Multi-level clustering with contexts via hierarchical nonparametric Bayesian inference

Long Nguyen

Department of Statistics
University of Michigan

Biostatistics Seminar, October 2016
Multi-level clustering analysis

I’d like to ...

- cluster the collection of documents into meaningful topics
- cluster images into meaningful categories
- cluster the users into typical profiles based on recorded activities
Multi-level clustering analysis

I’d like to ...

- cluster the collection of documents into meaningful topics
- cluster images into meaningful categories
- cluster the users into typical profiles based on recorded activities

I also want to exploit contextual information that may be available
Topic modeling
Topic modeling a popular tool for mining and analyzing patterns from texts in news articles, scientific papers, blogs, but also tweets, query logs, digital books, metadata records...
Topic modeling a popular tool for mining and analyzing patterns from texts in news articles, scientific papers, blogs, but also tweets, query logs, digital books, metadata records...

![News articles](image1.png) ![Scientific papers](image2.png) ![Tweets](image3.png)

News articles  Scientific papers  Tweets

also applicable to ther data formats (images, networks)
Topic modeling is a popular tool for mining and analyzing patterns from texts in news articles, scientific papers, blogs, but also tweets, query logs, digital books, metadata records...

also applicable to other data formats (images, networks)

in diverse domains in computer sciences, biomedical sciences, scientometrics, social and political science, and digital humanities.
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Take a document from the AP corpus (Blei, Ng, Jordan, 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

after feeding to Latent Dirichlet Allocation (LDA) model:

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>
This talk: modeling both content and context

"content data": words/documents/images
Modeling both content and context

“content data”: words/documents/images
“context data”: time, location, hashtags, etc
Goal: jointly discover clusters of contents and contexts, e.g., words and time/locations

Probabilistic modeling for jointly model both contents and document contexts Bayesian nonparametric approach

Multiple advantages:

- context-aware topic modeling of contents
- context clusters share content topics
- infer context given content and vice-versa
Mixture models
Mixture modeling

Mixture density:

\[ p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i) \]

\[ G = \sum_{i=1}^{k} p_i \delta_{\theta_i} \] is mixing measure
Mixture modeling

Mixture density:

\[ p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i) \]

\[ G = \sum_{i=1}^{k} p_i \delta_{\theta_i} \] is mixing measure

Nonparametric Bayesian inference

\[ G \sim \Pi, \]

\[ x_1, \ldots, x_n|G \sim p_G \]

Clusters are drawn from posterior distribution \( \Pi(G|x_1, \ldots, x_n) \)
Dirichlet process prior $G \sim \mathcal{D}_\alpha G_0$

(Ferguson, 1973)

- $\mathcal{D}_\alpha G_0$ (also, $\text{DP}(\alpha, G_0)$): Dirichlet distribution on the space of probability measure on $\Theta$

- $G$ is called a Dirichlet process (a random PM on $\Theta$)

- $G$ is discrete with probability one, and admit Sethuraman’s stick-breaking representation

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\eta_i},$$

where both $\pi_i$s and $\eta_i$s are random variables obeying suitable laws
Dirichlet process mixture


\[ G \sim D_{\alpha H} \]

\[ \theta_i | G \overset{iid}{\sim} G \]

\[ x_i | \theta_i \overset{indep}{\sim} f(\cdot | \theta_i) \]
Multi-level analysis

Data are naturally organized as a multi-level collection of data sets

- text corpus as collection of documents, document as collection of words
- image db as collection of images, image as collection of patches
- collection of users, user as collection of activities
Exchangeable collection of data sets

Each data set is a collection of exchangeable elements
⇒ mixture of mixture of distributions

This gives rise naturally to a hierarchical model

[courtesy M. Jordan’s slides]
Hierarchical Dirichlet Processes (HDP)

(Teh, Jordan, Blei and Beal, JASA 2006)

\[ G \sim D_{\gamma H} \]

\[ Q_1, \ldots, Q_m | G \overset{iid}{\sim} D_{\alpha G} \]

\[ Y_{i1}, \ldots, Y_{in} | Q_i \overset{iid}{\sim} p_{Q_i} \text{ for } i = 1, \ldots, m \]
Back to earth: topic modeling for documents

Documents: Bags of words

- PLSA (Hofmann et al. 1999)
- LDA (Blei, et al. 2003)
- HDP (Teh et al. 2005)

Topics: Multinomial distribution over words

- information 0.16
- retrieval 0.08
- search 0.07
- machine 0.16
- learning 0.08
- classifier 0.07
- data 0.16
- mining 0.08
- knowledge 0.07
- web 0.16
- semantic 0.08
- content 0.07
The hierarchical model from the previous slide is this:

- \( p_\beta \): Distribution of topics
- \( \beta_j \): Distribution of words within topics
- \( z_{il} \): Topic index
- \( w_{il} \): Observed word
- \( \theta_j \): Document topic proportions
- \( \pi \): Prior over topics

\( w_{il} \): (observed) word \( i \) in document \( i \)

\( z_{il} \): (latent) topic index that word \( w_{il} \) is associated with
Latent Dirichlet allocation model

Generative process:

- For each $j = 1, \ldots, k$, sample a vector of frequencies $\theta_j \in \Delta^{d-1}$
  - these are called “topics”, distributed by a Dirichlet
  - $d =$ vocabulary size
Latent Dirichlet allocation model

Generative process:

- For each \( j = 1, \ldots, k \), sample a vector of frequencies \( \theta_j \in \Delta^{d-1} \)
  - these are called “topics”, distributed by a Dirichlet
  - \( d \) = vocabulary size

- For each document \( i = 1, \ldots, m \),
  - sample a topic proportion \( \beta \in \Delta^{k-1} \) (e.g., another Dirichlet)
  - for each word position in document \( i \)
    - sample a topic label \( z \sim \text{Multinomial}(\beta) \);
    - given \( z \), sample a word \( w \sim \text{Multinomial}(\theta_z) \).

Inferential goal: given data of size \( m \times n \), estimate the topic vectors \( \theta_j \)'s
Feeding AP corpus of documents, e.g.:

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Feeding AP corpus of documents, e.g.:

The **William Randolph Hearst Foundation** will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

---

to LDA/HDP model, we obtain

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>
Back to our work

- HDP does not help us cluster documents (yet)
- nor does it help us handle contextual information (time/location/hashtags)
- since documents is associated with distribution over words, we need to be able to cluster over the space of distributions!
Clustering in space of distributions
Nested Dirichlet processes

[Rodriguez, Dunson and Gelfand, JASA 2008]

\[ Q_1, \ldots, Q_m \mid G \sim \text{iid } G, \]

\[ \text{DP}(\nu Q_0) \]

\[ \eta \]

\[ \alpha \]
Nested Dirichlet processes

[Rodriguez, Dunson and Gelfand, JASA 2008]

$$Q_1, \ldots, Q_m | G \overset{\text{iid}}{\sim} G,$$

where

$$G \sim D_{\alpha D_{\nu Q_0}}$$
Nested Dirichlet processes

[Rodriguez, Dunson and Gelfand, JASA 2008]

\[ S \]

\[ Q_0 \]

\[ \text{DP}(vQ_0) \]

\[ \alpha \]

\[ G \]

\[ \eta \]

\[ Q_j \]

\[ \varphi_{jj} \]

\[ N_j \]

\[ J \]

\[ Q_1, \ldots, Q_m | G \overset{iid}{\sim} G, \]

where

\[ G \sim \mathcal{D}_{\alpha \mathcal{D}_{vQ_0}} \]

E.g., \( Q_0 \) is a distribution over a space of atoms (words/image patches/human activities)
### HDP vs NDP

<table>
<thead>
<tr>
<th>HDP</th>
<th>NDP</th>
</tr>
</thead>
</table>
| \( G_j \sim \text{DP}(\alpha G_0) \)  
\( G_0 \sim \text{DP}(\beta H) \) | \( G_j \sim Q \)  
\( Q \sim \text{DP}(\alpha \text{DP}(\beta H)) \) |

- \( G_0 \)
  - ![Graph](image1)
- \( G_1 \)
  - ![Graph](image2)
- \( G_2 \)
  - ![Graph](image3)
- \( G_3 \)
  - ![Graph](image4)
  - \( \ldots \)
- \( G_1^* \)
  - ![Graph](image5)
- \( G_2^* \)
  - ![Graph](image6)
- \( G_3^* \)
  - ![Graph](image7)
- \( G_4^* \)
  - ![Graph](image8)

(Rodriguez et al, 2008)
Multi-level clustering with contexts: MC2
Multi-level clustering with contexts: MC2

(Nguyen et al, ICML, 2014; Huynh et al, UAI, 2016)

- pairing up context (document-level) with content (word-level) is unnatural since they lie on different levels of abstraction

- first idea: treat context as index for distributions over contents
  - but, raw contextual data are noisy (e.g., noisy tags, continuous location coordinates)
Multi-level clustering with contexts: MC2

(Nguyen et al, ICML, 2014; Huynh et al, UAI, 2016)

- pairing up context (document-level) with content (word-level) is unnatural since they lie on different levels of abstraction
- first idea: treat context as index for distributions over contents
  - but, raw contextual data are noisy (e.g., noisy tags, continuous location coordinates)
- second idea: make context indices random
  - context cluster acts as an index into a distribution of contents
  - this allows context (time/space) to influence both topics and document clusters.
- how to make this concrete?
Pairing up context atoms $\theta_i$ with content distributions $Q_j$:

$$(\theta_j, Q_j) | U \sim U,$$

where

$$U \sim \mathcal{D}_\gamma(H \times \mathcal{D}_{vQ_0})$$
form a product of base measure $H \times D_{vQ_0}$

use this as base measure in a nested DP fashion

$$U \sim D_{\gamma(H \times D_{vQ_0})}$$

marginalizing out content yields a DP mixture over context data

marginalizing out context yields a nested DP mixture over content
MC2 grounded in stick-breaking representation

(a) Generative view.

(b) Stick-breaking view.
Gibbs sampling for MC2

- **Sampling $z_j$**
  
  $p(z_j = k | \cdot) \propto p(z_j = k | z_{-j}, \alpha) \times p(x_j | z_j = k, z_{-j}, x_{-j}, H) \times p(l_{ji}^* | z_j = k, l_{-ji}^*, z_j, \epsilon, v)$

- **Sampling $l_{ji}$**
  
  $p(l_{ji} = m | \cdot) \propto p(w_{ji} | l, w_{-ji}, S) \times p(l_{ji} = m | l_{-ji}, z_j = k, z_{-j}, \epsilon, v)$

- **Sampling $\epsilon$**
  
  - $p(o_{km} = h | \cdot) \propto \text{Stirl}(h, n_{km}) (v \epsilon_m)^h, \quad h = 0, 1, \ldots, n_{km}$
  - $p(\epsilon | \cdot) \propto \epsilon_{\text{new}}^{\eta - 1} \prod_{m=1}^{M} \epsilon_m \sum_k o_{km}^{\eta - 1}$
Application 1: document modeling

- PNAS dataset
  - 79,800 documents (only titles and timestamps)
  - Vocabulary size is 36,782 (remove stop words)
  - Context observations are document timestamps (1915–2005)

- NIPS abstract dataset
  - 1740 documents; vocabulary size: 13,649 words
  - Three types of context information: timestamps, authors (2037 unique authors), article titles
### Perplexity (goodness of fit)

<table>
<thead>
<tr>
<th>Method</th>
<th>Perplexity (on words only)</th>
<th>Feature used</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PNAS</td>
<td>(K,M)</td>
</tr>
<tr>
<td>HDP (Teh et al., 2006b)</td>
<td>3027.5</td>
<td>(−, 86)</td>
</tr>
<tr>
<td>npTOT (Dubey et al., 2012; Phung et al., 2012)</td>
<td>2491.5</td>
<td>(−, 145)</td>
</tr>
<tr>
<td>MC² without context</td>
<td>1742.6</td>
<td>(40, 126)</td>
</tr>
<tr>
<td>MC² with titles</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>MC² with authors</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>MC² with timestamp</td>
<td><strong>895.3</strong></td>
<td>(12, 117)</td>
</tr>
</tbody>
</table>
Context-aware topics

![Graph showing the conditional distribution of timestamp context on word topic, with keywords like albinism, elution, losses, photoproducts, rubrum, anaplasma, coxsackievirus, did, distinguished, and don. The graph compares Google Scholar output to the authors' results, with a manual count statistic of the keyword “albinism” in Google Scholar.]
Context-aware topics

J. M. Jordan, Z. Ghahramani, T. Jaakkola, D. Cohn, D. Wolpert, M. Meila

- On the use of evidence in neural networks [1993]
- Supervised Learning from Incomplete Data via an EM [1994]
- Fast Learning by Bounding Likelihoods in ... Networks [1996]
- Factorial Hidden Markov Models [1997]
- Estimating Dependency Structure as a Hidden Variable [1998]
- Maximum Entropy Discrimination [1999]

Top word topics conditional on the author context:

- recognition
- hidden likelihood
- trained
- word
- data
- classifier
- propagation
- net
- em

- context
- recognition
- probability
- state
- images
- models
- clustering
- hmm
- mlp

- time
- methods
- approximation
- step
- learning
- update
- bound
- convergence
- bayesian
- input
Application 2: image clustering

Purity

Rand–Index

NMI

Fscore
Application 2: image clustering

t-SNE clustering projection on NUS-WIDE dataset
## Application 2: image clustering

<table>
<thead>
<tr>
<th>Missing(%)</th>
<th>Purity</th>
<th>NMI</th>
<th>RI</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0.407</td>
<td>0.298</td>
<td>0.901</td>
<td>0.157</td>
</tr>
<tr>
<td>25%</td>
<td>0.338</td>
<td>0.245</td>
<td>0.892</td>
<td>0.149</td>
</tr>
<tr>
<td>50%</td>
<td>0.32</td>
<td>0.236</td>
<td>0.883</td>
<td>0.137</td>
</tr>
<tr>
<td>75%</td>
<td>0.313</td>
<td>0.187</td>
<td>0.860</td>
<td>0.112</td>
</tr>
<tr>
<td>100%</td>
<td>0.306</td>
<td>0.188</td>
<td>0.867</td>
<td>0.119</td>
</tr>
</tbody>
</table>
Scaling up
Scaling up

- Wikipedia: 1.1 million documents from wikipedia.com  
  context: first author and top-level categories

- PubMed: 1.4 million documents from pubmed.gov  
  context: medical subject headings (MeSH)

- AUA (application user activities): > 1M users  
  context: background softwares
Stochastic mean-field approximation

- factorized posterior distribution into that of local and global variables
- gradient-based update for local variables via structured mean-field approximation (can be parallelized)
- update for global variables using natural gradient and via stochastic optimization
• Not possible to fit via a Gibbs sampler

• Run times on 8-node SPARK cluster

• Stochastic mean-field approximation take, resp., 17 hours, 18.5 hours, and 18 hours
Scaling up

<table>
<thead>
<tr>
<th>Context availability</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td></td>
</tr>
<tr>
<td><strong>Wikipedia - writer</strong></td>
<td><strong>2,167</strong></td>
</tr>
<tr>
<td><strong>Pubmed - MeSH</strong></td>
<td><strong>2,294</strong></td>
</tr>
<tr>
<td><strong>AUA - other products</strong></td>
<td><strong>142.3</strong></td>
</tr>
</tbody>
</table>

Table 2: Log perplexity of Wikipedia and PubMed data
Identifiability and posterior contraction
What is going on in the layers of latent variables?

![Diagram showing layers of latent variables with nodes labeled S, η, Q₀, v, H, γ, U, θ_j × Q_j, x_j, w_{ji}, N_j, and φ_{ji} connected with arrows representing the flow of information.](image)
What is going on in the layers of latent variables?

Battleship USS Texas
Suppose

\[ X_1, \ldots, X_n \overset{iid}{\sim} p_G(x) := \int f(x|\theta)G(d\theta) \]

\( f \) is known, while \( G = G_0 \) unknown discrete mixing measure

- **Consistency:** does the posterior distribution \( \Pi(G|X_1, \ldots, X_n) \) concentrate most of its mass around the “truth” \( G_0 \)?

- **Rate:** what is the rate of concentration (convergence) as \( n \to \infty \)?
Optimal transport distance
Optimal transportation problem (Monge-Kantorovich)

how to move the mass from one distribution to another?

Originally: how to transport goods from a collection of producers to a collection of consumers located in a common space

squares: locations of producers; circles: locations of consumers

The optimal cost of transportation defines a distance from “production density” — to — “consumption density”.

Wasserstein distance

Let $G, G'$ be two prob. measures on $\Theta$

A coupling $\kappa$ of $G, G'$ is a joint dist on $\Theta \times \Theta$ which induces marginals $G, G'$

**Definition**

Let $\rho$ be a distance function on $\Theta$, the Wasserstein distance is defined by:

$$d_\rho(G, G') = \inf_\kappa \int \rho(\theta, \theta')d\kappa.$$
Wasserstein distance

Let $G, G'$ be two prob. measures on $\Theta$

A coupling $\kappa$ of $G, G'$ is a joint dist on $\Theta \times \Theta$ which induces marginals $G, G'$

**Definition**

Let $\rho$ be a distance function on $\Theta$, the Wasserstein distance is defined by:

$$d_\rho(G, G') = \inf_{\kappa} \int \rho(\theta, \theta') d\kappa.$$  

When $\Theta = \mathbb{R}^d$, for $r \geq 1$, we obtain $L_r$ Wasserstein metric:

$$W_r(G, G') := \left[ \inf_{\kappa} \int \|\theta - \theta'\|^r d\kappa \right]^{1/r}.$$
Examples and Facts

Wasserstein distance $W_r$ metrizes weak convergence in the space of probability measures on $\Theta$. 

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$. 

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, then $W_1(G_0, G) = k \sum_{i=1}^k p_i \|\theta_0 - \theta_i\|$. 

If $G = \sum_{i=1}^k \delta_{\theta_i}$ and $G' = \sum_{j=1}^k \delta_{\theta'_j}$, then $W_1(G, G') = \inf_{\pi} \sum_{i=1}^k p_i \|\theta_i - \theta'_{\pi(i)}\|$, where $\pi$ ranges over the set of permutations on $(1, \ldots, k)$. 

Examples and Facts

Wasserstein distance $W_r$ metrizes weak convergence in the space of probability measures on $\Theta$.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$. 
Examples and Facts

Wasserstein distance $W_r$ metrizes weak convergence in the space of probability measures on $\Theta$.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^k p_i \|\theta_0 - \theta_i\|.$$
Examples and Facts

Wasserstein distance $W_r$ metrizes weak convergence in the space of probability measures on $\Theta$.

If $\Theta = \mathbb{R}$, then $W_1(G, G') = \|CDF(G) - CDF(G')\|_1$.

If $G_0 = \delta_{\theta_0}$ and $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, then

$$W_1(G_0, G) = \sum_{i=1}^k p_i \|\theta_0 - \theta_i\|.$$ 

If $G = \sum_{i=1}^k \frac{1}{k} \delta_{\theta_i}$, $G' = \sum_{j=1}^k \frac{1}{k} \delta_{\theta_j}$, then

$$W_1(G, G') = \inf_{\pi} \sum_{i=1}^k \frac{1}{k} \|\theta_i - \theta'_{\pi(i)}\|,$$

where $\pi$ ranges over the set of permutations on $(1, \ldots, k)$. 
Finite mixtures

(Nguyen, AOS 2013; Ho & Nguyen, EJS 2016)

For strongly identifiable mixture models the posterior \( \Pi(G|X_1, \ldots, X_n) \) contracts to true \( G_0 \) at the rate \( \epsilon_n \),

\[
\Pi(W_r(G, G_0) \leq \epsilon_n|X_1, \ldots, X_n) \xrightarrow{P} 1
\]

- if the number of mixing components known, \( \epsilon_n \asymp n^{-1/2} \) under \( W_1 \)
- if only an upper bound of the number of mixing component is known, \( \epsilon_n \asymp n^{-1/4} \) under \( W_2 \)

Strongly identifiable finite mixtures:
- location Gaussian mixtures, scale Gaussian mixtures, etc
Infinite mixtures

For infinite mixtures using Dirichlet process prior on a compact Euclidean space, the posterior $\Pi(G|X_1, \ldots, X_n)$ contracts to true $G_0$ at the rate $\epsilon_n$, 

$$\Pi(W_2(G, G_0) \leq \epsilon_n|X_1, \ldots, X_n) \overset{P}{\longrightarrow} 1$$

- if the mixture’s kernel is “ordinary smooth” (e.g., Laplace), then $\epsilon_n \asymp n^{-1/(4+\beta)}$, where $\delta$ is determined by the smoothness parameter
- if the mixture’s kernel is “supersmooth” (e.g., Gaussian), then $\epsilon_n \asymp (\log n)^{-1/\beta}$
Weakly identifiably models

(Ho & Nguyen, AOS 2016)

location-scale and finite Gaussian mixtures

The posterior of $G$ contracts very slowly, as the number of extra number of mixing components

- $n^{-1/8}$ if overfitting by one
- $n^{-1/12}$ if overfitting by two
- and so on
Weakly identifiably models

location-scale and finite Gaussian mixtures

The posterior of $G$ contracts very slowly, as the number of extra number of mixing components

- $n^{-1/8}$ if overfitting by one
- $n^{-1/12}$ if overfitting by two
- and so on

There is a more general theory behind this phenomenon based on the singularity structures of the mixture model’s parameter space
Posterior contraction in hierarchical models
Distance between nonparametric Bayesian hierarchies

Need a notion of distance between, say $\mathcal{D}_{\alpha G}$ and $\mathcal{D}_{\alpha' G'}$

Recall: for $G, G' \in \mathcal{P}(\Theta)$, space of Borel probability measures on $\Theta$,

$$W_r(G, G') := \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

$\mathcal{T}(G, G')$ is the space of all couplings of $G, G'$. 
Distance between nonparametric Bayesian hierarchies

Need a notion of distance between, say \( D_{\alpha G} \) and \( D_{\alpha' G'} \)

Recall: for \( G, G' \in \mathcal{P}(\Theta) \), space of Borel probability measures on \( \Theta \),

\[
W_r(G, G') := \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.
\]

\( \mathcal{T}(G, G') \) is the space of all couplings of \( G, G' \).

Distance between measures of measures in Bayesian hierarchy:

Let \( \mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta)) \) (the space of Borel probability measures on \( \mathcal{P}(\Theta) \)). Define Wasserstein distance between \( \mathcal{D}, \mathcal{D}' \)

\[
W_r(\mathcal{D}, \mathcal{D}') := \inf_{\kappa \in \mathcal{T}(\mathcal{D}, \mathcal{D}')} \left[ \int W_r(G, G') d\kappa(G, G') \right]^{1/r}.
\]

\( \mathcal{T}(\mathcal{D}, \mathcal{D}') \) is the space of all couplings of \( \mathcal{D}, \mathcal{D}' \in \mathcal{P}(\mathcal{P}(\Theta)) \).
Hierarchical Dirichlet processes

(Nguyen, Bernoulli 2016)

- rates of posterior contraction of the Dirichlet base measure residing at the top of the latent hierarchy
- there is a striking effect of “borrowing of strength” phenomenon, which can be quantified
  - parameteric rate of contraction can be achieved at individual group-level distributions if there are sufficiently many groups supported by data residing in the same level of the model’s hierarchy
Summary

- MC2: nonparametric Bayesian modeling for joint context/content inference
- scaling up via stochastic variational inference and parallel computing
- posterior contraction behavior of latent variables