

Case Study 5: Mixed Membership Modeling

Latent Dirichlet Allocation Collapsed Gibbs Sampler, Variational Methods

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

Emily Fox

March 11th, 2014

©Emily Fox 2014

1

Task 3: Mixed Membership Models

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

The image shows a screenshot of the New York Times website's Education section. A blue circle highlights the 'Education' link in the navigation bar. Below it, a blue arrow points from the 'Education' link to the word 'EDUCATION'. Another blue arrow points from the 'Education' link to the word 'FINANCE'. A third blue arrow points from the 'Education' link to the word 'TECHNOLOGY'. A blue bracket on the right side of these three words is labeled 'mixture of topics'. In the background, there is a colorful graphic for 'CALCULUS single variable' with 'CHAPTER 1 FUNCTIONS' and 'LECTURE 1 FUNCTIONS' written on it. A blue arrow points from this graphic to the 'FINANCE' label.

©Emily Fox 2014

2

Latent Dirichlet Allocation (LDA)

each topic k is a distribution over words in vocab, β_k , just as before

Topics

gene	0.04
dna	0.02
genetic	0.01
...	...

Life

evolve	0.02
organism	0.01
...	...

brain

neuron	0.04
nerve	0.02
...	...

data

number	0.02
computer	0.01
...	...

previously, each doc had one topic

now, each is a mixture of topics

Topic proportions and assignments

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 125 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, says Stephen Alexander, a University of Maryland biologist. "It's never more than a few percent difference. But coming up with reasonable numbers is a difficult task, and more attention has to be paid to the sequencing." "It may be a way of organizing any newly sequenced genome," explains Araceli Moshgogiani, a computational molecular biologist at the National Center for Biotechnology Information, in Bethesda, Maryland, Computer Science Division.

Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 312 • 24 MAY 1996

every word is assigned to a topic

each doc has its own prevalence of topics in that doc

©Emily Fox 2014

Latent Dirichlet Allocation (LDA)

Topics

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 125 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, says Stephen Alexander, a University of Maryland biologist. "It's never more than a few percent difference. But coming up with reasonable numbers is a difficult task, and more attention has to be paid to the sequencing." "It may be a way of organizing any newly sequenced genome," explains Araceli Moshgogiani, a computational molecular biologist at the National Center for Biotechnology Information, in Bethesda, Maryland, Computer Science Division.

Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 312 • 24 MAY 1996

All we see are words — β_k 's

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

©Emily Fox 2014

LDA Generative Model

- Observations: $w_1^d, \dots, w_{N_d}^d$
- Associated topics: $z_1^d, \dots, z_{N_d}^d$ ← *topic per word in doc d*
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

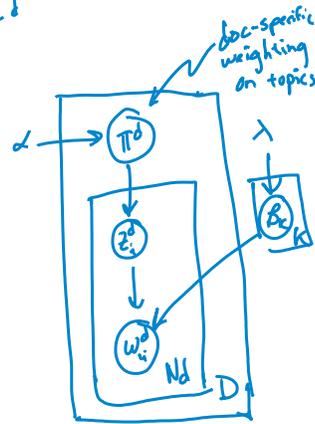
$$z_i^d \sim \pi^d \quad d=1, \dots, D \\ i=1, \dots, N_d$$

$$w_i^d | z_i^d \sim \beta_{z_i^d}$$

priors:

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

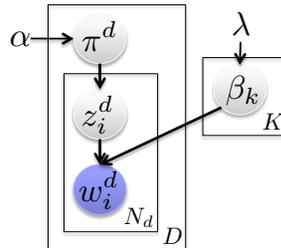
$$\beta_k \sim \text{Dir}(\lambda_1, \dots, \lambda_V) \quad k=1, \dots, K$$



©Emily Fox 2014

5

LDA Joint Probability



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

©Emily Fox 2014

6

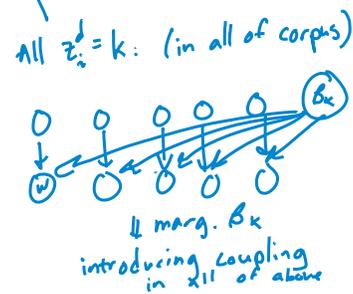
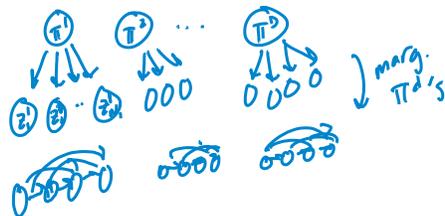
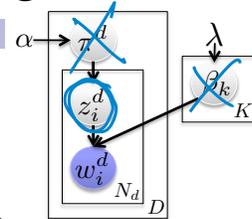
Collapsed LDA Sampling

- Marginalize parameters
 - Document-specific topic weights
 - Corpus-wide topic-specific word distributions

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

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

- Unplate to see dependencies induced



All $z_i^d = k$: (in all of corpus)

©Emily Fox 2014

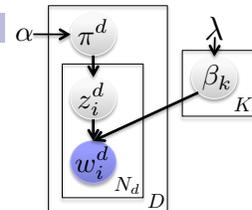
7

Collapsed LDA Sampling

- Sample topic indicators for each word
 - Algorithm:

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

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



©Emily Fox 2014

8

Select a Document

Etruscan	trade	price	temple	market

©Emily Fox 2014

9

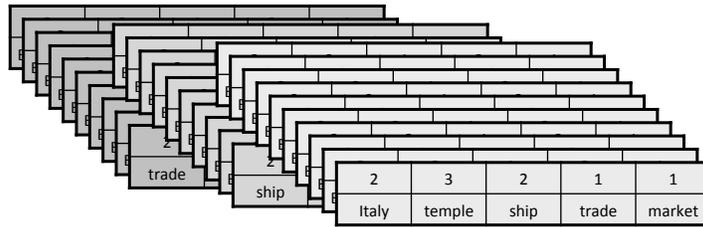
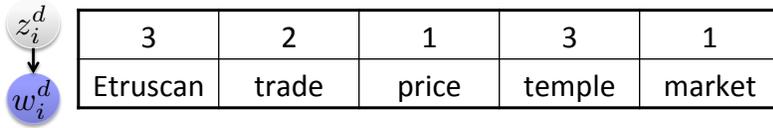
Randomly Assign Topics

z_i^d	3	2	1	3	1
w_i^d	Etruscan	trade	price	temple	market

©Emily Fox 2014

10

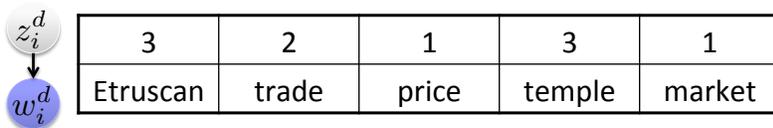
Randomly Assign Topics



©Emily Fox 2014

11

Maintain Local Statistics

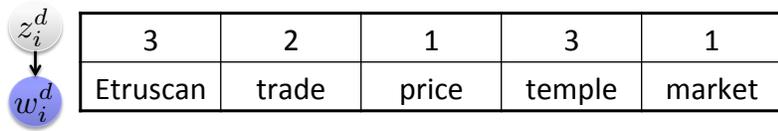


	Topic 1	Topic 2	Topic 3
Doc d			

©Emily Fox 2014

12

Maintain Global Statistics



	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
...			

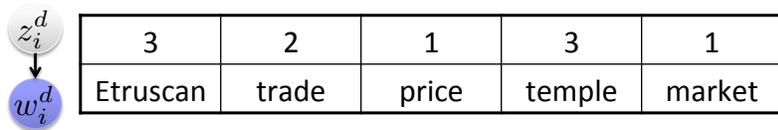
	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

Total counts from **all** docs

©Emily Fox 2014

13

Resample Assignments



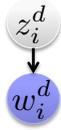
	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

©Emily Fox 2014

14

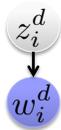
What is the conditional distribution for this topic?



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?



3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1

Topic 2

Topic 3

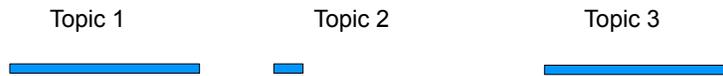


	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

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?

z_i^d	3	?	1	3	1
w_i^d	Etruscan	trade	price	temple	market



	Topic 1	Topic 2	Topic 3
trade	10	7	1

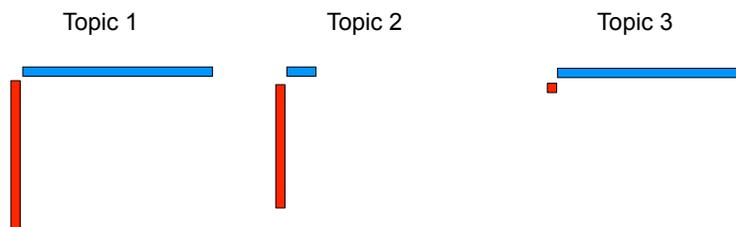
©Emily Fox 2014

17

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
w_i^d	Etruscan	trade	price	temple	market



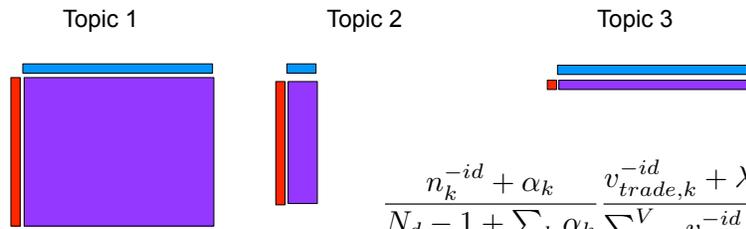
©Emily Fox 2014

18

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
w_i^d	Etruscan	trade	price	temple	market



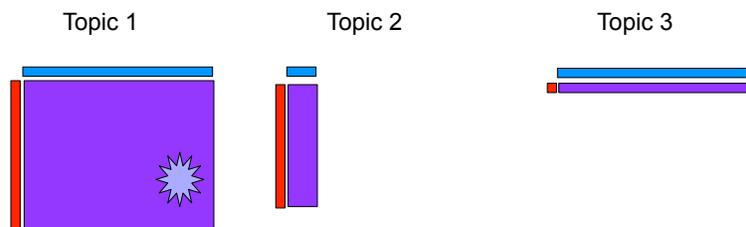
$$\frac{n_k^{-id} + \alpha_k}{N_d - 1 + \sum_k \alpha_k} \frac{v_{trade,k}^{-id} + \lambda_{trade}}{\sum_{\gamma=1}^V v_{\gamma,k}^{-id} + \lambda_{\gamma}}$$

©Emily Fox 2014

19

Sample a New Topic Indicator

z_i^d	3	?	1	3	1
w_i^d	Etruscan	trade	price	temple	market



©Emily Fox 2014

20

Update Counts

z_i^d
 w_i^d

3	?	1	3	1
Etruscan	trade	price	temple	market

	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	1
...			

	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

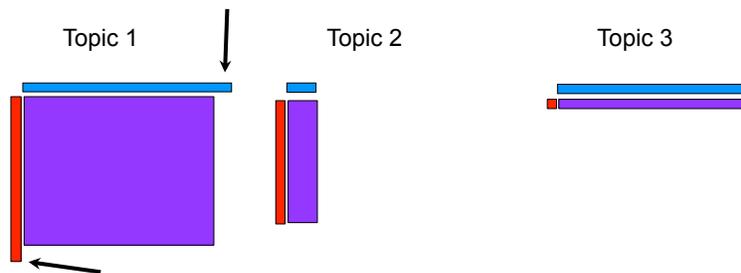
©Emily Fox 2014

21

Geometrically...

z_i^d
 w_i^d

3	1	1	3	1
Etruscan	trade	price	temple	market



©Emily Fox 2014

22

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

March 11th, 2014

Variational Methods Goal

- 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:

Variational Methods Cartoon

- Cartoon of goal:
- 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

- If $p = q$ for all θ

- Otherwise,

©Emily Fox 2014

27

A Measure of Closeness

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

- Not symmetric
- p determines where the difference is important:
 - $p(\theta)=0$ and $q(\theta)\neq 0$

 - $p(\theta)\neq 0$ and $q(\theta)=0$

- Want

- Just as hard as the original problem!

©Emily Fox 2014

28

Reverse Divergence

- Divergence $D(p \parallel q)$
 - true distribution p defines support of diff.
 - the “correct” direction
 - will be intractable to compute
- Reverse divergence $D(q \parallel p)$
 - approximate distribution defines support
 - tends to give overconfident results
 - will be tractable

Interpretations of Minimizing Reverse KL

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

- Similarity measure:

- Evidence lower bound (ELBO)

Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)

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

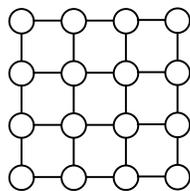
- Therefore,
 - ELBO provides a lower bound on marginal likelihood
 - Maximizing ELBO is equivalent to minimizing KL

©Emily Fox 2014

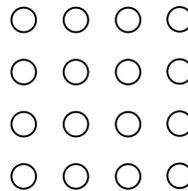
31

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



Original graph



Naive mean field

©Emily Fox 2014

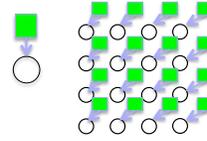
32

Mean Field

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

- Naïve mean field decomposition:

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



- Under this approximation, entropy term decomposes as

- Can (always) rewrite joint term as

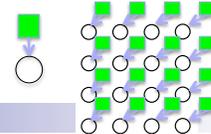
$$E_q[\log p(\theta, z, x)] = E_q[\log p(\theta | z, x)] + E_q[\log p(z, x)]$$

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

©Emily Fox 2014

33

Mean Field – Optimize γ



- Examine one free parameter, e.g., γ

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

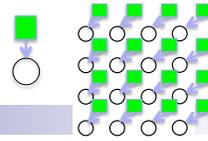
- Look at terms of ELBO just depending on γ

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

©Emily Fox 2014

34

Mean Field – Optimize ϕ^i



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

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

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

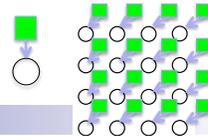
$$\mathcal{L}^{\phi^i} =$$

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

©Emily Fox 2014

35

Algorithm Outline



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

$$q_z^{(t)} = \arg \max_{q_z} \mathcal{L}(q_z, q_{\theta}^{(t-1)})$$
- **M-Step:** Given posterior distributions, find likely parameters

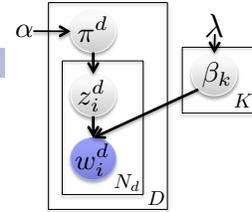
$$q_{\theta}^{(t)} = \arg \max_{q_{\theta}} \mathcal{L}(q_z^{(t)}, q_{\theta})$$
- **Iteration:** Alternate E-step & M-step until convergence

©Emily Fox 2014

36

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

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

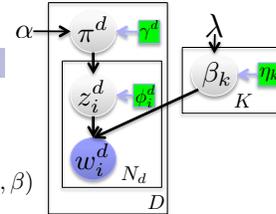
©Emily Fox 2014

37

Mean Field for LDA

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

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



- Examine the ELBO

$$\begin{aligned} \mathcal{L}(q) = & \sum_{k=1}^K E_q[\log p(\beta_k | \lambda)] + \sum_{d=1}^D E_q[\log p(\pi^d | \alpha)] \\ & + \sum_{d=1}^d \sum_{i=1}^{N_d} E_q[\log p(z_i^d | \pi^d)] + E_q[\log p(w_i^d | z_i^d, \beta)] \\ & - \sum_{k=1}^K E_q[\log q(\beta_k | \eta_k)] - \sum_{d=1}^D E_q[\log q(\pi^d | \gamma^d)] - \sum_{d=1}^d \sum_{i=1}^{N_d} E_q[\log q(z_i^d | \phi_i^d)] \end{aligned}$$

©Emily Fox 2014

38

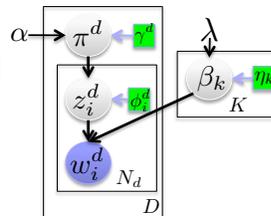
Mean Field for LDA

- Let's look at some of these terms

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

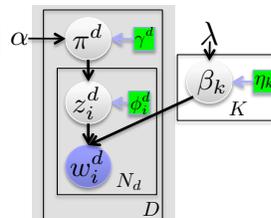
$$E_q[\log q(z_i^d | \phi_i^d)]$$

- Other terms follow similarly



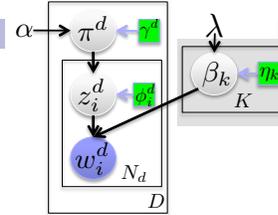
Optimize via Coordinate Ascent

- Algorithm:



Optimize via Coordinate Ascent

- Algorithm:



©Emily Fox 2014

41

Generalizing... $\log p(x) \geq \int q_z(z) q_\theta(\theta) \log \frac{p(x, z, \theta)}{q_z(z) q_\theta(\theta)} dz d\theta$

- Condition 1: Complete data likelihood is in exponential family
- Condition 2: Parameter prior is conjugate to complete data likelihood

EM for MAP estimation	Variational Bayesian EM
Goal: maximise $p(\theta x)$ w.r.t. θ E Step: compute $q_z^{(t+1)}(z) = p(z x, \theta^{(t)})$ M Step: $\theta^{(t+1)} = \arg \max_{\theta} \int q_z^{(t+1)}(z) \ln p(z, x, \theta) dz$	Goal: lower bound $p(x)$ VB-E Step: compute $\bar{\phi}^{(t)} = \mathbb{E}_{q_\theta^{(t)}}[\phi(\theta)]$ $q_z^{(t+1)}(z) = p(z x, \bar{\phi}^{(t)})$ VB-M Step: $q_\theta^{(t+1)}(\theta) \propto \exp \left[\int q_z^{(t+1)}(z) \ln p(z, x, \theta) dz \right]$

©Emily Fox 2014

42

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
 - Mean field for LDA

Reading

- **Mixed Membership Models: KM Sec. 27.3**
 - Basic LDA:
[Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 \(2003\): 993-1022.](#)
 - Introduction:
[Blei, David M. "Probabilistic topic models." Communications of the ACM, vol. 55, no. 4 \(2012\): 77-84.](#)
 - Sampling:
[Griffith, Thomas L. and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America, Volume: 101, Supplement: 1 \(2004\): Pages: 5228-5235](#)

Acknowledgements

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