Hierarchical Topic Models

Andrew Leverentz

Research Examination, Fall Quarter 2017
UC San Diego, Dept. of Computer Science and Engineering

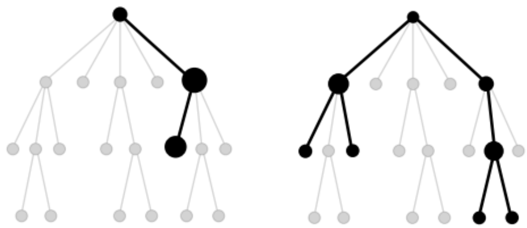


Image Source: Paisley et al [16]

Context

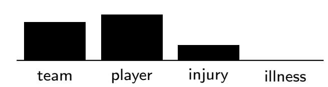
- Internet and digital archives → large collections of text data
- How can we navigate these collections efficiently?
- Typical task: find sets of documents that share the same topic or subject matter
- Natural language can be both redundant and ambiguous
- Superficial attributes of documents aren't enough
- We need a notion of *latent semantics*, or underlying meaning

Goals

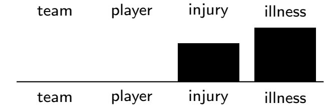
- General approach: documents are mixtures of topics, which are distributions over the vocabulary
- Probability provides a natural framework for this
- Topics can exist at different levels of abstraction (e.g., *baseball* and *basketball* are distinct subtopics under *sports*)
- Can we learn a hierarchy of topics based on a particular corpus?
- Similar to the Dewey Decimal System or Library of Congress Classification

Example: Documents as Mixtures of Topics

Topics: θ_1 = "sports"

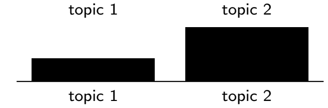


θ_2 = "medicine"



Documents: ϕ_1 = "document 1"

ϕ_2 = "document 2"

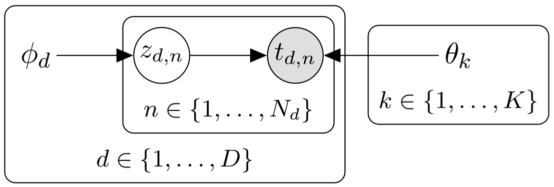


"Flat" Topic Models

Probabilistic Latent Semantic Analysis

- Idea: frequencies of words in documents determined by probabilities
- There are K latent topics, and each θ_k is a distribution over words in the vocabulary
- For document d, the vector ϕ_d is a distribution over topics
- For the n^{th} word in document d:

$$\begin{aligned} \text{Select a topic:} \quad & z_{d,n} \sim \text{Categorical}(\phi_d) \\ \text{Select a word:} \quad & t_{d,n} \sim \text{Categorical}(\theta_{z_{d,n}}) \end{aligned}$$



- Infer values of θ_k, ϕ_d using maximum likelihood

Latent Dirichlet Allocation

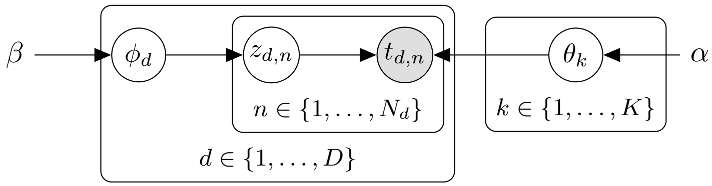
- Extension to PLSA: assume topic mixtures ϕ_d and topic vectors θ_k are drawn from Dirichlet distributions
- Dirichlet is a distribution over discrete probability distributions; density over simplex:

$$\text{Dirichlet}(\vec{x} \mid \vec{\alpha}) \propto \prod_{k=1}^{\text{len}(\vec{\alpha})} x_i^{\alpha_i - 1}$$

- Dirichlet distribution acts as a *regularizer*, reduces overfitting
- Allows Bayesian posterior inference

Latent Dirichlet Allocation: The Model

$$\begin{aligned} \theta_k &\sim \text{Dirichlet}(\alpha) && \text{for each topic } k \\ \phi_d &\sim \text{Dirichlet}(\beta) && \text{for each document } d \\ z_{d,n} &\sim \text{Categorical}(\phi_d) && \text{for the } n^{\text{th}} \text{ word in document } d \\ t_{d,n} &\sim \text{Categorical}(\theta_{z_{d,n}}) && \text{for the } n^{\text{th}} \text{ word in document } d \end{aligned}$$



Bayesian Inference Algorithms

Posterior Inference

- Latent-variable models contain *observed* and *latent* random variables
- Model specifies:
  - *Likelihood*: $p(\text{data} \mid \text{latent variables}, \text{fixed parameters})$
  - *Prior*: $p(\text{latent variables} \mid \text{fixed parameters})$
- Goal: try to estimate the *posterior* via Bayes' rule

$$p(\text{latent variables} \mid \text{data}, \text{fixed parameters}) = \frac{\text{Likelihood} \times \text{Prior}}{p(\text{data} \mid \text{fixed parameters})}$$

- Denominator: marginalization is often intractable
- Need approximate inference methods

Gibbs Sampling

- Markov Chain Monte Carlo (MCMC) method
  - *Monte Carlo*: Estimate a quantity by drawing samples from a random distribution
  - *Markov Chain*: Find stationary distribution of a stochastic process where update rules depend only on previous state
- State vector $\vec{z}$; each component corresponds to a latent variable
- Repeatedly update $\vec{z}$ by iterating through latent variables, updating z_k by sampling from its *complete conditional*:

  $$p(z_k \mid \vec{z}_{-k}, \vec{x})$$

  Here, $\vec{z}_{-k}$ denotes all components of $\vec{z}$ except z_k
- The distribution of the samples $\vec{z}$ approaches the true posterior $p(\vec{z} \mid \vec{x})$

Collapsed Gibbs Sampling

- For some models, we can eliminate some latent variables by marginalization
- For the remaining latent variables, we compute a modified form of the complete conditionals:

  $$p(z_k \mid \vec{z}_{\text{subset}-k}, \vec{x})$$

- Running Gibbs sampling based on these distributions yields an estimate for

  $$p(\vec{z}_{\text{subset}} \mid \vec{x})$$

Variational Inference

- Approximation technique: select an approximating family of distributions and search for best approximation
- Measure closeness using reversed Kullback-Leibler divergence

$$\mathsf{KL}(q,\, p(\cdot \mid \vec{x})) = E_{\vec{z}\sim q}[\log q(\vec{z}) - \log p(\vec{z} \mid \vec{x})]$$

- Mean-field approximation: consider parameterized functions which factor cleanly:

$$q(\vec{z}) = \prod_k q_k(z_k; \nu_k)$$

- Minimizing reversed KL corresponds to maximizing *evidence lower bound* (ELBO):

$$\mathsf{ELBO} = E_q[\log p(\vec{z}, \vec{x})] - E_q[\log q(\vec{z})]$$

Coordinate-Ascent Variational Inference

- Coordinate ascent: optimize one latent variable's parameters at a time
- Works best for exponential-family models, where conditional distributions can be written as

$$p(x \mid \theta) = h(x) \exp(\eta(\theta) \cdot T(x) - a(\theta))$$

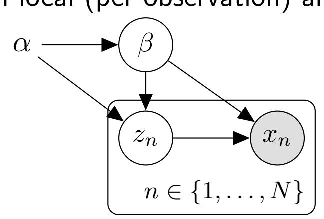
  ($\eta(\theta)$ = *natural parameters*, $T(x)$ = *sufficient statistics*)
- For exponential-family models, the update rule for z_k is

$$\nu_k = E_q[\eta_k(\vec{z}_{-k}, \vec{x})]$$

  where η_k denotes the natural parameters of the complete conditional of z_k

Stochastic Variational Inference: Context

- Generic model with local (per-observation) and global variables:



  x_n: observed data
  z_n: local variables (one per observation)
  β: global variable (shared for all observations)
  α: fixed parameters
- Complete conditional for local variables simplifies:

$$p(z_n \mid \alpha, \beta, z_{-n}, x_{1:N}) = p(z_n \mid \alpha, \beta, x_n)$$

- Complete conditional for global variable requires full dataset:

$$p(\beta \mid \alpha, z_{1:N}, x_{1:N})$$

Stochastic Variational Inference: Natural Gradient

- Euclidean distance on variational parameters may not reflect "true" distance between distributions
- Rather than standard gradient of the objective function ($\nabla\mathcal{L}$), use natural gradient $G^{-1}\nabla\mathcal{L}$
- G is a matrix (*metric tensor*) that encodes local information about "true" distances
- With a symmetric version of KL divergence and a model with exponential-family distributions, G cancels cleanly:

$$G^{-1}\nabla\mathcal{L} = E_q[\eta] - \nu$$

  where ν is the current value of the local variational params
- For local variables, the update rule is the same as in CAVI:

$$\nu^{\mathsf{local}} = E_q[\eta^{\mathsf{local}}]$$

Stochastic Variational Inference: Global Updates

- For global variables, repeatedly draw *mini-batches* b containing S observations
- Compute an *unbiased estimate* of the natural gradient $G^{-1}\nabla\mathcal{L}$ for each batch:

$$\mu = E_q[\eta_b^{\text{global}}] - \nu^{\text{global}}$$

  Here, η_b^{global} denotes the natural parameters of the complete conditional of the global variable, but with the true dataset replaced by N/S copies of the mini-batch b
- Update according to a decaying schedule of step sizes ρ_t:

$$\begin{aligned}\nu^{\text{global}} &\leftarrow \nu^{\text{global}} + \rho_t\, \mu \\ &= (1-\rho_t)\nu^{\text{global}} + \rho_t\, E_q[\eta_b^{\text{global}}]\end{aligned}$$

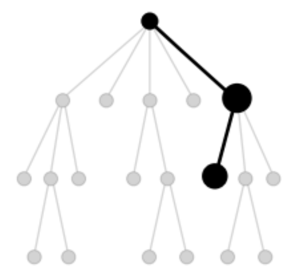
Learning Topic Hierarchies

Topic Modeling with Hierarchies

- Goal: extend LDA model so that:
  - Topics are arranged in a tree (root → abstract; leaves → concrete)
  - The size and structure of the tree can be determined in a data-driven way
- Documents can combine topics, but in a more constrained way
  - If a document draws words from one node, then it should also be somewhat likely to draw words from ancestor nodes
- We'll discuss two main models:
  - Nested Chinese Restaurant Process
  - Nested Hierarchical Dirichlet Process

Nested Chinese Restaurant Process

- Idea: Each document samples a path from an infinite tree
- Within each document, we can only select nodes (ie, topics) from the sampled path



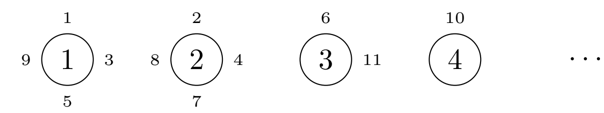
Source: Paisley et al [16]

Nested Chinese Restaurant Process: Preliminaries

- How to define distributions over paths in an infinite (or arbitrarily large) tree?
  Nested Chinese Restaurant Process
- How to define distributions over arbitrarily large partitions?
  Chinese Restaurant Process

Chinese Restaurant Process: Distribution Over Partitions

- Analogy: Sequence of customers entering a restaurant
- Infinitely many tables, each with infinite capacity
- First customer always sits at first table
- When $n \geq 1$ customers have been seated, the next customer follows these rules:
  - If the first k tables are occupied, with the i^{th} table containing m_i customers, sit at table i with probability $\frac{m_i}{n+\alpha}$
  - Sit at the next empty table with probability $\frac{\alpha}{n+\alpha}$



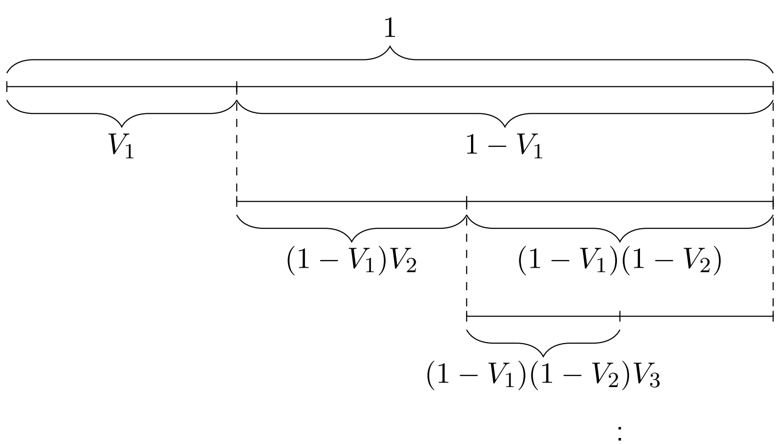
- Parameter α: As $\alpha \to \infty$, number of occupied tables increases

Chinese Restaurant Process: Distribution Over Partitions

- Stick-breaking construction:
  - Draw infinite sequence of beta-distributed variables:

$$V_k \sim \text{Beta}(1, \alpha) \qquad \text{for } k \geq 1$$

  - Draw table index k with probability $\pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$



Nested CRP: Distribution Over Paths

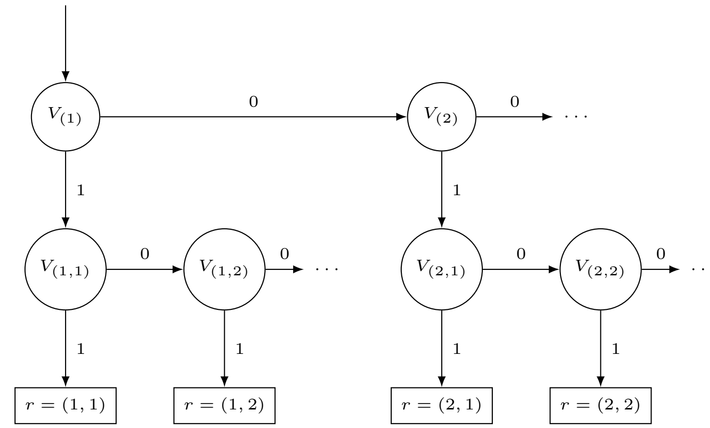
- Analogy: Infinitely many restaurants, arranged in a tree
- Customers enter the "root" restaurant and select a table
- Once seated, customers move to a restaurant indicated by a card at their table
- This process repeats indefinitely at new restaurants

Nested CRP: Distribution Over Paths

- A single draw from the NCRP is a distribution over infinite paths (finite-depth variant also exists)
- Stick-breaking construction: $T \sim \text{NCRP}(\alpha)$ denotes:

$$V_r \sim \text{Beta}(1, \alpha) \qquad \text{for any finite-length path } r$$
$$V_{()} = 1$$
$$\pi_{()} = 1$$
$$\pi_{r[1:\ell]} = \pi_{r[1:\ell-1]} \cdot \left(V_{r[1:\ell]} \prod_{j=1}^{r[\ell]-1} (1 - V_{r[1:\ell-1],j}) \right)$$
$$T = \sum_{r\text{:infinite path}} \pi_r \delta_r$$

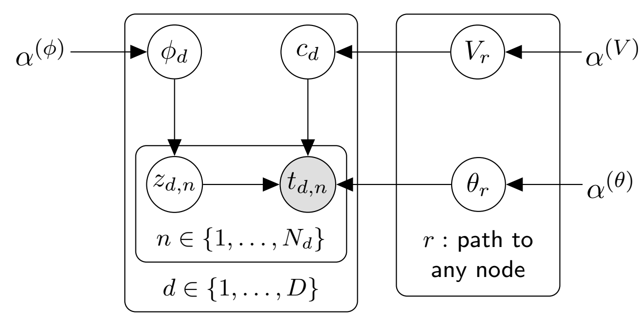
Nested CRP: A Finite-Depth Example



NCRP Topic Model

- Draw an infinite tree of topics, $\theta_r \sim \text{Dirichlet}(\alpha^{(\theta)})$
- Draw a global distribution over paths, $T \sim \text{NCRP}(\alpha^{(V)})$
- For each document d:
  - Draw a path $c_d \sim T$
  - Draw a stick-breaking distribution over depths ϕ_d
  - For each word-slot n:
    - Draw a depth $z_{d,n} \sim \text{Categorical}(\phi_d)$
    - Draw a vocabulary word $t_{d,n} \sim \text{Categorical}(\theta[c_d[1 : z_{d,n}]])$

NCRP Topic Model

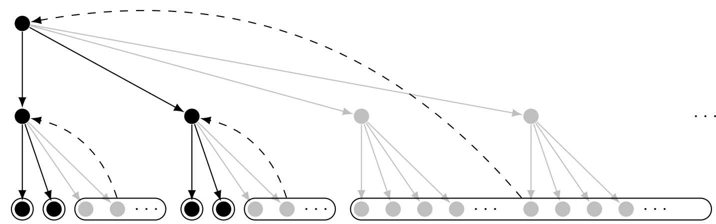


NCRP Topic Model: Gibbs Sampling

- Collapsed Gibbs sampling (marginalize out depth proportions ϕ_d and topic vectors θ_r)
- Griffiths et al [11]: Finite depth, uses order-dependent "restaurant analogy" formulation to avoid tracking infinitely many paths; each sampling step may grow or shrink the tree
- Blei et al [3]: Uses lazy evaluation; if final layer is ever sampled, then start tracking one extra layer

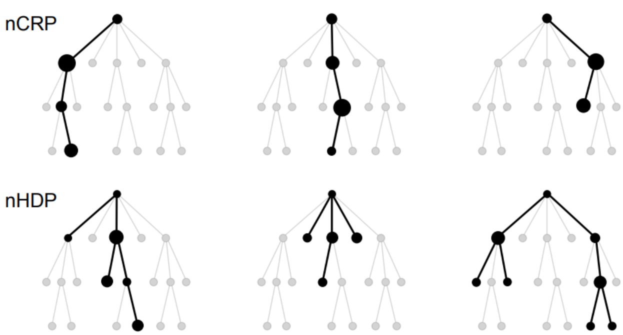
NCRP Topic Model: Variational Inference

- Infinitely many latent variables: need additional approximations
- Start with finite-depth, finite-width tree
- Depth stays constant, but width may change
- Outside of the finite truncation, variational distributions are assumed constant
- Divide infinite set of paths into equivalence classes
- If one equivalence class becomes sufficiently likely, add a representative path from it



Nested Hierarchical Dirichlet Process

- Idea: Global probability distribution over nodes, and each document samples a re-weighted version of that distribution
- Handles "hybrid" topics better than NCRP



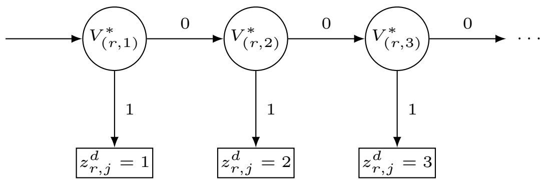
Source: Paisley et al [16]

NHDP Topic Model

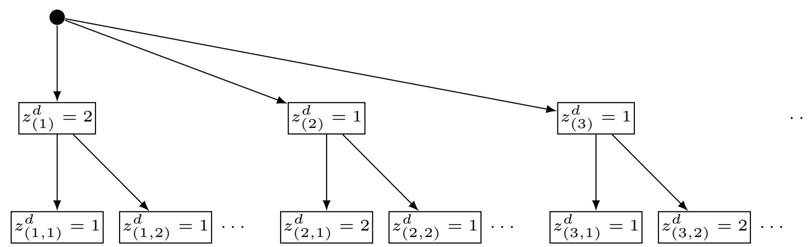
- Each node in infinite tree associated with a topic vector θ_r
- Global distribution over paths drawn from an NCRP distribution (i.e., draw global stick-breaking proportions $V^*_{r,j}$)
- Per-document:
  - Permutation of branches $z^d_{r,j}$
  - "Re-weighting" stick-breaking proportions $V^d_{r,j}$
  - "Path-propagation" proportions U^d_r
- Together, $z^d_{r,j}$, $V^d_{r,j}$, and U^d_r define a document-specific distribution over nodes

NHDP Topic Model: Selecting Indices

- For each document d, for each node r, and for each $j \geq 1$, select $z^d_{r,j}$:

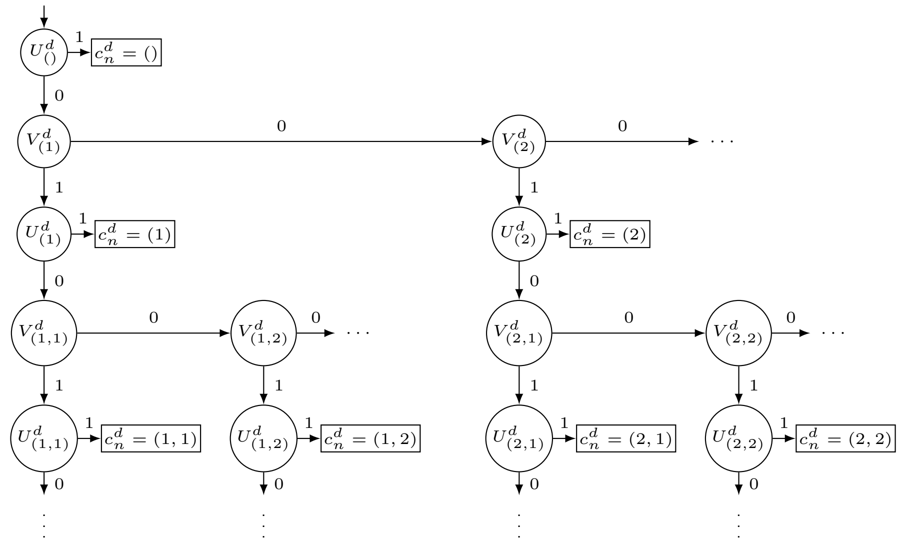


- Result defines how branches are permuted and copied (per document):



NHDP Topic Model: Path Propagation

Visualizing $V^d_{r,j}$ and U^d_r, ignoring branch permutations (i.e., assuming $z^d_{r,j} = j$)



NHDP Topic Model: Conditional Distributions

$$\theta_r \sim \text{Dirichlet}(\alpha^{(\theta)})$$

$$V^*_{r,j} \sim \text{Beta}(1, \alpha^{(V^*)})$$

$$V^d_{r,j} \sim \text{Beta}(1, \alpha^{(V)})$$

$$U^d_r \sim \text{Beta}(\alpha^{(U)}_1, \alpha^{(U)}_2)$$

$$z^d_{r,j} \sim \sum_{k \geq 1} \left(V^*_{r,k} \prod_{i=1}^{k-1} (1 - V^*_{r,i}) \right) \delta_k$$

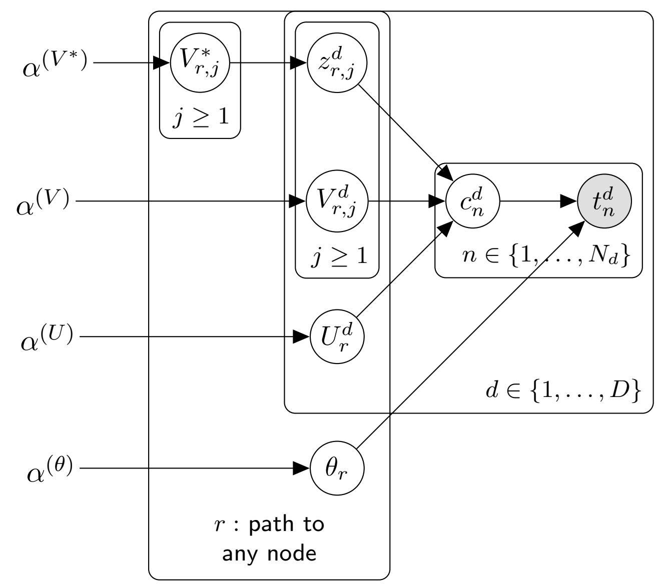
$$c^d_n \sim \sum_{r:\text{path}} A(r, V^d, z^d)\, B(r, U^d)\, \delta_r$$

$$A(r, V^d, z^d) = \prod_{m=0}^{\text{len}(r)-1} \sum_{k \geq 1} \mathbb{1}\left[z^d_{r[1:m],k} = r[m+1] \right] \left(V^d_{r[1:m],k} \prod_{i=1}^{k-1} \left(1 - V^d_{r[1:m],i}\right) \right)$$

$$B(r, U^d) = U^d_r \prod_{m=0}^{\text{len}(r)-1} \left(1 - U^d_{r[1:m]}\right)$$

$$t^d_n \sim \text{Categorical}(\theta_{c^d_n})$$

NHDP Topic Model: Plate Diagram



NHDP: Stochastic Variational Inference

- Use a finite-depth, finite-width tree
- Simplifications for document-specific indices $z_{r,j}^d$:
  - Use Dirac-δ variational distributions
  - For any d and any r, the indices $z_{r,j}^d$ do not repeat
  - For each document, greedy algorithm selects small number of nodes to include
- Greedy algorithm: start with root, add a node only if it increases ELBO by some threshold
- Remainder of algorithm is a standard application of stochastic variational inference

Directions for Future Research

- Scalable algorithms
- Interpreting models
- Incorporating human feedback
- Moving beyond the bag-of-words model
- Frameworks for Bayesian non-parametric inference

Acknowledgements

- **Advisor:** Sanjoy Dasgupta
- **LANL Mentor:** Kari Sentz
- **Research Exam Committee:** Vineet Bafna (chair), Yoav Freund, Julian McAuley

Thank You!

References

[1] Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu.
A practical algorithm for topic modeling with provable guarantees.
In *International Conference on Machine Learning*, pages 280–288, 2013.

[2] David Barber.
Bayesian reasoning and machine learning.
Cambridge University Press, 2012.
Online version available at `http://www.cs.ucl.ac.uk/staff/d.barber/brml/`; Draft dated 2017-02-02.

[3] David M Blei, Thomas L Griffiths, and Michael I Jordan.
The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies.
Journal of the ACM (JACM), 57(2):7, 2010.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe.
Variational inference: A review for statisticians.
Journal of the American Statistical Association, (just-accepted), 2017.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan.
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022, 2003.

[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman.
Indexing by latent semantic analysis.
Journal of the American society for information science, 41(6):391, 1990.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin.
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.

[8] Thomas S Ferguson.
A bayesian analysis of some nonparametric problems.
The annals of statistics, pages 209–230, 1973.

References

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
The elements of statistical learning.
Springer, 2001.

[10] Samuel J Gershman and David M Blei.
A tutorial on bayesian nonparametric models.
Journal of Mathematical Psychology, 56(1):1–12, 2012.

[11] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei.
Hierarchical topic models and the nested chinese restaurant process.
In *Advances in neural information processing systems*, pages 17–24, 2004.

[12] Thomas L Griffiths and Mark Steyvers.
Finding scientific topics.
Proceedings of the National academy of Sciences, 101(suppl 1):5228–5235, 2004.

[13] Gregor Heinrich.
Parameter estimation for text analysis.
Technical report, 2005.

[14] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley.
Stochastic variational inference.
The Journal of Machine Learning Research, 14(1):1303–1347, 2013.

[15] Thomas Hofmann.
Probabilistic latent semantic analysis.
In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[16] John Paisley, Chong Wang, David M Blei, and Michael I Jordan.
Nested hierarchical dirichlet processes.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(2):256–270, 2015.

References

[17] Rajesh Ranganath, Sean Gerrish, and David Blei.
Black box variational inference.
In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[18] Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei.
Operator variational inference.
In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

[19] Philip Resnik and Eric Hardisty.
Gibbs sampling for the uninitiated.
Technical report, University of Maryland, College Park, 2010.

[20] Jayaram Sethuraman.
A constructive definition of dirichlet priors.
Statistica sinica, pages 639–650, 1994.

[21] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei.
Sharing clusters among related groups: Hierarchical dirichlet processes.
In *Advances in neural information processing systems*, pages 1385–1392, 2005.

[22] Chong Wang and David M Blei.
Variational inference for the nested chinese restaurant process.
In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2009.

[23] Wikipedia.
Exponential family.
`https://en.wikipedia.org/w/index.php?title=Exponential_family&oldid=787816251`.
Accessed September 2017.

[24] Wikipedia.
Latent dirichlet allocation.
`https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=797823717`.
Accessed September 2017.

References

[25] Wikipedia.
Latent semantic analysis.
`https://en.wikipedia.org/w/index.php?title=Latent_semantic_analysis&oldid=798597246`.
Accessed September 2017.

[26] Wikipedia.
Probabilistic latent semantic analysis.
`https://en.wikipedia.org/w/index.php?title=Probabilistic_latent_semantic_analysis&oldid=783155225`.
Accessed September 2017.