Probabilistic Graphical Models

Bayesian nonparametrics: The Dirichlet process

Avinava Dubey
Lecture 18, March 23, 2015

Parametric vs nonparametric

Parametric model:
- Assumes all data can be represented using a fixed, finite number of parameters.
  - Mixture of $K$ Gaussians, polynomial regression.

Nonparametric model:
- Number of parameters can grow with sample size.
- Number of parameters may be random.
  - Kernel density estimation.

Bayesian nonparametrics:
- Allow an infinite number of parameters a priori.
- A finite data set will only use a finite number of parameters.
- Other parameters are integrated out.
Clustered data

- How to model this data?
- Mixture of Gaussians:
  \[ p(x_1, \ldots, x_N | \pi, \{ \mu_k \}, \{ \Sigma_k \}) = \prod_{n=1}^{\infty} \sum_{k=1}^{K} \pi_k N(x_k | \mu_k, \Sigma_k) \]
- Parametric model: Fixed finite number of parameters.

Bayesian finite mixture model

- How to choose the mixing weights and mixture parameters?
- Bayesian choice: Put a prior on them and integrate out:
  \[ p(x_1, \ldots, x_N) \]
  \[ = \int \int \int \left( \prod_{n=1}^{\infty} \sum_{k=1}^{K} \pi_k N(x_k | \mu_k, \Sigma_k) \right) p(\pi)p(\mu_{1:K})p(\Sigma_{1:K}) d\pi d\mu_{1:K} d\Sigma_{1:K} \]
- Where possible, use conjugate priors
  - Gaussian/inverse Wishart for mixture parameters
  - What to choose for mixture weights?
The Dirichlet distribution

- The Dirichlet distribution is a distribution over the \((K-1)\)-dimensional simplex.
- It is parametrized by a \(K\)-dimensional vector \((\alpha_1, \ldots, \alpha_K)\) such that \(\alpha_k \geq 0, k = 1, \ldots, K\) and \(\sum_k \alpha_k > 0\).
- Its distribution is given by

\[
\frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1}
\]

Samples from the Dirichlet distribution

- If \(\pi \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)\) then \(\pi_k \geq 0\) for all \(k\), and \(\sum_{k=1}^{K} \pi_k = 1\).
- Expectation: \(E\left[\langle \pi_1, \ldots, \pi_K \rangle\right] = \frac{(\alpha_1, \ldots, \alpha_K)}{\sum_k \alpha_k}\)
Conjugacy to the multinomial

- If \( \theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) and \( x_n \overset{iid}{\sim} \theta \)

\[
p(\pi|x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n|\pi)p(\pi) \\
= \left( \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} \right) \left( \frac{n!}{m_1! \ldots m_K! \pi_1^{m_1} \ldots \pi_K^{m_K}} \right) \\
\propto \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + m_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k + m_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k + m_k-1} \\
= \text{Dirichlet}(\pi|\alpha_1 + m_1, \ldots, \alpha_K + m_K)
\]

Distributions over distributions

- The Dirichlet distribution is a distribution over positive vectors that sum to one.
- We can associate each entry with a set of parameters
  - e.g. finite mixture model: each entry associated with a mean and covariance.
- In a Bayesian setting, we want these parameters to be random.
- We can combine the distribution over probability vectors with a distribution over parameters to get a distribution over distributions over parameters.
Example: finite mixture model

- Gaussian distribution: distribution over means.
  - Sample from a Gaussian is a real-valued number.

- Dirichlet distribution:
  - Sample from a Dirichlet distribution is a probability vector.
Example: finite mixture model

- Dirichlet prior
  - Each element of a Dirichlet-distributed vector is associated with a parameter value drawn from some distribution.
  - Sample from a Dirichlet prior is a probability distribution over parameters.

Properties of the Dirichlet distribution

- Relationship to gamma distribution: If \( \eta_k \sim \text{Gamma}(\alpha_k, 1) \),
  \[
  \frac{(\eta_1, \ldots, \eta_K)}{\sum_k \eta_k} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)
  \]

- If \( \eta_1 \sim \text{Gamma}(\alpha_1, 1) \) and \( \eta_2 \sim \text{Gamma}(\alpha_2, 1) \) then
  \[
  \eta_1 + \eta_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)
  \]

- Therefore, if \( (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \) then
  \[
  (\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K)
  \]
Properties of the Dirichlet distribution

- The beta distribution is a Dirichlet distribution on the 1-simplex.
- Let \((\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)\) and \(\theta \sim \text{Beta}(\alpha_1 b, \alpha_1 (1 - b)), 0 < b < 1\).
- Then
  \[ (\pi_1 \theta, \pi_1 (1 - \theta), \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 (1 - b_1), \alpha_2, \ldots, \alpha_K) \]
- More generally, if \(\theta \sim \text{Dirichlet}(\alpha_1 b_1, \alpha_1 b_2, \ldots, \alpha_1 b_N), \sum_i b_i = 1\), then
  \[ (\pi_1 \theta_1, \ldots, \pi_1 \theta_N, \pi_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 b_1, \ldots, \alpha_1 b_N, \alpha_2, \ldots, \alpha_K) \]

Renormalization:

If \((\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)\), then
\[
\frac{(\pi_2, \ldots, \pi_K)}{\sum_{k=1}^{K} \pi_k} \sim \\
\frac{(\pi_2, \ldots, \pi_K)}{\sum_{k=1}^{K} \pi_k} \sim \text{Dirichlet}(\alpha_2, \ldots, \alpha_K)
\]
Choosing the number of clusters

- Mixture of Gaussians – but how many components?
- What if we see more data – may find new components?

Bayesian nonparametric mixture models

- Make sure we always have more clusters than we need.
- Solution – infinite clusters *a priori*
  \[ p(x_n|\pi, \mu_k, \Sigma_k) = \sum_{k=1}^{\infty} \pi_k N(x_n|\mu_k, \Sigma_k) \]
- A finite data set will always use a finite – but *random* – number of clusters.
- How to choose the prior?
- We want something *like* a Dirichlet prior – but with an infinite number of components.
Constructing an appropriate prior

- Start off with $\pi^{(2)} = (\pi_1^{(2)}, \pi_2^{(2)}) \sim \text{Dirichlet} \left( \frac{\alpha}{2}, \frac{\alpha}{2} \right)$

- Split each component according to the splitting rule:

  $\theta^{(2)}_1, \theta^{(2)}_2 \sim \text{Beta} \left( \frac{\alpha}{2}, \frac{1}{2}, \frac{\alpha}{2}, \frac{1}{2} \right)$

  $\pi^{(4)} = (\theta^{(2)}_1, \pi^{(2)}_1, (1 - \theta^{(2)}_1)\pi^{(2)}_1, \theta^{(2)}_2, \pi^{(2)}_2, (1 - \theta^{(2)}_2)\pi^{(2)}_2)$

  $\sim \text{Dirichlet} \left( \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4}, \frac{\alpha}{4} \right)$

- Repeat to get $\pi^{(K)} \sim \text{Dirichlet} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right)$

- As $K \to \infty$, we get a vector with infinitely many components

The Dirichlet process

- Let $H$ be a distribution on some space $\Omega$ – e.g. a Gaussian distribution on the real line.

- Let $\pi \sim \lim_{K \to \infty} \text{Dirichlet} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right)$

- For $k = 1, \ldots, \infty$ let $\theta_k \sim H$.

- Then $G := \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ is an infinite distribution over $H$.

- We write $G \sim \text{DP}(\alpha, H)$
Samples from the Dirichlet process

- Samples from the Dirichlet process are discrete.
- We call the point masses in the resulting distribution, atoms.
- The base measure $H$ determines the locations of the atoms.

Samples from the Dirichlet process

- The concentration parameter $\alpha$ determines the distribution over atom sizes.
- Small values of $\alpha$ give sparse distributions.
Properties of the Dirichlet process

- For any partition \( A_1, \ldots, A_K \) of \( \Omega \), the total mass assigned to each partition is distributed according to \( \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K)) \)

Definition: Finite marginals

- A Dirichlet process is the unique distribution over probability distributions on some space \( \Omega \), such that for any finite partition \( A_1, \ldots, A_K \) of \( \Omega \),

\[
(P(A_1), \ldots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K)).
\]

[Ferguson, 1973]
Conjugacy of the Dirichlet process

- Let $A_1, \ldots, A_K$ be a partition of $\Omega$, and let $H$ be a measure on $\Omega$. Let $P(A_k)$ be the mass assigned by $G \sim DP(\alpha, H)$ to partition $A_k$. Then $(P(A_1), \ldots, P(A_K)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$.

- If we see an observation in the $J^{th}$ segment, then
  $$(P(A_1), \ldots, P(A_j), \ldots, P(A_K)|X_1 \in A_j) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_j) + 1, \ldots, \alpha H(A_K)).$$

- This must be true for all possible partitions of $\Omega$.
- This is only possible if the posterior of $G$, given an observation $x$, is given by
  $$G|X_1 = x \sim DP\left(\alpha + 1, \frac{\alpha H + \delta_x}{\alpha + 1}\right).$$

Predictive distribution

- The Dirichlet process clusters observations.
- A new data point can either join an existing cluster, or start a new cluster.
- Question: What is the predictive distribution for a new data point?
- Assume $H$ is a continuous distribution on $\Omega$. This means for every point $\theta$ in $\Omega$, $H(\theta) = 0$.
- First data point:
  - Start a new cluster.
  - Sample a parameter $\theta_1$ for that cluster.
Predictive distribution

- We have now split our parameter space in two: the singleton \( \theta_1 \), and everything else.
- Let \( \pi_1 \) be the atom at \( \theta_1 \).
- The combined mass of all the other atoms is \( \pi^* = 1 - \pi_1 \).
- A priori, \((\pi_1, \pi^*) \sim \text{Dirichlet}(0, \alpha)\)
- A posteriori, \((\pi_1, \pi^*|X_1 = \theta_1 \sim \text{Dirichlet}(1, \alpha)\)

Predictive distribution

- If we integrate out \( \pi_1 \), we get

\[
P(X_2 = \theta_k | X_1 = \theta_1) = \int P(X_2 = \theta_k | (\pi_1, \pi^*)) P((\pi_1, \pi^*|X_1 = \theta_1) d\pi_1
\]

\[
= \int \pi_k \text{Dirichlet}(\pi_1, 1 - \pi_1 | 1, \alpha) d\pi_1
\]

\[
= \mathbb{E}_{\text{Dirichlet}(1, \alpha)}[\pi_k]
\]

\[
= \begin{cases} 
\frac{\alpha}{1 + \alpha} & \text{if } k = 1 \\
\frac{1 + \alpha}{1 + \alpha} & \text{for new } k.
\end{cases}
\]
Predictive distribution

- Let's say we choose to start a new cluster, and sample a new parameter $\theta_2 \sim H$. Let $\pi_2$ be the size of the atom at $\theta_2$.
- A posteriori, $(\pi_1, \pi_2, \pi_\ast) | X_1 = \theta_1, X_2 = \theta_2 \sim \text{Dirichlet}(1, \alpha)$.
- If we integrate out $\pi = (\pi_1, \pi_2, \pi_\ast)$ we get

$$P(X_3 = \theta_k | X_1 = \theta_1, X_2 = \theta_2)$$

$$= \int P(X_3 = \theta_k | \pi) P(\pi | X_1 = \theta_1, X_2 = \theta_2) d\pi$$

$$= \mathbb{E}_{\text{Dirichlet}(1, \alpha)}[\pi_k]$$

$$= \begin{cases} \frac{1}{2 + \alpha} & \text{if } k = 1 \\ \frac{1}{2 + \alpha} + \frac{\alpha}{n + \alpha} & \text{if } k = 2 \\ \frac{\alpha}{n + \alpha} & \text{for new } k. \end{cases}$$

Predictive distribution

- In general, if $m_k$ is the number of times we have seen $X_i = k$, and $K$ is the total number of observed values,

$$P(X_{n+1} = \theta_k | X_1, \ldots, X_n) = \int P(X_{n+1} = \theta_k | \pi) P(\pi | X_1, \ldots, X_n) d\pi$$

$$= \mathbb{E}_{\text{Dirichlet}(m_1, \ldots, m_K, \alpha)}[\pi_k]$$

$$= \begin{cases} \frac{m_k}{n + \alpha} & \text{if } k \leq K \\ \frac{\alpha}{n + \alpha} & \text{for new cluster.} \end{cases}$$

- We tend to see observations that we have seen before – rich-get-richer property.
- We can always add new features – nonparametric.
Polya urn scheme

- The resulting distribution over data points can be thought of using the following urn scheme.
- An urn initially contains a black ball of mass $\alpha$.
- For $n=1,2,…$ sample a ball from the urn with probability proportional to its mass.
- If the ball is black, choose a previously unseen color, record that color, and return the black ball plus a unit-mass ball of the new color to the urn.
- If the ball is not black, record its color and return it, plus another unit-mass ball of the same color, to the urn.

[Blackwell and MacQueen, 1973]

Chinese restaurant process

- The distribution over partitions can be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
Chinese restaurant process

- The distribution over partitions can be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
- The $n^{th}$ customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where $m_k$ is the number of people sat at table $k$. He starts a new table with probability $\alpha/(n-1+\alpha)$.
Chinese restaurant process

- The distribution over partitions can be described in terms of the following restaurant metaphor:
- The first customer enters a restaurant, and picks a table.
- The $n$th customer enters the restaurant. He sits at an existing table with probability $\frac{m_k}{n-1+\alpha}$, where $m_k$ is the number of people sat at table $k$. He starts a new table with probability $\frac{\alpha}{n-1+\alpha}$.

© A. Dubey, S. Williamson, E. Xing @CMU, 2014
Exchangeability

- An interesting fact: the distribution over the clustering of the first $N$ customers does not depend on the order in which they arrived.
- Homework: Prove to yourself that this is true.
- However, the customers are not independent – they tend to sit at popular tables.
- We say that distributions like this are exchangeable.
- De Finetti’s theorem: If a sequence of observations is exchangeable, there must exist a distribution given which they are iid.
- The customers in the CRP are iid given the underlying Dirichlet process – by integrating out the DP, they become dependent.

Stick breaking construction

- We can represent samples from the Dirichlet process exactly.
- Imagine a stick of length 1, representing total probability.
- For $k=1,2,...$
  - Sample a beta$(1,\alpha)$ random variable $b_k$.
  - Break off a fraction $b_k$ of the stick. This is the $k^{th}$ atom size.
  - Sample a random location for this atom.
  - Recurse on the remaining stick.

$$G := \sum_{k=1}^{\infty} \pi_k b_{\theta_k}$$

$$\pi_k := b_k \prod_{j=1}^{k-1} (1 - b_j)$$

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

[Sethuraman, 1994]
Inference in the DP mixture model

\[ G := \sum_{k=1}^{\infty} \pi_k \delta_{\phi_n} \sim \text{DP}(\alpha, H) \]
\[ \phi_n \sim G \]
\[ x_n \sim f(\phi_n) \]

Inference: Collapsed sampler

- We can integrate out \( G \) to get the CRP.
- Reminder: Observations in the CRP are exchangeable.
- Corollary: When sampling any data point, we can always rearrange the ordering so that it is the last data point.
- Let \( z_n \) be the cluster allocation of the \( n \)th data point.
- Let \( K \) be the total number of instantiated clusters.
- Then
  \[ p(z_n = k|x_n, z_{-n}, \phi_{1:K}) \propto \begin{cases} m_k f(x_n|\phi_k) & k \leq K \\ \alpha \int_{\Omega} f(x_n|\phi)H(d\phi) & k = K + 1 \end{cases} \]
- If we use a conjugate prior for the likelihood, we can often integrate out the cluster parameters.
Problems with the collapsed sampler

- We are only updating one data point at a time.
- Imagine two “true” clusters are merged into a single cluster – a single data point is unlikely to “break away”.
- Getting to the true distribution involves going through low probability states \( \rightarrow \) mixing can be slow.
- If the likelihood is not conjugate, integrating out parameter values for new features can be difficult.
- Neal [2000] offers a variety of algorithms.
- Alternative: Instantiate the latent measure.

Inference: Blocked Gibbs sampler

- Rather than integrate out \( G \), we can instantiate it.
- Problem: \( G \) is infinite-dimensional.
- Solution: Approximate it with a truncated stick-breaking process:

\[
G^K := \sum_{k=1}^{K} \pi_k \delta \theta_k
\]

\[
\pi_k = b_k \prod_{j=1}^{k-1} (1 - b_j)
\]

\[
b_k \sim \text{Beta}(1, \alpha), \quad k = 1, \ldots, K - 1
\]

\[
b_K = 1
\]
Inference: Blocked Gibbs sampler

- Sampling the cluster indicators:
  \[ p(z_n = k | \text{rest}) \propto \pi_k f(x_n | \theta_k) \]

- Sampling the stick breaking variables:
  - We can think of the stick breaking process as a sequence of binary decisions.
  - Choose \( z_n = 1 \) with probability \( b_1 \).
  - If \( z_n \neq 1 \), choose \( z_n = 2 \) with probability \( b_2 \).
  - etc..

  \[ b_k | \text{rest} \sim \text{Beta}(1 + m_k, \alpha + \sum_{j=k+1}^{K} m_j) \]

Inference: Slice sampler

- Problem with batch sampler: Fixed truncation introduces error.
- Idea:
  - Introduce random truncation.
  - If we marginalize over the random truncation, we recover the full model.
- Introduce a uniform random variable \( u_n \) for each data point.
- Sample indicator \( z_n \) according to
  \[ p(z_n = k | \text{rest}) = I(\pi_k > u_n) f(x_n | \theta_k) \]

- Only a finite number of possible values.
Inference: Slice sampler

- The conditional distribution for $u_n$ is just:
  $$u_n | \text{rest} \sim \text{Uniform}[0, \pi_{z_n}]$$

- Conditioned on the $u_n$ and the $z_n$, the $\pi_k$ can be sampled according to the block Gibbs sampler.

- Only need to represent a finite number $K$ of components such that
  $$1 - \sum_{k=1}^{K} \pi_k < \min(u_n)$$

Topic models

- Topic models describe documents using a distribution over features.
- Each feature is a distribution over words
- Each document is represented as a collection of words (usually unordered – “bag of words” assumption).
- The words within a document are distributed according to a document-specific mixture model
  - Each word in a document is associated with a feature.
- The features are shared between documents.
- The features learned tend to give high probability to semantically related words – “topics”
Latent Dirichlet allocation

- For each topic $k=1,\ldots,K$
  - Sample a distribution over words, $\beta \sim \text{Dir}(\eta_1, \ldots, \eta_V)$
- For each document $m=1,\ldots,M$
  - Sample a distribution over topics, $\theta_m \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$
  - For each word $n=1,\ldots,N_m$
    - Sample a topic $z_{mn} \sim \text{Discrete}(\theta_m)$
    - Sample a word $w_{mn} \sim \text{Discrete}(\beta_{z_{mn}})$

“Topics” found by LDA

---

3/21/2015
Constructing a topic model with infinitely many topics

- LDA: Each distribution is associated with a distribution over $K$ topics.
- Problem: How to choose the number of topics?
- Solution:
  - Infinitely many topics!
  - Replace the Dirichlet distribution over topics with a Dirichlet process!
- Problem: We want to make sure the topics are shared between documents

Sharing topics

- In LDA, we have $M$ independent samples from a Dirichlet distribution.
- The weights are different, but the topics are fixed to be the same.
- If we replace the Dirichlet distributions with Dirichlet processes, each atom of each Dirichlet process will pick a topic independently of the other topics.
Sharing topics

- Because the base measure is continuous, we have zero probability of picking the same topic twice.
- If we want to pick the same topic twice, we need to use a discrete base measure.
- For example, if we chose the base measure to be
  \[ H = \sum_{k=1}^{K} \alpha_k \delta_{\beta_k} \]  
  then we would have LDA again.
- We want there to be an infinite number of topics, so we want an infinite, discrete base measure.
- We want the location of the topics to be random, so we want an infinite, discrete, random base measure.

Hierarchical Dirichlet Process (Teh et al, 2006)

- Solution: Sample the base measure from a Dirichlet process!
  \[ G_0 \sim \text{DP}(\gamma, H) \]
  \[ G_{m} \sim \text{DP}(\alpha, G_0) \]

\[ G_0 \]
\[ G_1 \]
\[ G_2 \]
\[ X_1 \]
\[ X_2 \]
Chinese restaurant franchise

- Imagine a franchise of restaurants, serving an infinitely large, global menu.
- Each table in each restaurant orders a single dish.
- Let $n_r$ be the number of customers in restaurant $r$ sitting at table $t$.
- Let $m_{rd}$ be the number of tables in restaurant $r$ serving dish $d$.
- Let $m_d$ be the number of tables, across all restaurants, serving dish $d$.

Chinese restaurant franchise

- Customers enter the restaurants, and sit at tables according to the Chinese restaurant process
  - The first customer enters a restaurant, and picks a table.
Chinese restaurant franchise

- Customers enter the restaurants, and sit at tables according to the Chinese restaurant process
  - The first customer enters a restaurant, and picks a table.
  - The $n^{th}$ customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where $m_k$ is the number of people sat at table $k$. He starts a new table with probability $\alpha/(n-1+\alpha)$. 

© A. Dubey, S. Williamson, E. Xing @CMU,2014-15
Chinese restaurant franchise

- Customers enter the restaurants, and sit at tables according to the Chinese restaurant process
  - The first customer enters a restaurant, and picks a table.
  - The $n^{th}$ customer enters the restaurant. He sits at an existing table with probability $m_k/(n-1+\alpha)$, where $m_k$ is the number of people sat at table $k$. He starts a new table with probability $\alpha/(n-1+\alpha)$.
Chinese restaurant franchise

- Each table in each restaurant picks a dish, with probability proportional to the number of times it has been served across all restaurants.

\[
p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T + \gamma} & \text{for an existing table} \\ \frac{\gamma}{T + \gamma} & \text{for a new table} \end{cases}
\]
Chinese restaurant franchise

- Each table in each restaurant picks a dish, with probability proportional to the number of times it has been served across all restaurants.

\[
p(\text{table } t \text{ chooses dish } d | \text{previous tables}) = \begin{cases} \frac{m_d}{T+\gamma} & \text{for an existing table} \\ \frac{\gamma}{T+\gamma} & \text{for a new table} \end{cases}
\]
An infinite topic model

- Restaurants = documents; dishes = topics.

- Let $H$ be a $V$-dimensional Dirichlet distribution, so a sample from $H$ is a distribution over a vocabulary of $V$ words.

- Sample a global distribution over topics,
  \[ G_0 := \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k} \sim \text{DP}(\alpha, H) \]

- For each document $m=1,\ldots,M$
  - Sample a distribution over topics, $G_m \sim \text{DP}(\gamma, G_0)$.
  - For each word $n=1,\ldots,N_m$
    - Sample a topic $\phi_{mn} \sim \text{Discrete}(G_0)$.
    - Sample a word $w_{mk} \sim \text{Discrete}(\phi_{mn})$.

The “right” number of topics