CS598JHM: Advanced NLP (Spring 2013)
http://courses.engr.illinois.edu/cs598jhm/

Lecture 14: Inference in Dirichlet Processes

(Blei & Jordan, *Variational inference for Dirichlet Process Mixture models,* Bayesian Analysis 2006)

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center
Office hours: by appointment

Dirichlet Process mixture models

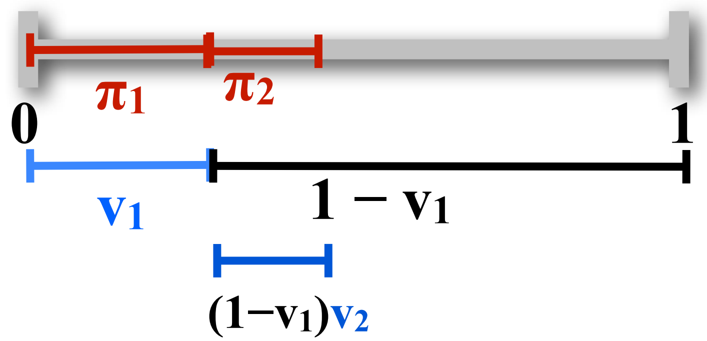
A mixture model with a DP as nonparametric prior:

‘Mixing weights’ (prior): $G \mid \{\alpha, G_0\} \sim DP(\alpha, G_0)$
The base distribution G_0 and G are distributions over the same probability space.

‘Cluster’ parameters: $\eta_n \mid G \sim G$
For each data point $n = 1,..., N$, draw a distribution η_n with value $\eta_c{*}$ over observations from G
(We can interpret this as clustering because G is discrete with probability 1; hence different η_n take on identical values $\eta_c{*}$ with nonzero probability. Data points are partitioned into $|\mathbf{C}|$ clusters: $\mathbf{c} = c_1...c_N$)

Observed data: $x_n|\eta_n \sim p(x_n \mid \eta_n)$
For each data point $n = 1,...,N$, draw observation x_n from η_n

Stick-breaking representation of DPMs



The component parameters $\boldsymbol{\eta}^*$: $\eta^*_i \sim G_0$

The mixing proportions $\pi_i(\mathbf{v})$ are defined by a stick-breaking process:

$$V_i \sim \mathit{Beta}(1, \alpha) \qquad \pi_i(\mathbf{v}) = v_i \prod_{j=1...i-1} (1-v_j)$$

also written as $\pi(\mathbf{v}) \sim \mathrm{GEM}(\alpha)$ (Griffiths/Engen/McCloskey)

Hence, if $G \sim \mathrm{DP}(\alpha, G_0)$:

$G = \sum_{i=1...\infty} \pi_i(\mathbf{v})\, \delta_{\eta i^*}$ with $\eta^*_i \sim G_0$

DP mixture models with $DP(\alpha, G_0)$

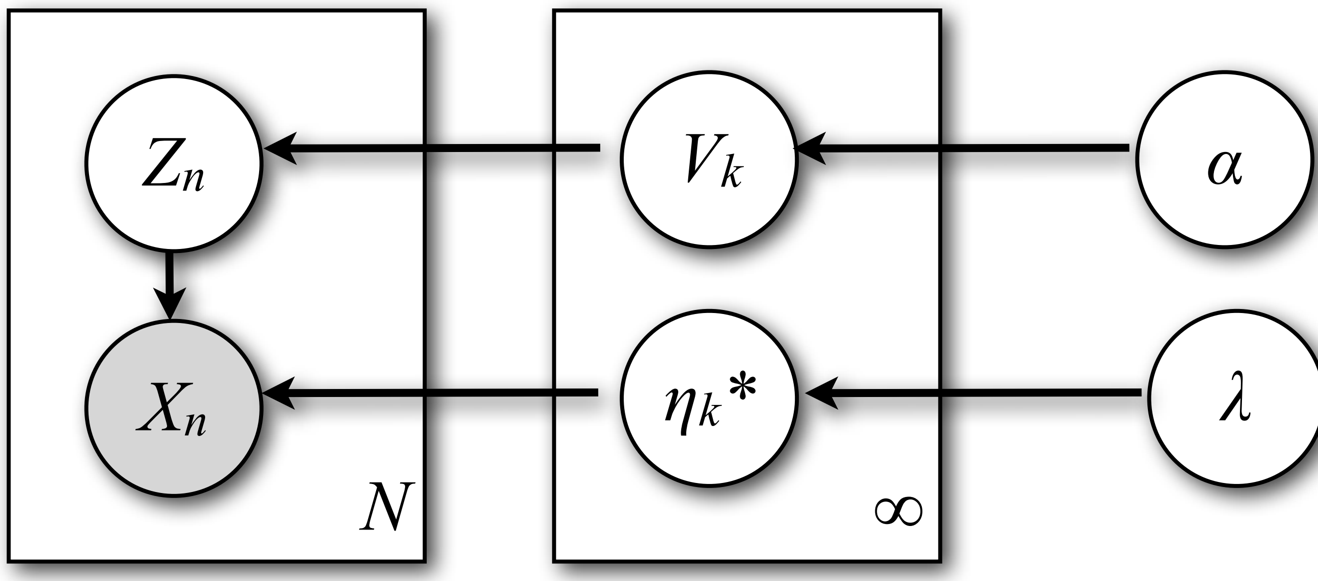
1. Define stick-breaking weights by drawing $V_i \mid \alpha \sim \text{Beta}(1, \alpha)$

2. Draw cluster $\eta_i^* \mid G_0 \sim G_0 \quad i = \{1, 2, ...\}$

3. For the nth data point:
   Draw cluster id $Z_n \mid \{v_1, v_2 ...\} \sim \text{Mult}(\pi(\mathbf{v}))$
   Draw observation $X_n \mid z_n \sim p(x \mid \eta_{z_n}^*)$

$p(x \mid \eta^*)$ is from an exponential family of distributions
G_0 is from the corresponding conjugate prior
e.g. $p(x \mid \eta^*)$ multinomial, G_0 Dirichlet

Stick-breaking construction of DPMs



Stick lengths $V_i \sim \text{Beta}(1, \alpha)$,
yielding **mixing weights** $\pi_i(\mathbf{v}) = v_i \prod_{j<i} (1 - v_j)$
Component parameters: $\eta_i^* \sim G_0$
(assume G_0 is conjugate prior with **hyperparameter** λ)
Assignment of data to components: $Z_n \mid \{v_1, ... \} \sim \text{Mult}(\pi(\mathbf{v}))$
Generating the observations: $X_n \mid z_n \sim p(x_n \mid \eta_{z_n}^*)$

Inference for DP mixture models

Given observed data $x_1, \ldots., x_n$, compute the **predictive density**:

$$p(x \mid x_1, \ldots., x_n, \alpha, G_0)$$
$$= \int p(x \mid \mathbf{w})\, p(\mathbf{w} \mid x_1, \ldots., x_n, \alpha, G_0)\, d\mathbf{w}$$

Problem: the posterior of the latent variables $p(\mathbf{w} \mid x_1, \ldots., x_n, \alpha, G_0)$ can't be computed in closed form

Approximate inference:

- **Gibbs sampling:**
  Sample from a Markov chain with equilibrium distribution $p(\mathbf{W} \mid x_1, \ldots., x_n, \alpha, G_0)$

- **Variational inference**:
  Construct a tractable variational approximation q of p with free variational parameters $\mathbf{v}$

Gibbs sampling

Gibbs sampling for DPMs

Two variants that differ in their definition of the Markov Chain

Collapsed Gibbs sampler:
Integrates out G and the distinct parameter values $\{\eta_1^* .... \eta_{|C|}^*\}$ associated with the clusters

Blocked Gibbs sampler:
Based on the stick-breaking construction.
This requires a truncated variant of the DP.

Collapsed Gibbs sampler for DPMs

Integrate out the random measure G and the distinct parameter values $\{\eta_1^* \dots \eta_{|C|}^*\}$ associated with each cluster

Given data $\mathbf{x} = x_1 \dots x_N$, each **state** of the Markov chain is a **cluster assignment** $\mathbf{c} = c_1 \dots c_N$ to each data point
Each **sample** is also a cluster assignment $\mathbf{c} = c_1 \dots c_N$

Given a cluster assignment $\mathbf{c}_b = c_1 \dots c_N$ with C distinct clusters, the **predictive density** is

$$p(x_{N+1} \mid \mathbf{c}_b, \mathbf{x}, \alpha, \lambda)$$
$$= \sum_{k \leq C+1} p(c_{N+1} = k \mid \mathbf{c}_b, \alpha)\, p(x_{N+1} \mid \mathbf{c}_b, c_{N+1} = k, \lambda)$$

Collapsed Gibbs sampler for DPMs

‘Macro-sample step’:
Assign a new cluster to all data points.

‘Micro-sample step’:
Sample assignment variables C_n for each data point conditioned on the assignment of the remaining points, $\mathbf{c}_{-n}$

C_n is either one of the values in $\mathbf{c}_{-n}$ or a new value:

$p(c_n = k \mid \mathbf{x}, \mathbf{c}_{-n}) \propto p(x_n \mid \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n{=}k, \lambda)\, p(c_n = k \mid \mathbf{c}_{-n}, \alpha)$

with $p(x_n \mid \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n{=}k, \lambda) = p(\mathbf{x}_n, \mathbf{c}_{-n}, c_n{=}k, \lambda) / p(\mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n{=}k, \lambda)$

and $p(c_n = k \mid \mathbf{c}_{-n}, \alpha)$ given by the Polya (Blackwell/McQueen) urn

Inference:
After burn-in, collect B sample assignments $\mathbf{c}_b$ and average across their predictive densities.

Blocked Gibbs sampling

Based on the stick-breaking construction.
States of the Markov chain consist of $(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})$

Problem: in the *actual* DPM model $\mathbf{V}, \boldsymbol{\eta}^*$ are infinite.

Instead, the blocked Gibbs sampler uses a *truncated DP* (TDP), which samples only a *finite* collection of T stick lengths (and hence clusters)

By setting $\mathrm{V}_{T\text{-}1} = 1$, $\boldsymbol{\pi}_i = 0$ for $i \geq T$:

$$\pi_i(\mathbf{v}) = \mathrm{v}_i \prod_{j<i} (1 - \mathrm{v}_j)$$

Blocked Gibbs sampling

The states of the Markov chain consist of

- the beta variables $\mathbf{V} = \{V_1 \ldots V_{T-1}\}$,
- the mixture component parameters $\boldsymbol{\eta}^* = \{\eta_1^* \ldots \eta_T^*\}$
- the indicator variables $\mathbf{Z} = \{Z_1 \ldots Z_N\}$

Sampling:

- For $n=1\ldots N$, sample Z_N from $p(z_n = k \mid \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \pi_k(\mathbf{v}) p(x_n \mid \eta_k^*)$
- For $k=1\ldots K$, sample V_k from $Beta(\gamma_{k2}, \gamma_{k1}\ \alpha + n_{k+1\ldots K})$

  $\gamma_{k1} = 1 + n_k$ with n_k : number of data points in cluster k

  $\gamma_{k2} = \alpha + n_{k+1\ldots K}$: with $n_{k+1\ldots K}$ the data points in clusters $k+1\ldots K$
- For $k=1\ldots K$, sample η_k^* from its posterior $p(\eta_k^* \mid \tau_k)$

  $\tau_k = (\lambda_1 + n_{-ik}(x_i)\ ,\ \lambda_2 + n_{-ik})$

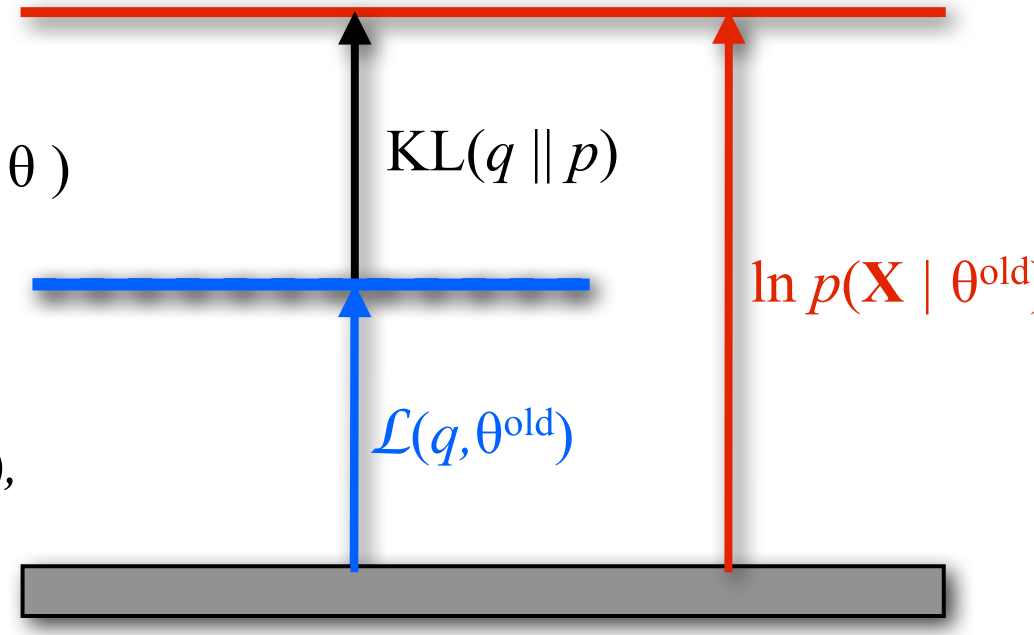
Predictive density for each sample:

$p(x_{n+1} \mid \mathbf{x}, \mathbf{z}, \alpha, \lambda) = \sum_k E[\pi_k(\mathbf{v}) \mid \gamma_1 \ldots \gamma_K]\ p(x_{n+1} \mid \tau_k)$

Variational inference (recap)

Standard EM

$\mathcal{L}(q, \theta) = \ln p(\mathbf{X} \mid \theta) - \mathrm{KL}(q \,||\, p)$
is a lower bound on the
incomplete log-likelihood $\ln p(\mathbf{X} \mid \theta)$



E-step:
With θ^{old} fixed, return q^{new}
that maximizes $\mathcal{L}(q, \theta^{old})$ wrt. $q(\mathbf{Z})$,
Now $\mathrm{KL}(q^{new} \,||\, p^{old}) = 0$.

M-step:
With q^{new} fixed, return θ^{new}
that maximizes $\mathcal{L}(q^{new}, \theta)$ wrt. θ.
If $\mathcal{L}(q^{new}, \theta^{new}) > \mathcal{L}(q^{new}, \theta^{old})$:
$\ln p(\mathbf{X} \mid \theta^{new}) > \ln p(\mathbf{X} \mid \theta^{old})$,
and hence $\mathrm{KL}(q^{new} \,||\, p^{new}) > 0$

Variational inference

Variational inference is applicable when you have to compute an *intractable* posterior over latent variables $p(\mathbf{W} \mid \mathbf{X})$

Basic idea:
Replace the exact, but intractable posterior $p(\mathbf{W} \mid \mathbf{X})$
with a ***tractable*** **approximate posterior** $q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$

$q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$ is from a family of simpler distributions over the latent variables $\mathbf{W}$ that is defined by a set of **free variational parameters** $\mathbf{V}$

Unlike in EM, $\mathrm{KL}(q \,\|\, p) > 0$ for any q, since q only approximates p

Variational EM

Initialization:
Define initial model θ^{old} and variational distribution $q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$

E-step:
Find $\mathbf{V}$ that maximize the variational distribution $q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$
Compute the expectation of true posterior $p(\mathbf{W} \mid \mathbf{X}, \theta^{old})$ under the new variational distribution $q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$

M-step:
Find model parameters θ^{new} that maximize the expectation of the $p(\mathbf{W}, \mathbf{X} \mid \theta)$ under the variational posterior $q(\mathbf{W} \mid \mathbf{X}, \mathbf{V})$

Set $\theta^{old} := \theta^{new}$

Blei and Jordan's mean-field variational inference for DP

Variational inference

Define a family of variational distributions $q_{\mathbf{v}}(\mathbf{w})$ with variational parameters $\boldsymbol{v} = v_1 \ldots . v_M$ that are specific to each observation x_i

Set $\boldsymbol{v}$ to minimze the KL-divergence between $q_{\mathbf{v}}(\mathbf{w})$ and $p(\mathbf{w} \mid \mathbf{x}, \theta)$:

$$D(q_v(\mathbf{w}) \| p(\mathbf{w} \mid \mathbf{x}, \theta)) = E_q[\log q_v(\mathbf{W})] - E_q[\log p(\mathbf{W}, \mathbf{x} \mid \theta)] + \log p(\mathbf{x} \mid \theta)$$

(Here, $\log p(\mathbf{x} \mid \theta)$ can be ignored when finding q)

This is equivalent to maximizing a lower bound on $\log p(\mathbf{x} \mid \theta)$:

$$\log p(\mathbf{x} \mid \theta) = E_q[\log p(\mathbf{W}, \mathbf{x} \mid \theta)] - E_q[\log q_v(\mathbf{W})] + D(q_v(\mathbf{w}) \| p(\mathbf{w} \mid \mathbf{x}, \theta))$$
$$\log p(\mathbf{x} \mid \theta) \geq E_q[\log p(\mathbf{W}, \mathbf{x} \mid \theta)] - E_q[\log q_v(\mathbf{W})]$$

$q_v(\mathbf{W})$ for DPMs

Blei and Jordan use again the stick-breaking construction.

Hence, the latent variables are $\mathbf{W} = (\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})$

- $\mathbf{V}$: $T-1$ truncated stick lengths
- $\boldsymbol{\eta}^*$: T component parameters
- $\mathbf{Z}$: cluster assignments of the N data points

Variational inference for DPMs

In general:

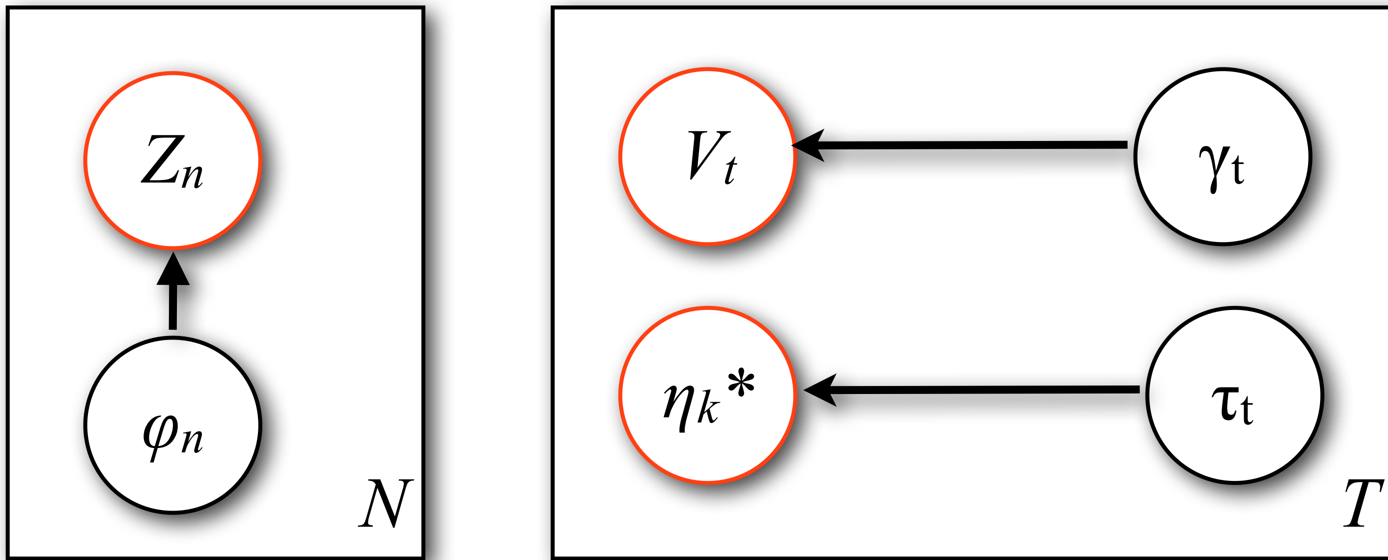
$$\log p(\mathbf{x} \mid \theta) \geq E_q [\log p(\mathbf{W}, \mathbf{x} \mid \theta)] - E_q [\log q_\nu(\mathbf{W})]$$

For DPMs: $\theta = (\alpha, \lambda)$; $\mathbf{W} = (\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})$

$$\log p(\mathbf{x} \mid \alpha, \lambda) \geq E_q [\log p(\mathbf{V} \mid \alpha)] + E_q [\log p(\boldsymbol{\eta}^* \mid \lambda)] + \sum_n [E_q[\log p(Z_n \mid \mathbf{V})] + E_q[\log p(x_n \mid Z_n)]] - E_q [\log q_\nu(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})]$$

Problem: $\mathbf{V} = \{V_1, V_2, \ldots\}$, $\boldsymbol{\eta}^* = \{\eta_1^*, \eta_2^*, \ldots\}$ are infinite.
Solution: use a truncated representation

Variational approximations $q_{\mathbf{v}}(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z})$



The variational parameters $\mathbf{v} = (\boldsymbol{\gamma}_{1..T-1}, \boldsymbol{\tau}_{1..T}, \boldsymbol{\varphi}_{1...N})$

$$q_{\mathbf{v}}(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z}) = \prod_{t<T} q_{\gamma t}(\mathrm{v}_t) \prod_{t<T} q_{\tau t}(\eta_t^*) \prod_{n \leq N} q_{\varphi_n}(\mathrm{z}_n)$$

$q_{\gamma t}(\mathrm{v}_t)$: Beta distributions with variational parameter γ_t

$q_{\tau t}(\eta_t^*)$: conjugate priors for η, with parameter τ_t

$q_{\varphi_n}(\mathrm{z}_n)$: multinomials with variational parameters φ_n