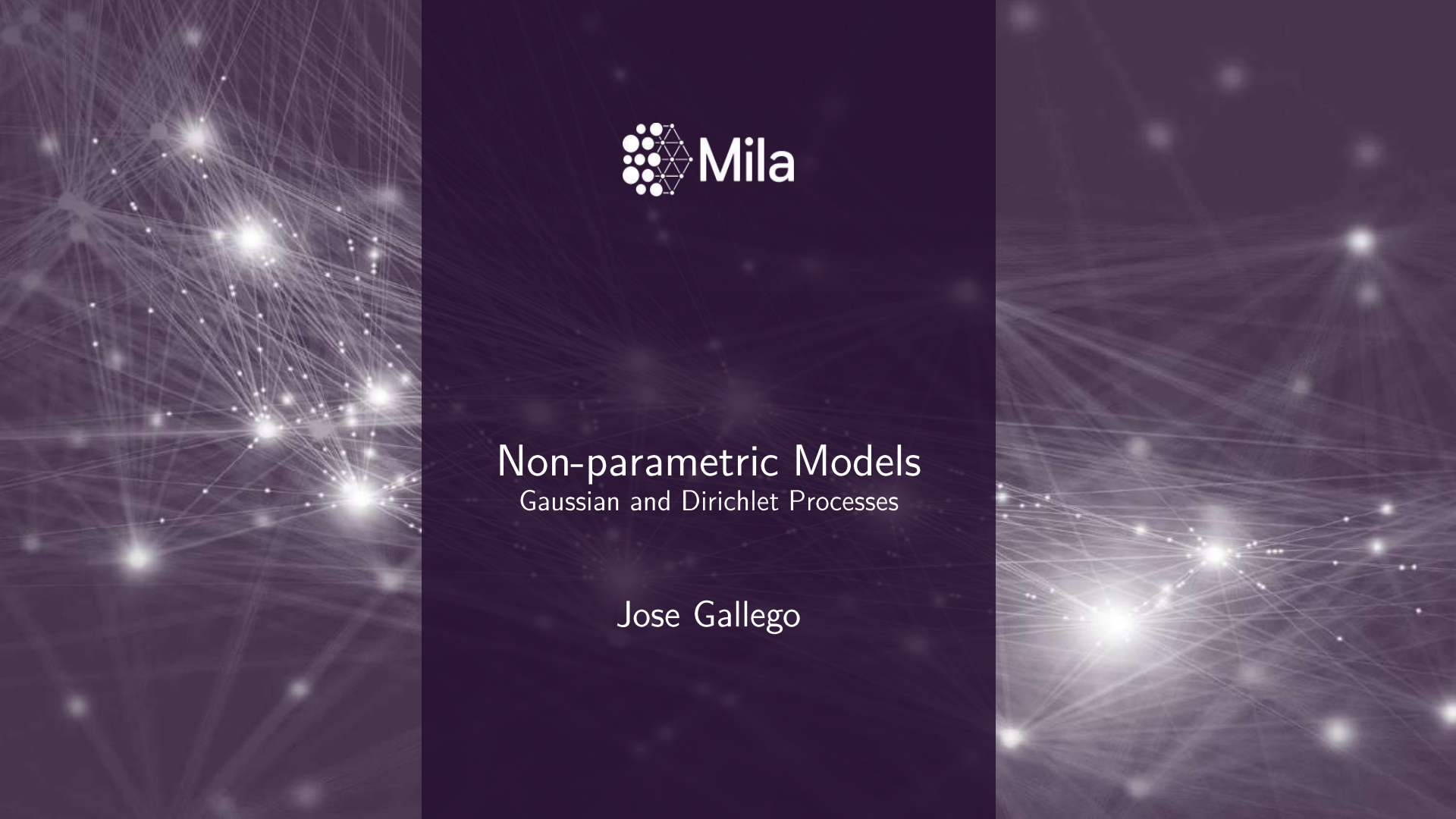
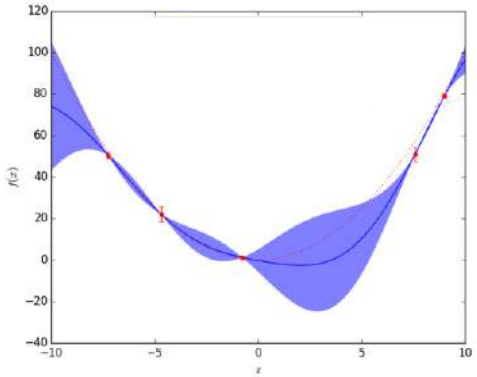


Non-parametric Models

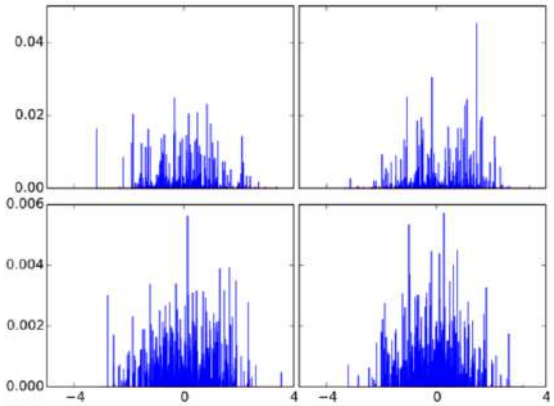
Gaussian and Dirichlet Processes

Jose Gallego

Today



Gaussian Processes



Dirichlet Processes

Gaussian Processes

Properties of Gaussians

- Normalization
- Marginalization
- Addition
- Conditioning
- Product of densities

$$\int p(x)\,dx = 1 \qquad \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

$$\int \ldots\, dx$$

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{bmatrix}\right) \longrightarrow x \sim \mathcal{N}(\mu_x, \Sigma_x)$$

$$x \sim \mathcal{N}(\mu_x, \Sigma_x) \qquad y \sim \mathcal{N}(\mu_y, \Sigma_y)$$

$$x + y \sim \mathcal{N}(\mu_x + \mu_y,\ \Sigma_x + \Sigma_y + 2\Sigma_{xy})$$

↳ characteristic function

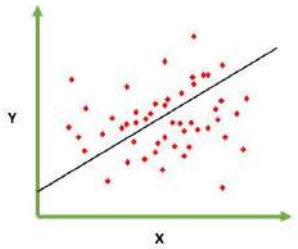
$$y \mid x \sim \mathcal{N}(\_\_)$$

$$\phi(x+y) = \phi(x) \cdot \phi(y)$$

$$X \sim \mathcal{N}(0,1) \longrightarrow X \cdot X \sim \chi^2(1)$$

$$\mathcal{N}(x \mid \mu_1, \Sigma_1) \cdot \mathcal{N}(x \mid \mu_2, \Sigma_2)$$

$$= \mathcal{N}\left(\Sigma_3 \Sigma_1^{-1}\mu_1 + \Sigma_3 \Sigma_2^{-1}\mu_2 \quad (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}\right) \qquad \Sigma_3$$



$y = f(x) = w^T x + \sigma\varepsilon$ $\quad \varepsilon \sim \mathcal{N}(0, \mathbb{I})$

$\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$

$p(y_i \mid x_i; w) = \mathcal{N}(w^T x_i, \sigma^2 \mathbb{I})$

MLE $\quad p(\mathcal{D} \mid w) = \prod_i p(y_i \mid x_i; w)\, p(x_i)$

$p(y_1, \dots, y_n \mid x_1, \dots, x_n, w) = \prod_i p(y_i \mid x_i, w)$

MAP $\quad p(w \mid \mathcal{D}) = \log \dots + \log p(w)$

$\log p(w) \hookrightarrow \|w\|^2$

$y = wx$

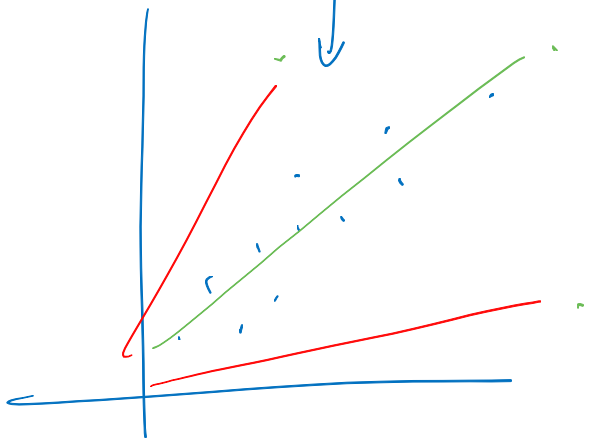
Bayesian Viewpoint (1)

$$p(y^* \mid x^*, \mathcal{D}) = \int_w p(y, w \mid x, \mathcal{D})\, dw = \int_w p(y \mid x, w) \cdot p(w \mid \mathcal{D})\, dw$$

$$= \int_w p(y \mid x, w) \cdot \frac{p(\mathcal{D} \mid w) \cdot p(w)}{p(\mathcal{D})}\, dw = \mathcal{N}(\_, \_)$$

Gaussian

"Use all of them"



GP Assumption

$$p\left(\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ y^* \end{bmatrix} \middle| [x_1, \dots, x_n], x^*\right) \overset{\text{some}}{\sim} \mathcal{N}(\mu, \Sigma).$$

$$p(y^* \mid \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, [x_1, \dots, x_n], x^*) \sim \mathcal{N}(?, ?)$$

$$p(y^* \mid y_1, \dots, y_n, x_1, \dots, x_n, x^*) \sim \mathcal{N}\left(K_*^T K^{-1} y, K_{xx} - K_x^T K^{-1} K_x\right)$$

The role of Σ

1	0.7	0
0.7	1	0
0	0	10

me
neigh
(wrong)

$\sigma_{12} = \rho_{12}\, \sigma_1 \cdot \sigma_2$

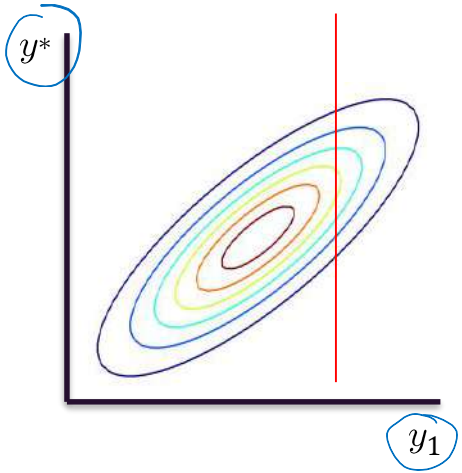
↳ $[-1, 1]$

Positive Definite Kernel

$k(x, x') \geq 0 \quad \forall x, x'$

$\forall l \quad \forall x_1, \ldots, x_l \quad [K]_{ij} = k(x_i, x_j)$

↳ positive definite



Isn't this just kernel regression?

$$p(y^* \mid y_1, \dots, y_n, x_1, \dots, x_n, x^*) \sim \mathcal{N}\left(K_*^T K^{-1} y,\ K_{xx} - K_x^T K^{-1} K_x\right)$$

$$\Sigma = \begin{bmatrix} K & K_x \\ K_*^T & K_{xx} \end{bmatrix}$$

K → $K(x_i, x_j) \quad \forall i, j \le n.$

K_x → x^* and training set

K_{xx} → Variance of x^*.

$$w^* = X^T (X^T X)^{-1} y$$

Demo

Demo

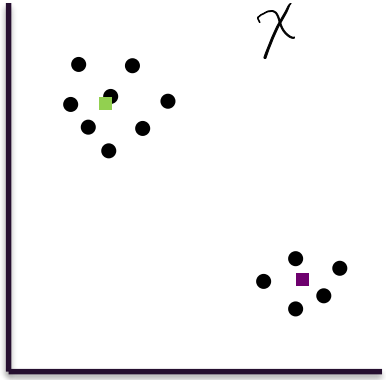
Problem: f is a function $\sim$ *infinte dimensional* vector! But, the multivariate Gaussian distributions is for finite dimensional random vectors.

Definition: A GP is a (potentially infinte) collection of rvs such that the **joint distribution of every finite subset of rvs** is multivariate Gaussian.

$$y^* \mid y_1, \dots, y_n, x_1, \dots, x_n, x^* \sim \mathcal{N}(\mathbf{K}_*^{\mathrm{T}} \mathbf{K}^{-1} \boldsymbol{y}, k_{**} - \mathbf{K}_*^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{K}_*)$$

Dirichlet Processes

Generative Model



$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$

$$x_n | z_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \quad \checkmark$$

$$\begin{cases} \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \quad \checkmark \\ \rho_1 \sim \text{Beta}(a_1, a_2) \\ \rho_2 = 1 - \rho_1 \end{cases}$$

Beta Distribution

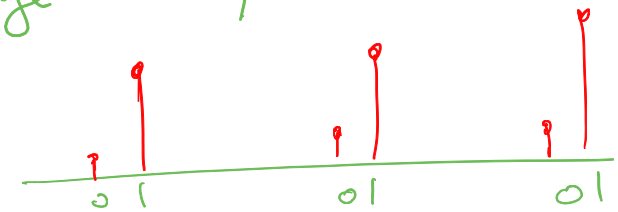
$$\text{Beta}(p_1 \mid a_1, a_2) = \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\cdot\Gamma(a_2)} p_1^{a_1-1}(1-p_1)^{a_2-1}$$

$p_1 \in (0,1)$ $\quad a_1, a_2 > 0$

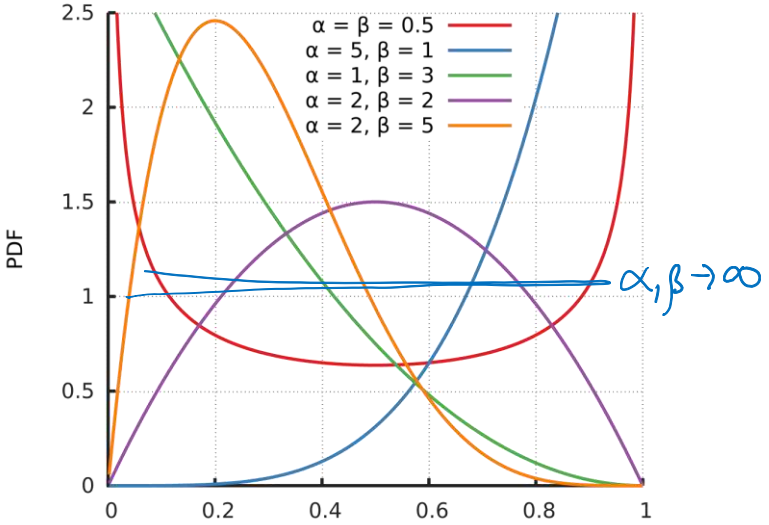
$\alpha = \beta \to 0$

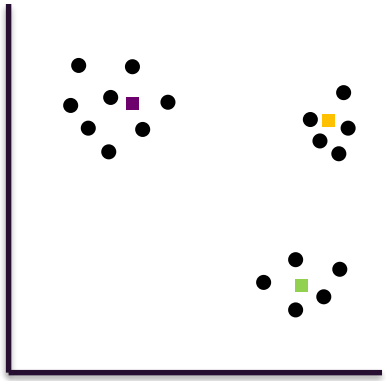
$[p_1\ p_2] = [\text{Beta},\ 1-p_1]$

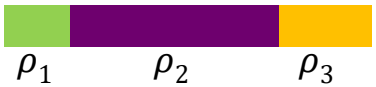
mass @ 0 large $\Rightarrow$ p_1 small



a_1 a_2







Same story

↳ Extension to Dirichlet distribution

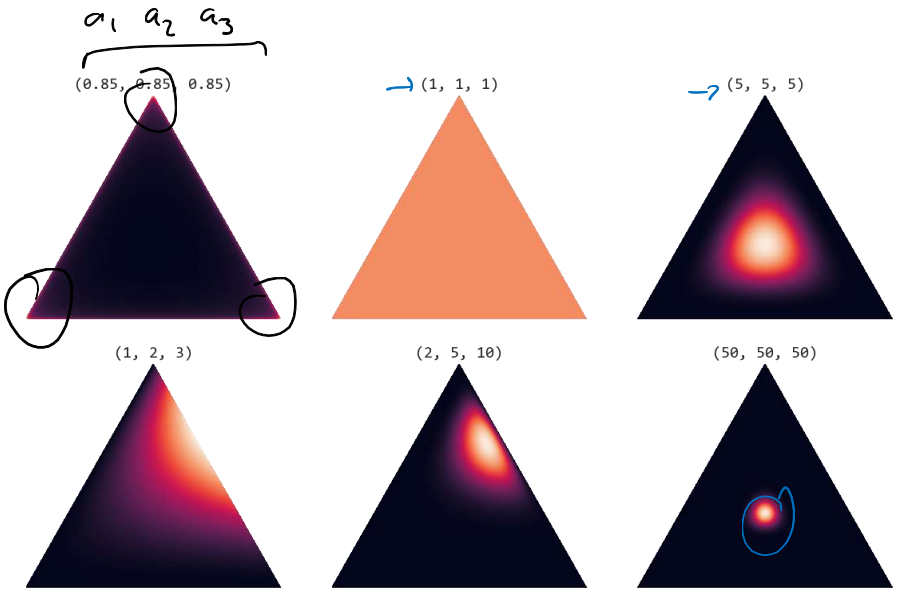
$\text{Cat}(\rho_1, \rho_2, \dots, \rho_K)$

$\rho \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$

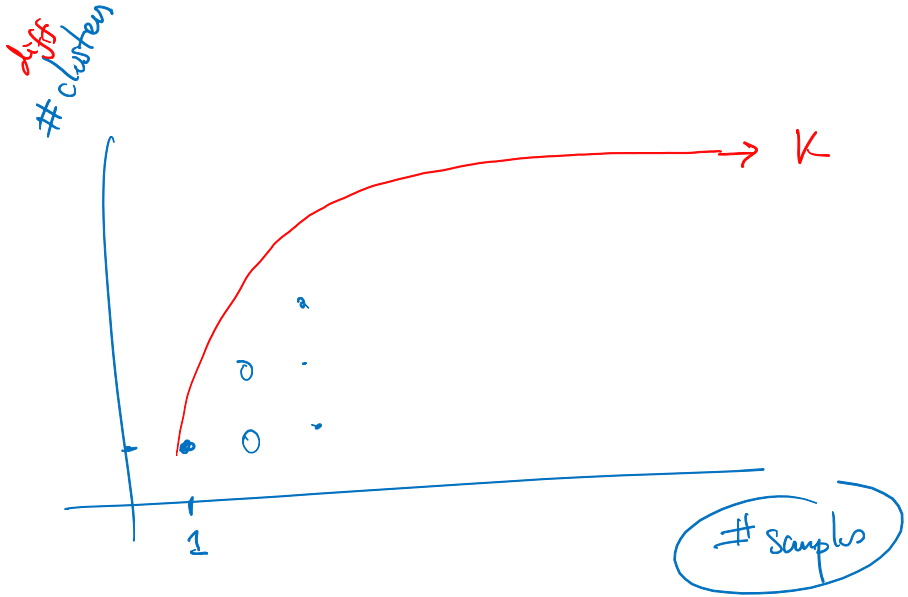
Dirichlet Distribution

$$P(\rho_1, \ldots, \rho_K \mid a_1, \ldots, a_K)$$

$$= \frac{\Gamma(\sum_i a_i)}{\prod_i \Gamma(a_i)} \prod_{i=1}^{K} \rho_i^{a_i - 1}$$



What if $K > N$?



components

vs

observed clusters

So, how do we choose K? We don't $K = \infty$

$\rho_1, \rho_2, \ldots$

$V_1 \sim \text{Beta}(a_1, b_1)$ $\quad 1 \cdot V_1 = \rho_1$

$V_2 \sim \text{Beta}(a_2, b_2)$ $\quad V_2 (1 - \rho_1) = \rho_2$

$$\rho_k = V_k \left[\prod_{i=1}^{k-1} (1 - \rho_i) \right]$$

$\{(a_i, b_i)\}_{i=1}^{\infty}$

Infinitely many parameters!

Stick Breaking Process

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

$\text{GEM}(\alpha)$

$\rho_1 = V_1$

$\rho_2 = V_2 (1 - \rho_1)$

$\vdots$

1

$\rho_1, \rho_2, \ldots \sim \text{GEM}(\alpha)$

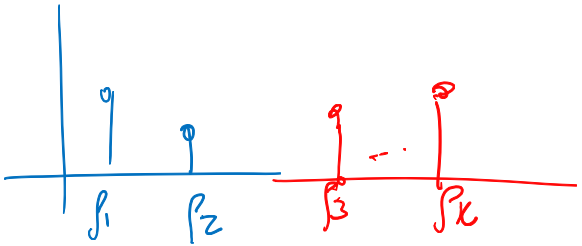
Distributions over Distributions over Distributions over Distributions...

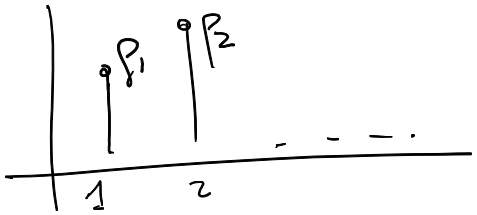
Beta → Distro $(\gamma_1, \gamma_2)_K$

Dir

GEM → Distro $(\mathbb{N})$

$1, 2, 3, \ldots$





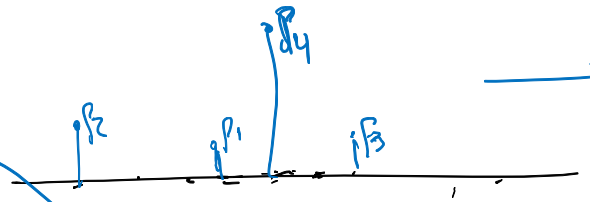
$\beta = (\beta_1, \beta_2, \ldots) \sim GEM(\alpha)$

$\phi_k \overset{iii}{\sim} H \rightarrow$ distro over Φ

$$\sum = \sum_{i=1}^{\infty} \beta_i \delta_{\phi_i}$$

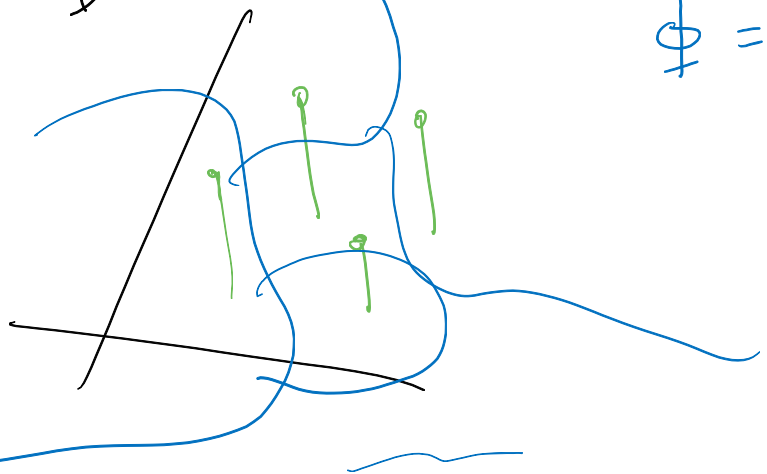
$\Phi = \mathbb{R}$

$H \sim \mathcal{N}$



$\sum_i \rho_i \delta_{\phi_i} = G$

$\mathcal{N}$ $\Phi = \mathbb{R}^2$



$\Phi = \bigsqcup_{i=1}^{\infty} A_i$ ← (Measurable) Partition of Φ

$G(A_i) \in [0,1]$

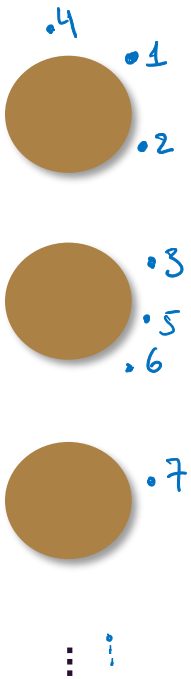
Marginal Cluster Assignments

$[\rho_1, \rho_2, \rho_3] \sim \mathrm{Dir}(1, 1, 1)$ ← Uniform over Δ_3 but not over ρ_1

$\rho_1 \sim \mathrm{Beta}(1, 2)$

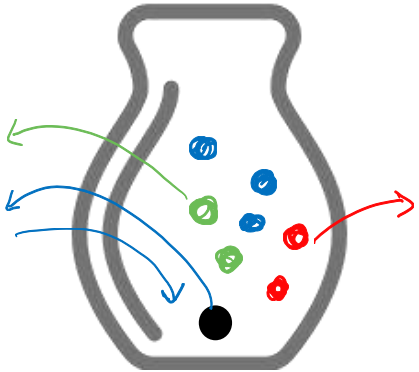
a_1 $\sum_{i \neq 1} a_i$

Equivalent Constructions



Chinese Restaurant Process

"Preferential Attachment"



Problem: $\rho_1, \rho_2, \ldots$ is an *infinte dimensional* probability vector! But, the Dirichlet distributions is defined for finite dimensional spaces.

Definition: Let H be a distribution over Φ and $\alpha > 0$. Then we say that G is a Dirichlet process with base distribution H and concentration parameter α if for all finite (measurable) partition $A_1, \ldots, A_r$,

$$(G(A_1), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha \mathrm{H}(A_1), \ldots, \alpha \mathrm{H}(A_1))$$

← Condition on finite objects

Compare to finite number of vectors in GP. ≃ Finite partition of Φ