# Introduction to Nonparametric Bayesian Modeling (Contd)

CS698X: Topics in Probabilistic Modeling and Inference

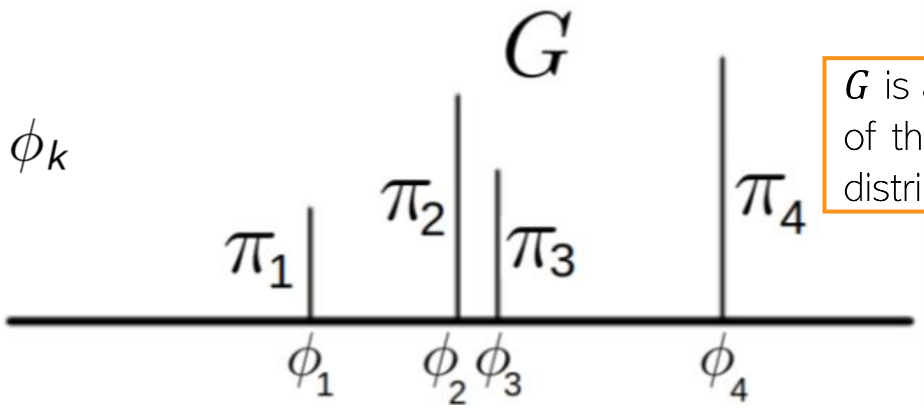
Piyush Rai

# Being Nonparametric using Models that have a Shrinkage Effect

# Mixture Models: Another Construction

- Consider a finite mixture model with $K$ components with params $(\mu_k, \Sigma_k)_{k=1}^{K}$

$(\mu_k, \Sigma_k) = \phi_k$




$G$ is a representation of this mixture distribution

Defined by $K$ locations or "atoms" with parameters $\{\phi_k\}_{k=1}^{K}$ with respective selection probabilities $\{\pi_k\}_{k=1}^{K}$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

- In the finite case, we can assume $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$ and $\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$
- We can make it a nonparametric model by making $\boldsymbol{\pi}$ an infinite-dimensional vector

In practice, only a finite of these will have nonzero values, and others will shrink to very small (or zero), as we will see

$$\pi_1, \pi_2, \pi_3, \dots, \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

Indeed. Called a "Dirichlet Process"

Related: "Stick-breaking Process"

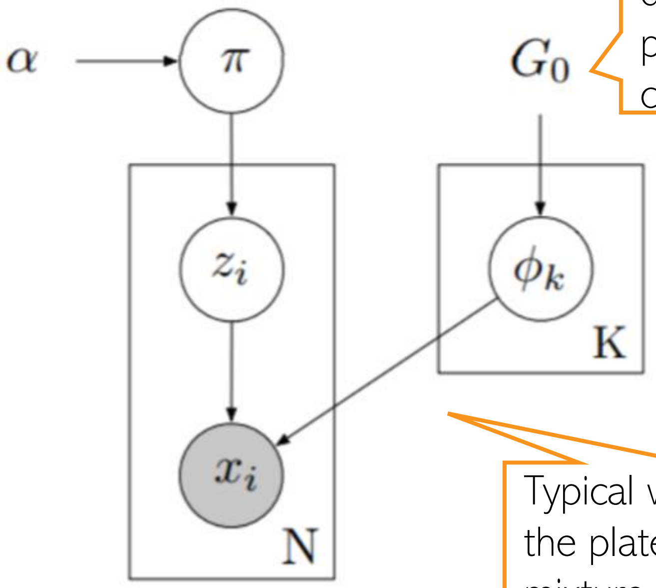
- How to construct such a vector? Is there an infinite dimensional Dirichlet distribution?

# Mixture Models: Two Equivalent Views

But how to construct such a $G$ distribution with potentially infinite components?

Prior (a.k.a. "base distribution" for the parameters of each mixture component

Example: $G_0$ can be NIW if each component is a Gaussian and $\phi_k = (\mu_k, \Sigma_k)$




Typical way of showing the plate notation of a mixture model

$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2,\dots,K$$

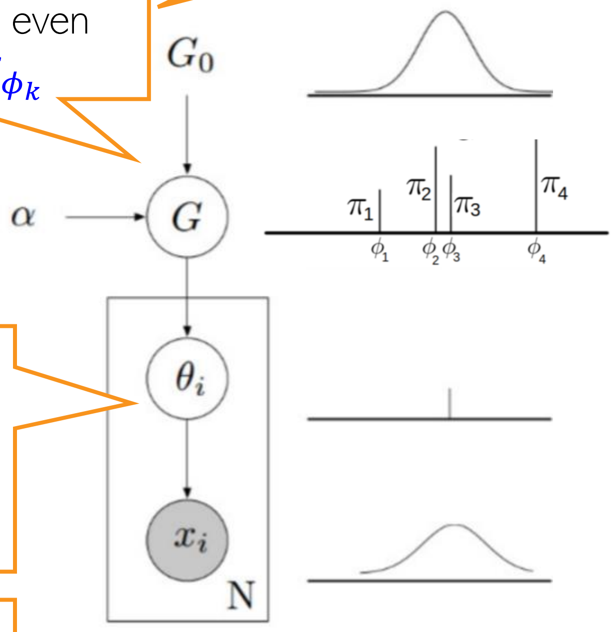
$$z_i \sim \text{multinoulli}(\boldsymbol{\pi}) \qquad i = 1,2,\dots,N$$

$$x_i \sim p(x|\phi_{z_i}) \qquad i = 1,2,\dots,N$$

Similar representation even when $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

No explicit cluster ids; instead, $\theta_i$ denotes the param of the distribution which will generate $x_i$

Since $G$ is discrete, there will at most be $K$ distinct $\theta_i$'s, thereby achieving clustering




$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2,\dots,K$$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

$$\theta_i \sim G \qquad i = 1,2,\dots,N$$

$$x_i \sim p(x|\theta_i) \qquad i = 1,2,\dots,N$$

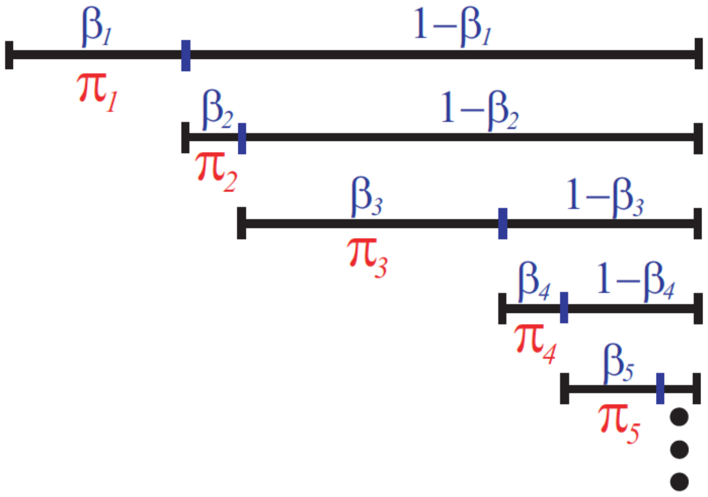
# Stick-Breaking Process (Sethuraman'94)

SBP gives us a way to construct infinite dimensional Dirichlet distribution known as the "Dirichlet Process"

A similar SBP construction exists for Beta Process/IBP

- Recursively break a length 1 stick into two pieces
- Assume breaking point in each round is drawn from a Beta distribution

$$\beta_k \sim \text{Beta}(1,\alpha) \qquad k = 1,\ldots,\infty$$
$$\pi_1 = \beta_1$$
$$\pi_k = \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) \qquad k = 2,\ldots,\infty$$




- Can show that $\sum_{k=1}^{\infty}\pi_k - 1 \to 0$ which is what we want
- We can now have a "nonparametric/infinite" mixture distribution $G = \sum_{k=1}^{\infty}\pi_k \delta_{\phi_k}$
- "Location/atoms" $\phi_k$ can be drawn from a "base" distr $G_0$, say NIW if $\phi_k = (\mu_k, \Sigma_k)$
- We basically replaced the Dirichlet prior on $\boldsymbol{\pi}$ by a Stick-Breaking Process (SBP) prior

# An Aside: Infinite Dimensional Dirichlet

- Drawing from an infinite-dim Dirichlet would give an infinite-dim prob. vector

$$\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \dots]$$

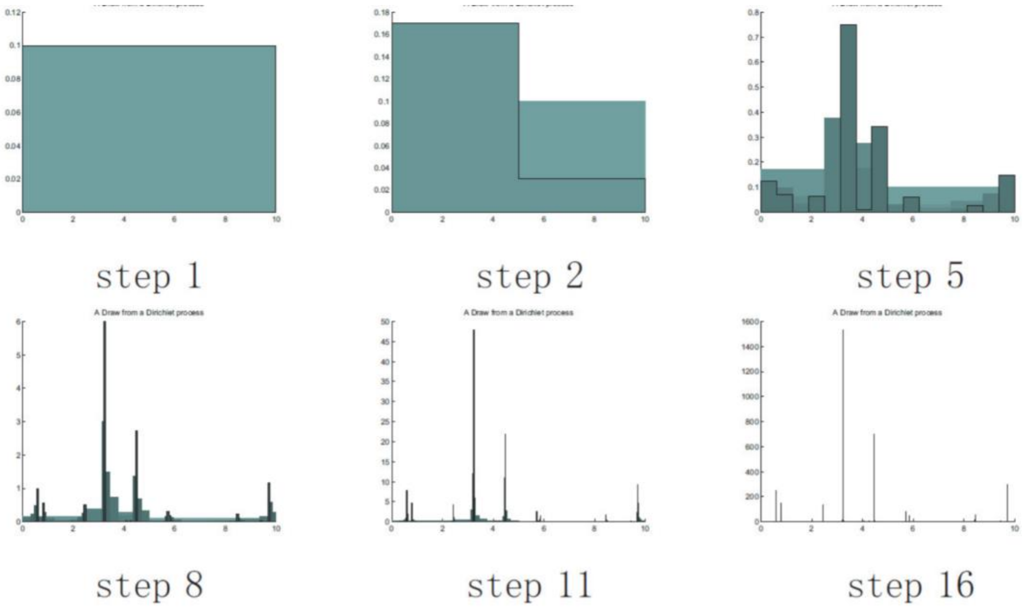
- We can construct this vector to have very few entries as nonzero
- Consider recursively drawing from a Dirichlet as defined below

$$\begin{aligned} 1 &\sim \text{Dirichlet}(\alpha) \\ (\pi_1, \pi_2) &\sim \text{Dirichlet}(\alpha/2, \alpha/2) \\ (\pi_1\pi_{11}, \pi_1\pi_{12}, \pi_2\pi_{21}, \pi_2\pi_{22}) &\sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4) \end{aligned}$$

As the concentration parameter gets smaller and smaller, the split of values in LHS get more and more skewed

Therefore, after doing the above a few times, the $\pi$ vector will only have a very few entries as nonzero and in the infinite-sized $\pi$, there will only be a finite many nonzero entries, and rest will be zero

This is basically what happens in the case of Dirichlet Process / Stick-Breaking Process



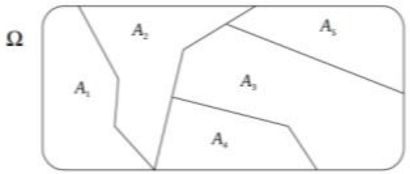
# An Aside: Dirichlet Process - Formally

SBP gives an explicit way to construct "Dirichlet Process"

- A Dirichlet Process $\text{DP}(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim \text{DP}(\alpha, G_0)$ will give us a distribution
  - $\alpha$ : concentration param, $G_0$: base distribution (=mean of DP)
  - Large $\alpha$ means $G \to G_0$
- **Fact 1:** If $G \sim \text{DP}(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed

$$[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)




- **Fact 2:** Any $G$ drawn from $\text{DP}(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)
- **Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant

# Another NPBayes Model with Shrinkage Construction

# Multiplicative Gamma Process

- Consider the SVD-style probabilistic model with an *a priori* unbounded $K$

$$\mathbf{X} = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$$

- Consider the following prior on each "singular values" $\lambda_k$

$$\begin{aligned} \lambda_k &\sim \mathcal{N}(0, \tau_k^{-1}) \\ \tau_k &= \prod_{\ell=1}^{k} \delta_\ell \\ \delta_\ell &\sim \text{Gamma}(\alpha, 1) \quad \text{where } \alpha > 1 \end{aligned}$$

Precision keeps on getting larger and larger as $k$ grows (thus variance keeps getting small and smaller)

Thus $\mathbb{E}[\delta_\ell] = \alpha$ (greater than 1 in expectation)

- In practice we can set $K$ to be a sufficiently very large
  - Due to the shrinkage property, only a finite many $\lambda_k$ will be nonzero
  - The nonzero $\lambda_k$'s will dictate the effective $K$

# Summary

- We saw three nonparametric Bayesian models (mainly used in unsup learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
  - Multiplicative Gamma Process: For SVD-like matrix factorization
- Many applications of these models to solve a wide range of problems
- Also saw GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models
  - Many other such nonparametric Bayesian models for other problems in machine learning
  - "A tutorial on Bayesian nonparametric models" Gershman and Blei, 2011) is a nice survey
- Rich theory based on stochastic processes (beyond the scope of this course)
- Inspired other non-probabilistic algos, e.g., Using Dirichlet Process Mixture Model to get a $K$-means like clustering algorithm (DP-means) which doesn't require $K$