Case Study 2: Document Retrieval

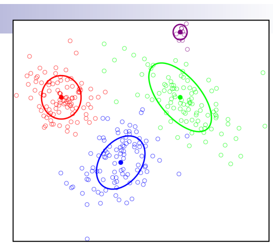
MAP EM, Latent Dirichlet Allocation, Gibbs Sampling

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

February 5th, 2013

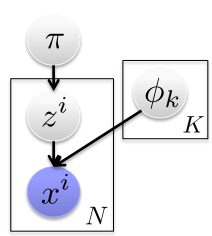
Gaussian Mixture Model

- Most commonly used mixture model
- Observations: $x^1, \ldots, x^N$



- Parameters: $\theta = \{\pi, \phi\}$

$$\pi = [\pi_1, \ldots, \pi_K]$$
$$\phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$$



- Likelihood:

$$p(x^i \mid \theta) = \sum_k \pi_k p(x^i \mid \phi_k)$$

- Ex. z^i = country of origin, x^i = height of ith person
  - kth mixture component = distribution of heights in country k

Motivates EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration *t*: $\hat{\theta}^{(t)}$

- **E-Step**

  Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$

- **M-Step**

  Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

MAP Estimation

- Bayesian approach:
  - Place **prior** $p(\theta)$ on parameters
  - Infer **posterior** $p(\theta \mid x)$

- Many, many, many motivations and implications
  - For the sake of this class, simplest motivation is to think of this as akin to regularization

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \log p(\theta \mid x)$$

  - Saw importance of regularization in logistic regression (ML estimate can overfit data and lead to poor generalization)

EM Algorithm – MAP Case

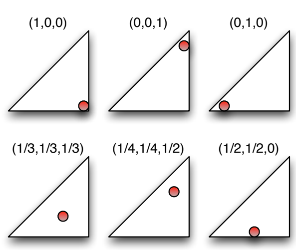
- Re-derive EM algorithm for $p(\theta \mid x)$
- Add $\log p(\theta)$ to $U(\theta, \hat{\theta}^{(t)})$
  - What must be computed in E-Step remains unchanged because this term does not depend on *y*.
  - M-Step becomes:

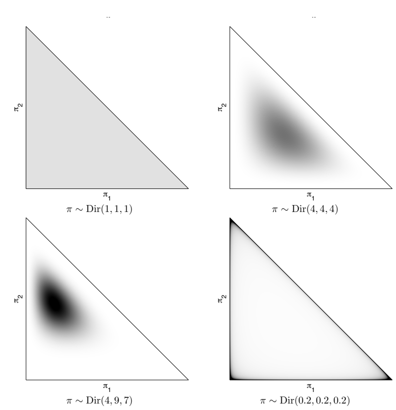
$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$$

MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$





MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \qquad p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

- Dirichlet posterior
  - Assume we condition on observations $z^i \sim \pi$
  - Count occurrences of $z^i = k$
  - Then,

$$p(\pi \mid \alpha, z^1, \ldots, z^N) \propto$$

  - Conjugacy: This **posterior** has same form as **prior**

MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \quad \{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(m_0, \kappa_0, \nu_0, S_0)$$

- Results in following M-Step:

$$\hat{\mu}_k = \frac{r_k \bar{x}_k + \kappa_0 m_0}{r_k + \kappa_0} \qquad \hat{\pi}_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}$$

$$\hat{\Sigma}_k = \frac{S_0 + r_k S_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)'}{\nu_0 + r_k + d + 2}$$

Posterior Computations

- MAP EM focuses on point estimation:

$$\hat{\theta}^{MAP} = \arg\max_{\theta} p(\theta \mid x)$$

- What if we want a full characterization of the posterior?
  - Maintain a measure of uncertainty
  - Estimators other than posterior mode (different loss functions)
  - Predictive distributions for future observations
- Often no closed-form characterization (e.g., mixture models)
- Alternatives:
  - Monte Carlo based estimates using samples from posterior
  - Variational approximations to posterior (more next time)

Gibb Sampling

- Want draws:
- Construct Markov chain whose steady state distribution is
- Simplest case:

Example – Mixture of Gaussians

- Recall model
  - Observations: $x^1, \dots, x^N$
  - Cluster indicators: $z^1, \dots, z^N$
  - Parameters: $\theta = \{\pi, \phi\}$ $\quad \pi = [\pi_1, \dots, \pi_K]$

    $\phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$
  - Generative model:

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

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

- Want to draw posterior samples of model parameters

$$\pi \sim p(\pi \mid \phi, x^1, \dots, x^N)$$

$$\phi \sim p(\phi \mid \pi, x^1, \dots, x^N)$$

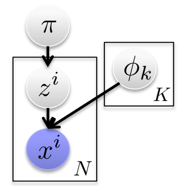
Auxiliary Variable Samplers

- Augment variables of interest θ with variables z to allow closed-form for sampling, just like in EM

- In both cases, simply looking at subchain $\{\theta^{(t)}\}$ converges to draws from marginal distribution $\pi(\theta)$

Example – Mixture of Gaussians

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



- Try auxiliary variable sampler
  - Introduce cluster indicators into sampler

Example – Clustering Results I

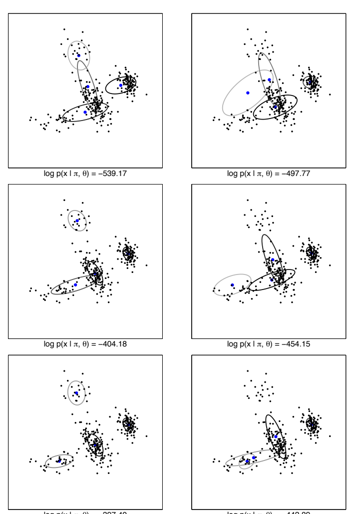


Figure courtesy of
Erik Sudderth

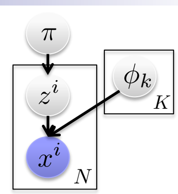
Collapsed Gibbs Samplers

- Marginalize a set of latent variables or parameters
  - Sometimes marginalized variables are nuisance parameters
  - Other times what gets marginalized are the variables
    - Make post-facto inferences on variables of interest based on sampled variables

- Can improve efficiency if marginalized variables are high-dim
  - Reduced dimension of search space
  - But, often introduces dependences!

Example – Collapsed MoG Sampling

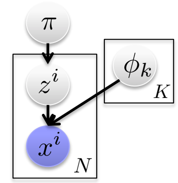
$$\pi \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K) \qquad z^i \sim \pi$$
$$\{\mu_k, \Sigma_k\} \sim F(\phi) \qquad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i})$$



- Collapsed sampler

Example – Collapsed MoG Sampling

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



- Derivation

- Important facts:

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

Example – Clustering Results II

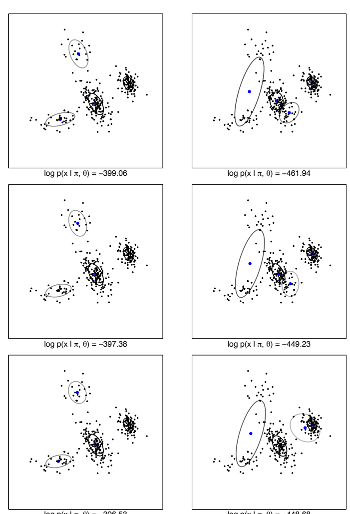


Figure courtesy of Erik Sudderth

Comparing Collapsed vs. Regular

Log Likelihood vs. Gibbs Iteration
(multiple chains)

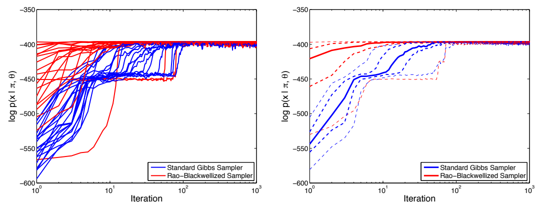


Figure courtesy of
Erik Sudderth

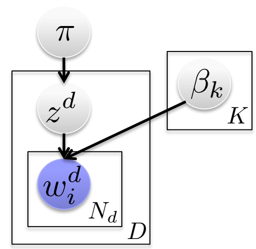
Task 2: Cluster Documents

- **Previously:**
  - Cluster documents based on topic

BBC WORLD NEWS

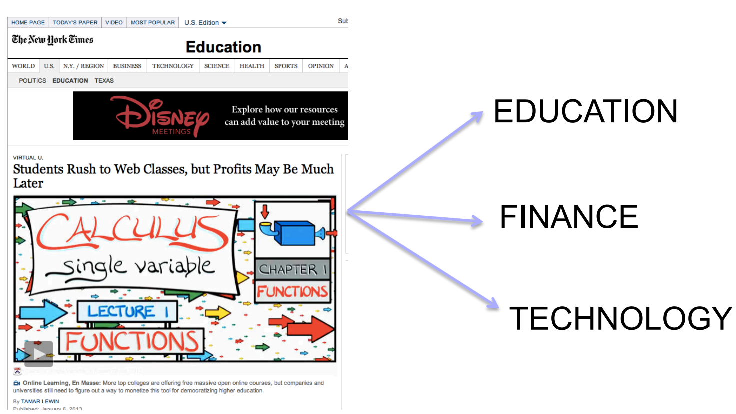
A Generative Model

- Documents: $x^1, \dots, x^D$
- Associated topics: $z^1, \dots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:



Task 2: Cluster Documents

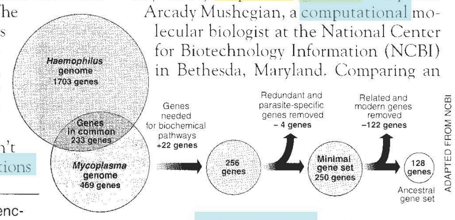
- **Now:** Document may belong to multiple clusters



Latent Dirichlet Allocation (LDA)

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 128 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.

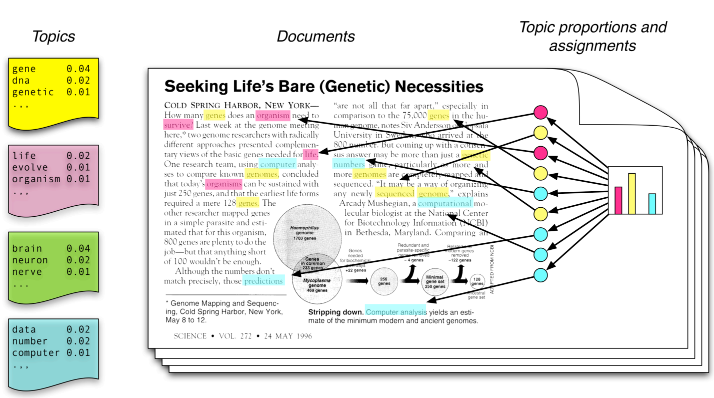
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



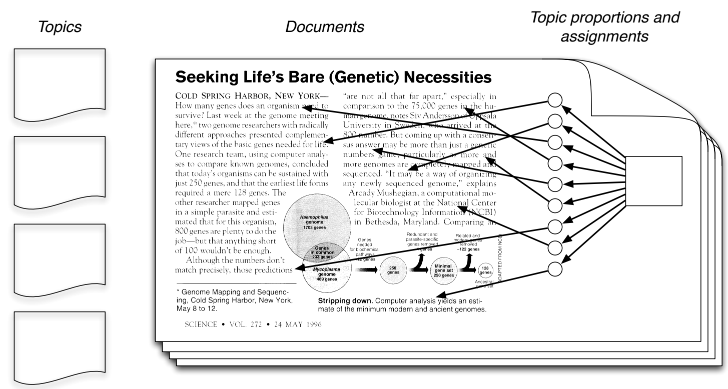
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

Latent Dirichlet Allocation (LDA)

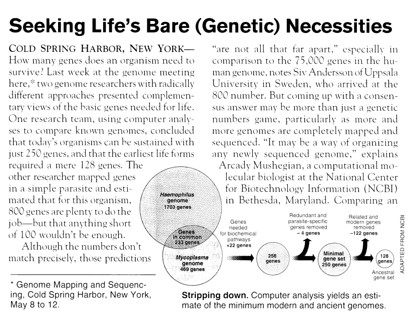


Latent Dirichlet Allocation (LDA)



Example Inference – Topic Weights

- **Data:** The OCR'ed collection of *Science* from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model



Seeking Life's Bare (Genetic) Necessities

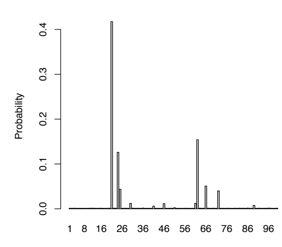
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 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, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

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

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

SCIENCE • VOL. 272 • 24 MAY 1996



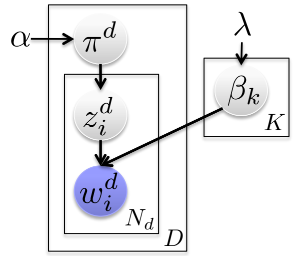
Example Inference – Topic Words

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA Generative Model

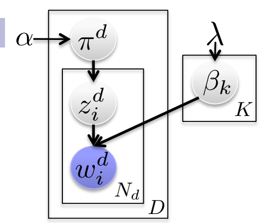
- Observations: $w_1^d, \ldots, w_{N_d}^d$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

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



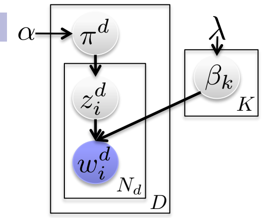
- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Derivation:

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

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

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

Collapsed LDA Sampling

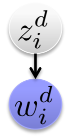


- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Algorithm:

Sample Document

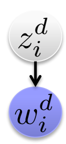
Etruscan	trade	price	temple	market

Randomly Assign Topics

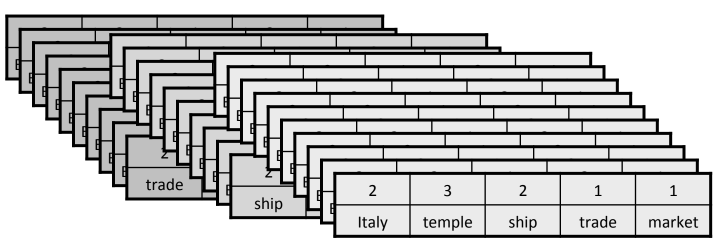


3	2	1	3	1
Etruscan	trade	price	temple	market

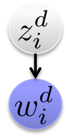
Randomly Assign Topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Maintain Global Statistics

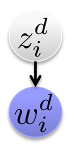


3	2	1	3	1
Etruscan	trade	price	temple	market

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
…			

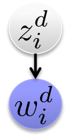
Resample Assignments



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
…			

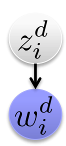
What is the conditional distribution for this topic?



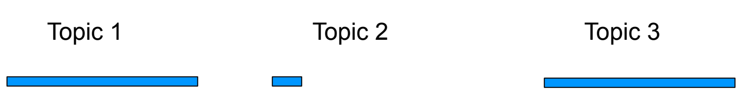
3	?	1	3	1
Etruscan	trade	price	temple	market

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



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

Topic 1 Topic 2 Topic 3

	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?

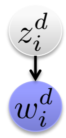
$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

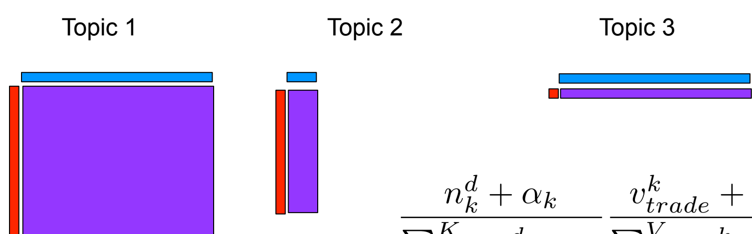
Topic 1 Topic 2 Topic 3

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?

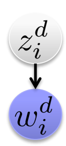


3	?	1	3	1
Etruscan	trade	price	temple	market

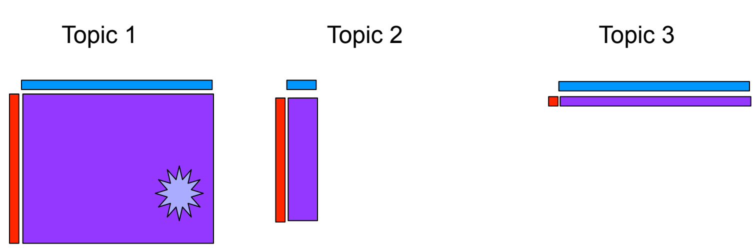


$$\frac{n_k^d + \alpha_k}{\sum_{j=1}^K n_j^d + \alpha_j} \frac{v_{trade}^k + \lambda_k}{\sum_{j=1}^V v_j^k + \lambda_j}$$

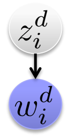
Sample a New Topic Indicator



3	?	1	3	1
Etruscan	trade	price	temple	market



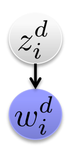
Update Counts



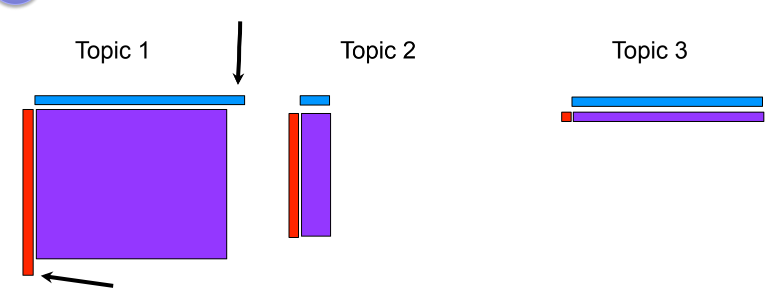
3	?	1	3	1
Etruscan	trade	price	temple	market

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

Geometrically…



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
- Next time: Variational methods instead of sampling

Acknowledgements

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

