Latent Dirichlet Allocation

David Blei

Andrew Ng

Michael Jordan

Outlines

- Notation and assumption
- Latent variable models: mixture of unigrams, probabilistic latent semantic indexing, latent Dirichlet allocation.
- A geometric interpretation
- Inference and estimation
- Experimental results

Outlines

- **Notation and assumption**
- Latent variable models
- A geometric interpretation
- Inference and estimation
- Experimental results

Notation and Terminology

- A *word* is the basic unit of discrete data, defined to be an item w from a vocabulary indexed by $\{1,\ldots,V\}$.
- A *document* is a sequence of N words denoted by $\mathbf{w}=\{w_1,w_2,\ldots,w_N\}$
- A *corpus* is a collection of M documents denoted by $D=\{\mathbf{w}_1,\mathbf{w}_2,\ldots\mathbf{w}_M\}$

Bag-of-words Assumption

- Word order is ignored
- "bag-of-words" – exchangeability
- A finite set of random variables $\{x_1, \ldots, x_N\}$ is said to be exchangeable if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(x_1, \ldots x_N) = p(x_{\pi(1)}, \ldots, x_{\pi(N)})$$

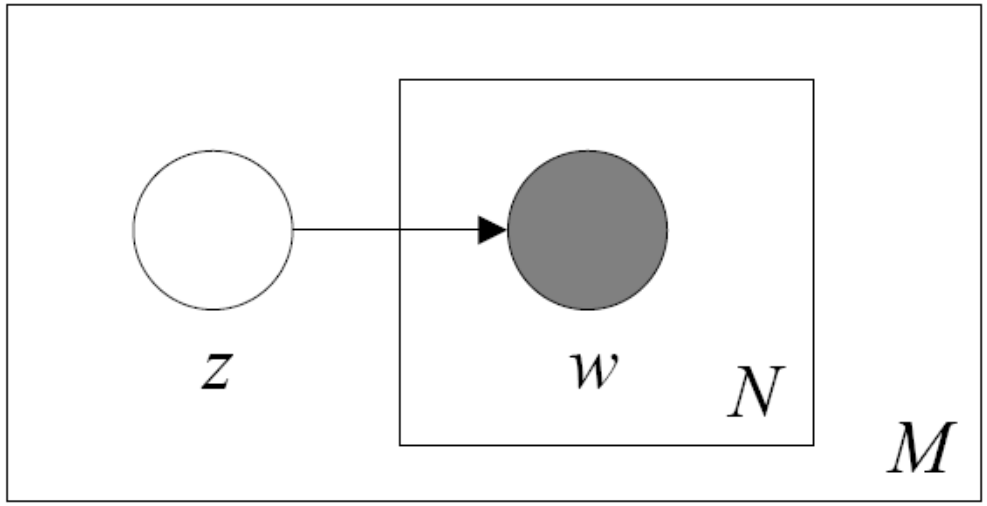
- **Theorem (De Finetti)** – if $(x_1, x_2, \ldots, x_N)$ are infinitely exchangeable, then the joint probability $p(x_1, x_2, \ldots, x_N)$ has a representation as a mixture:

$$p(x_1, x_2, \ldots, x_N) = \int p(\theta) \prod_{i=1}^{N} p(x_i | \theta) d\theta$$

Outlines

- Notation and assumption
- **Latent variable models**
- A geometric interpretation
- Inference and estimation
- Experimental results

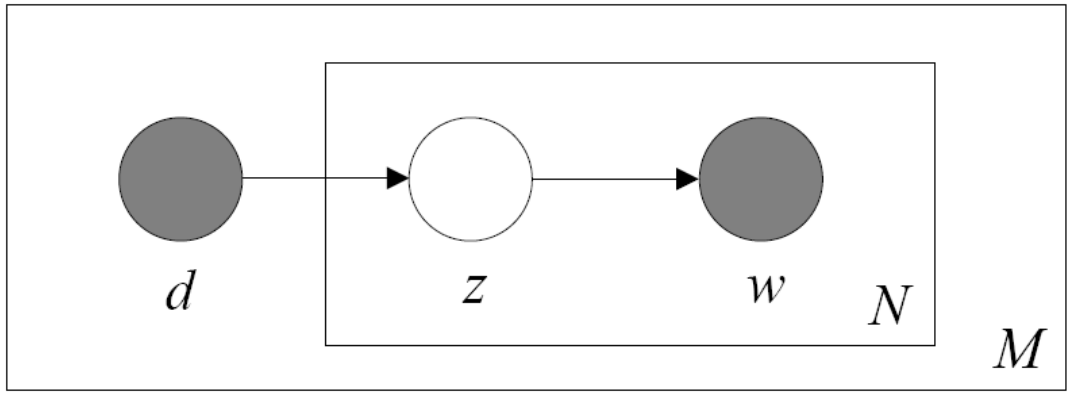
Mixture of Unigrams



$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$

- Each document exhibits only one topic

Probabilistic Latent Semantic Indexing



$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d)$$

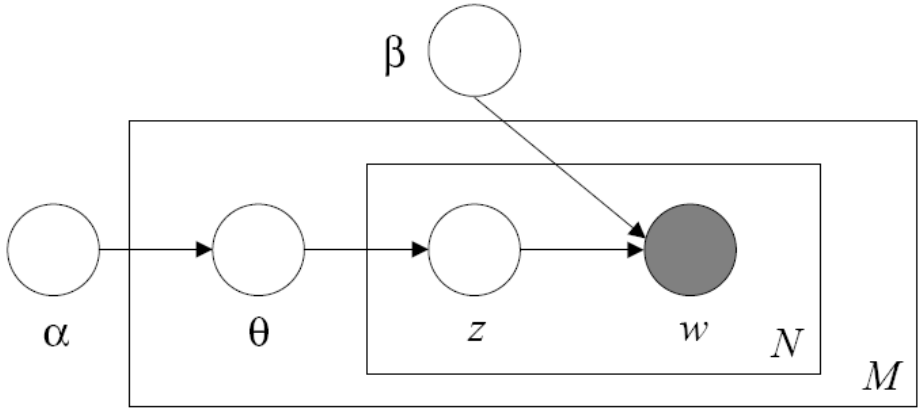
- Relaxes the assumption that each document is generated from only one topic

Probabilistic Latent Semantic Indexing

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d)$$

- *p(d)* is 0 for an unseen document.
- pLSI learns the topic mixture weights *p(z|d)* only for trained documents. It cannot assign probability to an unseen document. It is not a well defined generative model.
- *p(z|d)* needs *kM* parameters which is linearly grows with *M* – overfitting.

Latent Dirichlet Allocation



For each document:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$. $\quad p(\theta \mid \alpha) = \dfrac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$
3. For each of the N words w_n:

   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.

   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

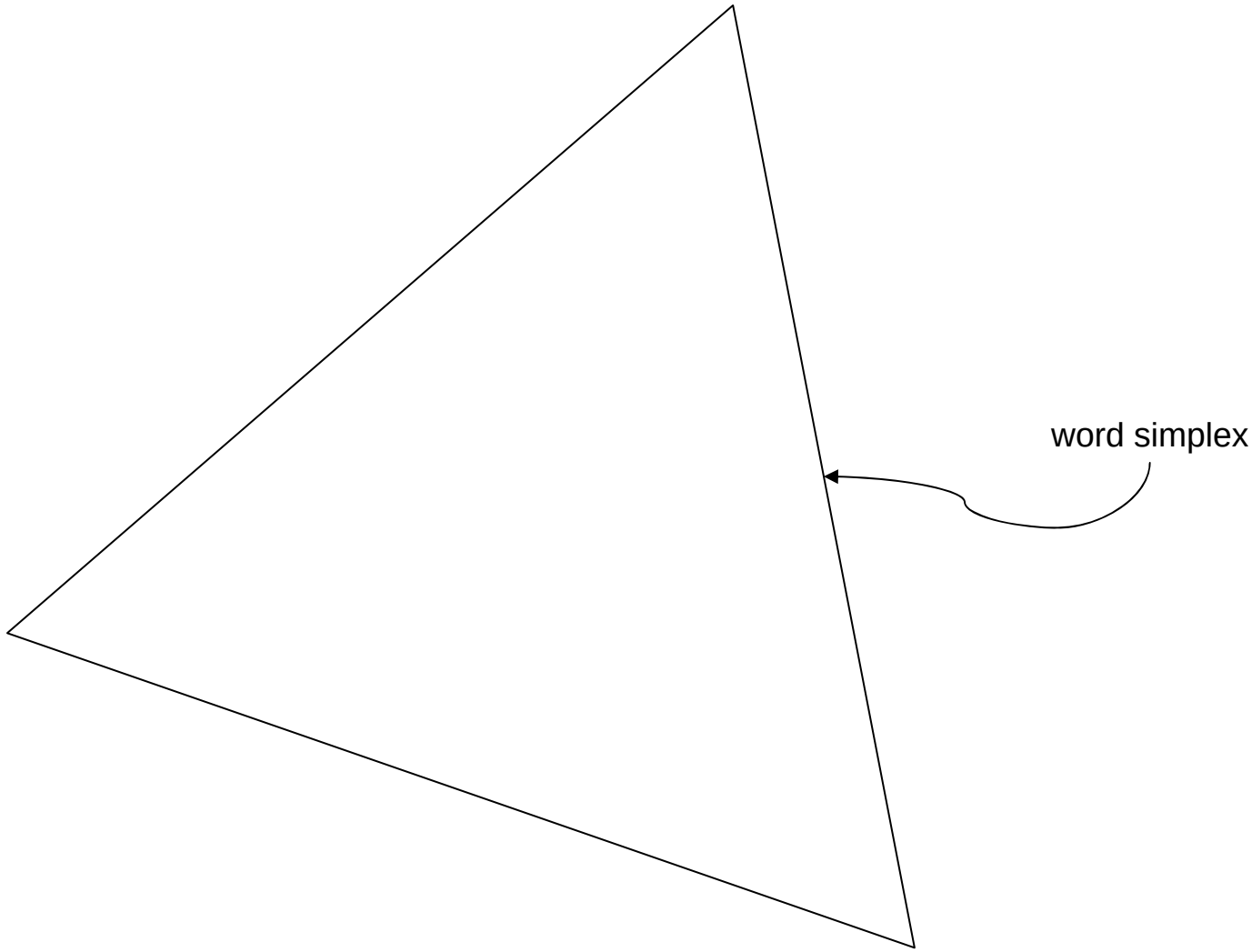
Latent Dirichlet Allocation

- LDA doesn't model documents d explicitly.
- LDA doesn't associate topics mixture weights with each document, instead, it treats them as a k-parameter hidden random variable (θ), and builds a Dirichlet distribution over it.
- A k-topic LDA needs $k+kV$ paprameters (k for α , kV for β), which doesn't increase with M.

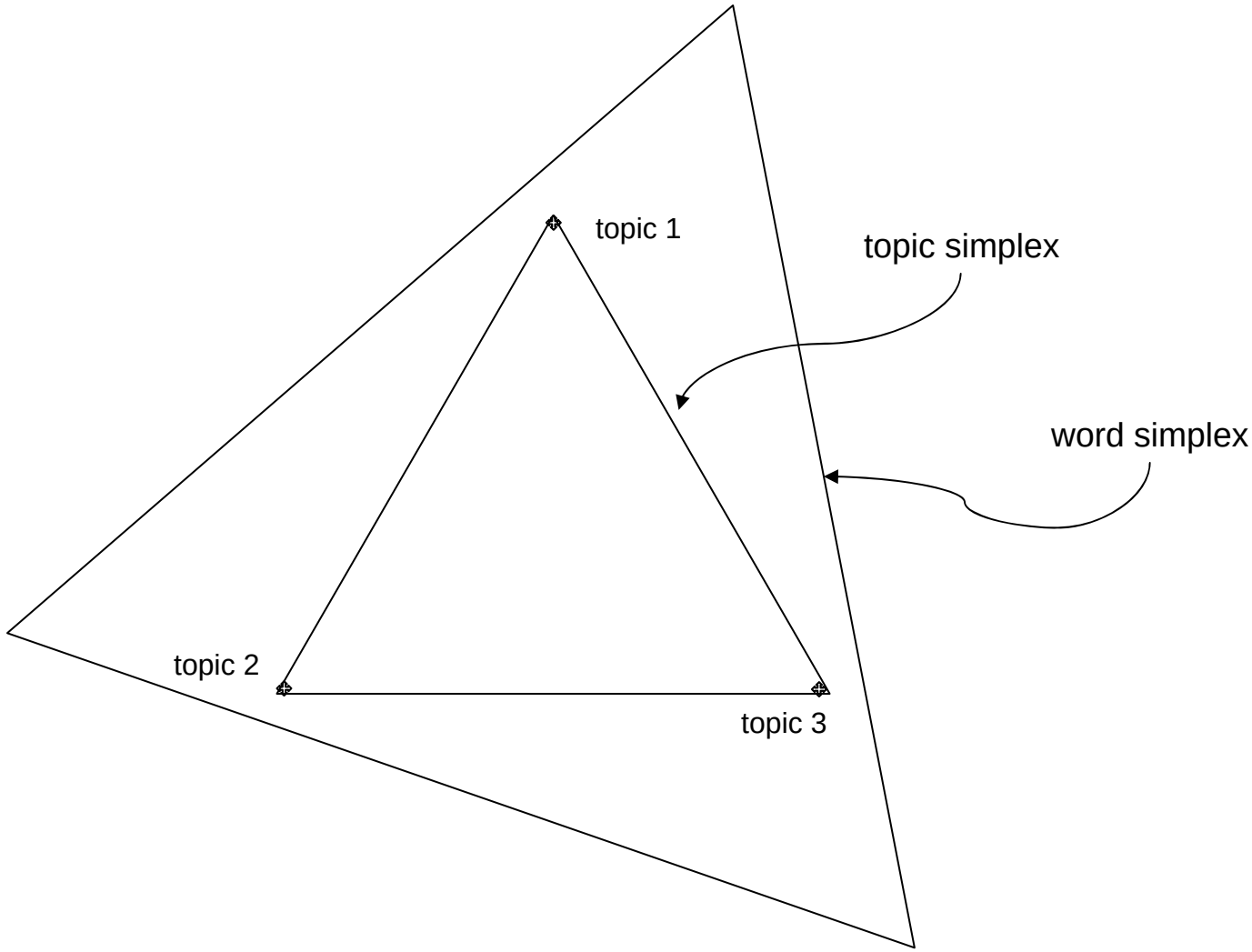
Outlines

- Notation and assumption
- Latent variable models
- **A geometric interpretation**
- Inference and estimation
- Experimental results

A Geometric Interpretation

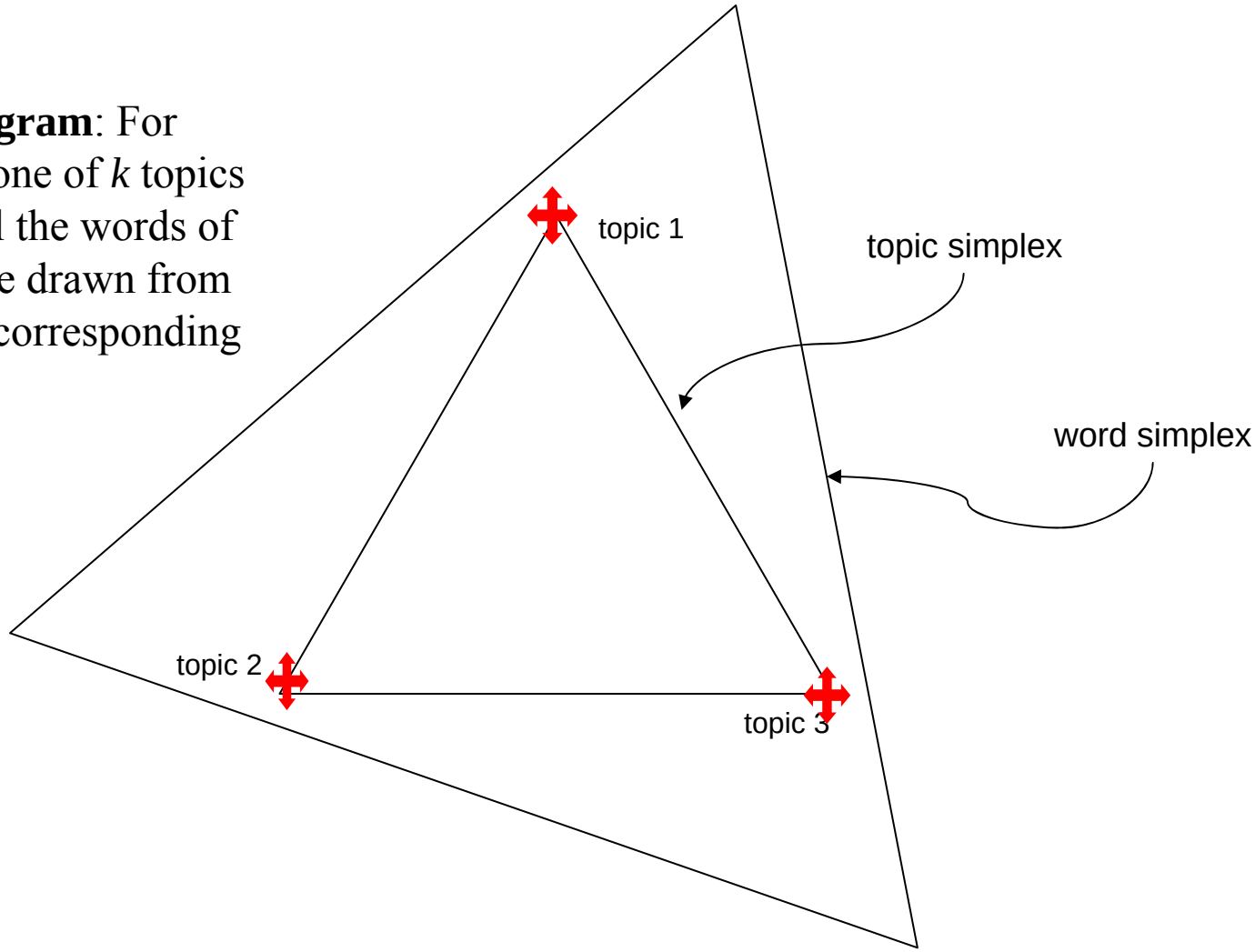


A Geometric Interpretation



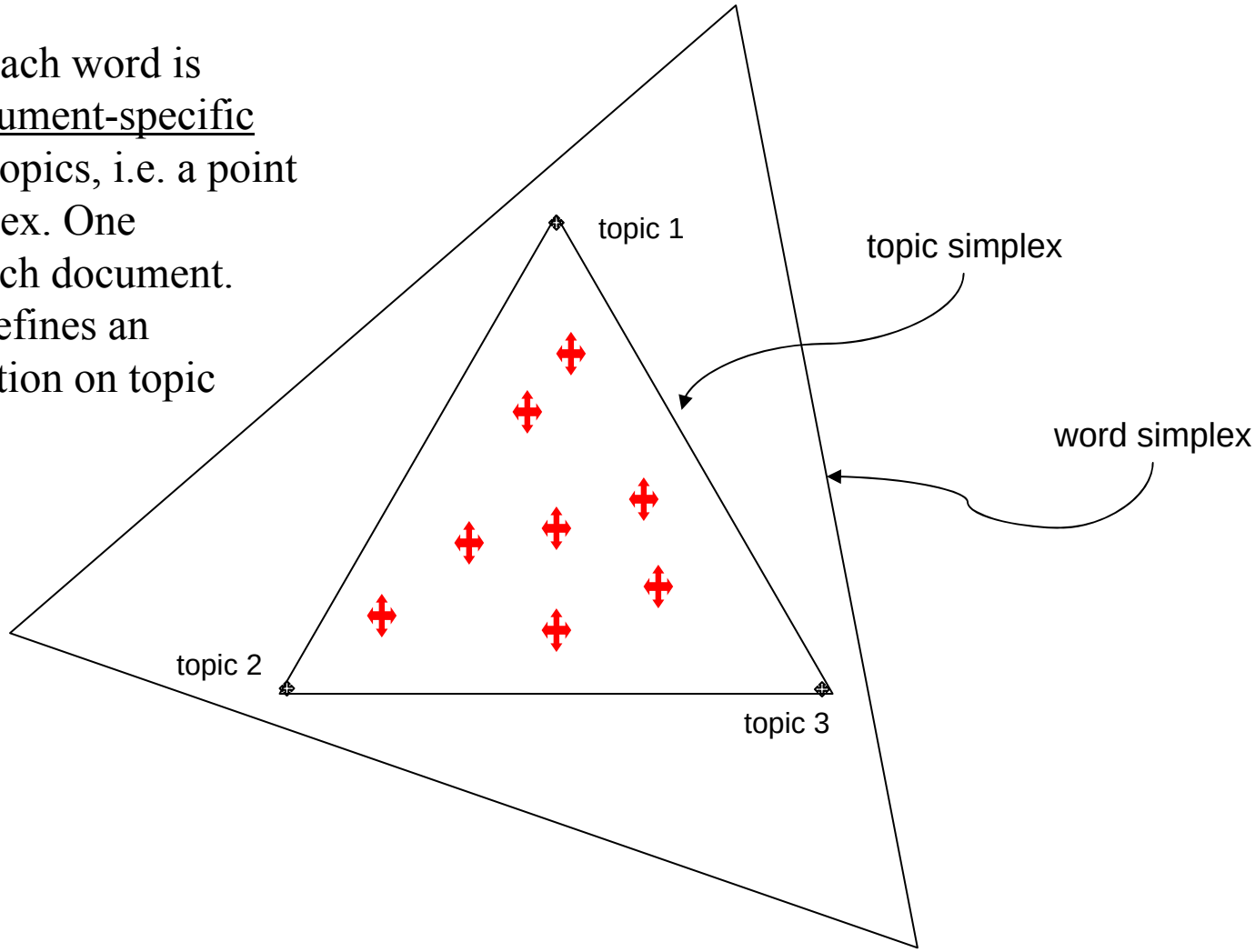
A Geometric Interpretation

Mixture of Unigram: For each document one of *k* topics is chosen and all the words of the document are drawn from the distribution corresponding to that point.



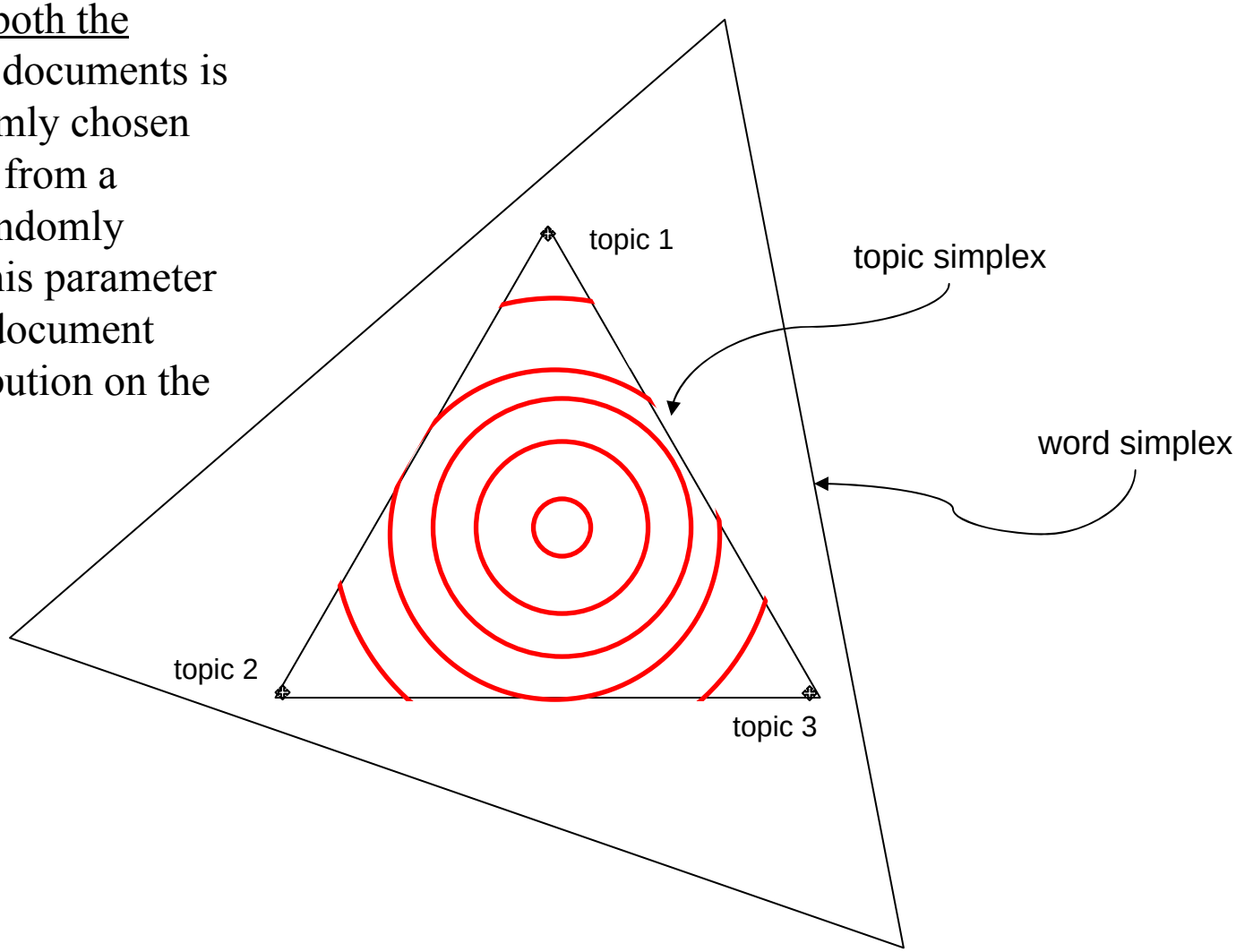
A Geometric Interpretation

pLSI: Topic for each word is drawn from a document-specific distribution over topics, i.e. a point on the topic simplex. One distribution for each document. The training set defines an empirical distribution on topic simplex



A Geometric Interpretation

LDA: Each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution on the topic simplex



Outlines

- Notation and assumption
- Latent variable models
- A geometric interpretation
- **Inference and estimation**
- Experimental results

Inference and Estimation

- Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$
\begin{aligned}
& p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= p(\boldsymbol{\theta} | \boldsymbol{\alpha}) p(\mathbf{z} | \boldsymbol{\theta}) p(\mathbf{w} | \mathbf{z}, \boldsymbol{\beta}) \\
&= p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n) p(w_n | z_n, \boldsymbol{\beta})
\end{aligned}
$$

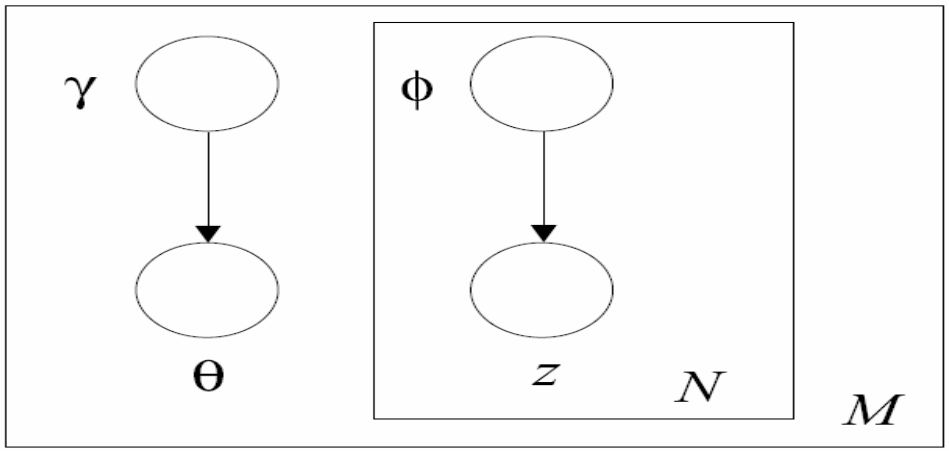
- Summering over **z** and **θ**

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta.$$

$$= \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\int\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j}\right)d\theta,$$

Variational Inference

- Basic idea: make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood.
- Introduce a family of distribution on the latent variables: $q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$



Variational Inference

- Obtain a lower bound for *log p(*$\mathbf{w}|\ \alpha, \beta$*)*:

$$
\begin{aligned}
&\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \log \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \\
&\geq \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \\
&= \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} - \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log q(\boldsymbol{\theta}, \mathbf{z}) d\theta \\
&= E_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})] - E_q[q(\boldsymbol{\theta}, \mathbf{z})] \\
&= L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})
\end{aligned}
$$

Variational EM Algorithm

- E-step: For each document, find the optimizing values of the variational parameters γ_d^*, ϕ_d^* , by maximizing the lower bound of *log p(***w**| α, β*)*.

- M-step: Maximizing the resulting lower bound of Σ *log p(***w**| α, β*)* to obtain the mode parameters α and β.

Outlines

- Notation and assumption
- Latent variable models
- A geometric interpretation
- Inference and estimation
- **Experimental results**

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Document modeling

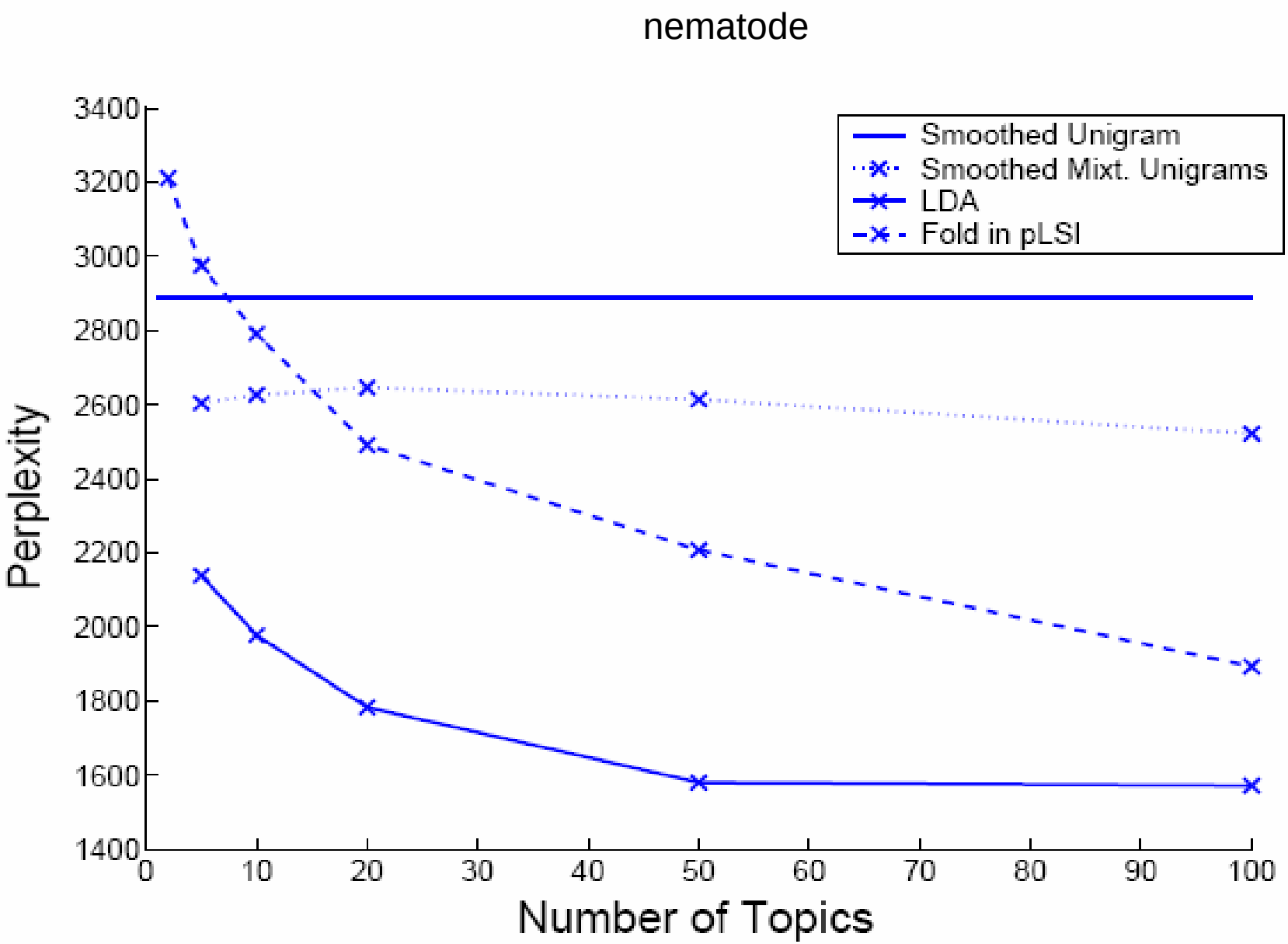
- Perplexity: How the model is “perplexed” by the data.

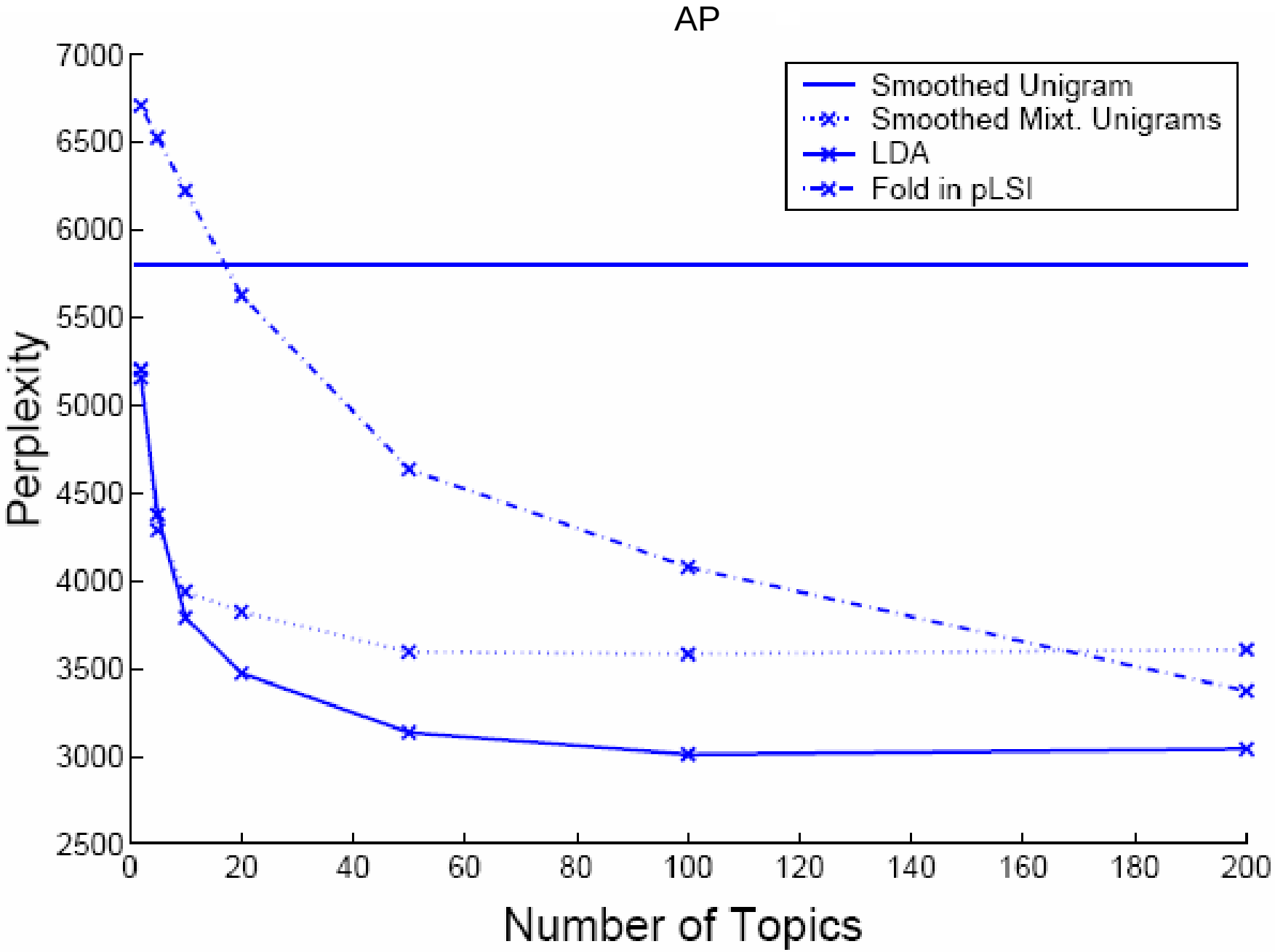
$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

- The lower the better.

Data sets

- C. Elegans Community abstracts
  - 5,225 abstracts
  - 28,414 unique terms
- TREC AP corpus (subset)
  - 16,333 newswire articles
  - 23,075 unique terms
- Held-out data – 10%
- Removed terms – 50 stop words, words appearing once (AP)


nematode
Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
Number of Topics
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
0
10
20
30
40
50
60
70
80
90
100


AP
Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
0
20
40
60
80
100
120
140
160
180
200
Number of Topics