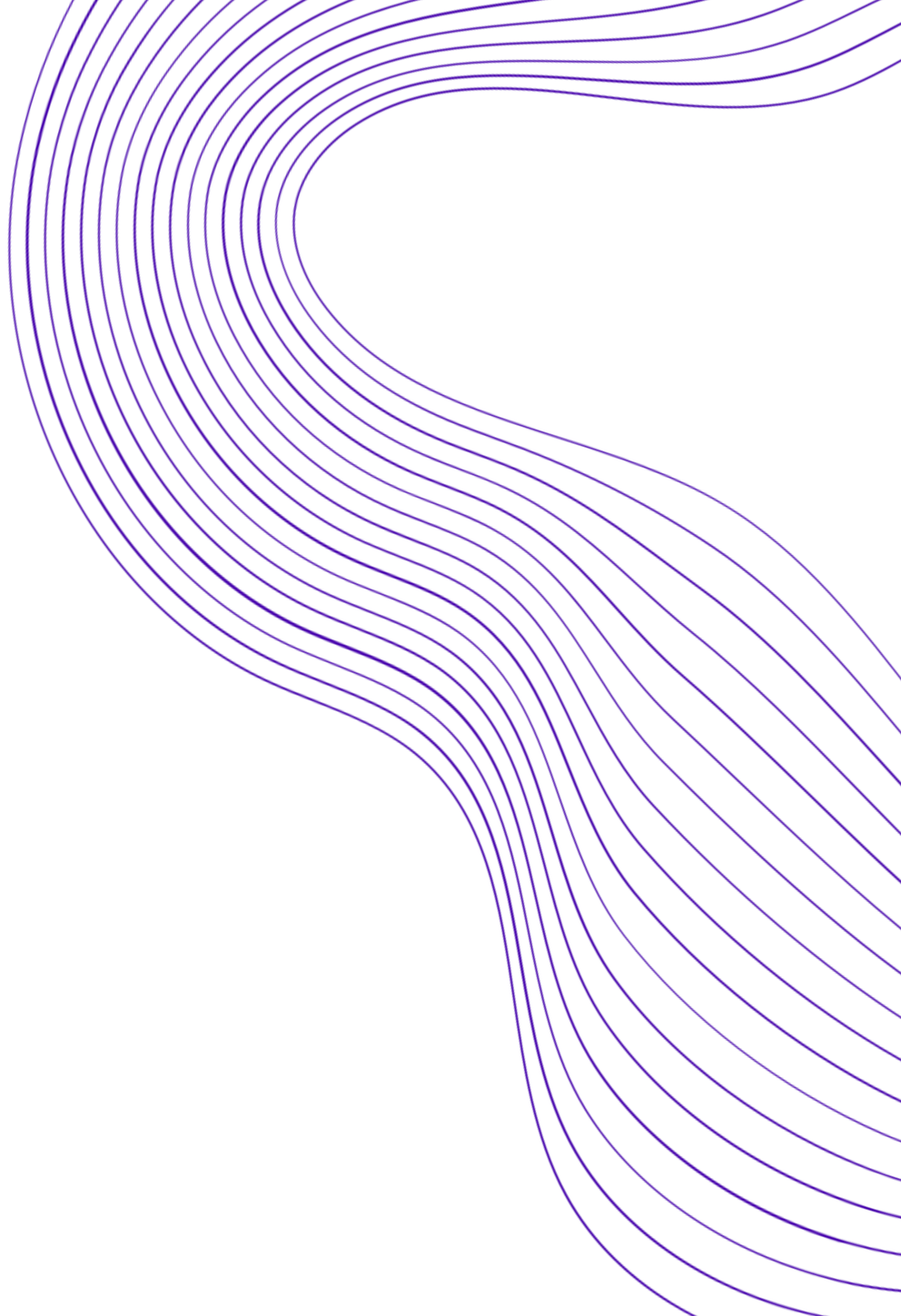
# Bayesian Non-Parametrics

Jose Gallego-Posada

IFT6269 | December 4, 2020



# Contents

- Stochastic Processes
- Gaussian Processes
- Dirichlet Processes

# Caveats

- Ideas over implementation/training details
- Hyperparameter selection
- Basic introduction - See suggested sources

- David MacKay
- Yee Whye Teh
- Kilian Weinberger
- Tamara Broderick
- Michael Jordan

# Inspiration

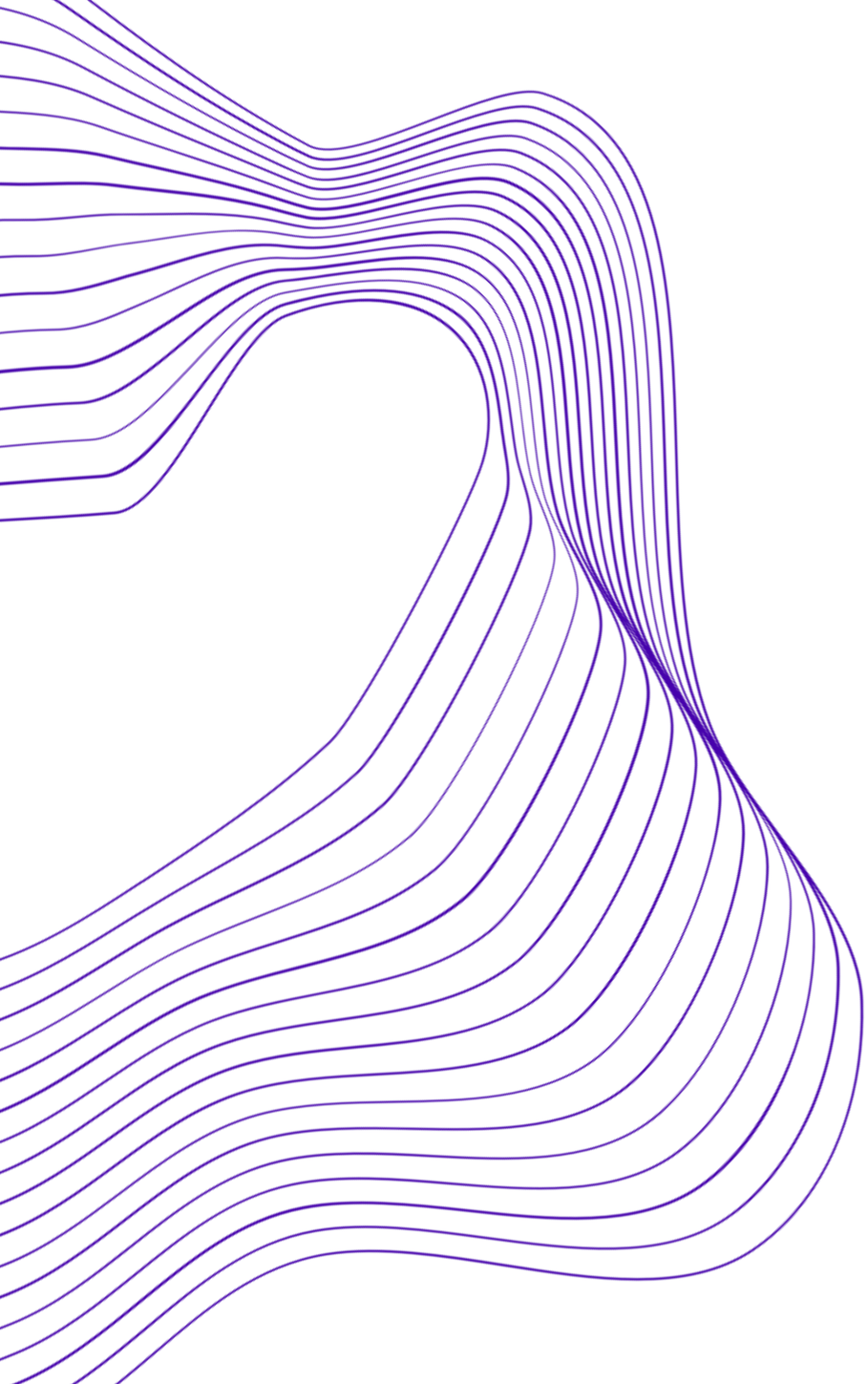
# Stochastic Process

A stochastic process is defined as a collection of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega$ is a sample space, $\mathcal{F}$ is a sigma-algebra, and $\mathbb{P}$ is a probability measure; and the random variables, indexed by some set $T$, all take values in the same space $S$, measurable with respect to some $\sigma$-algebra.

$$\{X(t) \mid t \in T\}$$

# Examples of Stochastic Process

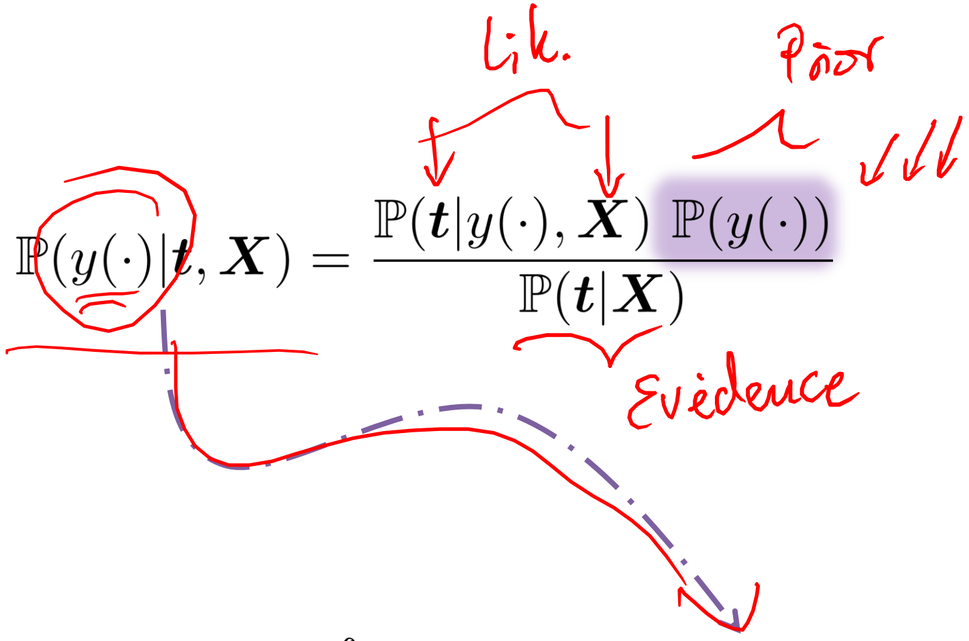
- Single random variable
- IID random variables
- Deterministic function
- {Wiener, Poisson, etc.} Process
- Gaussian Process
- Dirichlet Process



6

# Gaussian Processes

# Inference and Prediction




$$\mathbb{P}(y(\cdot)|\boldsymbol{t}, \boldsymbol{X}) = \frac{\mathbb{P}(\boldsymbol{t}|y(\cdot), \boldsymbol{X})\ \mathbb{P}(y(\cdot))}{\mathbb{P}(\boldsymbol{t}|\boldsymbol{X})}$$

$$\mathbb{P}(t^*|\boldsymbol{x}^*, \boldsymbol{t}, \boldsymbol{X}) = \int \mathbb{P}(t^*|y(\cdot), x^*)\mathbb{P}(y(\cdot)|\boldsymbol{t}, \boldsymbol{X})\mathrm{d}y$$

# Gaussian Properties

- (Tractable) Normalization $\int \tilde{p}(x)\mathrm{d}x = (2\pi)^{d/2}\,|\Sigma|^{\frac{1}{2}}$

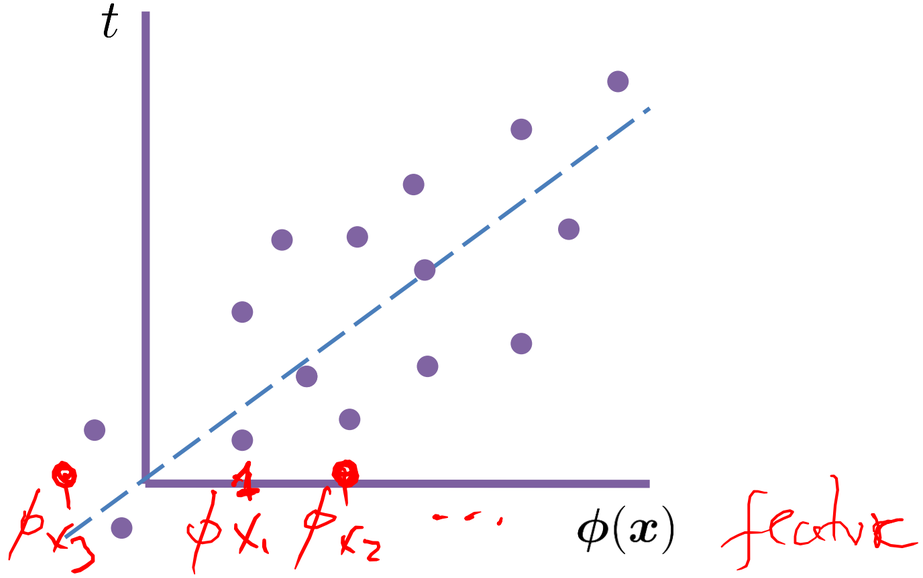
$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY}^\top & \boldsymbol{\Sigma}_Y \end{bmatrix}\right)$$

- Marginalization
  $X \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$
- Conditioning
  $X|Y = \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_Y), \boldsymbol{\Lambda}_X^{-1})$
- Addition
  $X+Y \sim \mathcal{N}(\boldsymbol{\mu}_X + \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_Y + 2\boldsymbol{\Sigma}_{XY})$
- Product of densities $\mathcal{N}(x \,|\, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \cdot \mathcal{N}(x \,|\, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \not\propto \mathcal{N}(x \,|\, \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$

# Linear Regression

Starting point:

- Locations $\{\boldsymbol{x}_i \,|\, i \in [1, N]\}$
- Basis functions $\{\phi_h(\boldsymbol{x})\}_{h=1}^{H}$
- Linear model $y(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$
- Prior $\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma_w^2 \boldsymbol{I}\right)$
- Observation noise $t|y \sim \mathcal{N}\left(y, \sigma_\nu^2\right)$




This leads to $\boldsymbol{y}_N \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top)$

$$\boldsymbol{t}_N \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \boldsymbol{\Phi}_N \boldsymbol{\Phi}_N^\top + \sigma_\nu^2 \boldsymbol{I})$$

(PSD) similarity matrix

$C(\boldsymbol{x}_i, \boldsymbol{x}_j)$

# Gaussian Process

The probability distribution of a function $y(\boldsymbol{x})$ is a Gaussian process if for any finite selection of points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, the vector $[y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_N)]$ follows a Gaussian distribution.

$\bigcup_{N=1}^{\infty} \mathcal{X}^N$

Index set $T$

$\mathbb{R}^l$

Value space $S$

Gaussians

RVs

(function values)

# Habemus Datam!

... we finally observe $\boldsymbol{t}_N = \{t_i \,|\, i \in [1, N]\}$ and would like to infer $\boldsymbol{t}_{N+1}$ for a new test point $\boldsymbol{x}_{N+1}$.

$$\boldsymbol{C}_{N+1} = \begin{bmatrix} \boldsymbol{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^\top & \kappa \end{bmatrix}$$

$$\mathbb{P}(t_{N+1}|\boldsymbol{t}_N) = \frac{\mathbb{P}(t_{N+1}, \boldsymbol{t}_N)}{\mathbb{P}(\boldsymbol{t}_N)} \propto \exp\left[-\frac{1}{2}\begin{bmatrix}\boldsymbol{t}_N & t_{N+1}\end{bmatrix}\boldsymbol{C}_{N+1}^{-1}\begin{bmatrix}\boldsymbol{t}_N \\ t_{N+1}\end{bmatrix}\right]$$

$$t^*|t_1, \ldots, t_N, x_1, \ldots, x_n, x^* \sim \mathcal{N}\left(\boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{t}_N, \kappa - \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{k}\right)$$

# Demo

$P(t^* = 1 \mid \ldots) = \sigma(\text{Regression})$

Logits

D. MacKay / C. Bishop $\Rightarrow$ Rassmussen '06

# GP Summary

**Problem:** $y$ is a function aka "infinite dimensional vector". But the multivariate Gaussian distribution is defined for **finite** dimensional vectors.

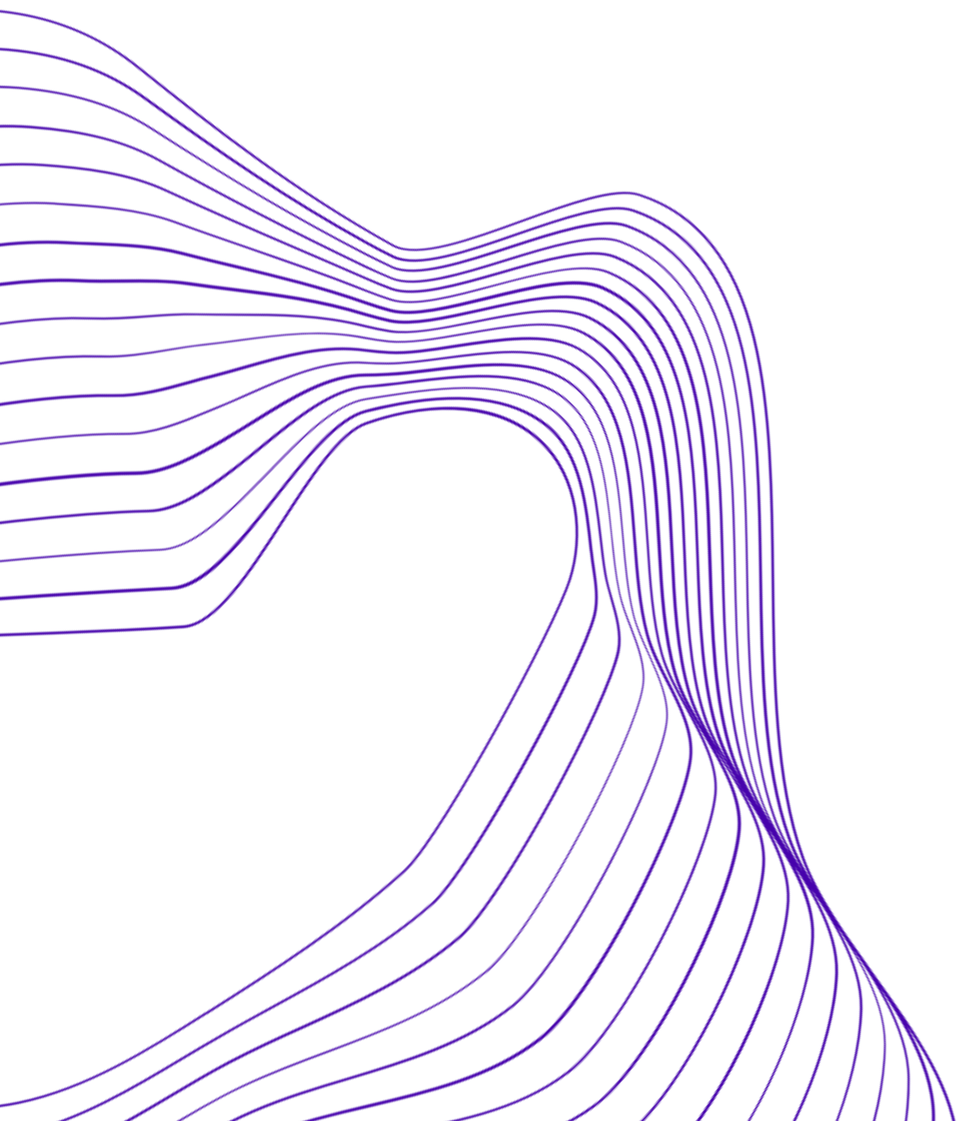
$y_N \sim$

**Definition:** A GP is a (potentially infinite) collection of random variables such that the joint distribution of **every finite subset** of them is a multivariate Gaussian.

$$t^* | t_1, \ldots, t_N, x_1, \ldots, x_n, x^* \sim \mathcal{N}\left(\boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{t}_N, \kappa - \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{k}\right)$$

$$[\, y(x_1) \ldots y(x_N) \,] \sim \mathcal{N}(0, C_N)$$

14

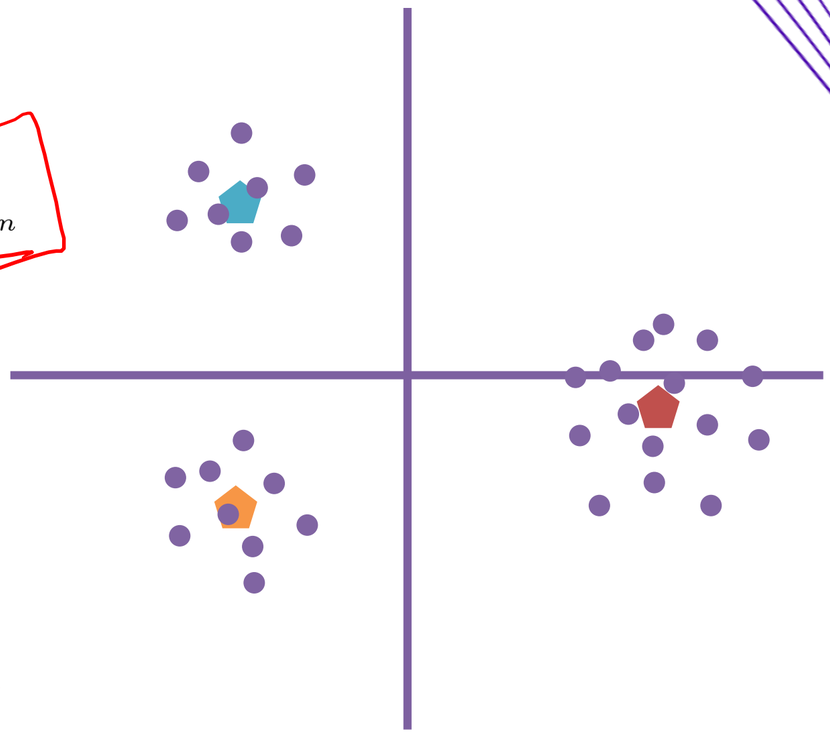


# Dirichlet Processes

# Generative Model

Model:

- Assignments $z_n|\boldsymbol{\rho} \sim \text{Categorical}(\rho_1, \ldots, \rho_K)$
- Class conditionals $\boldsymbol{x}_n|z_n \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}) = F_{z_n}$



Prior assumptions:

- Number of classes $K$
- Class proportions $\boldsymbol{\rho} \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$
- Class parameters $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma_0}) = H$

# Dirichlet Distribution

over Distributions

$\in \Delta_{K-1}$

$$\mathbb{P}(\rho_1, \ldots, \rho_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \rho_k^{\alpha_k - 1}$$

$\rho_i \sim \text{Gamma}(\alpha_i, \theta)$ ($\theta$ ↓)

$\vec{\rho} = \frac{\vec{\rho}}{\|\rho\|_1}$ $\quad \vec{\rho}_1$

$(n+1)! \quad (n+1)\, n!$

$$\Gamma(z+1) = z\Gamma(z)$$ ✓

$\mathbb{P}(\rho_1, \rho_2 \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \rho_1^{\alpha_1 - 1} \rho_2^{\alpha_2 - 1}$

Collapsibility

$$\rho_1, \ldots, \rho_i + \rho_j, \ldots, \rho_K \sim \text{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K)$$

Depends on all entries

$\alpha_0 = \sum_i \alpha_i$

$\alpha_0 = 4.5$

$\alpha_0 = 15$

$$(\rho_i, \sum_{j \neq i} \rho_j) \sim \mathrm{Dir}(\alpha_i, \sum_{j \neq i} \alpha_j)$$

$$\rho_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$$

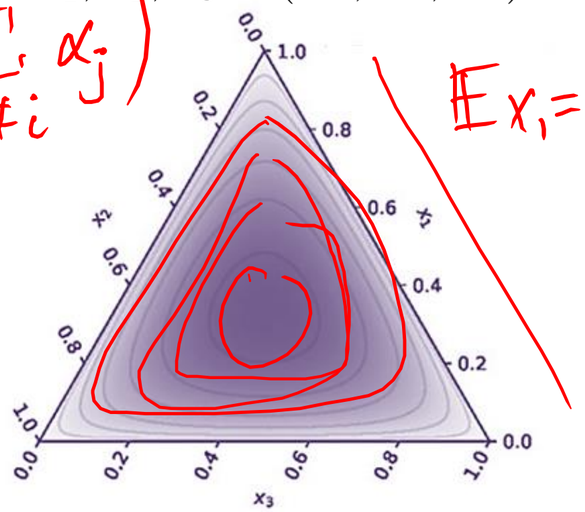
$W \sim \mathrm{Beta}(a, b) \Rightarrow \mathbb{E}W = \frac{a}{a+b} \quad \mathbb{V}W = \frac{ab}{(a+b)^2(a+b+1)}$

$$\mathbb{E}\rho_i = \frac{\alpha_i}{\alpha_i + \alpha_0 - \alpha_i} = \frac{\alpha_i}{\alpha_0}$$

$$\mathbb{V}\rho_i = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

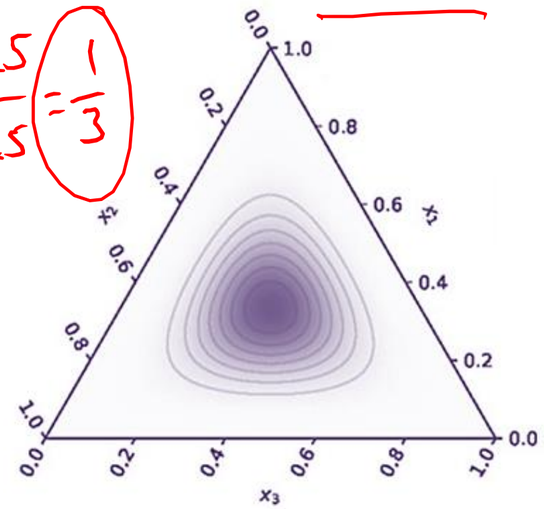
**17** $\mathrm{Cov}(\rho_i, \rho_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$

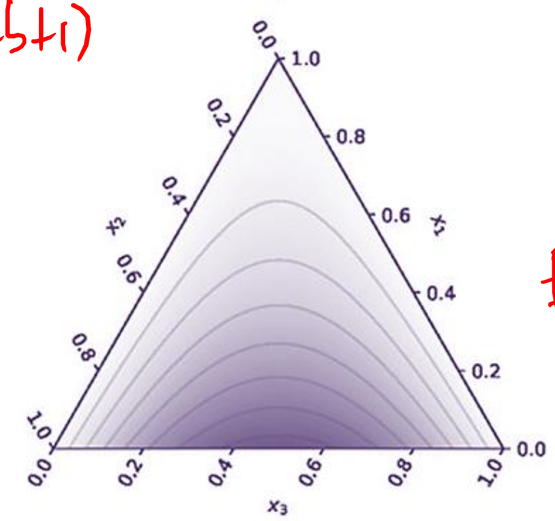
$\alpha_1, \alpha_2, \alpha_3 = (1.5, 1.5, 1.5)$




$\mathbb{E}x_1 = \frac{1.5}{4.5} = \frac{1}{3}$

$\alpha_1, \alpha_2, \alpha_3 = (5, 5, 5)$

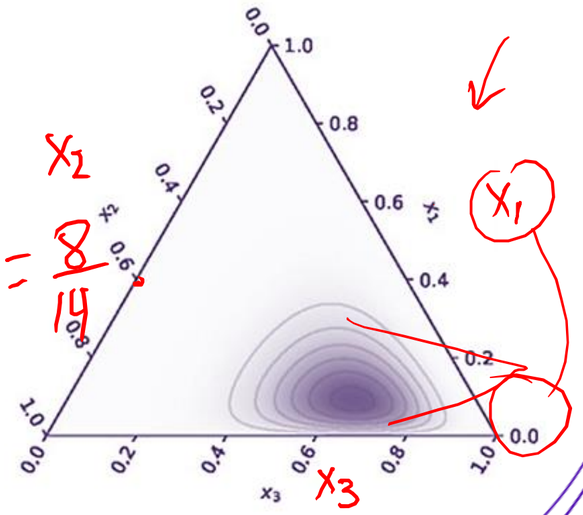



$\alpha_1, \alpha_2, \alpha_3 = (1, 2, 2)$

$\rho_1 \sim \mathrm{Beta}(2, 12)$

$\mathbb{E}\rho_1 = \frac{2}{14}$




$\mathbb{E}x_3 = \frac{8}{14}$

$\alpha_1, \alpha_2, \alpha_3 = (2, 4, 8)$

$\rho_3 \sim \mathrm{Beta}(8, 6)$

# Dirichlet Simulation

stick-breaking

$$\rho_1 \sim \text{Beta}(\alpha_1, \sum_{k=1}^{K} \alpha_k - \alpha_1) \quad \perp \quad \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \sim \text{Dir}(\alpha_2, \ldots, \alpha_K)$$

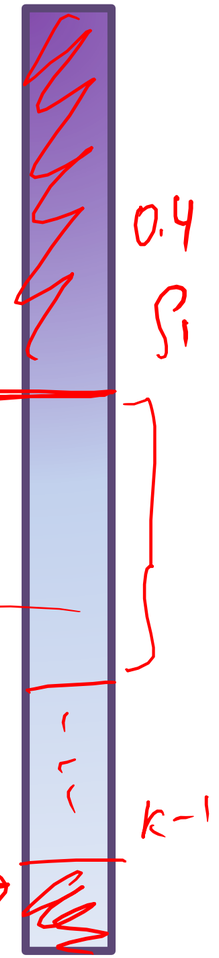
$$\nu_2 \sim \text{Beta}(\alpha_2, \sum_{k=2}^{K} \alpha_k - \alpha_2) \qquad \rho_2 = (1-\nu_1)\nu_2$$

$$\vdots$$

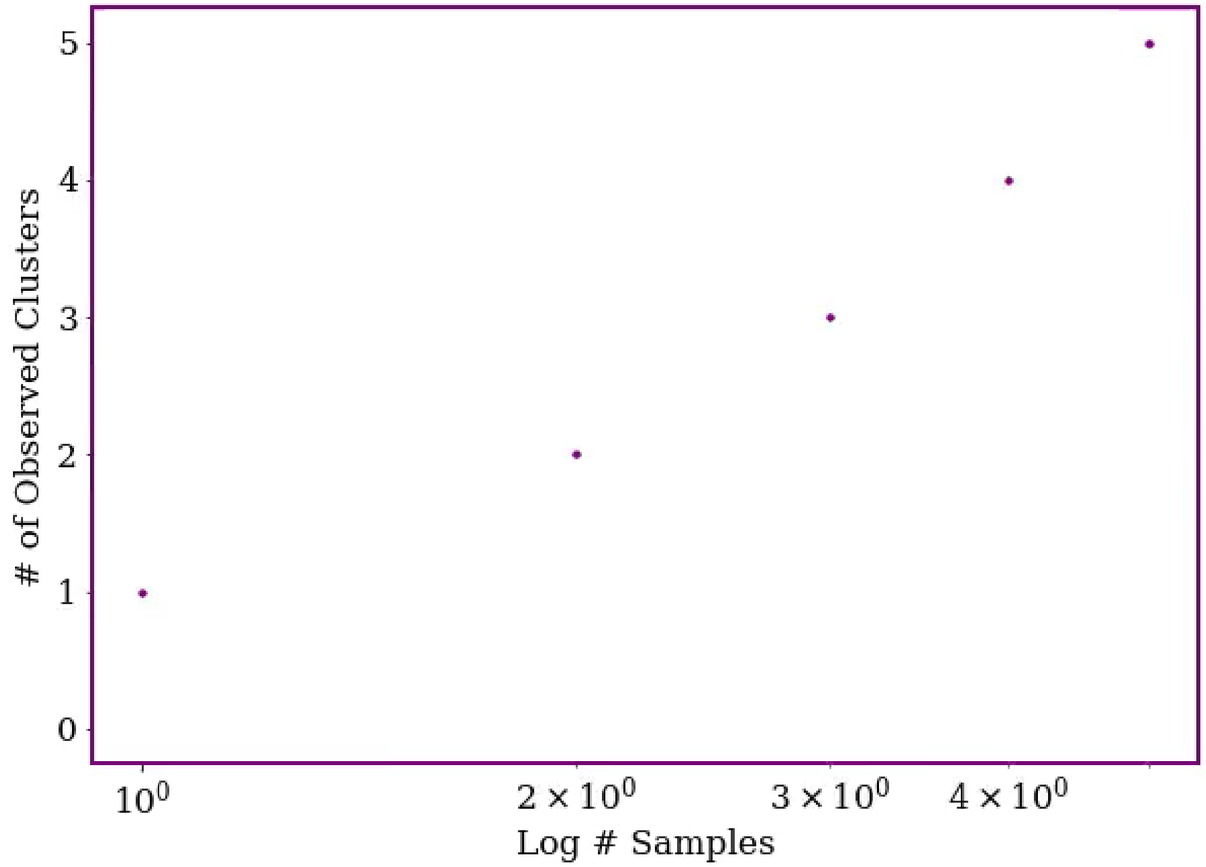
$$\nu_l \sim \text{Beta}(\alpha_l, \sum_{k=l}^{K} \alpha_k - \alpha_l) \qquad \rho_l = \left(\prod_{k=1}^{l-1}(1-\nu_k)\right)\nu_l$$

$$\vdots$$

$$\rho_K = 1 - \sum_{k=1}^{K-1} \rho_k$$



# Components vs Clusters



# How do we choose K?

**We don't!** Consider an infinite mixture model.

$$\nu_1 \sim \text{Beta}(\alpha_1, \beta_1) \qquad \rho_1 = \nu_1$$

$$\nu_2 \sim \text{Beta}(\alpha_2, \beta_2) \qquad \rho_2 = (1 - \nu_1)\nu_2$$

$$\vdots \qquad\qquad \vdots$$

$$\nu_l \sim \text{Beta}(\alpha_l, \beta_l) \qquad \rho_l = \left(\prod_{k=1}^{l-1}(1 - \nu_k)\right)\nu_l$$

$$\vdots \qquad\qquad \vdots$$

Conditions on $\{\alpha_l, \beta_l\}$ for proper normalization?

$\mathbb{E}\,\nu_l = \frac{1}{1+\alpha}$ $\quad \alpha \ll 1 \Rightarrow 1$ $\quad \alpha \gg 1 \Rightarrow 0$

Griffiths, Engen & McCloskey

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

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

# Random Measures

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

$$\phi_1, \phi_2, \ldots \overset{\text{iid}}{\sim} H$$

e.g. $\mathcal{N}(\mu_0, \Sigma_0)$

Two sources of randomness

Distro on $S$

$$G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k}$$

Distro on $S$

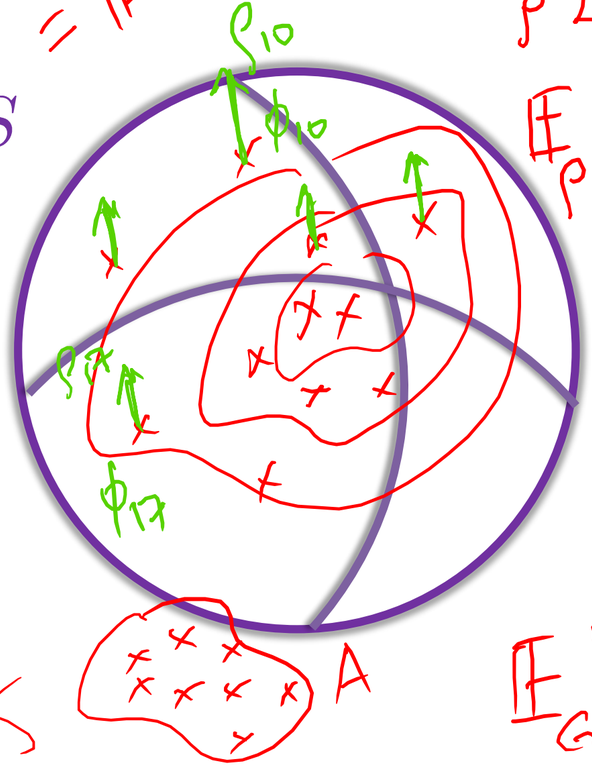
$$z_n \mid G \sim G$$

$$\boldsymbol{x}_n \mid z_n \sim F_{z_n}$$

$\leftarrow \mathcal{N}(\mu_{z_n}, \Sigma_i)$

Realizations of $z_n$ are $\phi_k$'s $\leftarrow (\mu_k, \Sigma_i)$

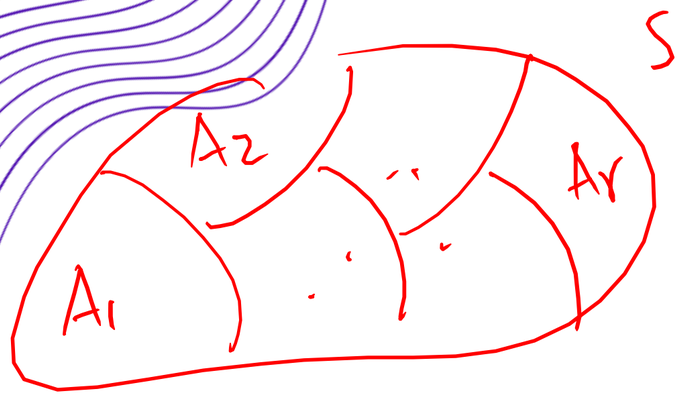
$S = \mathbb{R}^d \times \{\Sigma_i\}$




$$(*) \quad \mathbb{E}_{\rho}\left[\sum_k \rho_k \mathbb{E}_{\phi_k}[\delta_{\phi_k}(A)]\right] = \mathbb{E}_{\rho}\left[\sum_k \rho_k H(A)\right] = H(A)$$

$H(A)$ random

$$\mathbb{E}_G[G(A)] \overset{[0,1]}{=} \mathbb{E}_G\left[\sum_k \rho_k \delta_{\phi_k}(A)\right] \quad (*)$$




# Dirichlet Process

Given a measurable space $(S, \mathcal{M})$, a base probability distribution $H$ and $\alpha > 0$, the probability distribution of a measure $G$ is a Dirichlet process if for any finite partition $\{A_1, \ldots, A_r\}$ of $S$, the vector $[G(A_1), \ldots, G(A_r)]$ $\sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$

Concentration parameter

$\mathcal{M}$ — Index set $T$

Dirichlet — RVs

Ex Behavior
DP($\alpha$, H)
$\alpha \to 0$ ?
$\alpha \to \infty$ ?

# DP Summary

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

**Definition:** A DP is a distribution over probability measures such that for **every finite partition**, the probabilities of the partition elements follow a joint Dirichlet distribution.

$$[G(A_1), \ldots, G(A_r)] \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r))$$

# Questions?