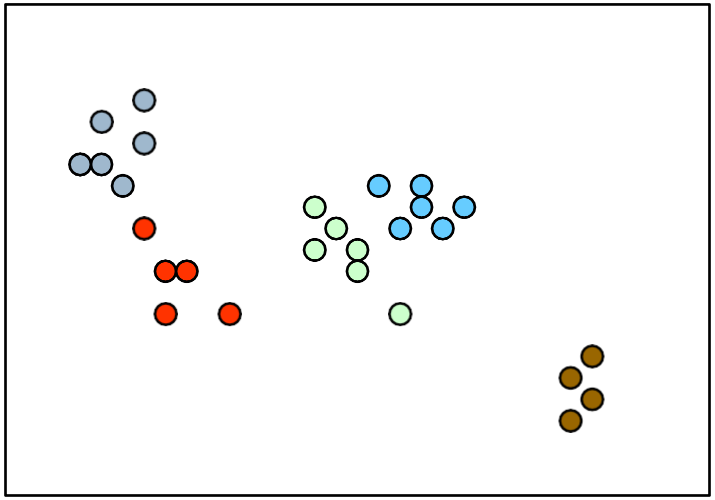
# Statistical Machine Learning

## Lecture 06
## Dirichlet Process

Spring 2021
Sharif University of Technology

# Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

# Motivation

## Recall Clustering with GM

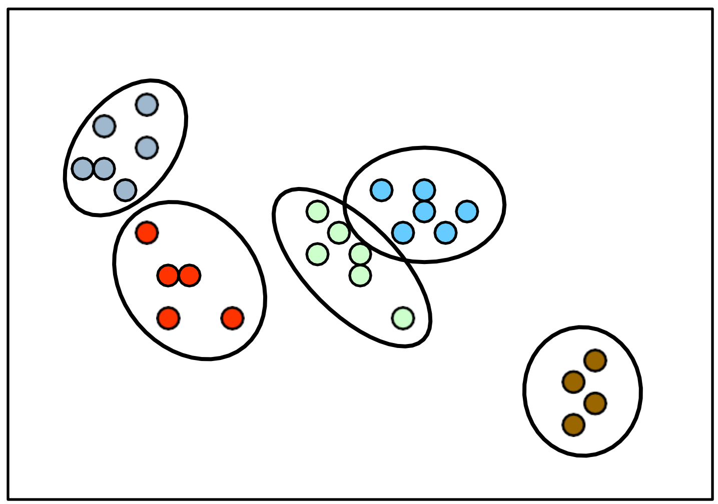
- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances: $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities: $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$
- Gaussian mixture marginal likelihood:

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

# Motivation

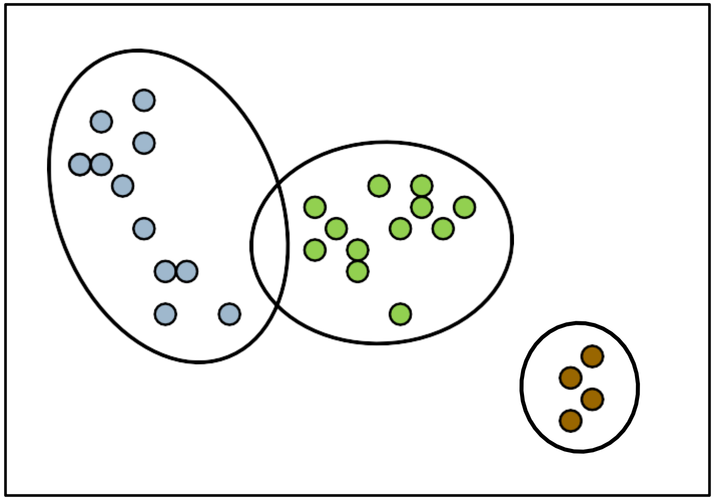
- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

# Motivation

- We are given a data set, and are told that it was generated from a mixture of Gaussian distributions.



- Unfortunately, no one has any idea *how many* Gaussians produced the data.

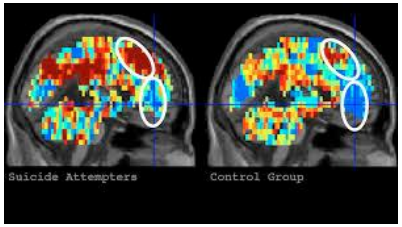
# What to do?

- We can guess the number of clusters, run Expectation Maximization (EM) for Gaussian Mixture Models, look at the results, and then try again.
- We can run hierarchical agglomerative clustering, and cut the tree at a visually appealing level…

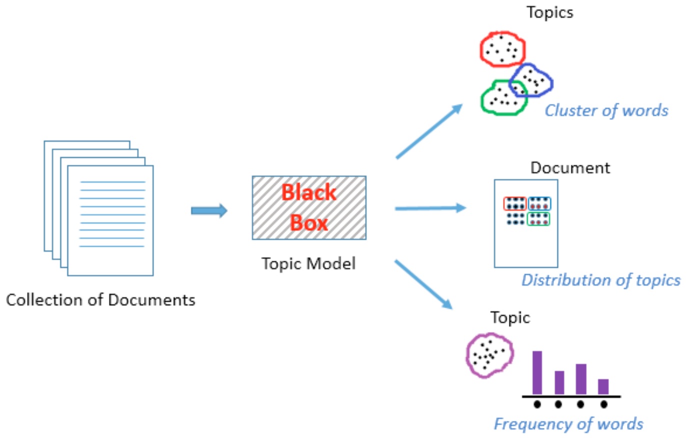
- We want to cluster the data in a statistically principled manner.

# Other motivating examples

- Brain Imaging: Model an unknown number of spatial activation patterns in fMRI images.




- Topic Modeling: Model an unknown number of topics across several corpora of documents.
- …



# Alternative approach: The Dirichlet Process

## Dirichlet process: What is this good for?

- Principled, Bayesian method for fitting a mixture model with an unknown number of clusters.
- Because it is Bayesian, can build hierarchies (e.g. HDPs) and integrate with other random variables in a principled way.

# The Dirichlet Distribution (DD)

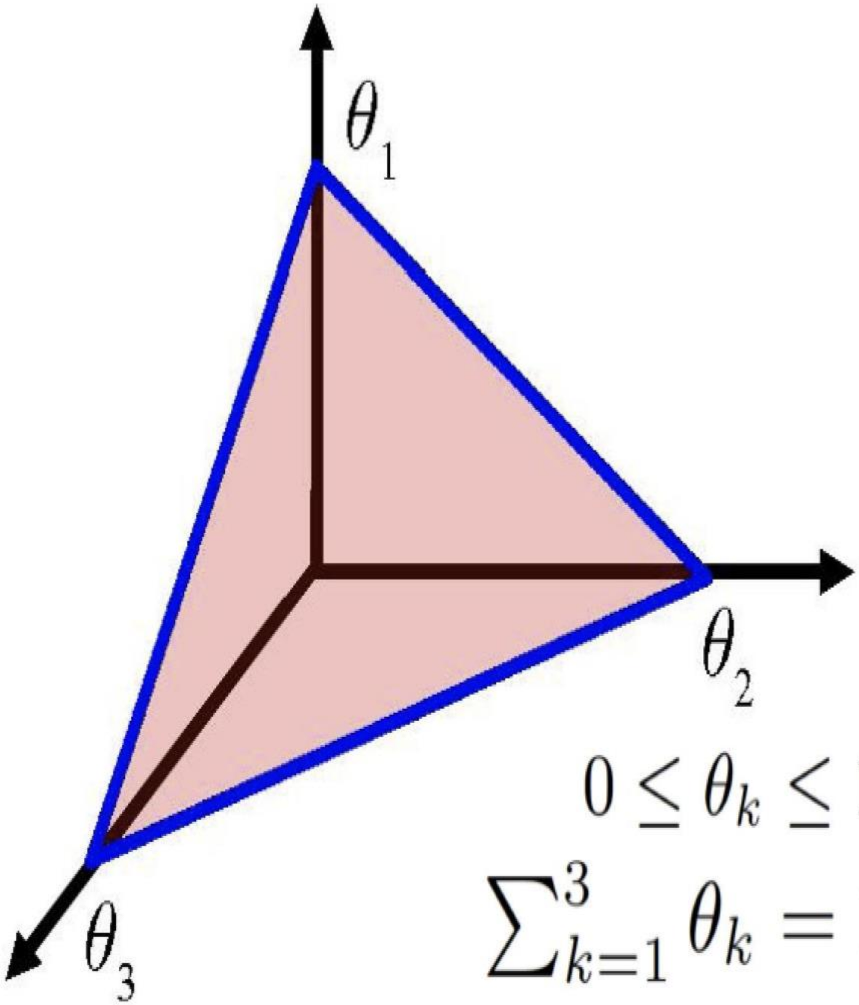
- Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\Theta \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

$$P(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

- Samples from the distribution lie in the *m-1* dimensional probability simplex

# Multinomial Simplex




$$0 \leq \theta_k \leq 1$$

$$\sum_{k=1}^{3} \theta_k = 1$$

# The Dirichlet Distribution

- Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$
- We write:

$$\Theta \sim \mathrm{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_m)$$

- Dirichlet Distribution over possible parameter vectors for a multinomial distribution, and is the conjugate prior for the multinomial.
- Beta distribution is the special case of a Dirichlet for 2 dimensions.
- Thus, it is in fact DD is a distribution over distributions.

# Dirichlet Distribution

Flip a coin with prob $q$ (if head shows up)

prob of $x$ heads in $n$ flips:

Binomial dist.

$$p(x|q,n) = \binom{n}{x} q^x (1-q)^{n-x}$$

what if you don't know $q$?

Bayesian represent uncertainty about $q$

prior dist.

# Dirichlet Distribution

We assume Beta dist. on $q$

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

hyperparameter

Why Beta?

- it is defined over [0, 1]
- Beta & Binomial are conjugate

Beta prior $\longrightarrow$ Binomial likelihood $\longrightarrow$ posterior Beta

# Dirichlet Distribution

- multinomial extends binomial to a multiclass ~~pr~~ problem.
- Consider rolling a die:

$Z \triangleq$ multinomial R.V.

$$p(Z_k = 1) = q_k$$

→ Here we use a Dirichlet dist.

Dirichlet dist → Multinomial likelihood → DD Posterior

# Dirichlet Distribution

$$p(q|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} q_1^{\alpha_1 - 1} \cdots q_m^{\alpha_m - 1}$$

$q = (q_1, \ldots q_m)$ is a point on $(m-1)$-simplex $0 < q_i < 1$

$$\sum_{i=1}^{m} q_i = 1$$

$\alpha = (\alpha_1, \alpha_2, \ldots \alpha_m)$ $\xrightarrow{\alpha_i > 0}$ Parameters

Posterior $= \mathrm{Dir}(\alpha_1 + z_1, \ldots \alpha_m + z_m)$

# Dirichlet Distribution

① $\text{Dir}(\alpha_1, \ldots, \alpha_m) = \left( \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1) + \Gamma(\alpha_2) + \cdots + \Gamma(\alpha_m)} \right) \ldots = \Sigma\sigma$

$\left( \frac{\Gamma(\alpha_m)}{\Sigma\sigma} \right)$

② Sum of Gamma RV with n
Common Scale parameter $\triangleq$ Gamma RV

# Dirichlet Process

- A *Dirichlet Process* is also a distribution over distributions.
- Let G be Dirichlet Process distributed:

$$G \sim DP(\alpha, G_0)$$

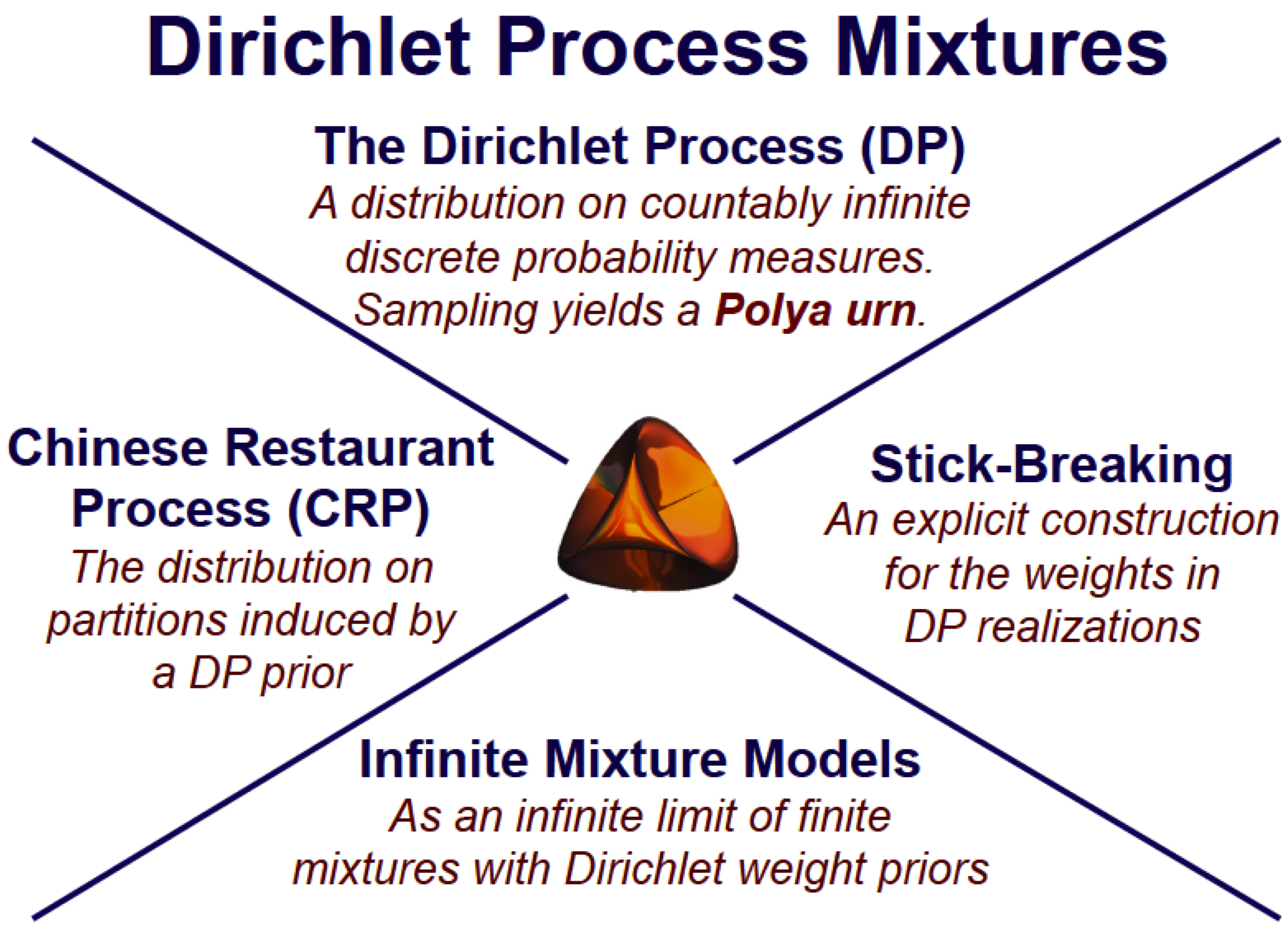
  - $G_0$ is a base distribution
  - $\alpha$ is a positive scaling parameter
- G is a random probability measure that has the same support as $G_0$
- The Dirichlet process can also be seen as the infinite-dimensional generalization of the Dirichlet distribution.
- A particularly important application of Dirichlet processes is as a prior probability distribution in infinite mixture models.

# Dirichlet Process

- In the same way as the Dirichlet distribution is the conjugate prior for the multinomial distribution, the Dirichlet process is the conjugate prior for infinite, nonparametric discrete distributions.

Dirichlet processes (DPs) are a class of Bayesian nonparametric models.

# Dirichlet Process



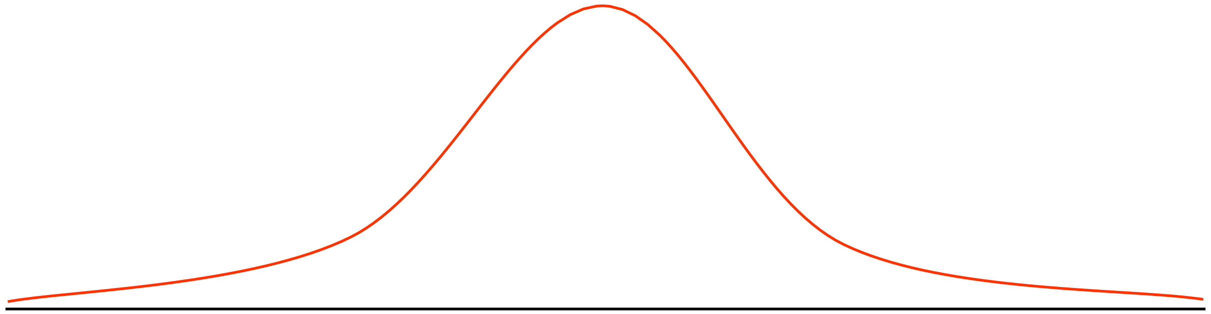
# Dirichlet Processes: Big Picture

There are many ways to derive the Dirichlet Process:

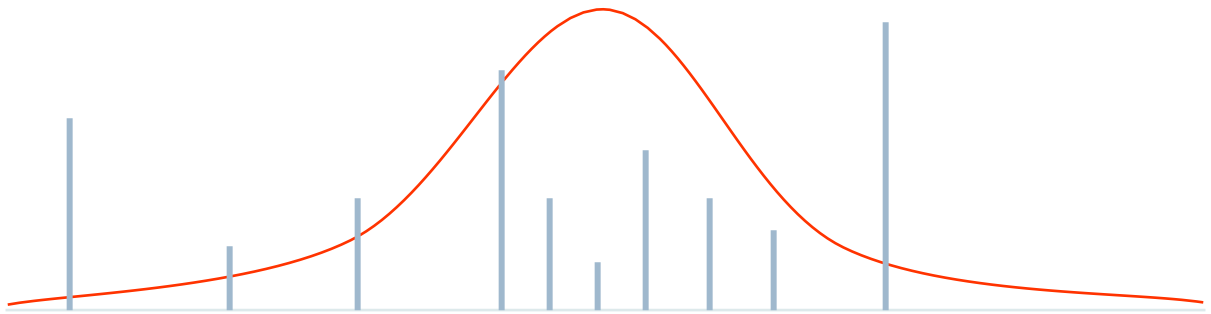
- Dirichlet distribution
- Urn model
- Chinese restaurant process
- Stick breaking
- Gamma process

# Dirichlet Process

- Consider Gaussian $G_0$

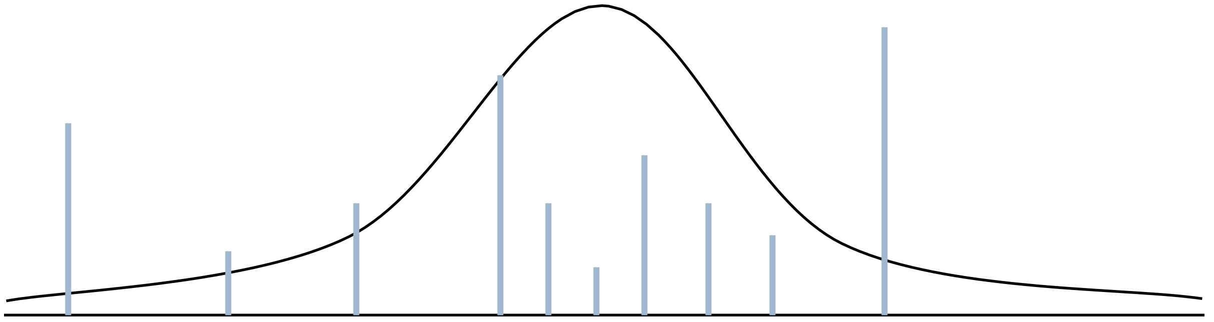


- $G \sim DP(\alpha, G_0)$



# Dirichlet Process

- $G \sim DP(\alpha, G_0)$



- $G_0$ is continuous, so the probability that any two samples are equal is precisely zero.
- However, G is a discrete distribution, made up of a countably infinite number of point masses.
  - Therefore, there is always a non-zero probability of two samples colliding