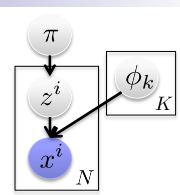
Case Study 2: Document Retrieval

Collapsed Gibbs and Variational Methods for LDA

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox
February 7th, 2013

Example – Collapsed MoG Sampling

$$\pi \sim \mathrm{Dir}(\alpha_1,\dots,\alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k,\Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i;\mu_{z^i},\Sigma_{z^i})$$

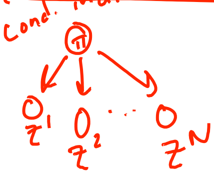


- Collapsed sampler

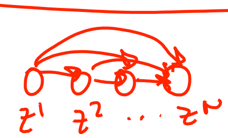
For $i=1,\dots,N$

$$z^{i(t)} \sim p(z^i \mid \underbrace{z^{1(t)},\dots,z^{i-1(t)},z^{i+1(t-1)},\dots,z^{N(t-1)}}_{z^{(t)}_{\setminus i}}, x_{1:N},\alpha,\lambda)$$

cond. ind.



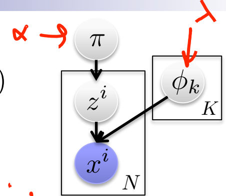
marg. π



Example – Collapsed MoG Sampling

$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ $\qquad z^i \sim \pi$

$\{\mu_k, \Sigma_k\} \sim F(\cancel{\phi})$ $\qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$



- Derivation

$p(z^i \mid z_{\setminus i}, x_{1:N}, \alpha, \lambda) \propto p(z^i \mid z_{\setminus i}, \alpha)\, p(x^i \mid z^i, z_{\setminus i}, x_{\setminus i}, \lambda)$

$p(z^i = k \mid z_{\setminus i}, \alpha) = \int p(z^i = k \mid \pi) \underbrace{p(\pi \mid z_{\setminus i}, \alpha)}_{\text{Dir post.}} d\pi = \dfrac{n_k^{-i} + \alpha_k}{N - 1 + \sum \alpha_k}$

$p(x^i \mid z_{1:N}, x_{\setminus i}, \lambda)$ = student-t pred. likelihood

- Important facts:

$z^i \sim \pi$, $\pi \sim \mathrm{Dir}$

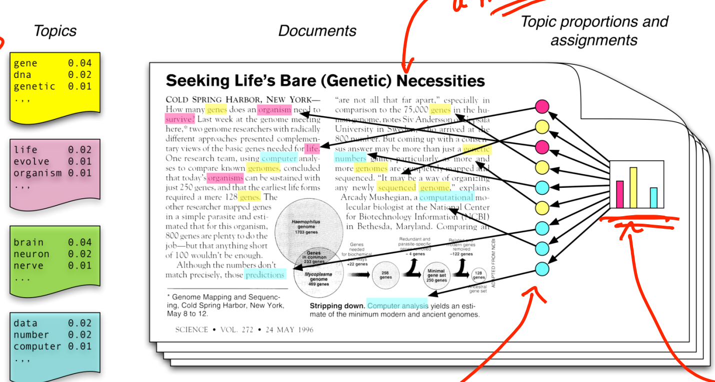
$$p(z_{1:N} \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_k + \alpha_k)}{\Gamma(\sum_k n_k + \alpha_k)} \qquad \frac{\Gamma(m+1)}{\Gamma(m)} = m$$

Gamma fcn

Latent Dirichlet Allocation (LDA)

each doc is a mixture of these corpus-wide topics

each topic as a dist. over words $\{\beta_k\}$



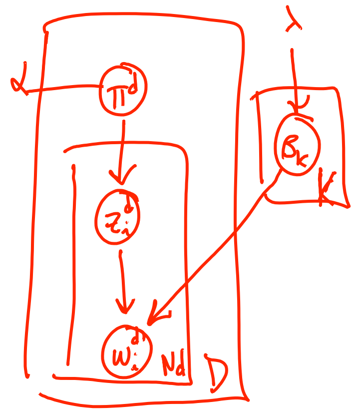
every word is assigned to a topic

each doc has its own prevelance of topics in doc

LDA Generative Model

- Observations: $w_1^d, \dots, w_{N_d}^d$ $d=1,\dots,D$
- Associated topics: $z_1^d, \dots, z_{N_d}^d$ corpus-wide topic
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model: doc-specific preferences of topics

$z_i^d \sim \pi^d$ $d=1,\dots,D$
$w_i^d \mid z_i^d \sim \beta_{z_i^d}$ $i=1,\dots,N$

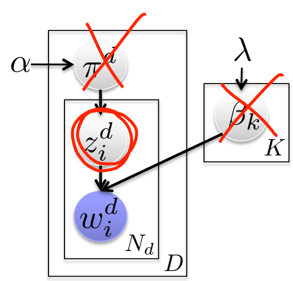


Priors:

$\pi^d \sim Dir(\alpha_1, \dots, \alpha_K)$ $d=1,\dots,D$

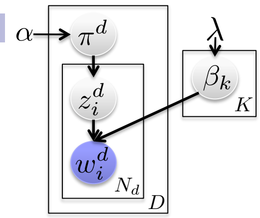
$\beta_k \sim Dir(\lambda_1, \dots, \lambda_V)$ $k=1,\dots,K$

LDA Generative Model



$$p(\cdot) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \left(\prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta) \right)$$

Collapsed LDA Sampling



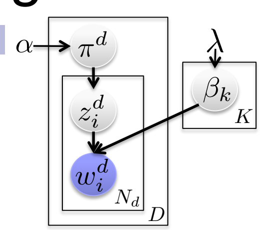
- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Derivation:

$$p(z^d_{1:N_d} \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_k^d + \alpha_k)}{\Gamma(\sum_k n_k^d + \alpha_k)}$$

$$p(\{w_i^d \mid z_i^d = k\}, \lambda) = \frac{\Gamma(\sum_\nu \lambda_\nu)}{\prod_\nu \Gamma(\lambda_\nu)} \frac{\prod_\nu \Gamma(v_\nu^k + \lambda_\nu)}{\Gamma(\sum_\nu v_\nu^k + \lambda_\nu)}$$

$$p(z \mid \alpha) = \prod_{d=1}^{D} p(z^d_{1:N_d} \mid \alpha) \qquad p(w \mid z, \lambda) = \prod_{k=1}^{K} p(\{w_i^d \mid z_i^d = k\}, \lambda)$$

Collapsed LDA Sampling

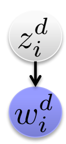


- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Algorithm:

Sample Document

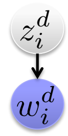
Etruscan	trade	price	temple	market

Randomly Assign Topics

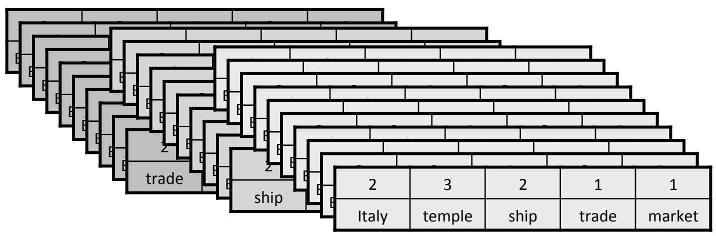


3	2	1	3	1
Etruscan	trade	price	temple	market

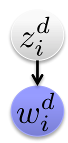
Randomly Assign Topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Maintain Global Statistics



3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
…			

Resample Assignments

z_i^d

$\downarrow$

w_i^d

3	2	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
…			

What is the conditional distribution for this topic?

z_i^d

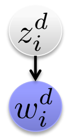
$\downarrow$

w_i^d

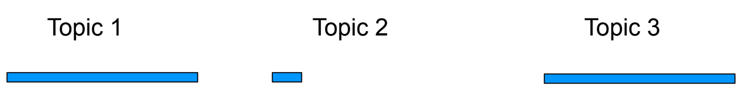
3	?	1	3	1
Etruscan	trade	price	temple	market

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?

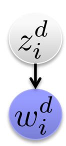


3	?	1	3	1
Etruscan	trade	price	temple	market

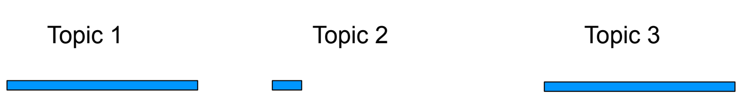


What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?



3	?	1	3	1
Etruscan	trade	price	temple	market



	1	2	3
trade	10	7	1

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

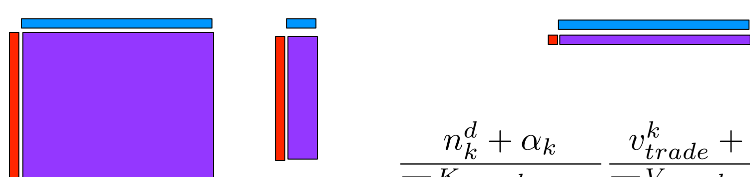
What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

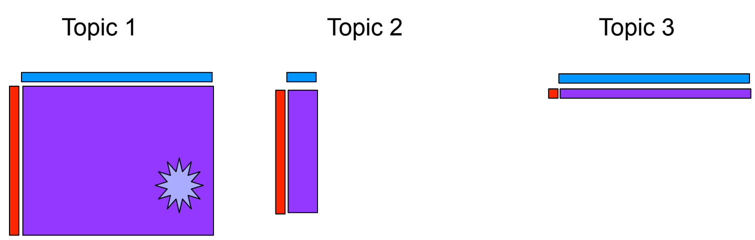


$$\frac{n_k^d + \alpha_k}{\sum_{j=1}^K n_j^d + \alpha_j} \frac{v_{trade}^k + \lambda_k}{\sum_{j=1}^V v_j^k + \lambda_j}$$

Sample a New Topic Indicator

$z_i^d \to w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market



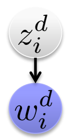
Update Counts

$z_i^d \to w_i^d$

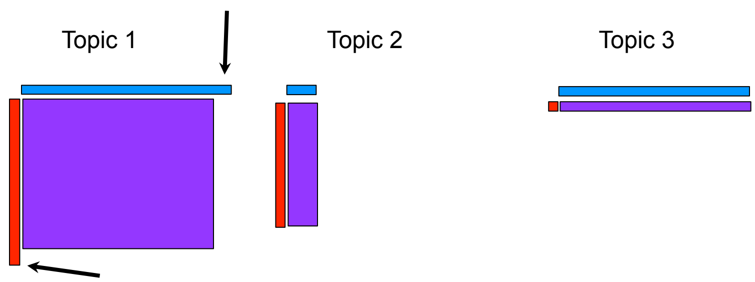
3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
…			

Geometrically…



3	**1**	1	3	1
Etruscan	trade	price	temple	market



Issues with Generic LDA Sampling

- Slow mixing rates → Need many iterations
- Each iteration cycles through sampling topic assignments for *all* words in *all* documents
- Modern approaches:
  - Large-scale LDA. For example,
    Mimno, David, Matthew D. Hoffman and David M. Blei. "Sparse stochastic inference for latent Dirichlet allocation." International Conference on Machine Learning, 2012.
  - Distributed LDA. For example,
    Ahmed, Amr, et al. "Scalable inference in latent variable models." Proceedings of the fifth ACM international conference on Web search and data mining (2012): 123-132
- Alternative: Variational methods instead of sampling
  - Approximate posterior with an optimized variational distribution

Variational Methods

- Recall task: Characterize the posterior
- Turn posterior inference into an optimization task
- Introduce a "tractable" family of distributions over parameters and latent variables
  - Family is indexed by a set of "free parameters"
  - Find member of the family closest to:
- Questions:
  - How do we measure "closeness"?
  - If the posterior is intractable, how can we approximate something we do not have to begin with?

A Measure of Closeness

- Kullback-Leibler (KL) divergence
  - Measures "distance" between two distributions *p* and *q*
- Not symmetric
- *p* determines where the difference is important:
  - $p(x)=0$ and $q(x)\neq 0$
  - $p(x)\neq 0$ and $q(x)=0$
- Want
- Just as hard as the original problem!

Reverse Divergence

- Divergence D(*q* || *p*)
  - true distribution *p* defines support of diff.
  - the "correct" direction
  - will be intractable to compute
- Reverse divergence D(*q* || *p*)
  - approximate distribution defines support
  - tends to give overconfident results
  - will be tractable

Interpretations of Minimizing Reverse KL

- Similarity measure:
- Evidence lower bound (ELBO)
- Therefore, minimizing KL is equivalent to maximizing a lower bound on the marginal likelihood:
  - Max $\mathcal{L}(q)$ = min $D(q||p)$ = max lower bound of $\log p(x)$

Mean Field

- How do we choose a *Q* such that the following is tractable?

- Simplest case = mean field approximation
  - Assume each parameter and latent variable is conditionally independent given the set of free parameters

- Then, entropy term decomposes as

Mean Field

- Examine one free parameter, *e.g.*, γ
  - Can rewrite joint as

$$E_q[\log p(\theta, z, x)] = E_q[\log p(\theta \mid z, x)] + E_q[\log p(z, x)]$$

  - Look at terms of ELBO just depending on γ

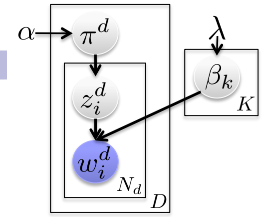
$$\mathcal{L}^{\gamma} =$$

- Likewise,

$$\mathcal{L}^{\phi^n} =$$

- This motivates using a coordinate ascent algorithm for optimization
  - Iteratively optimize each free parameter holding all others fixed

Mean Field for LDA



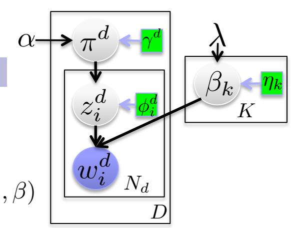
- In LDA, our parameters are $\theta = \{\pi^d\}, \{\beta_k\}$
  $z = \{z_i^d\}$

- The variational distribution factorizes as

- The joint distribution factorizes as

$$p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta)$$

Mean Field for LDA



$$q(\pi, \beta, z) = \prod_{k=1}^{K} q(\beta_k \mid \eta_k) \prod_{d=1}^{D} q(\pi^d \mid \gamma^d) \prod_{i=1}^{N_d} q(z_i^d \mid \phi_i^d)$$

$$p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta)$$

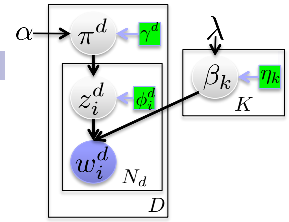
- Examine the ELBO

$$\mathcal{L}(q) = \sum_{k=1}^{K} E_q[\log p(\beta_k \mid \lambda)] + \sum_{d=1}^{D} E_q[\log p(\pi^d \mid \alpha)]$$

$$+ \sum_{d=1}^{d} \sum_{i=1}^{N_d} E_q[\log p(z_i^d \mid \pi^d)] + E_q[\log p(w_i^d \mid z_i^d, \beta)]$$

$$- \sum_{k=1}^{K} E_q[\log q(\beta_k \mid \eta_k)] - \sum_{d=1}^{D} E_q[\log q(\pi^d \mid \gamma^d)] - \sum_{d=1}^{d} \sum_{i=1}^{N_d} E_q[\log q(z_i^d \mid \phi_i^d)]$$

Mean Field for LDA



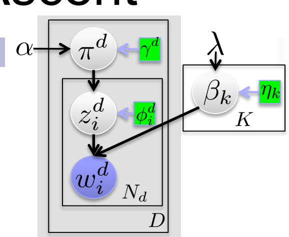
- Let's look at some of these terms

$$E_q[\log p(z_i^d \mid \pi^d)]$$

$$E_q[\log q(z_i^d \mid \phi_i^d)]$$

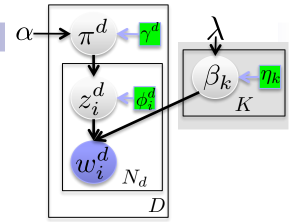
- Other terms follow similarly

Optimize via Coordinate Ascent



- Algorithm:

Optimize via Coordinate Ascent



- Algorithm:

Alternative Optimization Schemes

- Inefficient:
  - Start from randomly initialized η_k (topics)
  - Analyze whole corpus before updating η_k again
  - If streaming data scenario, can't compute even one iteration!

- Didn't have to do coord. ascent. Could have used gradient ascent.

Alternative Optimization Schemes

- Recall stochastic gradient ascent:
  - Assume M = 1
  - Unbiased, but noisy
- Here,

$$\mathcal{L} = E_q[\log p(\beta)] - E_q[\log q(\beta)] + \sum_{d=1}^{D} E_q[\log p(\pi^d)] - E_q[\log q(\pi^d)]$$
$$+ \sum_{d=1}^{D} E_q[\log p(z^d, x^d \mid \pi^d, \beta)] - E_q[\log q(z^d)]$$

$$\mathcal{L}_t = E_q[\log p(\beta)] - E_q[\log q(\beta)] + D\left(E_q[\log p(\pi^t)] - E[\log q(\pi^t)]\right)$$
$$+ D\left(E_q[\log p(z^t, x^t \mid \pi^t, \beta)] - E_q[\log q(z^t)]\right)$$

Stochastic Variational Inference for LDA

- Initialize $\eta^{(0)}$ randomly.
- Repeat (indefinitely):
  - Sample a document ***d*** uniformly from the data set.
  - For all *k*, initialize $\gamma_k^d = 1$
  - Repeat until converged
    - For i=1,…,N_d

      $$\phi_{ik}^d \propto \exp\{E[\log \pi_k^d] + E[\log \beta_{k,w_i^d}]\}$$
    - Set $\gamma^d = \alpha + \sum_{i=1}^{N_d} \phi_i^d$
  - Take a stochastic gradient step $\eta^{(t)} = \eta^{(t-1)} + \rho_t \nabla_\eta \mathcal{L}_d$

Acknowledgements

- Thanks to Dave Blei, David Mimno, and Jordan Boyd-Graber for some material in this lecture relating to LDA