

Nonparametric Bayesian Models

CS772A: Probabilistic Machine Learning

Piyush Rai

Nonparametric Bayesian Methods

- Need for nonparametric Bayesian modeling
- Some motivating problems
- NPBayes modeling mixture models (clustering)
- Some standard ways of constructing NPBayes models
 - Stick-breaking process, Dirichlet process
 - Some metaphors: Chinese Restaurant Process



Motivating Problem: Mixture Models

- Suppose each observation is generated from a K component mixture model

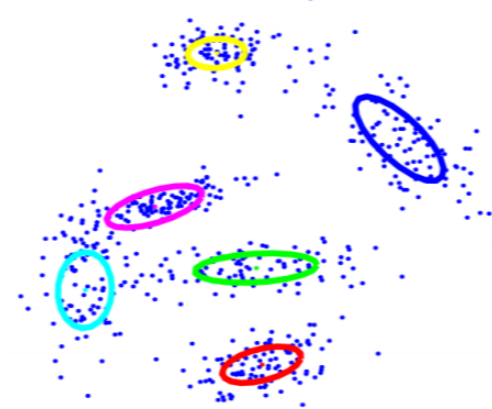
Cluster id of the n^{th} observation

K -dim probability vector of component mixing proportions

$$\mathbf{z}_n \sim \text{multinoulli}(\boldsymbol{\pi})$$

n^{th} observation

$$\mathbf{x}_n \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}_n}, \boldsymbol{\Sigma}_{\mathbf{z}_n})$$



- How to learn K , i.e., the number of components (clusters) for such a mixture model?
- Can use marginal-likelihood based model selection but is expensive
 - Need to train the model several times for each possible value of K
- Also difficult if the data is streaming (hard to know beforehand how many clusters)
- How about a prior over $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]$ (or $\boldsymbol{\pi}$) that allows learning the “right” K ?



Motivating Problem: Latent Feature Models

- Suppose each observation is a subset sum of K “basis vectors” (or “latent features”*)

Is k^{th} latent feature present in the n^{th} observation?

$$z_{nk} \sim \text{Bernoulli}(\pi_k) \quad k = 1, \dots, K$$

An example: Each text document in a collection being a subset sum of K “latent” themes or topics present in the collection

The n^{th} observation ($D \times 1$) expressed as a subset sum of the K latent features (each $D \times 1$), plus some noise

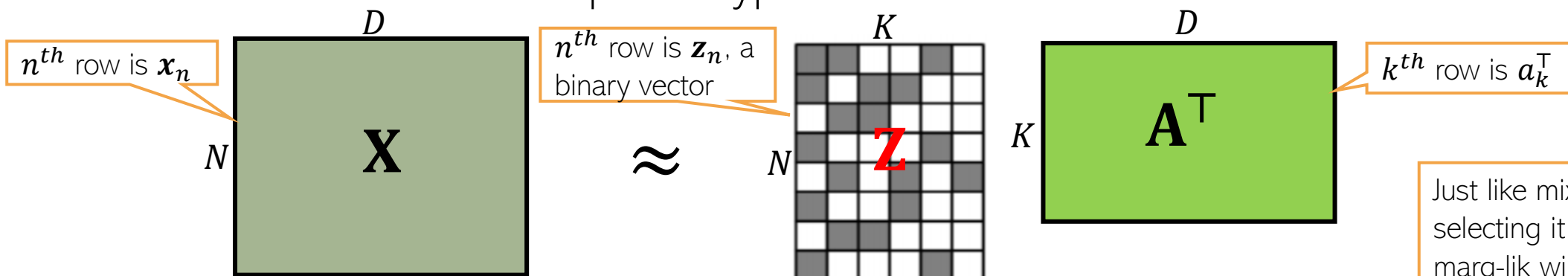
$$\mathbf{x}_n = \sum_{k=1}^K z_{nk} \mathbf{a}_k + \epsilon_n = \mathbf{A} \mathbf{z}_n + \epsilon_n$$

Noise (e.g., zero mean Gaussian)

The k^{th} latent feature ($D \times 1$)

A binary sparse matrix

- This can also be seen as special type of matrix factorization $\mathbf{X} = \mathbf{Z} \mathbf{A}^T + \mathbf{E}$



Just like mixture models, selecting it based on marg-lik will be expensive

- How about a prior over \mathbf{Z} (or \mathbf{A} or $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$) that allows learning the “right” K ?

Motivating Problem: SVD-style Matrix Factorization⁵

- Consider the following SVD-style decomposition for an $N \times M$ matrix \mathbf{X}

$$\mathbf{X} = \sum_{k=1}^K \lambda_k \mathbf{u}_k \mathbf{v}_k^T + \mathbf{E} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T + \mathbf{E}$$

Rank 1 matrix

Zero mean Gaussian noise

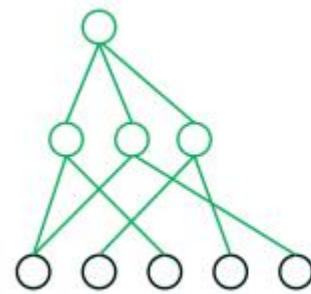
- Each $\mathbf{u}_k \in \mathbb{R}^N$, $\mathbf{v}_k \in \mathbb{R}^M$, $\lambda_k \in \mathbb{R}$, and $\mathbf{\Lambda}$ is a $K \times K$ diag matrix with λ_k 's on diags
- This is basically a weighted sum of K rank-1 matrices
 - λ_k 's are the weights
 - λ_k 's are akin to the singular values in SVD
- How to learn K , i.e., the “rank” of the above factorization?
- How about a prior on $\mathbf{\Lambda}$, or \mathbf{U} or \mathbf{V} , that allows us to learn the “right” K ?



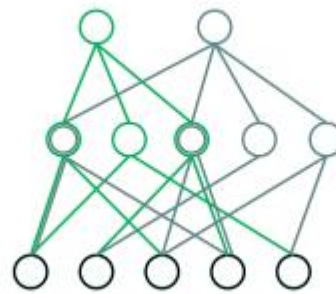
*Sparse Bayesian infinite factor models (Bhattacharya and Dunson, 2011)

Motivating Problem: “Lifelong” Learning

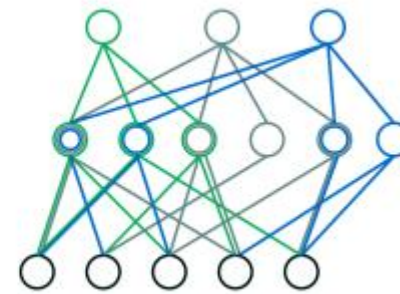
- In many scenarios, we want models that can learn in a “lifelong” fashion
 - Also known as “**continual learning**”
 - Goal is to **learn new knowledge** but also to **not forget previously acquired knowledge**
- In such cases we want models that do not have a fixed size decided beforehand
- Instead we prefer models with size that can grow/adapt with time as it
 - Sees more training data
 - Sees new tasks



Task 1



Task 1 + Task 2



Task 1 + Task 2 + Task 3



Nonparametric Bayesian Modeling

A vast area of research in ML and statistics. We will only be looking at a basic flavor of some approaches



- Enables constructing models that do not have an *a priori* fixed size

- Nonparametric does not mean no parameters

- Instead, have a possibly infinite (unbounded) number of parameters
 - Note: We've already seen Gaussian Processes which is a nonparametric Bayesian model

And can potentially grow as we see more and more data (actual number will depend on the amount/properties of data)

- Usually constructed via one of the following ways

- Take a finite model (e.g., a finite mixture model) and consider its “infinite limit”
 - Have a model that allows very large number of params but has a “shrinkage” effect, e.g.,

$$\mathbf{X} = \sum_{k=1}^K \lambda_k \mathbf{u}_k \mathbf{v}_k^T + \mathbf{E} \quad \lambda_k \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty$$

- We will look at some examples of both these approaches



Being Nonparametric by Taking Infinite Limit of Finite Models



A Finite Mixture Model

- Data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, cluster assignments $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]$, K clusters
- Denote the mixing proportion by a vector $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$, $\sum_{k=1}^K \pi_k = 1$

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

$$p(\mathbf{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

$$p(\mathbf{X}|\boldsymbol{\pi}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k p(\mathbf{x}_n|\mathbf{z}_n = k)$$

a.k.a. “collapsing” a variable; one less variable to infer now

- Integrating out $\boldsymbol{\pi}$, the marginal prior probability of cluster assignments

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \frac{\prod_{k=1}^K \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \quad (\text{verify})$$

Number of points with $\mathbf{z}_n = k$



A Finite Mixture Model

- The prior distribution of \mathbf{z}_n given cluster assignment \mathbf{Z}_{-n} of other points?

A discrete distribution (multinoulli) since \mathbf{z}_n can take one of K possibilities

$$p(\mathbf{z}_n | \mathbf{Z}_{-n}, \alpha) = \frac{p(\mathbf{z}_n, \mathbf{Z}_{-n} | \alpha)}{p(\mathbf{Z}_{-n} | \alpha)} = \frac{p(\mathbf{Z} | \alpha)}{p(\mathbf{Z}_{-n} | \alpha)}$$

This “conditional” prior is needed since we have integrated out $\boldsymbol{\pi}$ and thus \mathbf{z}_n 's become coupled

- Using $p(\mathbf{Z} | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\prod_{k=1}^K \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}$ we have

Number of points in cluster j , not counting \mathbf{x}_n

$$p(\mathbf{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \frac{p(\mathbf{z}_n = j, \mathbf{Z}_{-n} | \alpha)}{p(\mathbf{Z}_{-n} | \alpha)} = \frac{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(m_j + \frac{\alpha}{K}) \prod_{k \neq j} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}}{\frac{\Gamma(\alpha)}{\Gamma(N-1+\alpha)} \frac{\Gamma(m_j - 1 + \frac{\alpha}{K}) \prod_{k \neq j} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}} = \frac{m_{-n,j} + \frac{\alpha}{K}}{N - 1 + \alpha}$$

- Note: Can also get this result using $p(\mathbf{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \int p(\mathbf{z}_n = j | \boldsymbol{\pi}) p(\boldsymbol{\pi} | \mathbf{Z}_{-n}, \alpha) d\boldsymbol{\pi}$
- Thus prior prob. of $\mathbf{z}_n = j$ is proportional to how many other points are in cluster j
- Note that it also implies that mixture models have a **rich-gets-richer** property
 - Meaning: *a priori*, a cluster with more points is likely to attract more points



Taking the Infinite Limit..

- Since $p(\mathbf{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j} + \frac{\alpha}{K}}{N-1+\alpha}$, as $K \rightarrow \infty$, $p(\mathbf{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha}$
- Suppose only K_+ clusters are currently occupied (i.e., have at least one data point)
- Total prob. of \mathbf{x}_n going to any of these K_+ clusters $= \sum_{j=1}^{K_+} \frac{m_{-n,j}}{N-1+\alpha} = \frac{N-1}{N-1+\alpha}$
- Probability of \mathbf{x}_n going to a new (i.e., so far unoccupied) cluster $= \frac{\alpha}{N-1+\alpha}$
- Therefore in the limit of an unbounded number of clusters, we have

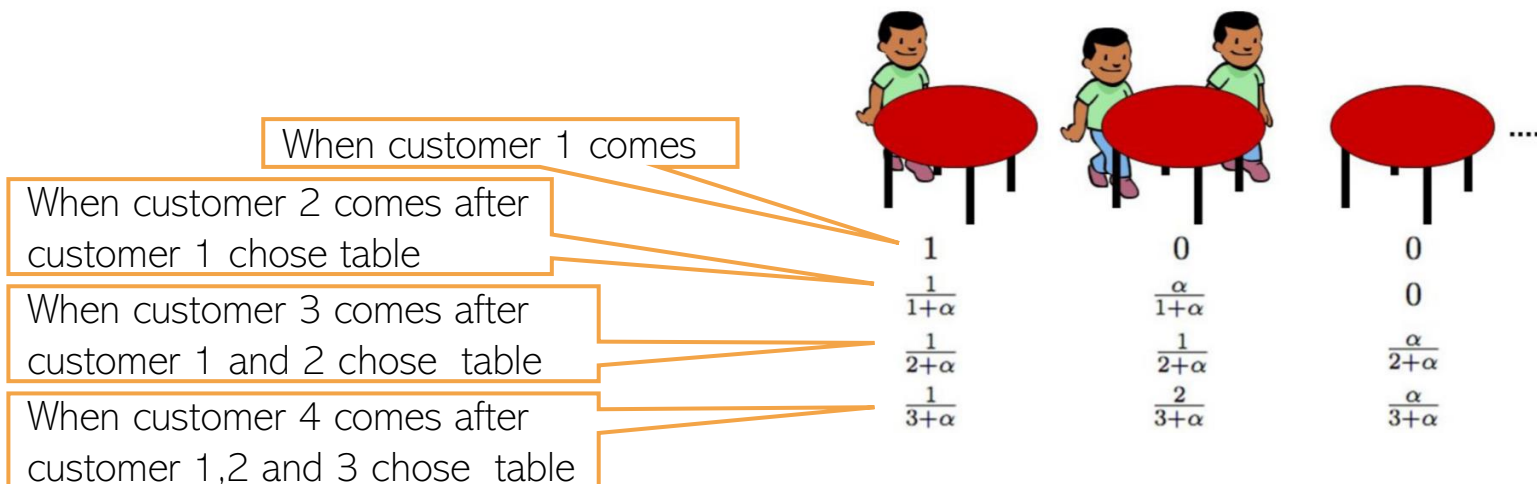
$$p(\mathbf{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} & \text{(prob. of going to } j = 1, \dots, K_+) \\ \frac{\alpha}{N-1+\alpha} & \text{(prob. of creating a new cluster } K_+ + 1) \end{cases}$$

- The above gives us a prior distribution for mixture models with unbounded K
 - Can combine it now with the suitable likelihood to infer the posterior* of \mathbf{Z}
- Note: Prob. of starting a new cluster is prop. to Dirichlet hyperparam α (can learn it)



A Metaphor: Chinese Restaurant Process (CRP)

- Assume a restaurant with infinite number of tables (each table denotes a cluster)
- Customer 1 sits at a randomly chosen table (all tables are equivalent to begin with)
- Each subsequent customer $n > 1$ sits using the following scheme
 - Sits at an already occupied table k with probability $\frac{m_k}{n-1+\alpha}$
 - Sits at a new table with probability $\frac{\alpha}{n-1+\alpha}$

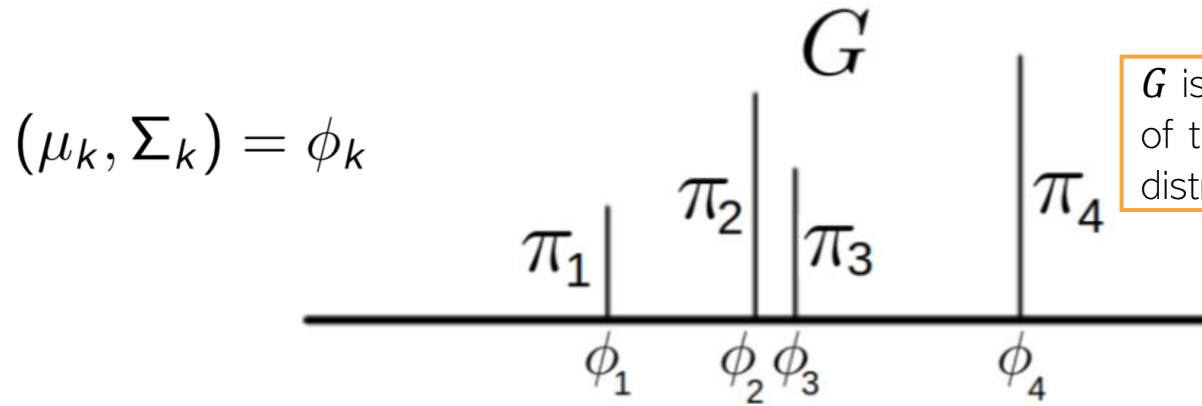


Being Nonparametric using Models that have a Shrinkage Effect



Mixture Models: Another Construction

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



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

$\pi_1, \pi_2, \pi_3, \dots,$

$$\sum_{k=1}^{\infty} \pi_k = 1$$

Indeed. Called a "Dirichlet Process"

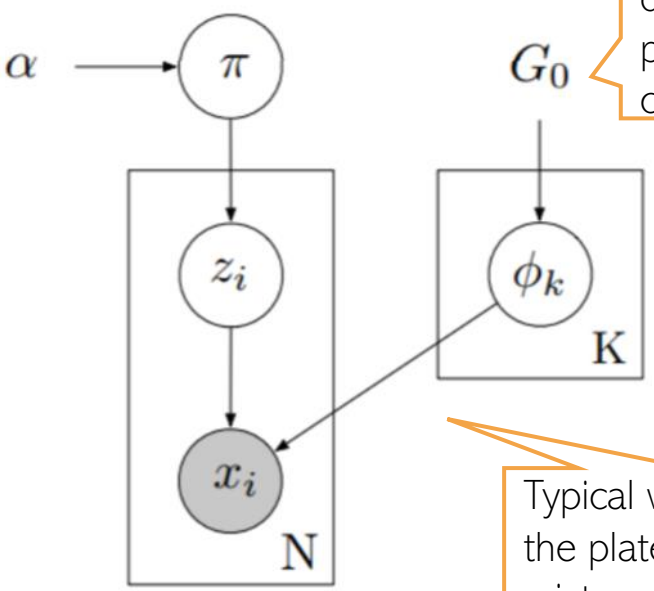
Related: "Stick-breaking Process"



- 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

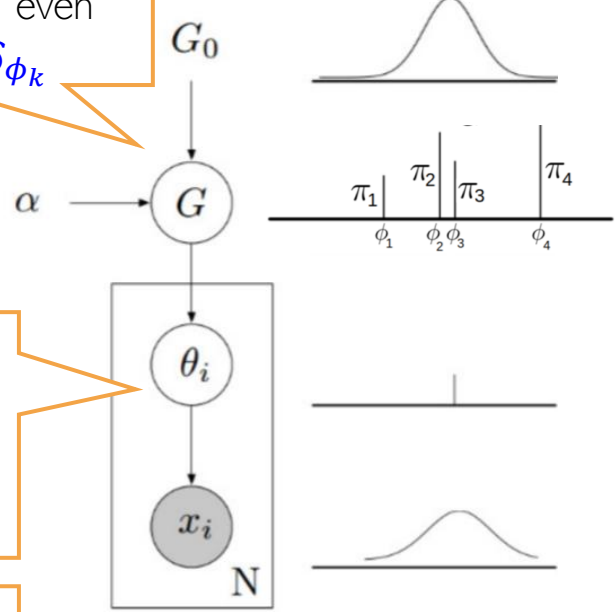
Example: G_0 can be NIW if each component is a Gaussian and $\phi_k = (\mu_k, \Sigma_k)$

Typical way of showing the plate notation of a mixture model

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

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

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



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

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

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

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

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

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

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

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

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



Stick-Breaking Process (Sethuraman'94)

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

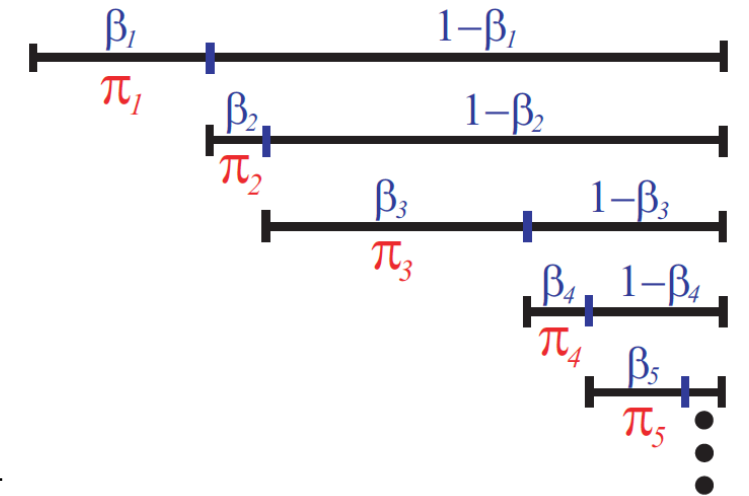


- Recursively break a length 1 stick into two pieces
- Assume breaking point in each round is drawn from a Beta distribution

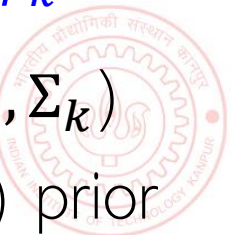
$$\beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, \dots, \infty$$

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \quad k = 2, \dots, \infty$$



- Can show that $\sum_{k=1}^{\infty} \pi_k = 1 \rightarrow 0$ 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" ϕ_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



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

$\mathbf{1} \sim \text{Dirichlet}(\alpha)$

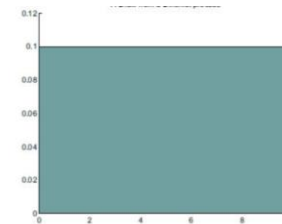
$(\pi_1, \pi_2) \sim \text{Dirichlet}(\alpha/2, \alpha/2)$

$(\pi_1\pi_{11}, \pi_1\pi_{12}, \pi_2\pi_{21}, \pi_2\pi_{22}) \sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4)$

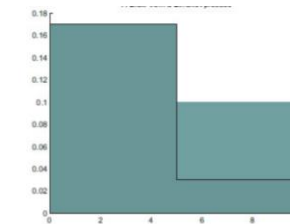
As the concentration parameter gets smaller and smaller, the split of values in LHS get more and more skewed

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

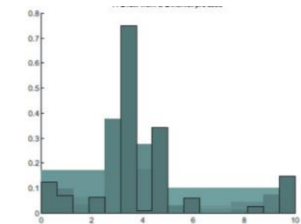
This is basically what happens in the case of Dirichlet Process / Stick-Breaking Process



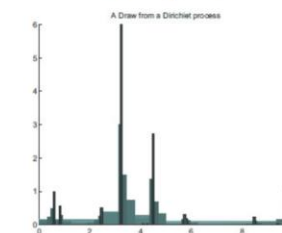
step 1



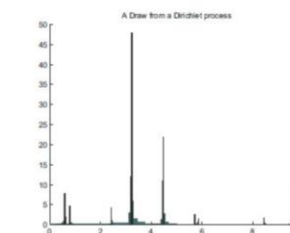
step 2



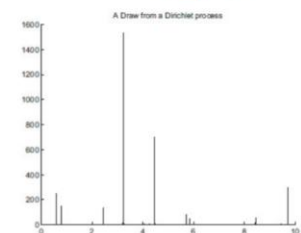
step 5



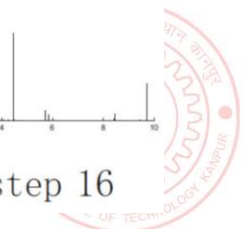
step 8



step 11



step 16



Dirichlet Process - Formally

SBP gives an explicit way to construct “Dirichlet Process”

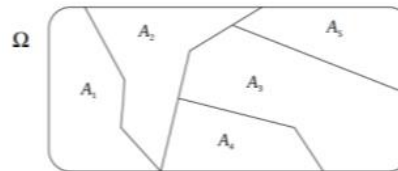


- A Dirichlet Process $DP(\alpha, G_0)$ defines a **distribution over distributions**
 - So $G \sim DP(\alpha, G_0)$ will give us a distribution
 - α : **concentration param**, G_0 : **base distribution** (=mean of DP)
 - Large α means $G \rightarrow G_0$

- **Fact 1:** If $G \sim DP(\alpha, G_0)$ then any **finite dim. marginal** of G is Dirichlet distributed

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

for any finite partition A_1, \dots, A_K of the space Ω (Ferguson, 1973)



- **Fact 2:** Any G drawn from $DP(\alpha, G_0)$ 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 π_k 's will be significant



Another NPBayes Prior: Multiplicative Gamma Process ¹⁹

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

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

- Consider the following prior on each “singular values” λ_k

$$\lambda_k \sim \mathcal{N}(0, \tau_k^{-1})$$

$$\tau_k = \prod_{\ell=1}^k \delta_{\ell}$$

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

$$\delta_{\ell} \sim \text{Gamma}(\alpha, 1) \quad \text{where } \alpha > 1$$

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 λ_k will be nonzero
 - The nonzero λ_k 's will dictate the effective K



Summary of NPBayes

- We saw some nonparametric Bayesian models (mainly used in unsup learning)
 - CRP/Dirichlet Process: For clustering problems
 - 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

