Case Study 2: Document Retrieval

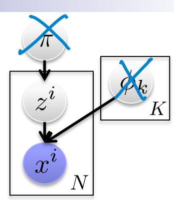
Collapsed Gibbs and Variational Methods for LDA

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox

February 7th, 2013

Example – Collapsed MoG Sampling

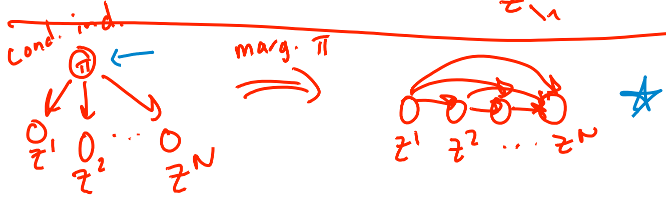
$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$



- Collapsed sampler

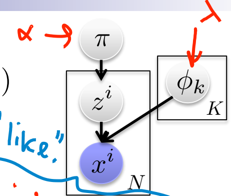
For $i=1,\dots,N$

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



Example – Collapsed MoG Sampling

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



- Derivation

$$p(z^i \mid z_{\setminus i}, x_{1:N}, \alpha, \lambda) \propto \underbrace{p(z^i \mid z_{\setminus i}, \alpha)}_{\text{"prior"}} \underbrace{p(x^i \mid z^i, z_{\setminus i}, x_{\setminus i}, \lambda)}_{\text{"like."}}$$

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

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

- Important facts: $z^i \sim \pi$, $\pi \sim \mathrm{Dir}$

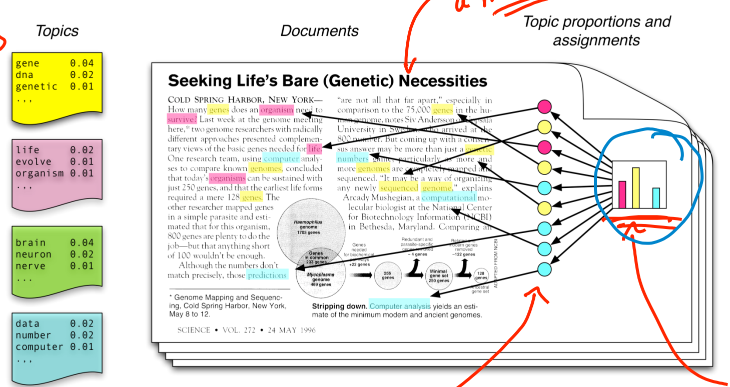
$$p(z_{1:N} \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_k + \alpha_k)}{\Gamma(\sum_k n_k + \alpha_k)} \qquad \frac{\Gamma(m+1)}{\Gamma(m)} = m$$

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

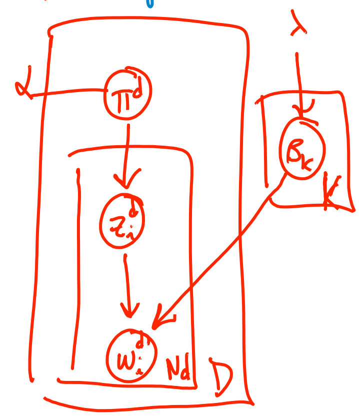
each doc has its own prevelance of topics in doc

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 "global param"
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model: doc-specific preferences of topics

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

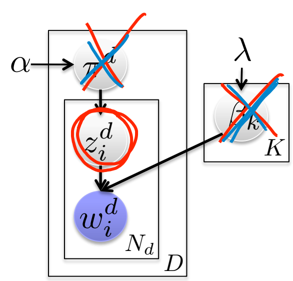
$w_i^d \mid z_i^d \sim \beta_{z_i^d}$ $\quad i=1,\dots,N$



Priors:

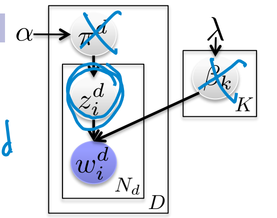
$$\begin{cases} \pi^d \sim \mathrm{Dir}(\alpha_1,\dots,\alpha_K) & d=1,\dots,D \\ \beta_k \sim \mathrm{Dir}(\lambda_1,\dots,\lambda_V) & k=1,\dots,K \end{cases}$$

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:

$z_i^d \sim \pi^d \qquad \pi^d \sim \mathrm{Dir}$

# of assign. to topic k in doc d

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

$w_i^d \mid z_i^d = k \sim \beta_k \qquad \beta_k \sim \mathrm{Dir}$

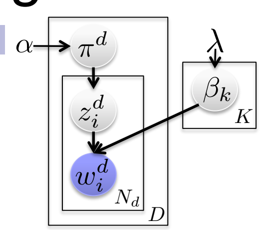
$\forall i, \forall d$

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

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

# of assign. of word v to topic k

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

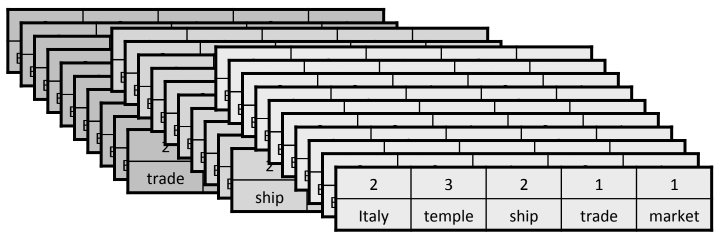
$z_i^d \rightarrow w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market

Randomly Assign Topics

$z_i^d \rightarrow w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market



Maintain Global Statistics

$z_i^d \rightarrow w_i^d$

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
...			

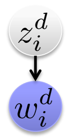
What is the conditional distribution for this topic?

$$z_i^d \downarrow w_i^d$$

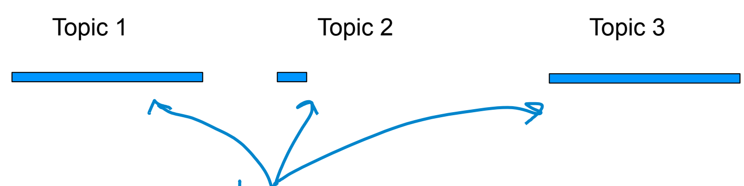
3	?	1	3	1
Etruscan	trade	price	temple	market

What is the conditional distribution for this topic?

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



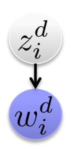
3	?	1	3	1
Etruscan	trade	price	temple	market



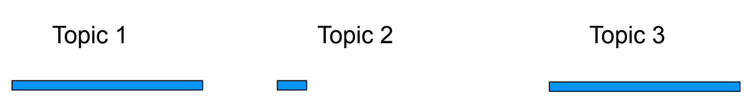
$$\frac{n_k^d + \alpha_k}{\sum_j n_j^d + \alpha_j}$$

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



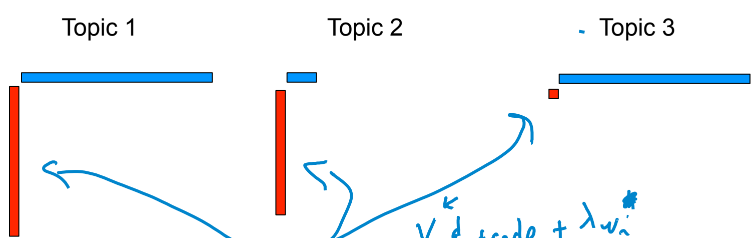
	1	2	3
trade	10	7	1

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



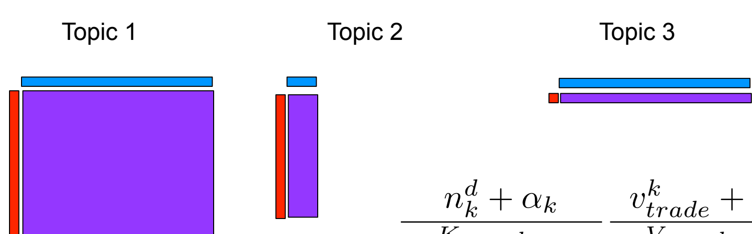
$$\frac{v^k_{w_i^d = \text{trade}} + \lambda_{w_i^d}}{\sum_w v^k_w + \lambda_v}$$

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

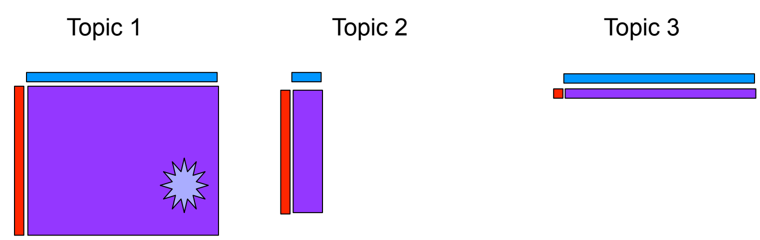


$$\frac{n_k^d + \alpha_k}{\sum_{j=1}^K n_j^d + \alpha_j} \frac{v_{trade}^k + \lambda_k}{\sum_{j=1}^V v_j^k + \lambda_j}$$

Sample a New Topic Indicator

$z_i^d \rightarrow w_i^d$

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



Update Counts

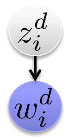
$z_i^d \rightarrow w_i^d$

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

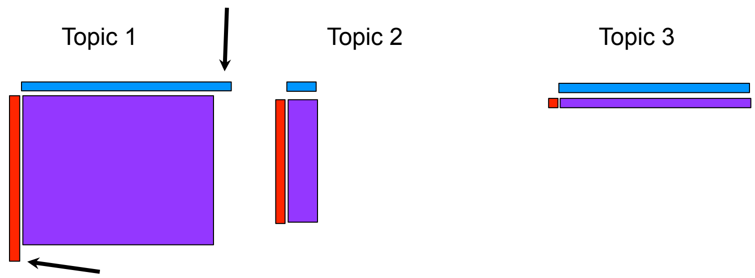
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	11 ~~10~~	7	1
…			

Geometrically…

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



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

- 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*

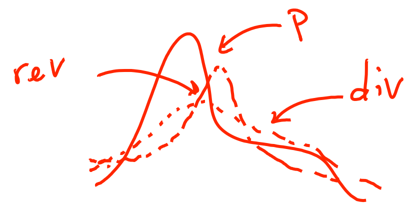
$$KL(p\|q) \triangleq D(p\|q) = E_p\left[\log \frac{p}{q}\right] \quad \left(\int_\theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta\right)$$

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

  If $D(p\|q)$ finite, $\mathrm{supp}(q) \supseteq \mathrm{supp}(p)$
- Want $\hat{q} = \arg\min_{q \in \mathcal{Q}} D(p\|q)$
- 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



Interpretations of Minimizing Reverse KL

- Similarity measure:

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

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

- Evidence lower bound (ELBO)

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

add to a const

- 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 $\log p(x)$ ← entropy of q

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

Mean Field

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

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

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

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

"free params"

- Then, entropy term decomposes as

$$-E_q[\log q(z,\theta)] = -E_{\underline{q(\theta|\gamma)}}[\log q(\theta \mid \gamma)]$$

decouples across γ, ϕ^n

$$- \sum_n E_{q(z^n|\phi^n)}[\log q(z^n \mid \phi^n)]$$

Mean Field

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

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

  - Look at terms of ELBO just depending on γ

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

under $q(\cdot)$ $z^i \perp\!\!\!\perp \theta$

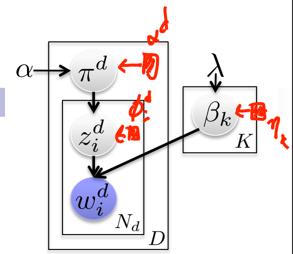
"full cond."

- Likewise,

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

- 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

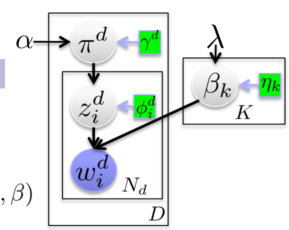
$$q(\pi,\beta,z) = \prod_{k=1}^{K} \underbrace{q(\beta_k \mid \eta_k)}_{Dir(\eta_{k1},\dots,\eta_{kV})} \prod_{d=1}^{D} \underbrace{q(\pi^d \mid \alpha^d)}_{Dir(\alpha^d_1,\dots,\alpha^d_K)} \prod_{i=1}^{N_d} \underbrace{q(z_i^d \mid \phi_i^d)}_{Mult(\phi_i^d)}$$

$\sum_k \phi^d_{ik} = 1$ need to enforce this

- 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

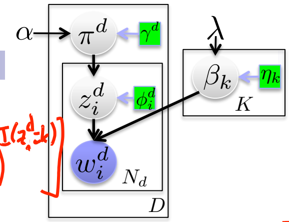
$$\mathcal{L}(q) = \sum_{k=1}^{K} E_q[\log p(\beta_k \mid \lambda)] + \sum_{d=1}^{D} E_q[\log p(\pi^d \mid \alpha)]$$
$$+ \sum_{d=1}^{d} \sum_{i=1}^{N_d} E_q[\log p(z_i^d \mid \pi^d)] + E_q[\log p(w_i^d \mid z_i^d, \beta)]$$

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



- Let's look at some of these terms

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

$= \sum_{k=1}^{K} E_q[I(z_i^d=k) \log \pi_k^d] = \sum_{k=1}^{K} \underbrace{E_q[I(z_i^d=k)]}_{\phi_{ik}^d} \underbrace{E_q[\log \pi_k^d]}_{\Psi(\gamma_k^d) - \Psi(\sum_k \gamma_k^d)}$

$z_i^d \perp\!\!\!\perp \pi_k^d$ given free params under $q(\cdot)$

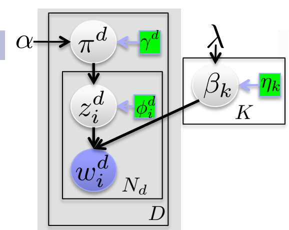
$\Rightarrow$ why mean fied is so important

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

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

- 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 $i=1,\dots,N_d$

$\frac{\partial \mathcal{L}}{\partial \phi_i^d} = 0 \rightarrow \phi_i^d \propto \exp\{\Psi(\gamma^{d(t+1)}) + \Psi(\eta^{(t)}_{\cdot w_i^d}) - \Psi(\sum_v \eta^{(t)}_{\cdot v})\}$

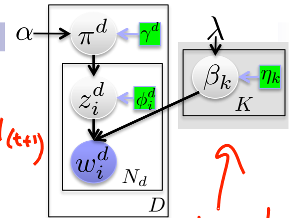
use Lagrange multipliers to enforce pmf

DATA PARALLEL

Optimize via Coordinate Ascent

- Algorithm:

$$\frac{\partial \mathcal{L}}{\partial \eta_k} = 0 \rightarrow \eta_k^{(t+1)} = \lambda + \underbrace{\sum_{d=1}^{D}\sum_{i=1}^{N_d} w_i^d \phi_i^{d(t+1)}}_{\text{aggregate}}$$



hard

Map Reduce

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.

$$\theta^{(t+1)} = \theta^{(t)} + \rho_t \underbrace{\nabla_\theta \mathcal{L}(\theta)}_{\text{again, need to touch all docs}}$$

$$\nabla_\theta \mathcal{L} = E_x[\nabla_\theta \mathcal{L}(\theta, x)] \approx \frac{1}{M}\sum_{t=1}^{M} \nabla_\theta \mathcal{L}(\theta, x^t)$$

x^t sampled iid

Alternative Optimization Schemes

- Recall stochastic gradient ascent:
  - Assume M = 1

$$\nabla_\theta \mathcal{L}(\theta) \approx \nabla_\theta \mathcal{L}(\theta, x^t) \triangleq \nabla_\theta \mathcal{L}_t$$

  - Unbiased, but noisy $\quad E_x[\nabla_\theta \mathcal{L}_t] = \nabla_\theta \mathcal{L}(\theta)$

- Here, LDA

ELBO

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

just doc t

t-ELBO

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

as if we saw doc t D times

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 $\gamma^d = \alpha + \sum_{i=1}^{N_d} \phi_i^d$

just like in coord. asc. for this doc

  - Take a stochastic gradient step $\eta^{(t)} = \eta^{(t-1)} + \rho_t \nabla_\eta \mathcal{L}_d$

$$\left(\lambda + D\sum_{i=1}^{N_d} \phi_{i,w_i}^d - \eta^{(t-1)}\right)$$

$$\eta^{(t)} = (1-\rho_t)\eta^{(t-1)} + \rho_t\left(\quad\right)$$

looks exactly like coord. asc. update for doc t D times

Acknowledgements

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