Assorted Topics in Probabilistic ML (2)

CS772A: Probabilistic Machine Learning

Piyush Rai

Plan for today

- Assorted Topics
  - Conformal Prediction (simple and fast way to get prediction uncertainty/set)
  - Nonparametric Bayesian methods (learning the right model size/complexity)

Conformal Prediction

- A simple technique to easily obtain confidence intervals
  - In classification, such an interval may refer to the set of highly likely classes for a test input

  - For more difficult test inputs, the set would typically be larger
- In a way, conformal prediction gives predictive uncertainty
  - However, unlike Bayesian ML, we don't get model uncertainty
  - Only one model is learned in the standard way and we construct the set of likely classes
  - It's like a black-box method; no change to training procedure for the model

Conformal Prediction

Assume it's a classification model which produces softmax scores

Conformal prediction can be used for regression problems too*

- Assume we already have a trained model $\hat{f}$ using some labelled data
- Idea: Use a calibration set of n examples to generate a prediction set $\mathcal{C}(X_{test})$ s.t.

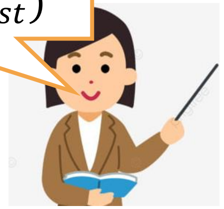
α is a user chosen error rate

Its true label

Another fresh test input

With high prob., the true label is in set $\mathcal{C}(X_{test})$

$$1 - \alpha \leq p\big(Y_{test} \in \mathcal{C}(X_{test})\big) \leq 1 - \alpha + \frac{1}{n+1}$$

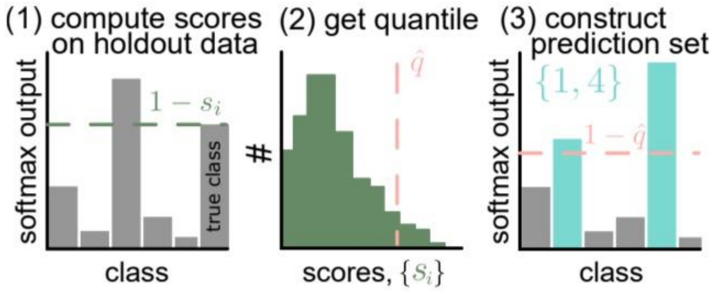


- The approach* to construct the prediction set $\mathcal{C}(X_{test})$ is as follows:
  - Assuming classification task, for each example in the calibration set, compute

high means low-confidence prediction by the model

Conformal score: one minus the softmax score of the correct class

$$s_i = 1 - \hat{f}(x_i)_{y_i}$$



  - Compute the $1 - \alpha$ quantile of $s_1, s_2, \ldots, s_n$. Call it $\hat{q}$
  - Now the calibration set for a new test input X_{test} can be defined as

Set of all classes whose predicted softmax values are "high enough"

$$\mathcal{C}(X_{test}) = \{y : \hat{f}(X_{test})_y \geq 1 - \hat{q}\}$$

*A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (Angelopoulos and Bates, 2022)

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

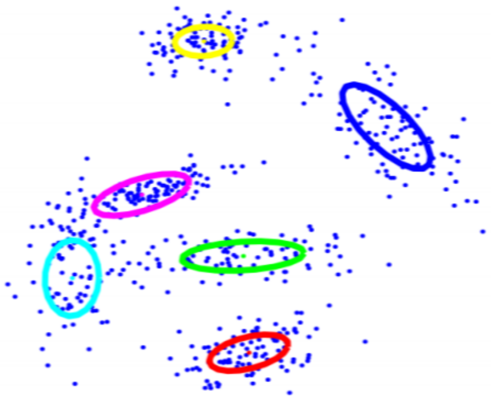
- Suppose each observation is generated from a K component mixture model

Cluster id of the n^{th} observation

K-dim probability vector of component mixing proportions

n^{th} observation

$$\mathbf{z}_n \sim \text{multinoulli}(\pi)$$
$$\mathbf{x}_n \sim \mathcal{N}(\mu_{\mathbf{z}_n}, \Sigma_{\mathbf{z}_n})$$



- 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 <u>subset sum</u> of K “basis vectors” (or “latent features”*)

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

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

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

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

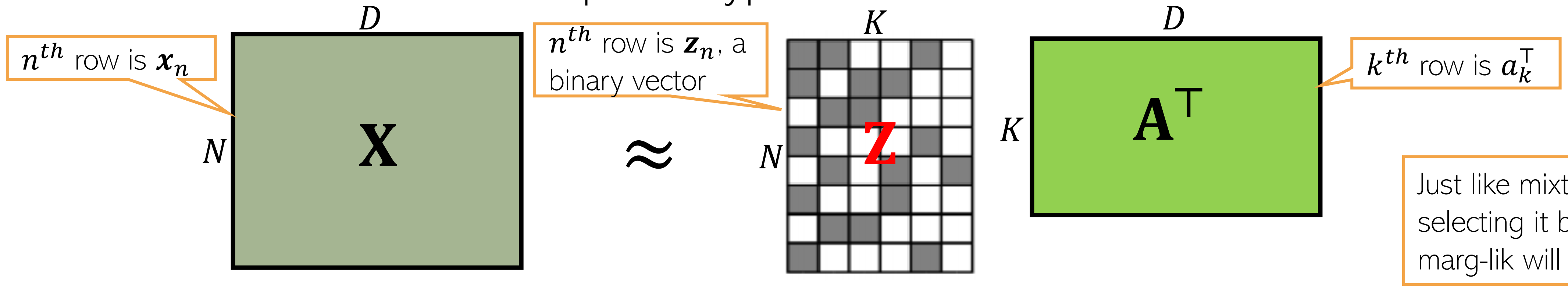
$$\boldsymbol{x}_n = \sum_{k=1}^{K} z_{nk} \boldsymbol{a}_k + \epsilon_n = \mathbf{A}\boldsymbol{z}_n + \epsilon_n$$

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 $\mathbf{X} = \mathbf{Z}\mathbf{A}^\top + \mathbf{E}$



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

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

* Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)

Motivating Problem: SVD-style Matrix Factorization

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

Rank 1 matrix

Zero mean Gaussian noise

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

- Each $\boldsymbol{u}_k \in \mathbb{R}^N$, $\boldsymbol{v}_k \in \mathbb{R}^M$, $\lambda_k \in \mathbb{R}$, and Λ is a $K \times K$ diag matrix with λ_k's on diags

- This is basically a <u>weighted</u> 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 Λ, or $\mathbf{U}$ or $\mathbf{V}$, that allows us to learn the "right" K ?

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

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.,

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

- We will look at some examples of both these approaches

A tutorial on Bayesian nonparametric models (Gershman and Blei, 2012)

Being Nonparametric by Taking Infinite Limit of Finite Models

A Finite Mixture Model

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

$$
\begin{aligned}
p(\boldsymbol{\pi}|\alpha) &= \text{Dirichlet}\left(\frac{\alpha}{K}, \frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) \\
p(\boldsymbol{z}_n|\pi) &= \prod_{k=1}^{K} \pi_k^{z_{nk}} \\
p(\mathbf{X}|\boldsymbol{\pi}) &= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\boldsymbol{z}_n = k)
\end{aligned}
$$

a.k.a. "collapsing" a variable; one less variable to infer now

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

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

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

A Finite Mixture Model

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

A discrete distribution (multinoulli) since $\boldsymbol{z}_n$ 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 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 $\boldsymbol{j}$, not counting $\boldsymbol{x}_n$

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

- Note: Can also get this result using $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n},\alpha) = \int p(\boldsymbol{z}_n = j|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{Z}_{-n},\alpha)d\boldsymbol{\pi}$
- Thus prior prob. of $\boldsymbol{z}_n = j$ 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 $p(z_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}+\frac{\alpha}{K}}{N-1+\alpha}$, as $K \to \infty$, $p(z_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha}$
- Suppose only K_+ clusters are currently occupied (i.e., have at least one data point)
- Total prob. of $\boldsymbol{x}_n$ 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 $\boldsymbol{x}_n$ 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

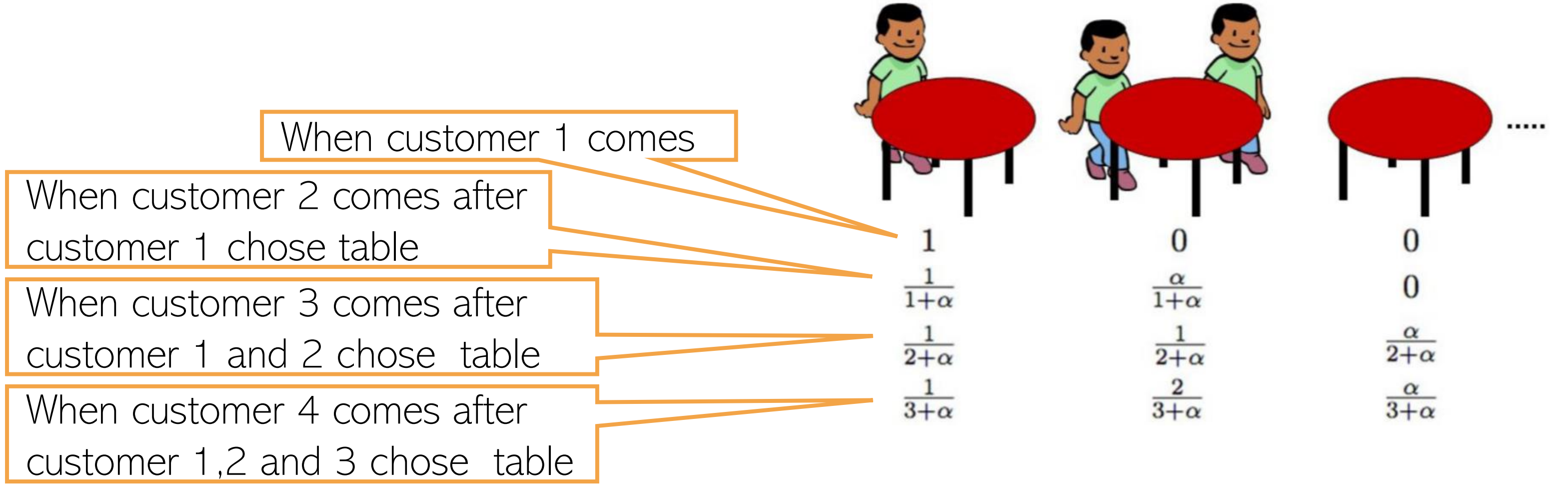
$$p(z_n = j|\mathbf{Z}_{-n}, \alpha) = \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} & \text{(prob. of going to } j = 1, \ldots, K_+\text{)} \\ \frac{\alpha}{N-1+\alpha} & \text{(prob. of creating a new cluster } K_+ + 1\text{)} \end{cases}$$

- 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 $\boldsymbol{Z}$
- Note: Prob. of starting a new cluster is prop. to Dirichlet hyperparam α (can learn it)

*Markov chain sampling methods for Dirichlet process mixture models, (Neal, 2000), Variational inference for Dirichlet process mixtures (Blei and Jordan, 2006)

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 $\frac{\alpha}{n-1+\alpha}$

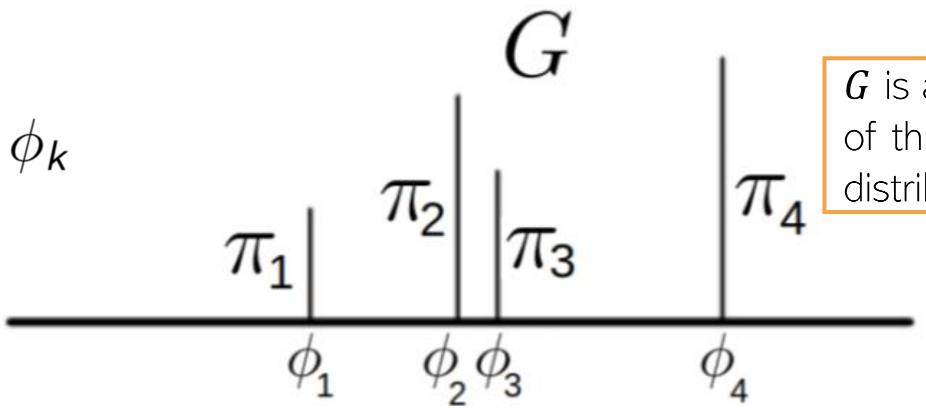


Being Nonparametric using Models that have a Shrinkage Effect

Mixture Models: Another Construction

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

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



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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ and $\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
- 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

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

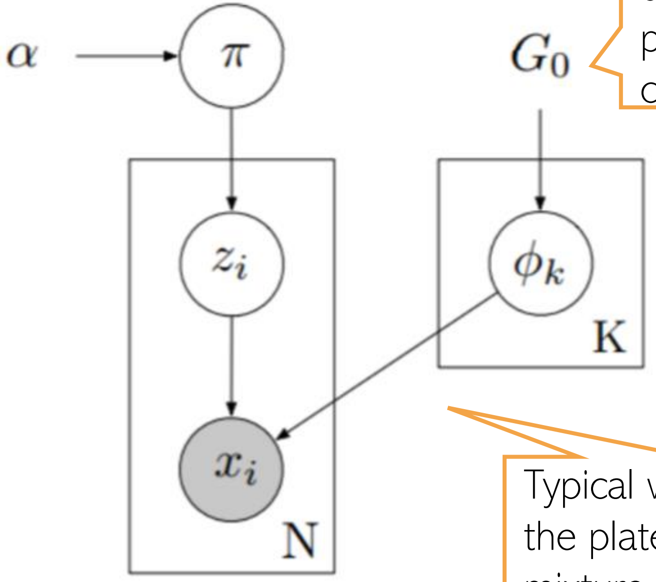
Indeed. Called a "Dirichlet Process"

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

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

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

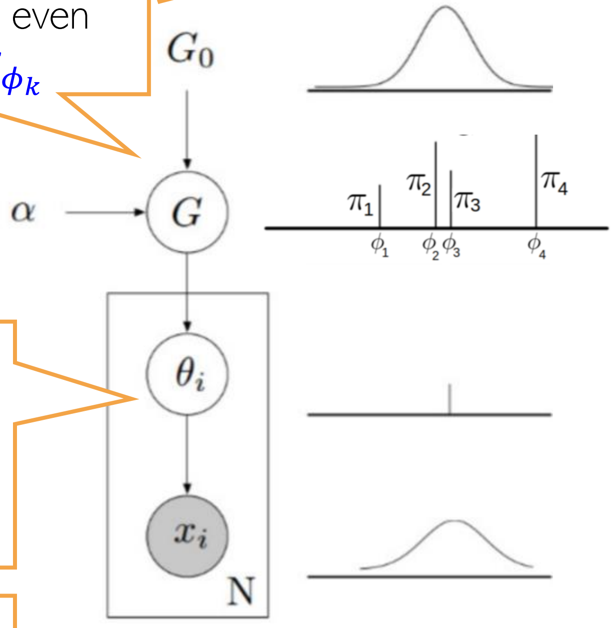
$$z_i \sim \text{multinoulli}(\boldsymbol{\pi}) \qquad i = 1,2, \dots, N$$

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

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

No explicit cluster ids; instead, 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



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

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

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

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

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

Stick-Breaking Process (Sethuraman'94)

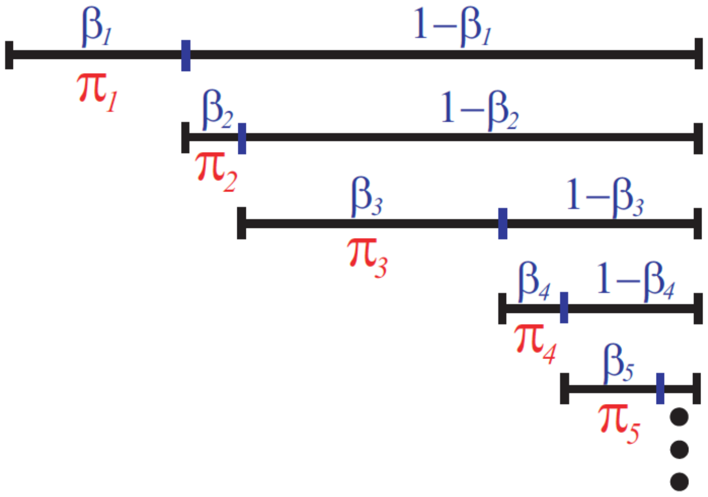
SBP gives us a way to construct infinite dimensional Dirichlet distribution known as the "Dirichlet Process"

- 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" 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

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

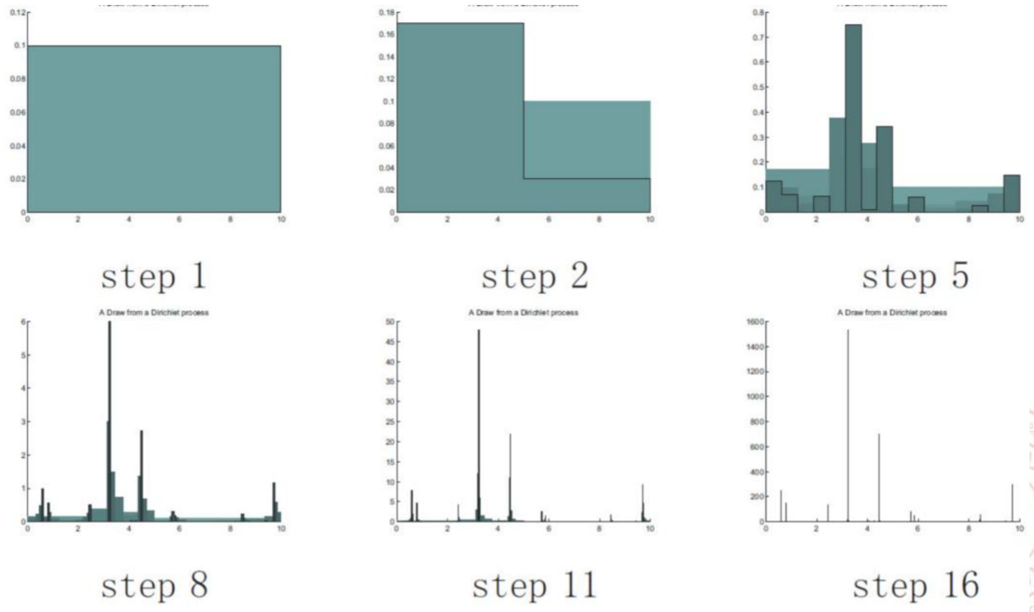
- We can construct this vector to have very few entries as nonzero
- Consider recursively drawing from a Dirichlet as defined below

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

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 π vector will only have a very few entries as nonzero and in the infinite-sized π, 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



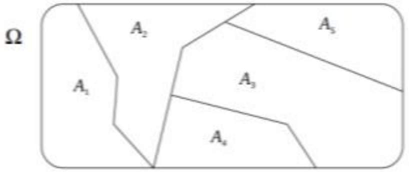
Dirichlet Process - Formally

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
  - α : concentration param, G_0: base distribution (=mean of DP)
  - Large α 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 Ω (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 π_k's will be significant

Summary

- We saw an example of a nonparametric Bayesian model
  - CRP/Dirichlet Process: For clustering problems
- NPBayes models exist for many other problems, e.g., 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)
- 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