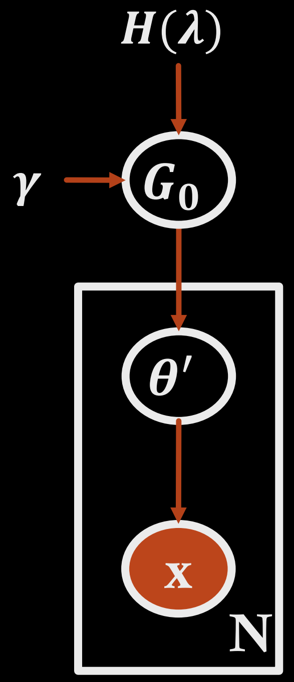
GMM

- Common usage of Dirichlet process : Prior on parameters of a mixture model
  - Like $P(z_k = 1) = \pi_k$
    - $z_k \in \{0,1\}, \sum_k z_k = 1, P(z_k = 1) = \pi_k, \sum_{k=1}^{K} \pi_k = 1, 0 \le \pi_k \le 1$
- Indicator representation of GMM with infinite K
  - $\beta|\gamma \sim GEM(\gamma)$, $\theta_k|H,\lambda \sim H(\lambda)$, $z_i|\beta \sim \beta$, $x_i|\{\theta_k\}_{k=1}^{\infty}, z_i \sim F(\theta_{z_i})$
    - $\beta \sim GEM(\alpha) \rightarrow k = 1,2,\ldots,\infty, v_k|\alpha \sim Beta(1,\alpha)$, $\beta_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$
- Alternative representation of GMM with infinite K
  - $G_0|H,\gamma \sim DP(\gamma, H)$, $\theta_i'|G_0 \sim G_0$, $x_i|\theta_i' \sim F(\theta_i')$
    - $\theta_n|\theta_1 \ldots \theta_{n-1}, \gamma, H \sim DP\left(\gamma + n - 1, \frac{\gamma}{\gamma+n-1} H + \frac{n-1}{\gamma+n-1} \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{n-1}\right)$
- Continuously updating the assignment of an instance
  - Learning concept
    - de Finetti's theorem + Chinese restaurant process + Gibbs Sampling
  - Each assignment
    - Surely updates the parameter of each cluster
    - May create a new cluster

Indicator View Graphical Model



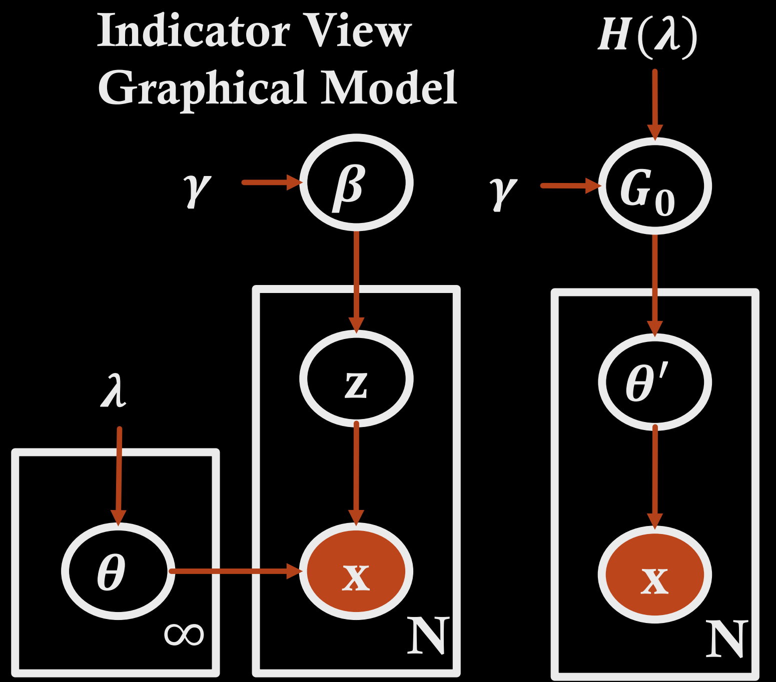
Alternative Representation For Mixture Models



Implementation Details of DPMM

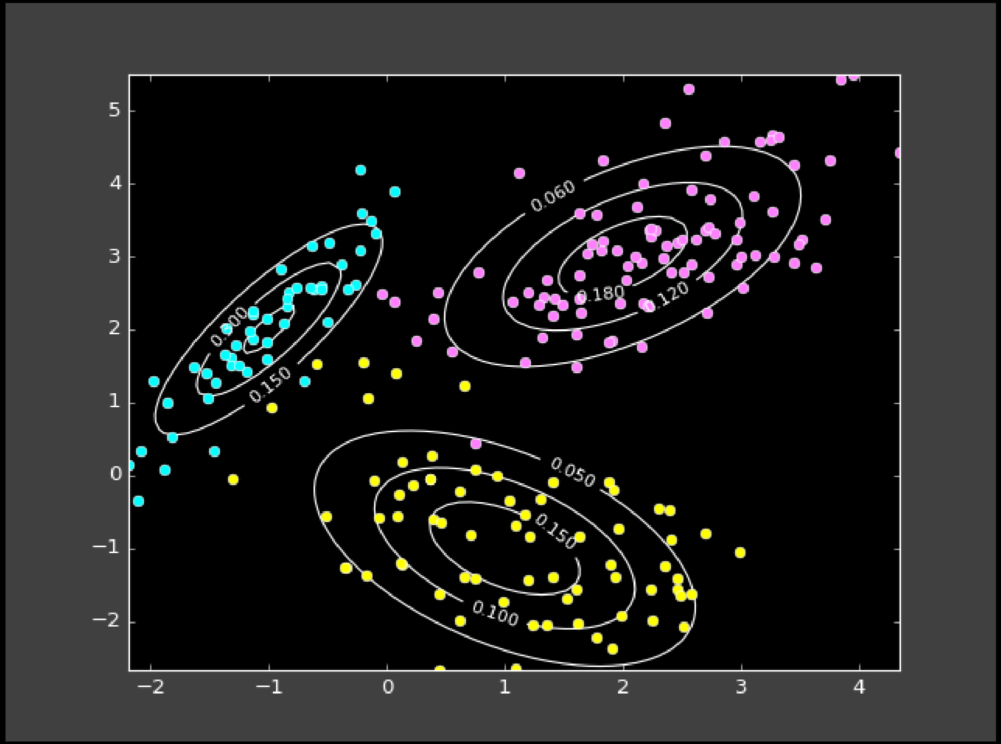
- Online update of the component parameter
  - $G_0|H,\gamma \sim DP(\gamma, H)$, $\theta_i'|G_0 \sim G_0$, $x_i|\theta_i' \sim F(\theta_i')$
    - $\theta_n|\theta_1 \dots \theta_{n-1}, \gamma, H \sim DP\left(\gamma + n - 1, \frac{\gamma}{\gamma+n-1} H + \frac{n-1}{\gamma+n-1} \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{n-1}\right)$
  - $F(x_i|\theta_i') = N(x_i|\mu_{\theta_i'}, \Sigma_{\theta_i'})$
  - $\mu_{\theta_i'}$ and $\Sigma_{\theta_i'}$ are the component parameters given that the component follows the Gaussian distribution
- **DPMM**
  - **Initial table assignments**
  - **While sampling iterations**
    - **While each data instance in the dataset**
      - **Remove the instance from the assignment**
      - **Calculate the prior :** $\theta_n|\theta_1 \dots \theta_{n-1}, \gamma, H \sim DP$
      - **Calculate the likelihood :** $N(x_i|\mu_{\theta_i'}, \Sigma_{\theta_i'})$
      - **Calculate the posterior**
      - **Sample the cluster assignment from the posterior**
      - **Update the component parameter**
- Truncated Dirichlet process mixture model
  - Finish the sampling of stick-breaking with the limit on the number of atoms
    - Same as limiting the table numbers

Alternative Representation For Mixture Models

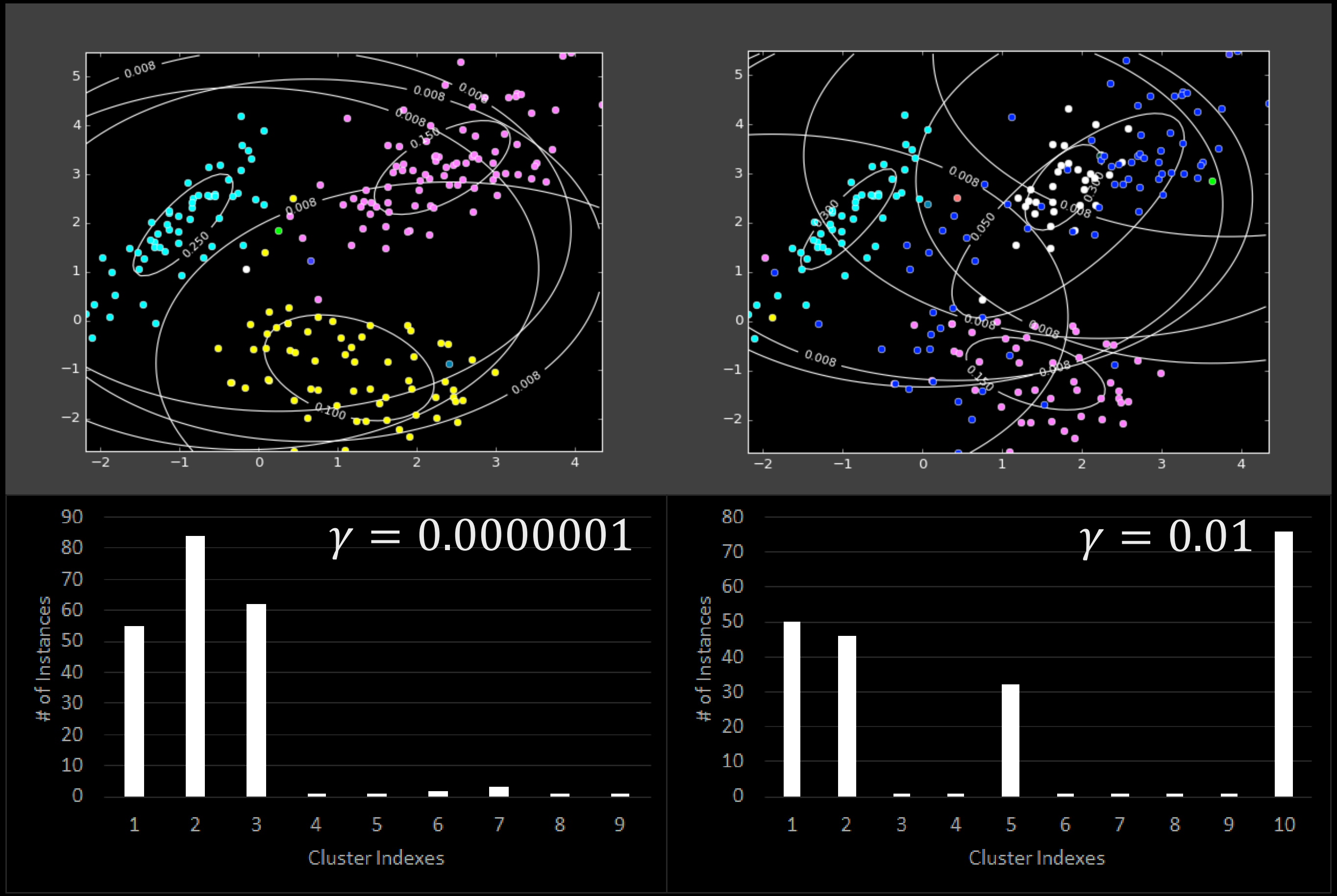


DPMM Sampling Process

- The Sampling process produces the different clustering results per iterations
  - γ can determine the sensitivity of the cluster generation
- $\theta_n | \theta_1 \ldots \theta_{n-1}, \gamma, H \sim DP\left(\gamma + n - 1, \frac{\gamma}{\gamma+n-1} H + \frac{n-1}{\gamma+n-1} \frac{\sum_{i=1}^{n-1} \delta_{\theta_i}}{n-1}\right)$



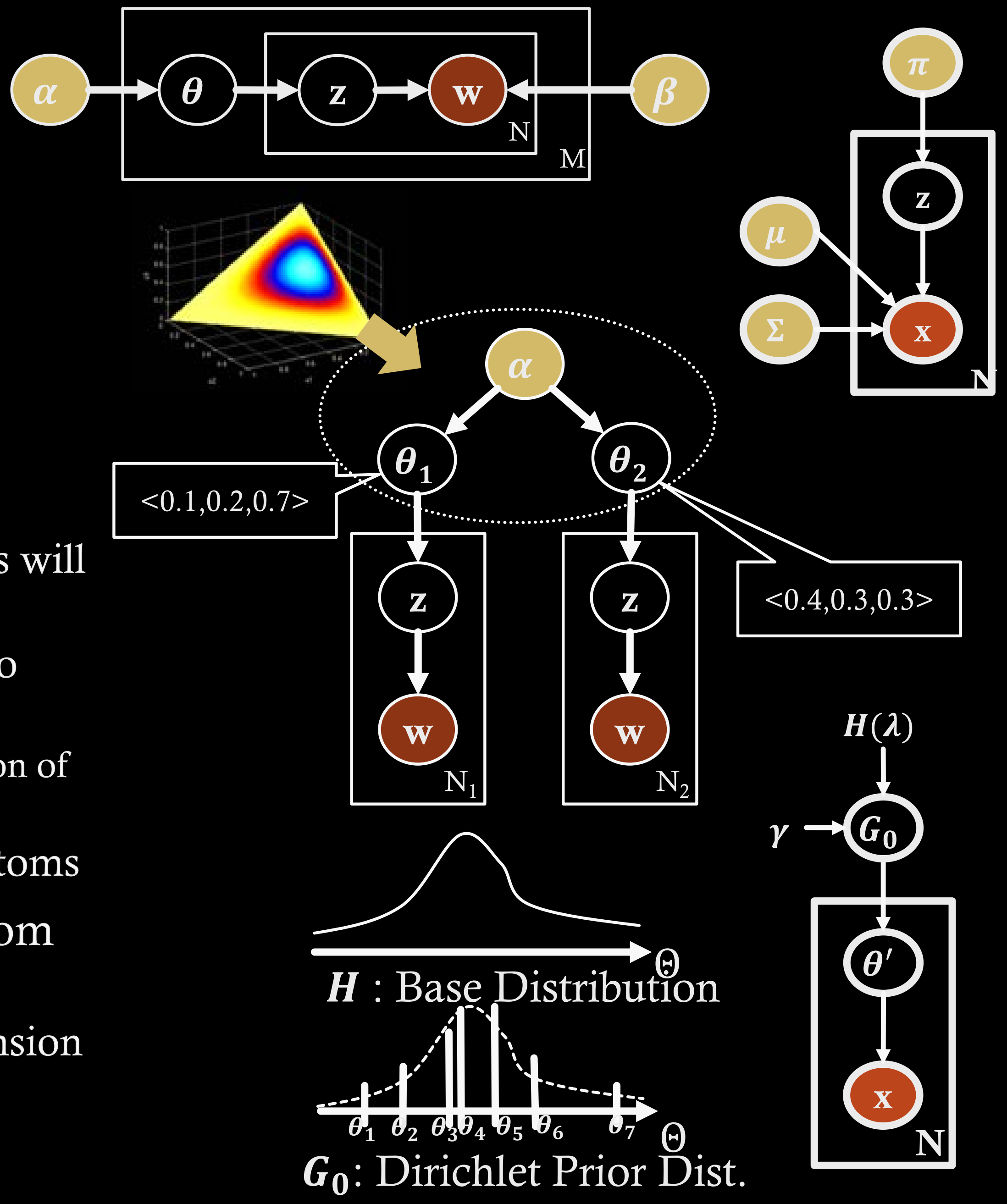
Synthesized True Dataset



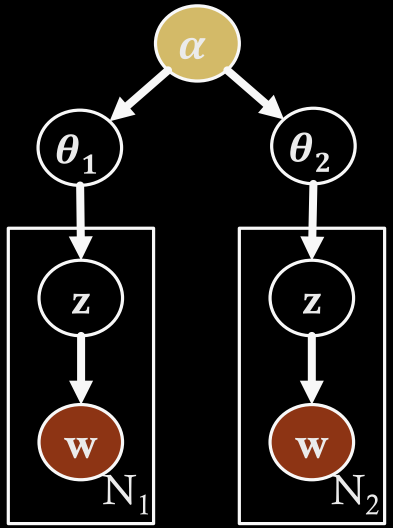
Hierarchical Dirichlet Process

Problem of Separate Prior

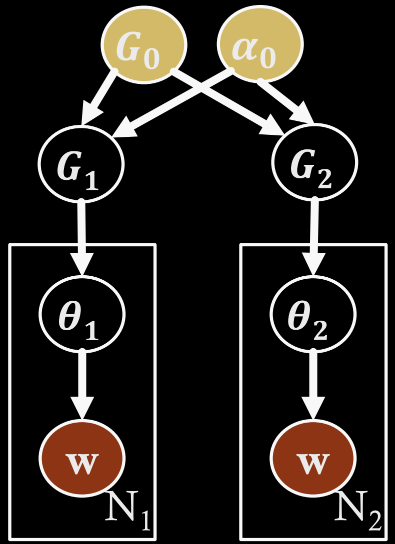
- Datasets are often structured
  - LDA : Corpus-Document structure
  - Hierarchical structure
- Finite dimension of clusters
  - Choice is finite, and the atoms will overlap
  - Infinite model might have zero overlap in atoms
    - Smooth continuous distribution of the base distribution
  - Need to enforce sharing the atoms
- Clustering result is different from one branch to another
  - Need to share the same dimension of clusters
  - How to correlate θ_1 and θ_2



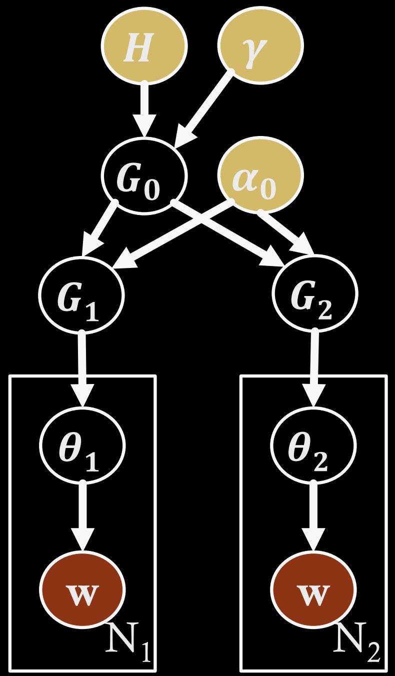
Solution of Atom Sharing



Parametric LDA

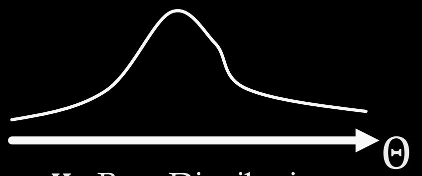


Non-Parametric LDA
without Atom Sharing

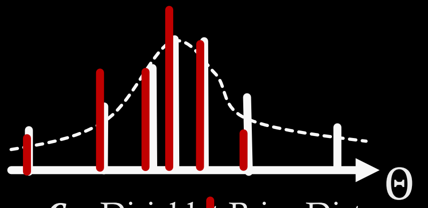


Non-Parametric LDA
with Atom Sharing

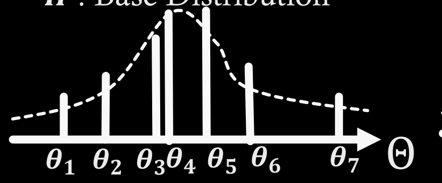
- Hierarchical structure of Dirichlet processes
  - H : the continuous base distribution
  - G_0 : a draw from $G_0 \sim \mathrm{DP}(\mathrm{H}, \gamma)$
  - G_i : a draw from $G_i | G_0 \sim DP(G_0, \alpha_0)$
- Here, G_0 is a discrete distribution
  - so G_i will only sample from the atoms of G_0



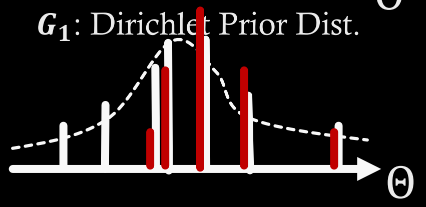
$\boldsymbol{H}$: Base Distribution



$\boldsymbol{G_1}$: Dirichlet Prior Dist.



$\boldsymbol{G_0}$: Dirichlet Prior Dist.

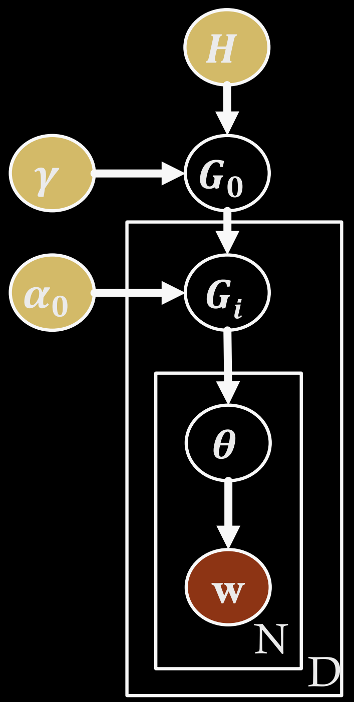


$\boldsymbol{G_2}$: Dirichlet Prior Dist.

Stick Breaking Construction

- A hierarchical Dirichlet process with a corpus with D documents
    - Can be applied to domains other than texts
  - $G_0 \sim \mathrm{DP}(\mathrm{H}, \gamma)$
  - $G_i | G_0 \sim DP(G_0, \alpha_0)$
- Stick breaking *(prior distribution)* construction of HDP
  - $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$
  - $\phi_k \sim H$
  - $\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l)$
  - $\beta'_k | \gamma \sim Beta(1, \gamma)$
  - $G_i = \sum_{k=1}^{\infty} \pi_{ik} \delta_{\phi_k}$
  - $\pi_{ik} = \pi'_{ik} \prod_{l=1}^{k-1} (1 - \pi'_{il})$
  - $\pi'_{ik} | \gamma \sim Beta\left(\alpha_0 \beta_k, \alpha_0 \left(1 - \sum_{i=1}^{k} \beta_i\right)\right)$

$\phi_k \sim H$ is shared



Hierarchical Dirichlet Process

Chinese Restaurant Franchise

- $G_0 \sim \mathrm{DP}(\mathrm{H}, \gamma)$
- $G_i | G_0 \sim DP(G_0, \alpha_0)$
  - $\theta_{in} \sim G_i$: a θ_{in}'s seating on a ψ_{it} table of each restaurant
  - $\psi_{it} \sim G_0$: a ψ_{it}'s table serves a ϕ_k menu of the franchise