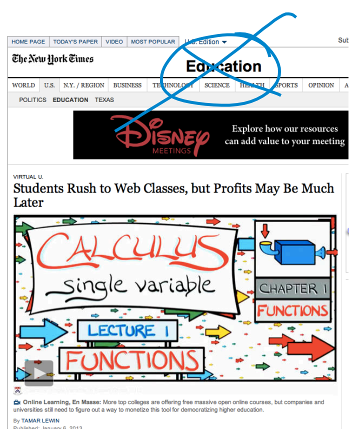
Case Study 5: Mixed Membership Modeling

Latent Dirichlet Allocation
Collapsed Gibbs Sampler, Variational Methods

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
March 11th, 2014

Task 3: Mixed Membership Models

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



EDUCATION

FINANCE

TECHNOLOGY

mixture of topics

Latent Dirichlet Allocation (LDA)

"Global" params

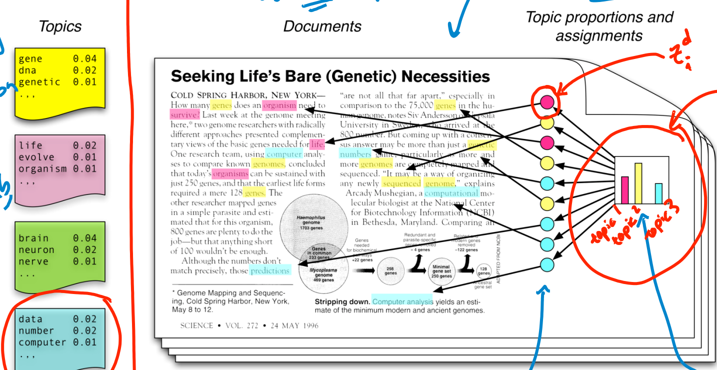
each topic k is a distribution over words in vocab β_k, just as before

β_k

1 2 ... V

previously, each doc had one topic

now, each is a mixture of topics



z_i^d

π^d

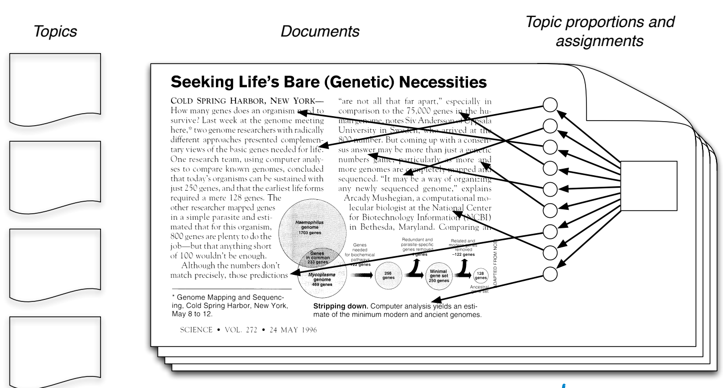
topic 1 topic 2 topic 3

β_k

every word is assigned to a topic

each doc has its own prevalence of topics in that doc

Latent Dirichlet Allocation (LDA)



Obs: All we see are words

Want: posterior p(topics, doc prop. of topics, assign. vars. | words in docs)

β_k's

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:

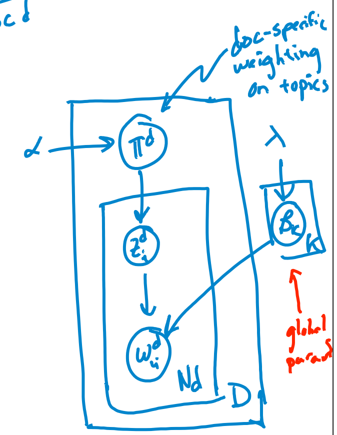
$z_i^d \sim \pi^d \quad d=1,\dots,D \quad i=1,\dots,N_d$

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

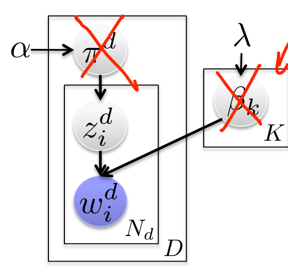
Priors:

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

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



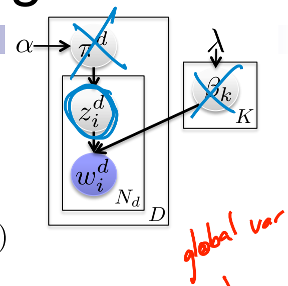
LDA Joint Probability



marginalize params and just sample topic indicators z_i^d

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

Collapsed LDA Sampling



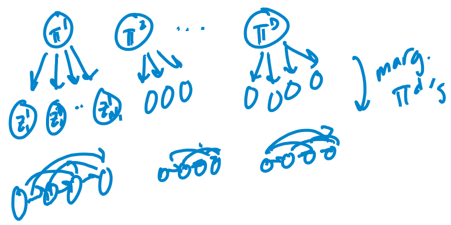
- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

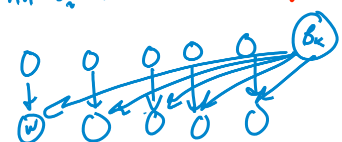
$$p(z_i^d = k \mid z_{\backslash id}, \{w_i^d\}, \alpha, \lambda)$$
$$\propto p(z_i^d = k \mid z_{\backslash id}, \alpha) p(w_i^d \mid z_i^d = k, z_{\backslash id}, w_{\backslash id}, \lambda)$$

- Unplate to see dependencies induced

separately for each doc (π^d = local var)

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

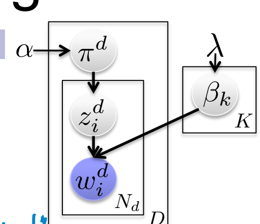




⇓ marg. β_k

introducing coupling in all of above

Collapsed LDA Sampling



- Sample topic indicators for each word
  - Algorithm:

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

"prior" "likelihood"

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

$$\propto \frac{n_k^{-id} + \alpha_k}{\underbrace{N_d - 1 + \sum \alpha_k}} \cdot \frac{m_{w_i^d,k}^{-id} + \lambda_{w_i^d}}{\underbrace{\sum_\gamma (m_{\gamma,k}^{-id} + \lambda_\gamma)}_{\text{over whole corpus}}}$$

words assigned to topic k in doc d not counting i^{th} word

normalize within doc

$m_{\gamma,k}^{-id}$ = # of times word γ appears in topic k (not counting w_i^d)

Select a Document

Etruscan	trade	price	temple	market

all words in doc d

Randomly Assign Topics

$z_i^d \rightarrow w_i^d$

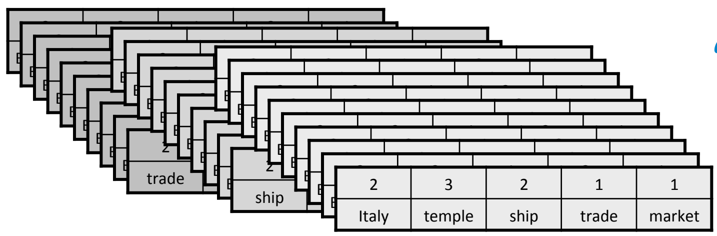
3	2	1	3	1
Etruscan	trade	price	temple	market

one approach
to initialize sampler

Randomly Assign Topics

$z_i^d \downarrow w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market



do for all docs in corpus

Maintain Local Statistics

$z_i^d \downarrow w_i^d$

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

$z_i^d \to w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

	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
…			

$m_{\text{trade},2}$ ← word, topic

Total counts from **all** docs

©Emily Fox 2014 13

Resample Assignments

$z_i^d \to w_i^d$

3	$\cancel{2}$ z_2^d	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	2	0 $\cancel{1}$	2

n_k^{-id}

	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 $\cancel{8}$	1
…			

$m_{\text{word,topic}}^{-id}$

©Emily Fox 2014 14

What is the conditional distribution for this topic?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

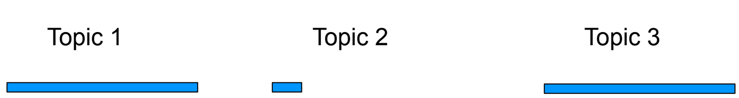
$p(z_i^d \mid \text{everything else})$

What is the conditional distribution for this topic?

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

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market



$$\frac{n_k^{-id} + \alpha_k}{N_d - 1 + \sum \alpha_k}$$

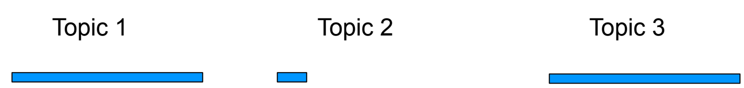
"prior" term

	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



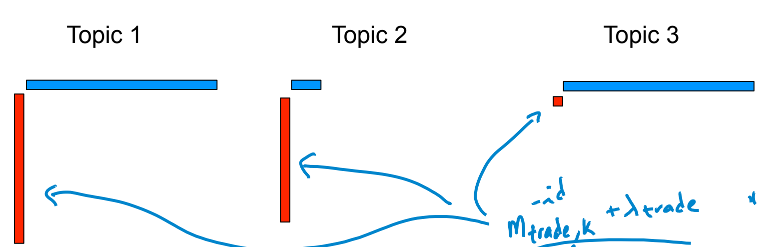
row from global table

	Topic 1	Topic 2	Topic 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	3	?	1	3	1
w_i^d	Etruscan	trade	price	temple	market



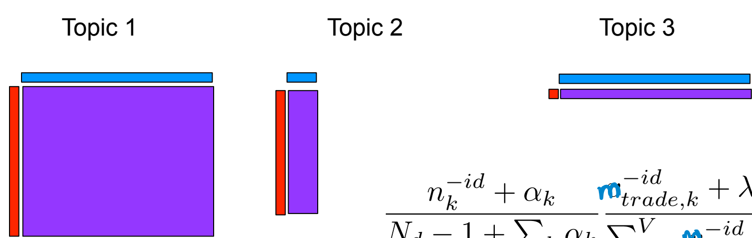
$$\frac{m_{trade,k}^{-id} + \lambda_{trade}}{\sum_{\gamma}(m_{\gamma,k} + \lambda_{\gamma})}$$

"likelihood" term

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



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

Sample a New Topic Indicator

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

happened to sample topic 1

Update Counts

$z_i^d \rightarrow w_i^d$

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

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

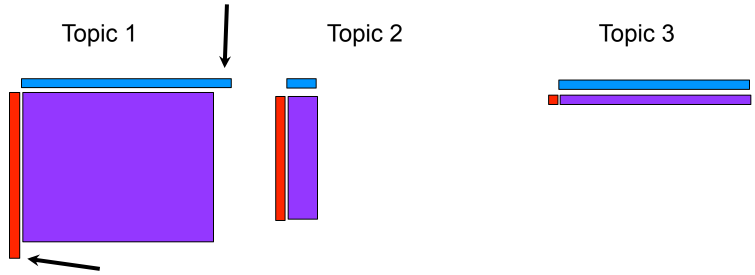
	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	11 ~~10~~	7	1
…			

Geometrically…

inc. popularity of topic 1 in doc d
and word prevalence for topic 1 in corpus

$z_i^d \rightarrow w_i^d$

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
March 11th, 2014

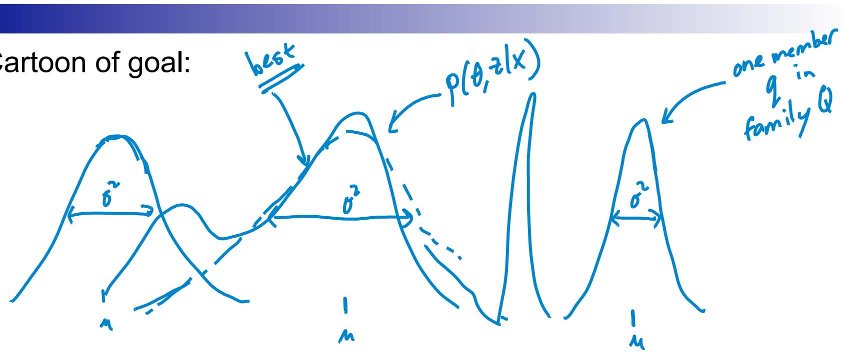
Variational Methods Goal

- Recall task: Characterize the posterior $p(\theta, z \mid x)$ obs. params latent vars
- 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 the family Q and want $q \in Q$ that is closest to $p(\theta, z \mid x)$

Variational Methods Cartoon

- Cartoon of goal:



e.g., Q: all Gaussians

- 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 $D(p\|q) \neq D(q\|p)$ ← $\int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$ → not a true distance metric
- p determines where the difference is important:
  - $\exists\theta$ p(θ)=0 and q(θ)≠0 $0 \log 0 = 0$
  - $\exists\theta$ p(θ)≠0 ($=\epsilon$) and q(θ)=0 $\epsilon \log \frac{\epsilon}{0} = \infty$

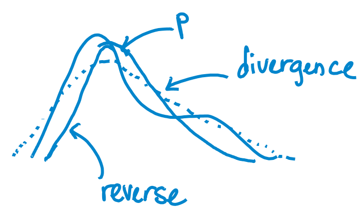
If $D(p\|q)$ finite, $\mathrm{supp}(q) \supseteq \mathrm{supp}(p)$

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

Reverse Divergence

- Divergence D(p || q)
  - 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 ← q now less diffuse than p
  - will be tractable

what we have control over



Interpretations of Minimizing Reverse KL

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

- Similarity measure:

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

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

- Evidence lower bound (ELBO)

$$\log p(x) = \underbrace{D(q(z,\theta)||p(\theta,z|x))}_{\geq 0} + \mathcal{L}(q) \geq \mathcal{L}(q)$$

"ELBO"

Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)

← log marginal likelihood or "evidence"

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

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 of $\log p(x)$

what we can control

depends on what we don't know

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

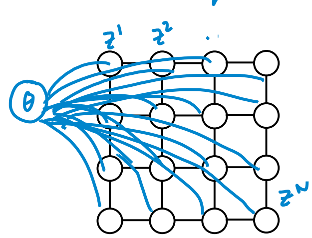
- How do we choose a Q such that the following is tractable?

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

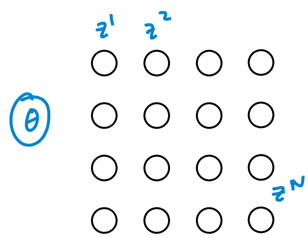
- Simplest case = mean field approximation $\theta, z = \{z^1, \ldots, z^N\}$
  - Assume each parameter and latent variable is conditionally independent given the set of free parameters

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

$\gamma, \{\phi^i\}$ are "free params" = control knobs in getting q close to p



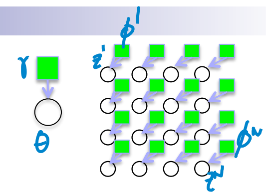
Original graph



Naïve mean field

← (can also look at "structured" mean field approx (breaking only some depend.))

Mean Field $\mathcal{L}(q) = E_q[\log p(z,\theta,x)] - \overbrace{E_q[\log q(z,\theta)]}^{\text{entropy}}$



- Naïve mean field decomposition:

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

- Under this approximation, entropy term decomposes as

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

decouples across γ, ϕ^i

- Can (always) rewrite joint term as   full cond. of θ

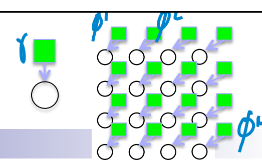
$$E_q[\log p(\theta,z,x)] = E_q[\log p(\theta \mid z,x)] + E_q[\log p(z,x)]$$

OR   full cond. of z^i

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

©Emily Fox 2014 33

Mean Field – Optimize γ



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

$$\mathcal{L}(q) = \underbrace{E_q[\log p(\theta \mid z,x)] + E_q[\log p(z,x)]}_{\text{consider } \theta\text{-full-cond. form}} - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

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

  - Look at terms of ELBO just depending on γ

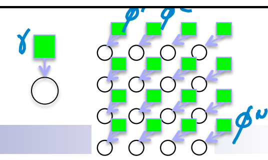
$$\mathcal{L}^\gamma = E_q[\log p(\theta \mid z,x)] - E_q[\log q(\theta \mid \gamma)] + \text{const.}$$

w.r.t. γ

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

©Emily Fox 2014 34

Mean Field – Optimize ϕ^i



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

$$\mathcal{L}(q) = \underbrace{E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)]}_{\text{consider the } z^i\text{-full-cond. form}} + E_q[\log p(z_{\setminus i}, \theta, x)] - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

const. wrt ϕ^i

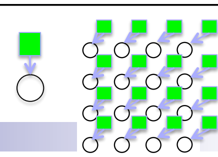
  - Look at terms of ELBO just depending on ϕ^i

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

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

- 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 (latent vars z)

optimize ϕ $\rightarrow$ $$q_z^{(t)} = \arg\max_{q_z} \mathcal{L}(q_z, q_\theta^{(t-1)})$$

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

optimize γ $\rightarrow$ $$q_\theta^{(t)} = \arg\max_{q_\theta} \mathcal{L}(q_z^{(t)}, q_\theta)$$

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

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

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