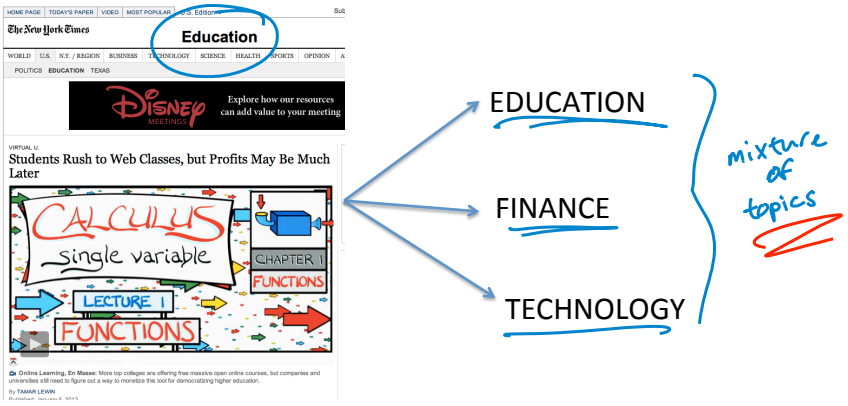
Case Study 5: Mixed Membership Modeling

LDA Collapsed Gibbs Sampler, Variational Inference

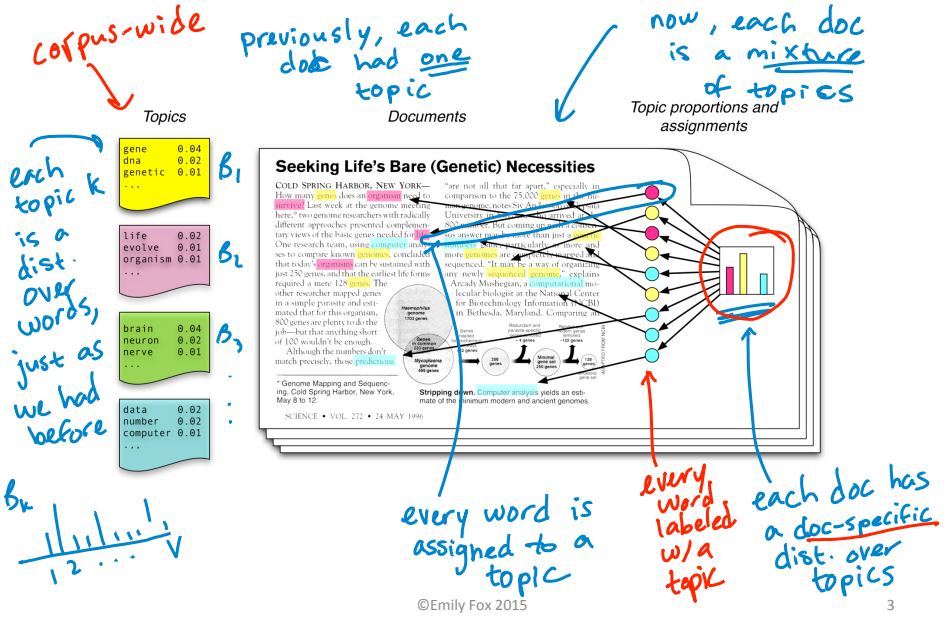
Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 28th, 2015

Task 3: Mixed Membership Models

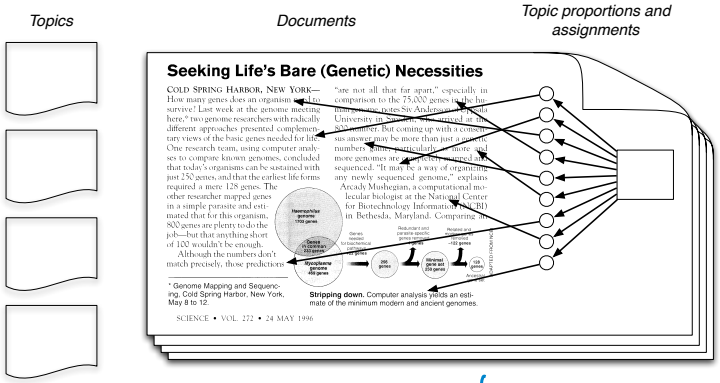
- **Now:** Document may belong to multiple clusters



Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation (LDA)



All we see are words $\{w_{dn}\}$

★ Want: posterior $p(\text{topics}, \text{doc. proportions}, \text{assign vars} \mid \text{words in docs})$

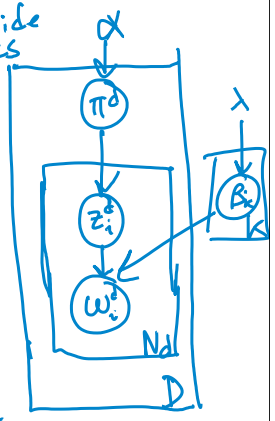
LDA Generative Model

- Observations: $w_1^d, \ldots, w_{N_d}^d$ $d=1,\ldots,D$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$ ← assign var. per word
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$ (β_k: corpus-wide topics; π^d: doc-specific topic weights)
- Generative model:

per word topic ind.

$z_i^d \sim \pi^d \quad d=1,\ldots,D$

$w_i^d \mid z_i^d \sim \beta_{z_i^d} \quad i=1,\ldots,N_d$



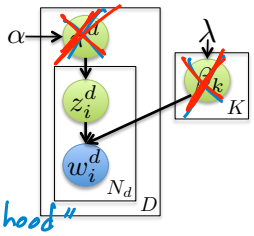
Priors:

$\pi^d \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \quad d=1,\ldots,D$

$\beta_k \sim \mathrm{Dir}(\lambda_1, \ldots, \lambda_V) \quad k=1,\ldots,K$

Collapsed LDA Sampling

- Sample topic indicators for each word
  - Algorithm:



$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$

"likelihood"

$$\propto \underbrace{p(z_i^d = k \mid \{z_j^d, j \neq i\}, \alpha)}_{\text{"prior"}} p(w_i^d \mid \{w_j^c : z_j^c = k, (j,c) \neq (i,d)\}, \lambda)$$

$\propto \dfrac{n_k^{-id} + \alpha_k}{N_d - 1 + \sum \alpha_k} \cdot \dfrac{m_{w_i^d,k}^{-id} + \lambda_{w_i^d}}{\sum_\gamma m_{\gamma,k}^{-id} + \lambda_\gamma}$

# of words assigned to topic k in doc d (not counting ith word)

normalize within doc

examine entire corpus

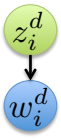
# of times word γ appears in topic k (not including w_i^d)

Select a Document

Etruscan	trade	price	temple	market

all words in doc d

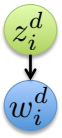
Randomly Assign Topics



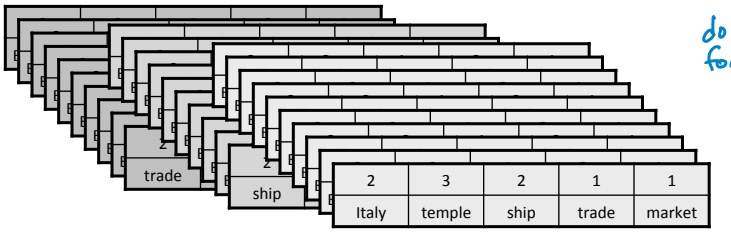
3	2	1	3	1
Etruscan	trade	price	temple	market

to initialize sampler (one approach)

Randomly Assign Topics

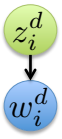


3	2	1	3	1
Etruscan	trade	price	temple	market



do this for each doc in corpus

Maintain Local Statistics

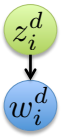


3	2	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

n_1^d n_2^d n_3^d

Maintain Global Statistics



3	2	1	3	1
Etruscan	trade	price	temple	market

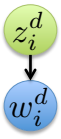
	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

$n^{word}_{price,2}$ ← word, topic

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

Total counts from **all** docs

Resample Assignments



z_2^d

3	~~2~~	1	3	1
Etruscan	trade	price	temple	market

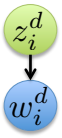
	Topic 1	Topic 2	Topic 3
Doc d	2	0 → 1	2

n^{-id}_{k}

	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7 → 8	1
…			

$m^{-id}_{word,topic}$

What is the conditional distribution for this topic?

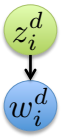


3	?	1	3	1
Etruscan	trade	price	temple	market

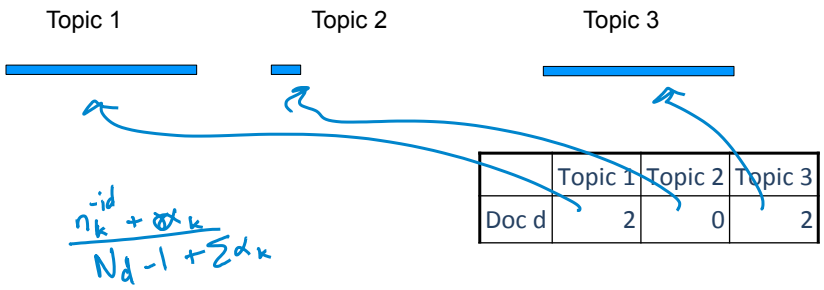
$p(z_i^d \mid \text{everything else})$

What is the conditional distribution for this topic?

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



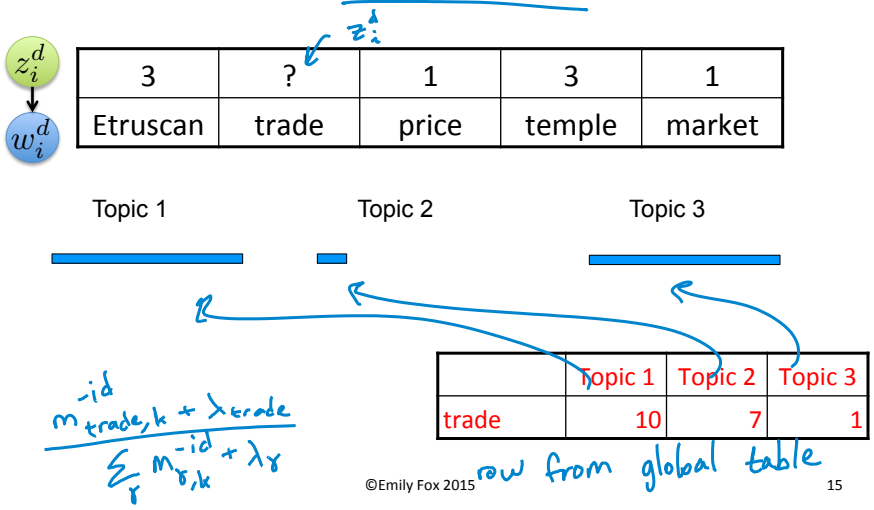
3	?	1	3	1
Etruscan	trade	price	temple	market



	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

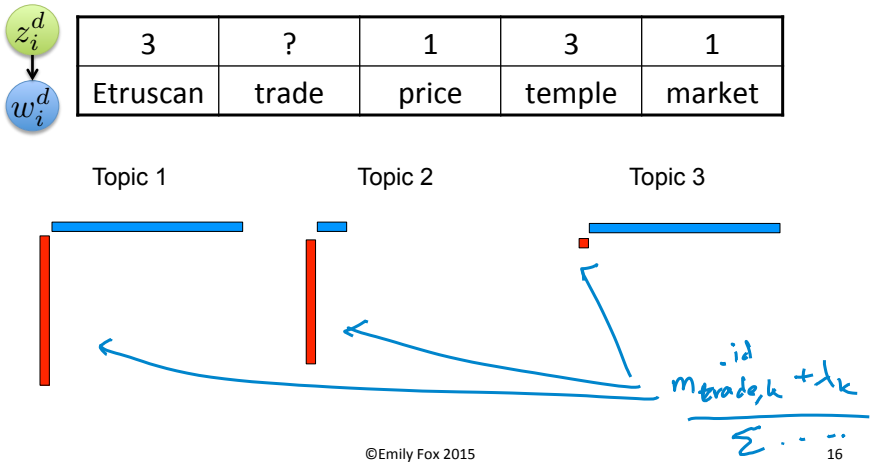
What is the conditional distribution for this topic?

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



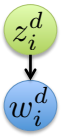
What is the conditional distribution for this topic?

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



What is the conditional distribution for this topic?

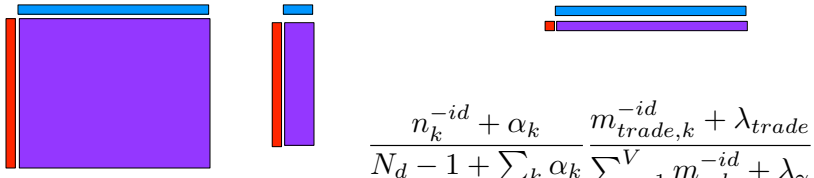
- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?



$z_i^d = ?$

3	?	1	3	1
Etruscan	trade	price	temple	market

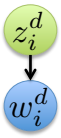
Topic 1 Topic 2 Topic 3



$$\frac{n_k^{-id} + \alpha_k}{N_d - 1 + \sum_k \alpha_k} \frac{m_{trade,k}^{-id} + \lambda_{trade}}{\sum_{\gamma=1}^{V} m_{\gamma,k}^{-id} + \lambda_\gamma}$$

throw a dart

Sample a New Topic Indicator

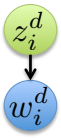


3	1 ~~?~~	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

lands here ⇒ happened to sample $z_i^d = 1$

Update Counts



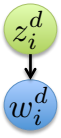
3	1 ~~7~~	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	3 ~~2~~	0	2

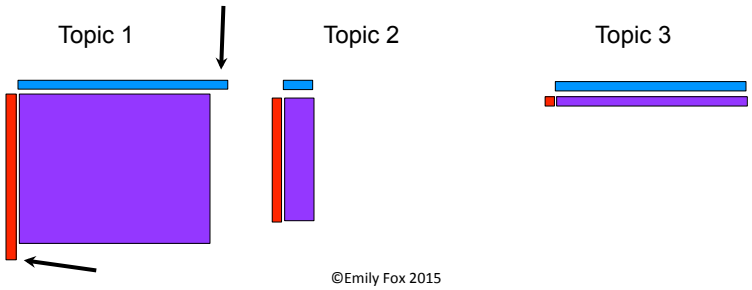
	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	11 ~~10~~	7	1
…			

Geometrically…

inc. popularity of topic 1 in doc d and prevalence of word "trade" in topic 1 (corpus wide)



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 include:
  - 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
  - And many, many more!
- Alternative: Variational methods instead of sampling
  - Approximate posterior with an optimized variational distribution

Case Study 5: Mixed Membership Modeling

Variational Methods

Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

May 28th, 2015

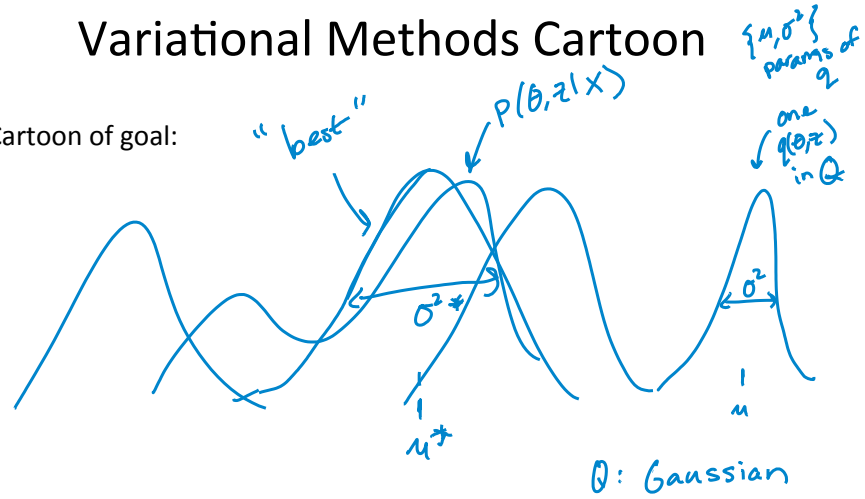
Variational Methods Goal

- Recall task: Characterize the posterior $p(\theta, z \mid x)$ (params, latent vars, obs.)
- 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: $p(\theta, z \mid x)$

Call family Q and want $q \in Q$ that is closest to $p(\theta, z \mid x)$

Variational Methods Cartoon

- Cartoon of goal:



- Questions:
  1. How do we measure "closeness"?
  2. 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*

$$KL(p||q) \triangleq D(p||q) = E_p\left[\log \frac{p}{q}\right] = \int_\theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

- If $p = q$ for all ϑ

$$D(p||q) = \int p(\theta) \log 1 \, d\theta = 0$$

- Otherwise, $D(p||q) > 0$

A Measure of Closeness

$$\mathrm{KL}(p||q) \triangleq D(p||q) = \int_\theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

- Not symmetric $D(p||q) \neq D(q||p)$ … not a true distance metric
- *p* determines where the difference is important:
  - $\exists\theta$ $p(\theta)=0$ and $q(\theta)\neq 0$ $0 \log 0 = 0$
  - $\exists\theta$ $p(\theta)\neq 0$ and $q(\theta)=0$ $\epsilon \log \frac{\epsilon}{0} = \infty$ (ϵ $p(\theta)=\epsilon$)

If $D(p||q)$ Finite, $\text{supp}(q) \supseteq \text{supp}(p)$

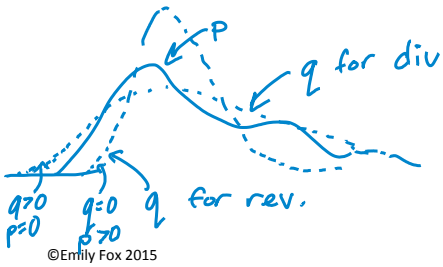
- Want $\hat{q} = \arg\min_{q \in Q} D(p||q)$
- Just as hard as the original problem! $E_p[\dots]$

Reverse Divergence

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

q: what we have control over

q now less diffuse than p



Interpretations of Minimizing Reverse KL

$$D(q||p) = E_q\left[\log \frac{q}{p}\right]$$

$p(\theta, z, x) / p(x)$

- Similarity measure:

$$D(q(\theta,z)\,||\,p(\theta,z|x)) = E_q[\log q(\theta,z)] - E_q[\log p(\theta,z|x)]$$

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

$-\mathcal{L}(q)$

- Evidence lower bound (ELBO)

log marginal likelihood = "evidence"

$$\log p(x) = D(q(z,\theta)\,||\,p(\theta,z|x)) + \mathcal{L}(q) \geq \mathcal{L}(q)$$

≥ 0 "ELBO"

Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)

$$\log p(x) = D(q(z,\theta)||p(z,\theta|x)) + \mathcal{L}(q) \geq \mathcal{L}(q)$$

const. add to a const. ELBO

- Therefore,
  - ELBO provides a lower bound on marginal likelihood
  - Maximizing ELBO is equivalent to minimizing KL

$\max \mathcal{L}(q) = \min D(q\|p) = \max$ lower bound on $\log p(x)$

what we can control; depends on what we don't know

Mean Field

$$\mathcal{L}(q) = E_q[\log p(z,\theta,x)] - E_q[\log q(z,\theta)]$$

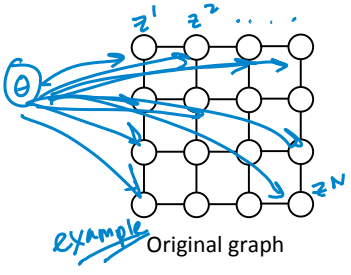
- How do we choose a Q such that the following is tractable?

$\hat{q} = \arg\max_{q \in Q} \mathcal{L}(q)$ ← new objective

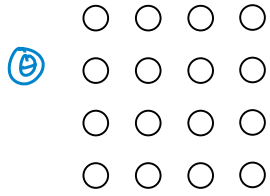
- Simplest case = mean field approximation
  - Assume each parameter and latent variable is conditionally independent given the set of free parameters

$\theta, z = \{z^1, \ldots, z^N\}$

$q(\theta, z) = q(\theta|\gamma) \prod_{i=1}^{N} q(z^i|\phi^i)$ ("free param")



Original graph



Naïve mean field

our focus, but there also structured approx.

example

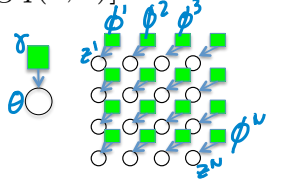
Mean Field

$$\mathcal{L}(q) = E_q[\log p(z, \theta, x)] - E_q[\log q(z, \theta)]$$

joint term; entropy term; free params

- Naïve mean field decomposition:

$$q(z, \theta) = q(\theta \mid \gamma) \prod_{i=1}^{N} q(z^i \mid \phi^i)$$



- Under this approximation, entropy term decomposes as

$$-E_q[\log q(\theta, z)] = -E_q[\log q(\theta \mid \gamma)] - \sum_{i=1}^{N} E_q[\log q(z^i \mid \phi^i)]$$

decouples across γ, ϕ^i

- Can (always) rewrite joint term as

full conditional of θ

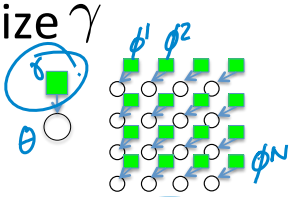
$$E_q[\log p(\theta, z, x)] = E_q[\log p(\theta \mid z, x)] + E_q[\log p(z, x)]$$

OR

full cond. of z^i

$$E_q[\log p(\theta, z, x)] = E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)] + E_q[\log p(z_{\setminus i}, \theta, x)]$$

Mean Field – Optimize γ



- Examine one free parameter, *e.g.*, γ

$$\mathcal{L}(q) = E_q[\log p(\theta \mid z, x)] + E_q[\log p(z, x)] - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

consider the θ full cond. form of joint

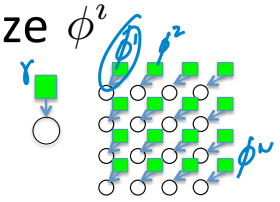
doesn't depend on γ because under q, $\theta \perp\!\!\!\perp z^i$!

  - Look at terms of ELBO just depending on γ

$$\mathcal{L}^\gamma = E_q[\log p(\theta \mid z, x)] - E_q[\log q(\theta \mid \gamma)] + \text{const}$$

really just $q_\theta = q(\theta \mid \gamma)$ needed here

Mean Field – Optimize ϕ^i



- Examine another free parameter, *e.g.*, ϕ^i

$$\mathcal{L}(q) = E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)] + E_q[\log p(z_{\setminus i}, \theta, x)] - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

z^i full cond. form of joint

const. wrt ϕ^i

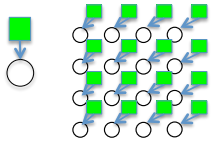
  - Look at terms of ELBO just depending on ϕ^i

$$\mathcal{L}^{\phi^i} = E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)] - E_q[\log q(z^i \mid \phi^i)]$$

really just $q_{\phi_i} = q(z^i \mid \phi^i)$

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

Algorithm Outline



- **Initialization:** Randomly select starting distribution $q_\theta^{(0)}$
- **E-Step:** Given parameters, find posterior of hidden data

optimize ϕ → $q_z^{(t)} = \arg\max_{q_z} \mathcal{L}(q_z, q_\theta^{(t-1)})$ fixing prev. $\gamma^{(t-1)}$

- **M-Step:** Given posterior distributions, find likely parameters

optimize γ → $q_\theta^{(t)} = \arg\max_{q_\theta} \mathcal{L}(q_z^{(t)}, q_\theta)$ fixing prev. $\phi^{(t)}$

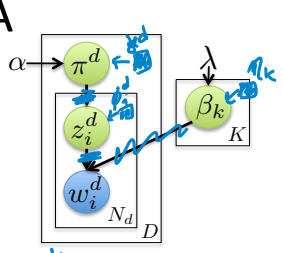
- **Iteration:** Alternate E-step & M-step until convergence

Case Study 5: Mixed Membership Modeling

Variational Inference for LDA

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 28th, 2015

Mean Field for LDA



- In LDA, our parameters are $\theta = \{\pi^d\}, \{\beta_k\}$
  $z = \{z_i^d\}$
- The variational distribution factorizes as

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

$\beta_k \sim \text{Dir}(\eta_{k1}, \ldots, \eta_{kV})$, $\pi^d \sim \text{Dir}(\gamma_1^d, \ldots, \gamma_K^d)$, $z_i^d \sim \text{Mult}(\phi_i^d)$, $\sum_{k=1}^{K} \phi_{ik}^d = 1$

- The joint distribution factorizes as

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

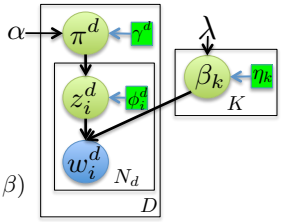
need to enforce this constraint

Mean Field for LDA

mean field approx.

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

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



- Examine the ELBO

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

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

from joint

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

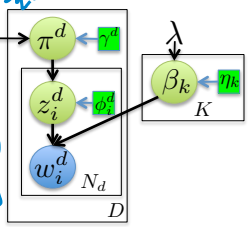
all terms from q

Mean Field for LDA

$E_{q(r,s)}[f(r)f(s)] = q(r)q(s)$

$= \int\int q(r)q(s)f(r)f(s)\,dr\,ds = \int_r q(r)f(r)dr \int_s q(s)f(s)ds = E_q[f(r)]\,E_q[f(s)]$

- Let's look at some of these terms



$$E_q[\log p(z_i^d \mid \pi^d)] = E_q[\log \pi^d_{z_i^d}] = E_q\Big[\log \prod_{k=1}^{K} (\pi_k^d)^{I(z_i^d=k)}\Big]$$

$$= \sum_{k=1}^{K} E_q[I(z_i^d = k) \log \pi_k^d] = \sum_{k=1}^{K} E_q[I(z_i^d=k)]\, E_q[\log \pi_k^d]$$

$E_q[I(z_i^d=k)] = \phi_{ik}^d$ ← $p(z_i^d=k)$ under q

$E_q[\log \pi_k^d] = \Psi(\gamma_k^d) - \Psi(\sum_k \gamma_k^d)$

$z_i^d \perp\!\!\!\perp \pi_k^d$ under q

⇒ why mean field approx. is so important ☆

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

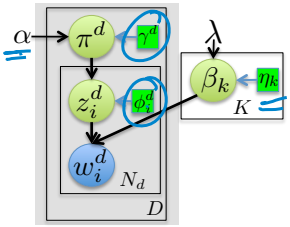
$$= \sum_k E_q[I(z_i^d = k) \log \phi_{ik}^d] = \sum_k \phi_{ik}^d \log \phi_{ik}^d$$

given

- Other terms follow similarly

Optimize via Coordinate Ascent

- Algorithm:



for $d=1,\dots,D$

$$\frac{\partial \mathcal{L}}{\partial \gamma^d} = 0 \rightarrow \gamma^{d(t+1)} = \alpha + \sum_{i=1}^{N_d} \phi_i^{d(t)}$$

for $i=1,\dots,N_d$

$$\frac{\partial \mathcal{L}}{\partial \phi_i^d} = 0 \rightarrow \phi_i^d \propto \exp\left\{\Psi(\gamma_{1:k}^{d(t+1)}) + \Psi(\eta_{1:k,w_i^d}^{(t)}) - \Psi\left(\sum_v \eta_{1:k,v}^{(t)}\right)\right\}$$

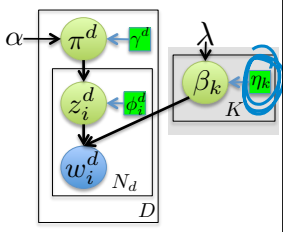
(ϕ_i^d: pmf over topics $1,\dots,K$)

use Lagrange multiplier to enforce $\sum \phi_i^d = 1$

DATA PARALLEL across ϕ_i^d

Optimize via Coordinate Ascent

- Algorithm:



What you need to know…

- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler
- Variational methods
  - Overall goal
  - Interpretation in terms of minimizing (reverse) KL
  - Mean field approximation

Acknowledgements

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

