

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

©Emily Fox 2013

1

Example – Collapsed MoG Sampling

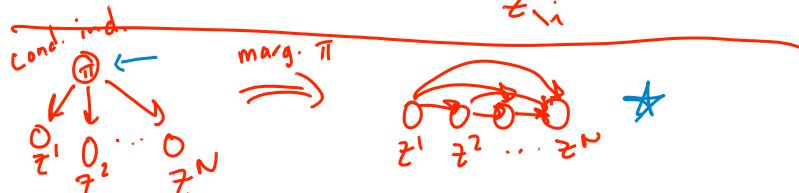
$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad z^i \sim \pi$$

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

- Collapsed sampler

For $i=1, \dots, N$

$$z^i \sim p(z^i | z^{1:(i-1)}, \dots, z^{i-1}, z^{i+1:(N)}, \dots, z^N, x_{1:N}, \alpha, \mu, \Sigma)$$



©Emily Fox 2013

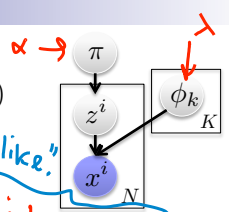
2

Example – Collapsed MoG Sampling

$$\pi \sim \text{Dir}(\alpha_1, \dots, \alpha_K) \quad z^i \sim \pi$$

$$\{\mu_k, \Sigma_k\} \sim F(\phi)$$

$$x^i | z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$



Derivation

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

$$p(z^i = k | z_{-i}, \alpha) = \int p(z^i = k | \pi) p(\pi | z_{-i}, \alpha) d\pi = \frac{n_k^i + \alpha_k}{N - 1 + \sum \alpha_k}$$

$$p(x^i | z_{-i}, x_{-i}, \lambda) = \text{student-t pred. likelihood}$$

Important facts:

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

$$\frac{\Gamma(m+1)}{\Gamma(m)} = m$$

©Emily Fox 2013

3

Latent Dirichlet Allocation (LDA)

each doc is a mixture of these corpus-wide topics

each topic as a dist. over words {br?}

every word is assigned to a topic

each doc has its own prevalence of topics in doc

Topics

gene	0.04
dna	0.02
genetic	0.01
...	...
life	0.02
evolve	0.01
organism	0.01
...	...
brain	0.04
neuron	0.02
nerve	0.01
...	...
data	0.02
number	0.02
computer	0.01
...	...

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 **computational** analyses to compare known **organisms**, concluded that today's **organisms** can be sustained with just 250 genes, small that the earliest life forms required a mere 125 genes. The other researcher mapped genes to 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 25,000 genes in the human genome, notes Steve Anderson, a University of Maryland researcher. "The coming of such estimates may be more than just a **milestone** in genomics," he says. "It may be a way of organizing any newly **sequenced genomes**," explains Arechi Mishigatan, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. **Computational** analyses of genomes are becoming more **available** and **accurate**, and more **genomes** are being **sequenced**.

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

SCIENCE • VOL. 317 • 24 MAY 1996

Topic proportions and assignments

©Emily Fox 2013

4

LDA Generative Model

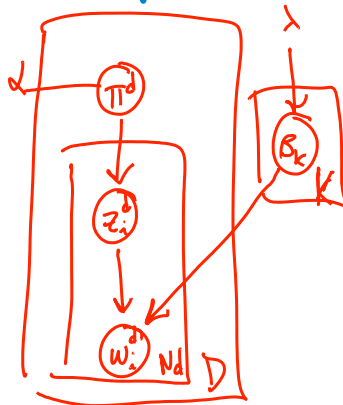
- Observations: $w_1^d, \dots, w_{N_d}^d$ $d=1, \dots, D$
- Associated topics: $z_1^d, \dots, z_{N_d}^d$ *corpus-wide topic "global param"*
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model: *doc-specific preferences of topics*

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

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

Priors:

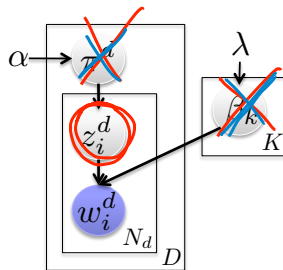
$$\begin{cases} \pi^d \sim \text{Dir}(\alpha_1, \dots, \alpha_K) & d=1, \dots, D \\ \beta_k \sim \text{Dir}(\lambda_1, \dots, \lambda_V) & k=1, \dots, K \end{cases}$$



©Emily Fox 2013

5

LDA Generative Model



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

©Emily Fox 2013

6

Collapsed LDA Sampling

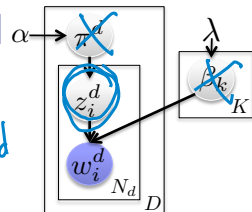
- Marginalize parameters
 - Document-specific topic weights
 - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
 - Derivation:

Handwritten notes: $z_i^d \sim \pi^d$, $\pi^d \sim \text{Dir}$, $w_i^d | z_i^d = k \sim \beta_k$, $\beta_k \sim \text{Dir}$.
Annotations: "# of assign. to topic k in doc d" (pointing to n_k^d), "# of assign. of word v to topic k" (pointing to v^k).

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

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

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

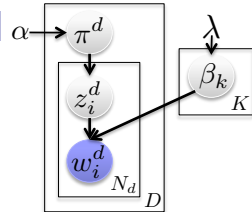


©Emily Fox 2013

7

Collapsed LDA Sampling

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



©Emily Fox 2013

8

Sample Document

Etruscan	trade	price	temple	market

©Emily Fox 2013

9

Randomly Assign Topics

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

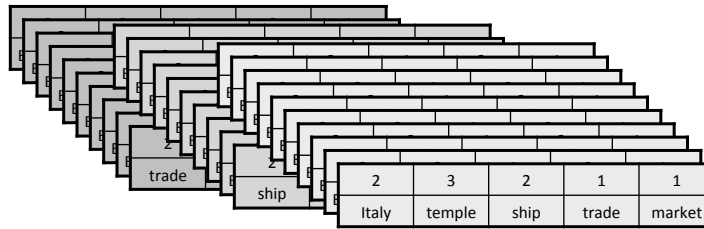
©Emily Fox 2013

10

Randomly Assign Topics

z_i^d
 w_i^d

3	2	1	3	1
Etruscan	trade	price	temple	market



©Emily Fox 2013

11

Maintain Global Statistics

z_i^d
 w_i^d

3	2	1	3	1
Etruscan	trade	price	temple	market

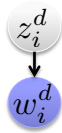
Total counts from all docs

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

©Emily Fox 2013

12

Resample Assignments




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
...			

©Emily Fox 2013

13

What is the conditional distribution for this topic?



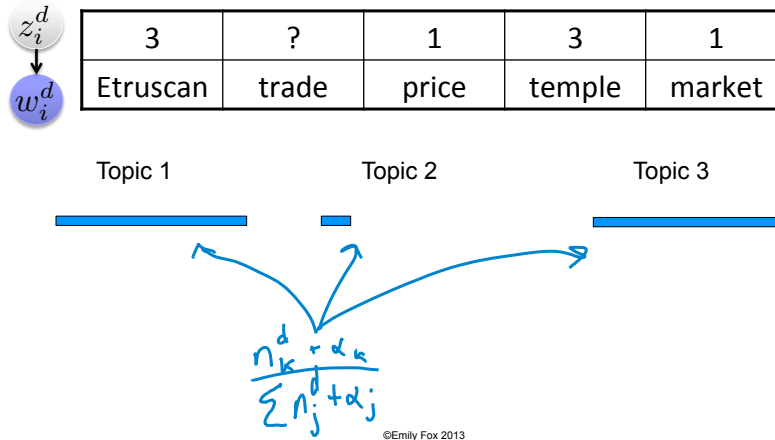
3	?	1	3	1
Etruscan	trade	price	temple	market

©Emily Fox 2013

14

What is the conditional distribution for this topic?

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

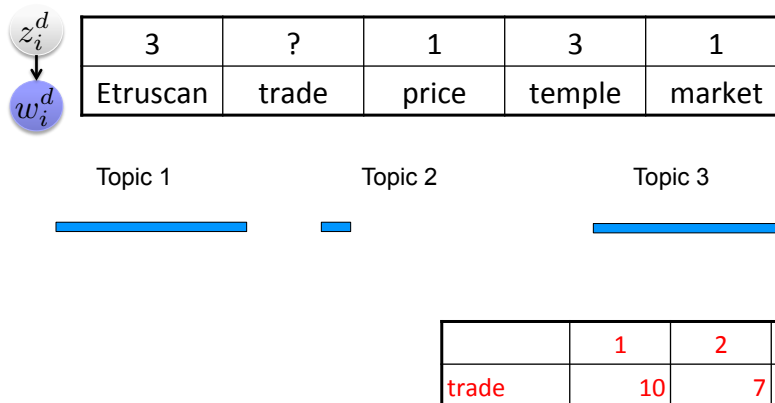


©Emily Fox 2013

15

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?



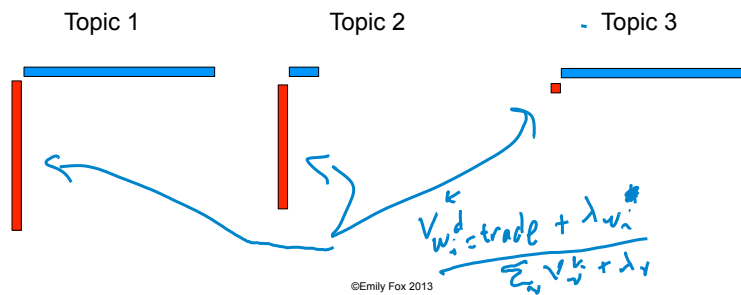
©Emily Fox 2013

16

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



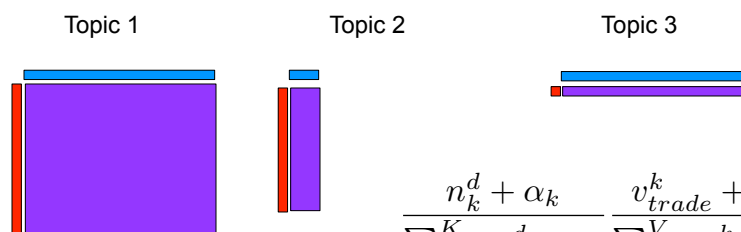
©Emily Fox 2013

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

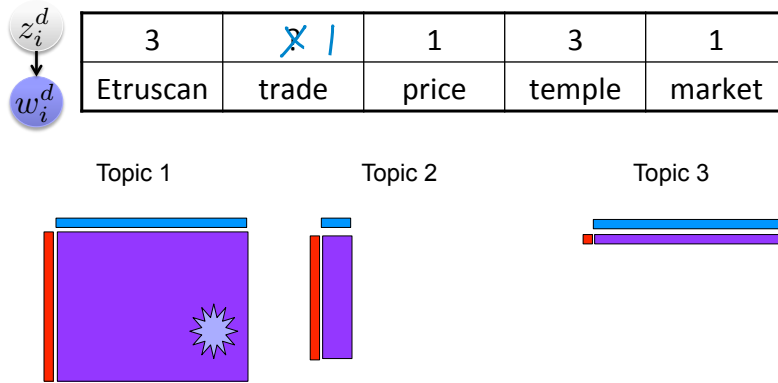


$$\frac{n_k^d + \alpha_k}{\sum_{j=1}^K n_j^d + \alpha_j} \frac{v_{trade}^k + \lambda_k}{\sum_{j=1}^V v_j^k + \lambda_j}$$

©Emily Fox 2013

18

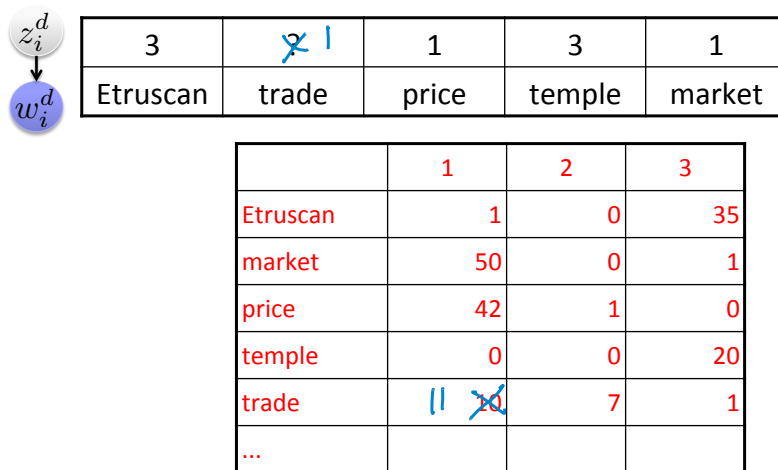
Sample a New Topic Indicator



©Emily Fox 2013

19

Update Counts

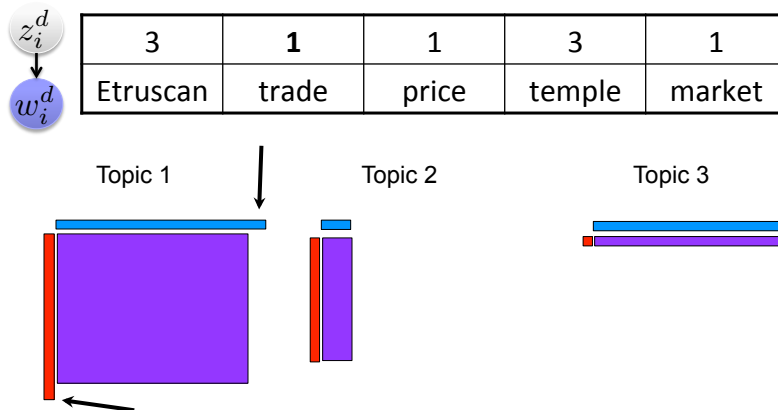


©Emily Fox 2013

20

Geometrically...

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



©Emily Fox 2013

21

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

©Emily Fox 2013

22

Variational Methods

- Recall task: Characterize the posterior $p(\theta, z | x)$
 - 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: $p(\theta, z | x)$

call the family Q and want $q \in Q$ that is closest to $p(\theta, z | x)$
- Questions:
 - How do we measure “closeness”?
 - If the posterior is intractable, how can we approximate something we do not have to begin with?

©Emily Fox 2013

23

A Measure of Closeness

- Kullback-Leibler (KL) divergence
 - Measures “distance” between two distributions p and q

$$KL(p||q) \triangleq D(p||q) = E_p[\log \frac{p}{q}] \quad \left(\int_{\theta} p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta \right)$$
- Not symmetric $D(p||q) \neq D(q||p)$... not a true distance
- p determines where the difference is important:
 - $\exists x \square p(x)=0$ and $q(x) \neq 0$ $0 \log 0 = 0$
 - $\exists x \square p(x) \neq 0$ and $q(x)=0$ $\in \log \frac{\infty}{0} = \infty$

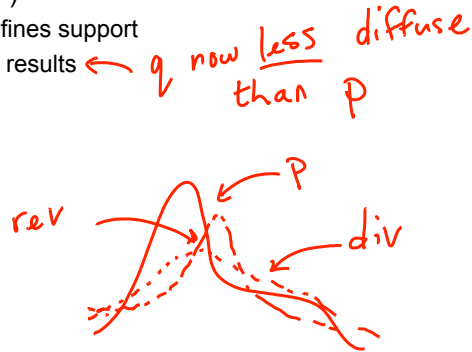
If $D(p||q)$ finite, $\text{supp}(q) \supseteq \text{supp}(p)$
- Want $\hat{q} = \underset{q \in Q}{\text{argmin}} D(p||q)$
- Just as hard as the original problem! $E_p[\dots]$

©Emily Fox 2013

24

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



©Emily Fox 2013

25

Interpretations of Minimizing Reverse KL

- Similarity measure:

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

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

$- \mathcal{L}(q)$ $\neq \log p(x)$

- Evidence lower bound (ELBO)

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

$\underbrace{\log p(x)}_{\text{const.}}$ $\xrightarrow{\text{add to a const}}$ $\underline{\underline{\mathcal{L}(q)}}$

- Therefore, minimizing KL is equivalent to maximizing a lower bound on the marginal likelihood:

- Max $\mathcal{L}(q) = \min D(q \parallel p) = \max$ lower bound of $\log p(x)$

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

\leftarrow entropy of q

©Emily Fox 2013

26

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

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

- Then, entropy term decomposes as

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

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

decouples across γ, ϕ^n

"free params"

Mean Field

- Examine one free parameter, e.g., γ

- Can rewrite joint as *always*

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

- Look at terms of ELBO just depending on γ

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

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

"full cond."

- Likewise,

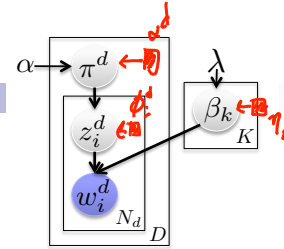
$$\mathcal{L}^{\phi^n} = -E_q[\log q(z^n | \phi^n)] + E_q[\log p(z^n | z_{-n}, x, \theta)] + \text{const.}$$

- 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

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

$\text{Dir}(\eta_{k1}, \dots, \eta_{kV})$ $\text{Dir}(\alpha^d_1, \dots, \alpha^d_K)$ $\text{Mult}(\phi_i^d)$

$\sum_k \phi_{ik}^d = 1$
 need to enforce this

- The joint distribution factorizes as

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

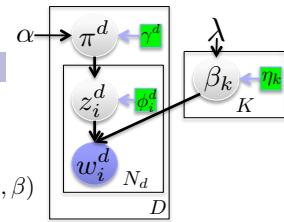
©Emily Fox 2013

29

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

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

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

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

} from joint

all terms from q

©Emily Fox 2013

30

Mean Field for LDA

- Let's look at some of these terms

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

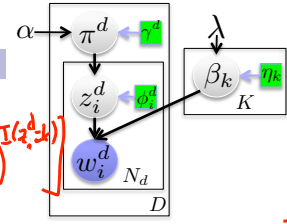
$$= \sum_{k=1}^K E_q[I(z_i^d=k) \log \pi_k^d] = \sum_{k=1}^K E_q[I(z_i^d=k)] E_q[\log \pi_k^d]$$

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

\Rightarrow why mean field is so important

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

ϕ_{ik}^d given



Optimize via Coordinate Ascent

- Algorithm:

For $d=1, \dots, D$

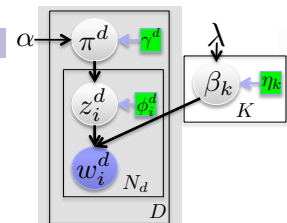
$$\frac{\partial \mathcal{L}}{\partial \gamma^d} = 0 \rightarrow \gamma^{d(t+1)} = \alpha + \sum_{i=1}^{N_d} \phi_i^d$$

For $i=1, \dots, N_d$

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

use Lagrange multipliers to enforce pmf

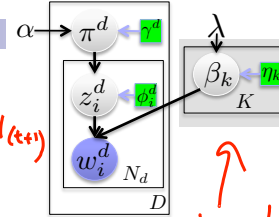
DATA PARALLEL



Optimize via Coordinate Ascent

- Algorithm:

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



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.

$$\theta^{(t+1)} = \theta^{(t)} + \rho_t \nabla_{\theta} \mathcal{L}(\theta)$$

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 $E_x[\nabla_{\theta} \mathcal{L}_t] = \nabla_{\theta} \mathcal{L}(\theta)$

- Here, **LDA**

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

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

ELBO (pointing to the first two terms)

just doc t (pointing to the last two terms)

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

$$+ D (E_q[\log p(z^t, x^t | \pi^t, \beta)] - E_q[\log q(z^t)])$$

t-ELBO (pointing to the first two terms)

as if we saw doc t D times (pointing to the last two terms)

©Emily Fox 2013

35

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 $i=1, \dots, N_d$

$$\phi_{ik}^d \propto \exp\{E[\log \pi_k^d] + E[\log \beta_{k,w_i^d}]\}$$

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

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

just like in coord. asc. for this doc

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

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

©Emily Fox 2013

36

Acknowledgements

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