IX.3 Latent topic models

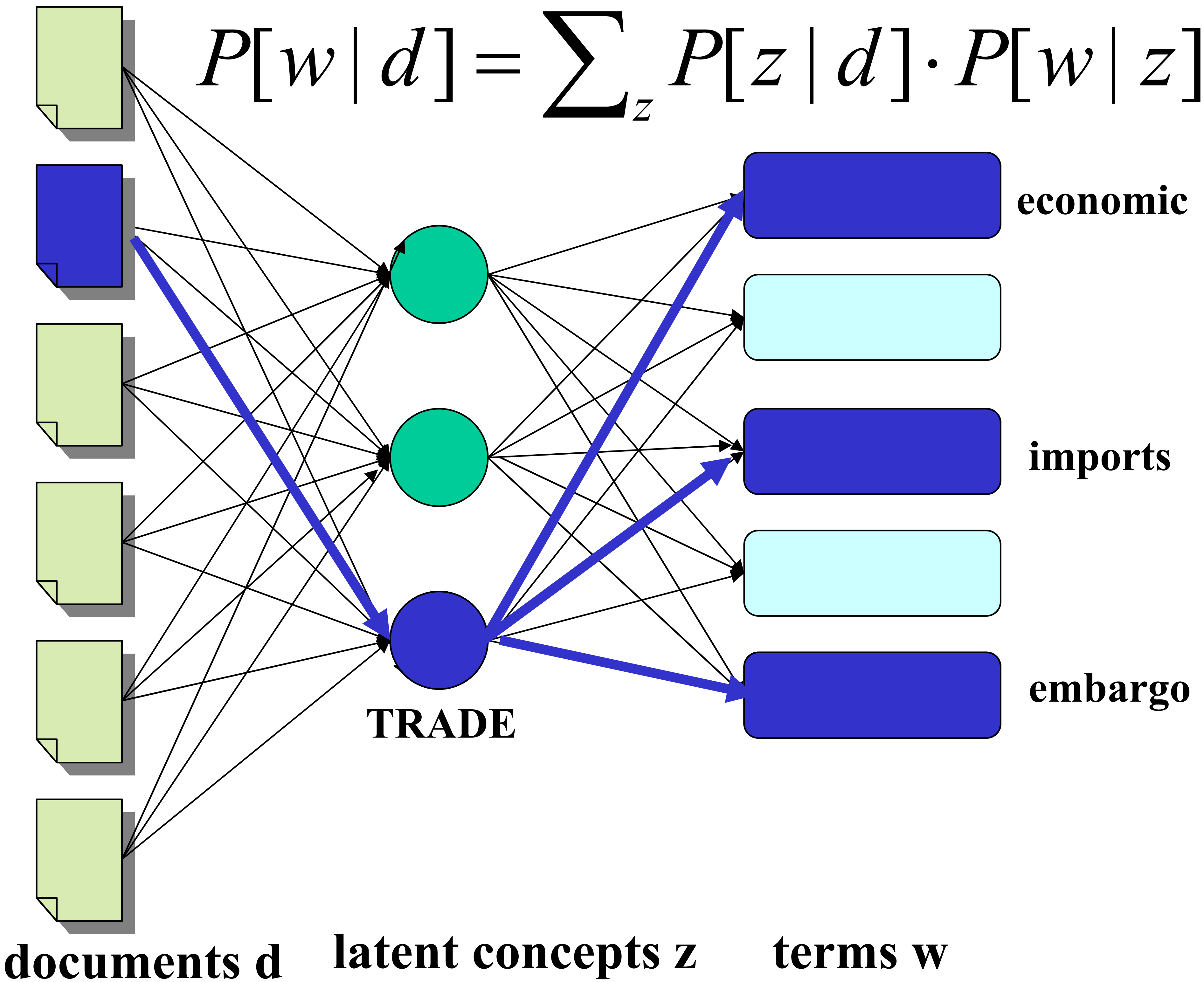
1. Basic idea
2. Latent semantic indexing (LSI)
3. Probabilistic latent semantic indexing (pLSI)
4. Latent Dirichlet allocation (LDA)

Probabilistic latent semantic indexing (pLSI)

- We model documents as (probabilistic) mixtures of topics (a.k.a. aspects)
- Each topic generates words with *topic-specific probabilities*
- We assume *conditional independence of word w and document d given topic t*:
  - $\Pr[w \wedge d \wedge t] = \Pr[w \wedge d \mid t] \Pr[t] = \Pr[w \mid t]\Pr[d \mid t]\Pr[t]$
  - $\Pr[w \wedge d] = \sum_t \Pr[w \mid t] \Pr[d \mid t] \Pr[t]$
- Generative model:
  - $\Pr[w \mid d] = \sum_t \Pr[t \mid d] \Pr[w \mid t]$

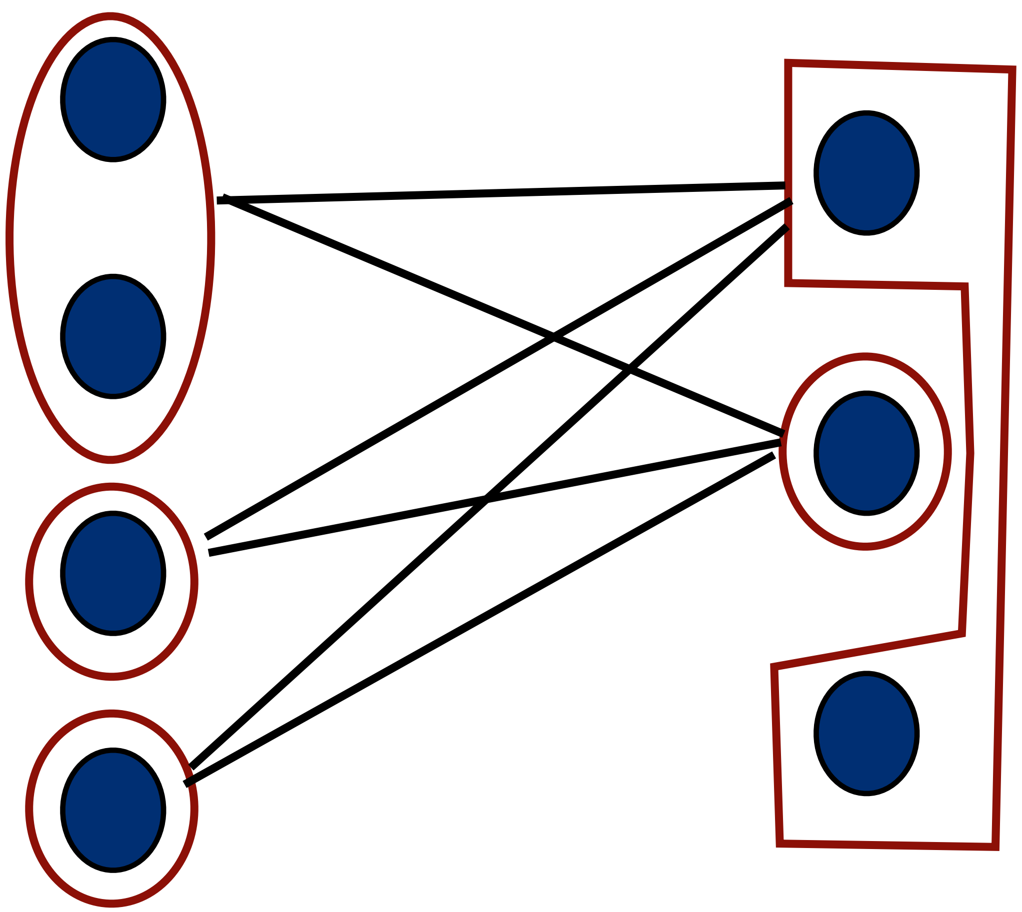
pLSI example

$$P[w \mid d] = \sum_z P[z \mid d] \cdot P[w \mid z]$$

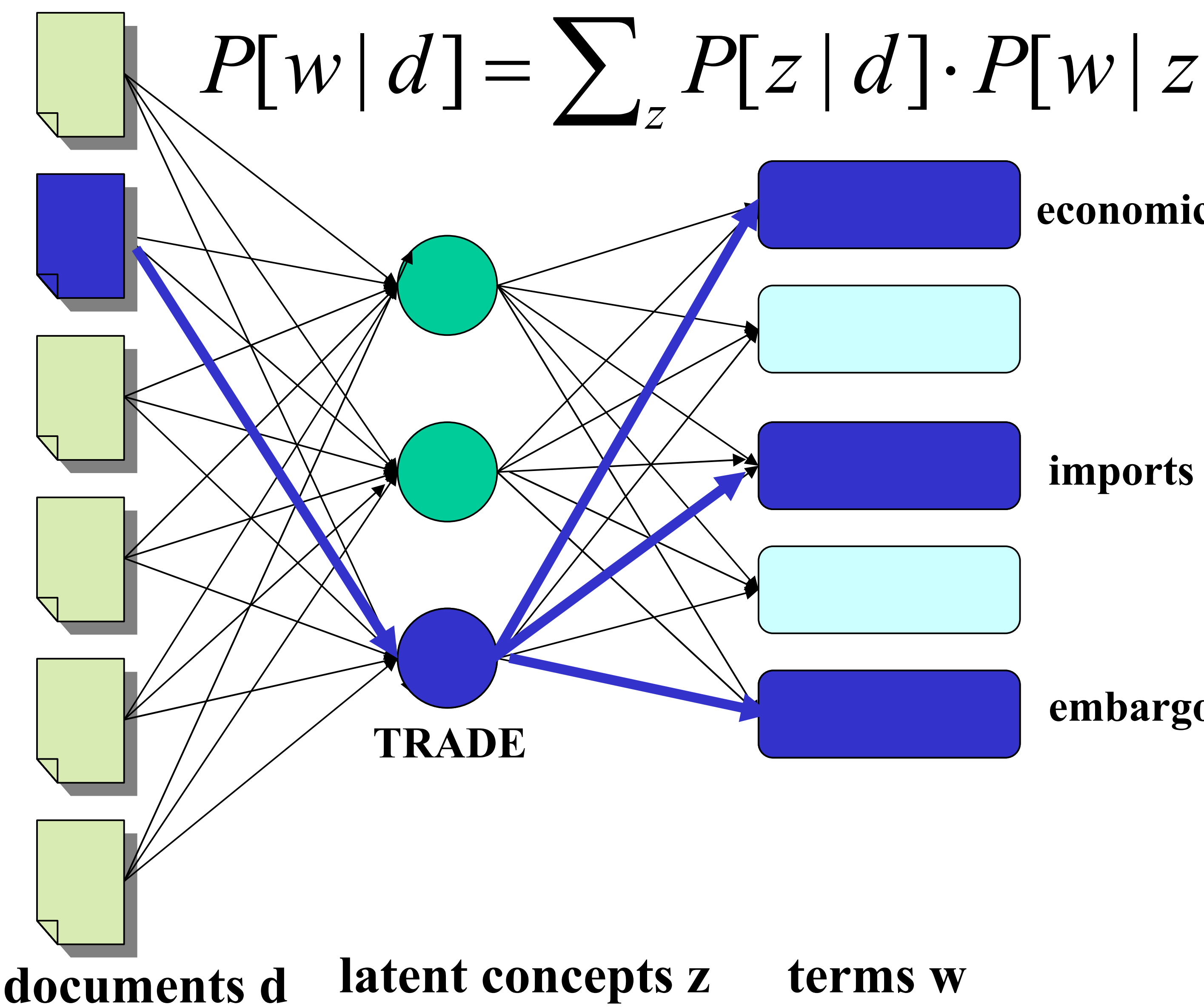


Relationship of pLSI to co-clustering

Co-clustering clusters documents and terms – no overlapping
Co-cluster mean μ is the ”strength of words in these documents”

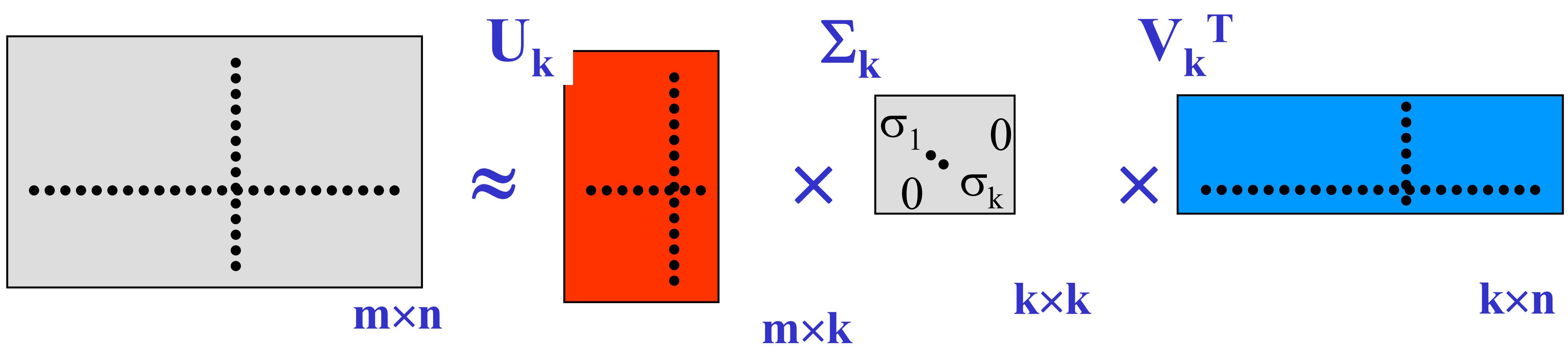


$$P[w \mid d] = \sum_z P[z \mid d] \cdot P[w \mid z]$$



Relationship of pLSI to LSI

$$\Pr[d,w] = \Sigma_t \Pr[d \mid t] \times \Pr[t] \times \Pr[w \mid t]$$



doc probs per concept **concept probs** **term probs per concept**

Differences to SVD:

- Probabilities are nonnegative (NMF!) and normalized
- Loss function is not squared loss, but Kullback–Leibler divergence

Geometry of pLSI

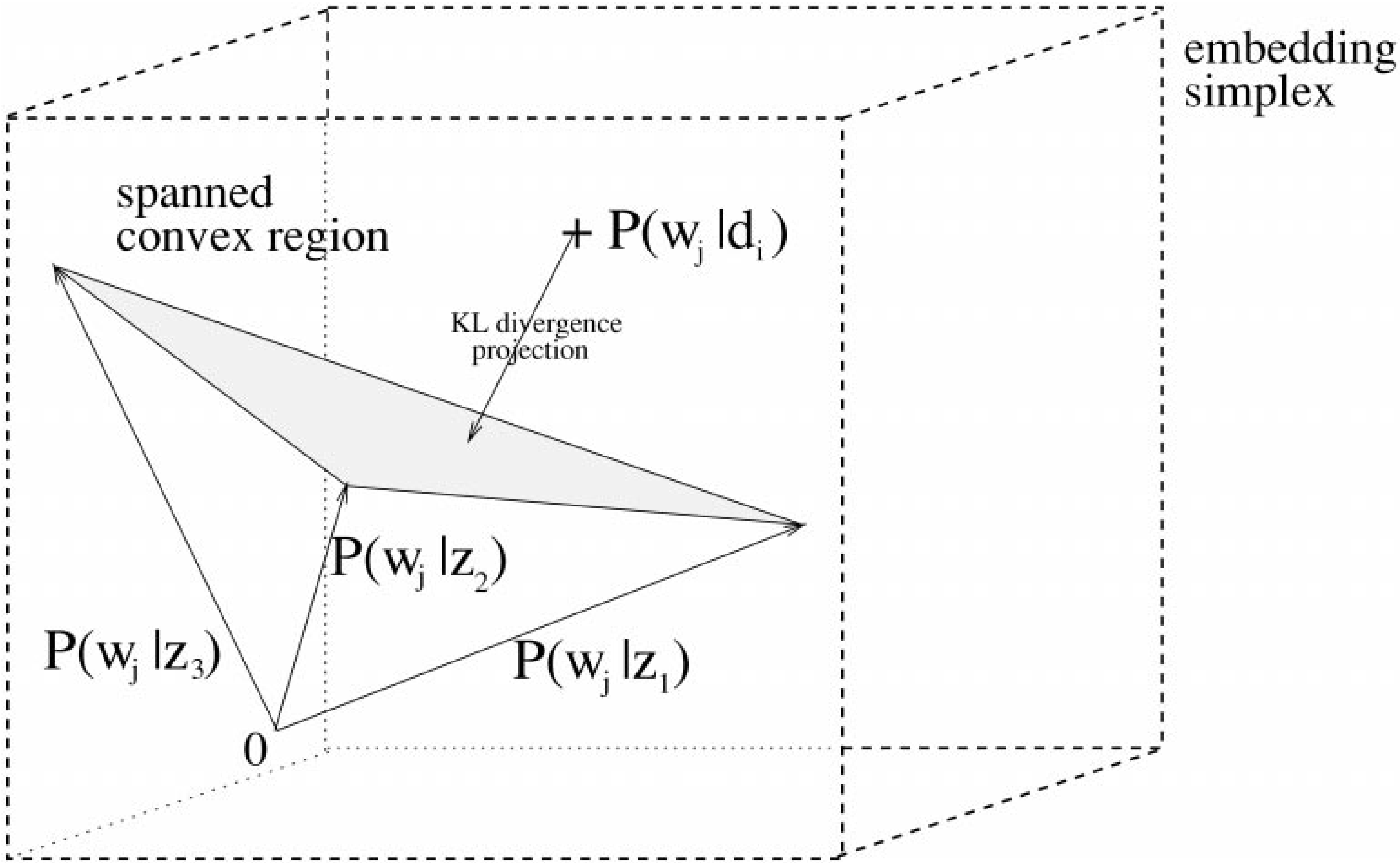


Image: T. Hofmann *Unsupervised learning by probabilistic latent semantic analysis*. 2001

Relationship of pLSI to NMF

- pLSI is equivalent to NMF that
  - tries to minimize KL-divergence, not squared loss
  - has factors normalized (probabilities must sum to 1)
  - [Ding, Li & Peng, 2008]
- Equivalency means that they try to minimize the same loss function
  - Typical algorithmic approaches differ

EM for pLSI

- Data: $n(d,w)$ – absolute freq. of word w in doc d
- Parameters: $\Pr[t \mid d]$, $\Pr[w \mid t]$
- Log-likelihood: $\sum_d \sum_w n(d,w) \log \Pr[d,w]$
- E-step: $$\Pr[t \mid d, w] = \frac{\Pr[t \mid d] \Pr[w \mid t]}{\sum_y \Pr[y \mid d] \Pr[w \mid y]}$$

  M-step:
  $$\Pr[w \mid t] \propto \sum_d n(d, w) \Pr[t \mid d, w]$$
  $$\Pr[t \mid d] \propto \sum_w n(d, w) \Pr[t \mid d, w]$$

In addition uses 'tempered' method to avoid overfitting.

EM for pLSI

- Data: $n(d,w)$ – absolute freq. of word w in doc d
- Parameters: $\Pr[t \mid d]$, $\Pr[w \mid t]$
- Log-likelihood: $\sum_d \sum_w n(d,w) \log \Pr[d,w]$
- E-step:

$$\Pr[t \mid d, w] = \frac{\Pr[t \mid d] \Pr[w \mid t]}{\sum_y \Pr[y \mid d] \Pr[w \mid y]}$$

$\boxed{\text{freq. of } w \text{ associated with } t}$

M-step:

$$\Pr[w \mid t] \propto \sum_d n(d,w) \Pr[t \mid d, w]$$

$$\Pr[t \mid d] \propto \sum_w n(d,w) \Pr[t \mid d, w]$$

In addition uses 'tempered' method to avoid overfitting.

EM for pLSI

- Data: $n(d,w)$ – absolute freq. of word w in doc d
- Parameters: $\Pr[t \mid d]$, $\Pr[w \mid t]$
- Log-likelihood: $\sum_d \sum_w n(d,w) \log \Pr[d,w]$
- E-step:

$$\Pr[t \mid d, w] = \frac{\Pr[t \mid d] \Pr[w \mid t]}{\sum_y \Pr[y \mid d] \Pr[w \mid y]}$$

  M-step:

$$\Pr[w \mid t] \propto \sum_d n(d,w) \Pr[t \mid d, w]$$

$$\Pr[t \mid d] \propto \sum_w n(d,w) \Pr[t \mid d, w]$$

freq. of *d* associated with *t*

In addition uses ‘tempered’ method to avoid overfitting.

EM for pLSI

- Data: $n(d,w)$ – absolute freq. of word w in doc d
- Parameters: $\Pr[t \mid d]$, $\Pr[w \mid t]$
- Log-likelihood: $\sum_d \sum_w n(d,w) \log \Pr[d,w]$
- E-step: $$\Pr[t \mid d, w] = \frac{\Pr[t \mid d] \Pr[w \mid t]}{\sum_y \Pr[y \mid d] \Pr[w \mid y]}$$

  M-step:
  $$\Pr[w \mid t] \propto \sum_d n(d,w) \Pr[t \mid d,w]$$
  $$\Pr[t \mid d] \propto \sum_w n(d,w) \Pr[t \mid d,w]$$

In addition uses 'tempered' method to avoid overfitting.

Folding-in of queries

- Keep all estimated parameters fixed
- Treat a query as a 'new document' to be explained
  - Find topics that most likely generate the query
    - Query = document; Pr[w | t] is kept fixed
  - EM for query parameters

$$\Pr[t \mid q, w] = \frac{\Pr[t \mid q]\hat{p}[w \mid t]}{\sum_y \Pr[y \mid q]\hat{p}[w \mid y]}$$

$$\Pr[t \mid q] = \frac{\sum_w n(q, w) \Pr[t \mid q, w]}{\sum_{w,y} n(q, w) \Pr[y \mid q, w]}$$

Query processing

- Documents and queries are both represented as *probability distributions over k topics*
  - k-dimensional vectors with $x_i \geq 0$ and $\sum x_i = 1$
- Any convenient vector-space similarity measure works
  - scalar product
  - cosine
  - KL divergence
  - …

Experimental results: example

- Concepts (10 of 128) extracted from Science Magazine articles (12K)

P(w|z)

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148

P(w|z)

bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824

Source: Thomas Hofmann, Tutorial at ADFOCS 2004

On perplexity

- How well does the model generalize to unseen data?
  - **The** question in statistics/machine learning
  - Many measures
    - But the proof of the pudding is in the eating…
- Perplexity is one measure of generalization performance
  - Log-averaged inverse probability of unseen data:

$$\mathcal{P} = \exp\left\{ -\frac{\sum_{d,w} n'(d,w) \log \Pr[w \mid d]}{\sum_{d,w} n'(d,w)} \right\}$$

    - *n'*(*d,w*) = frequency of word *w* in doc *d in test-data*

Experimental results: perplexity

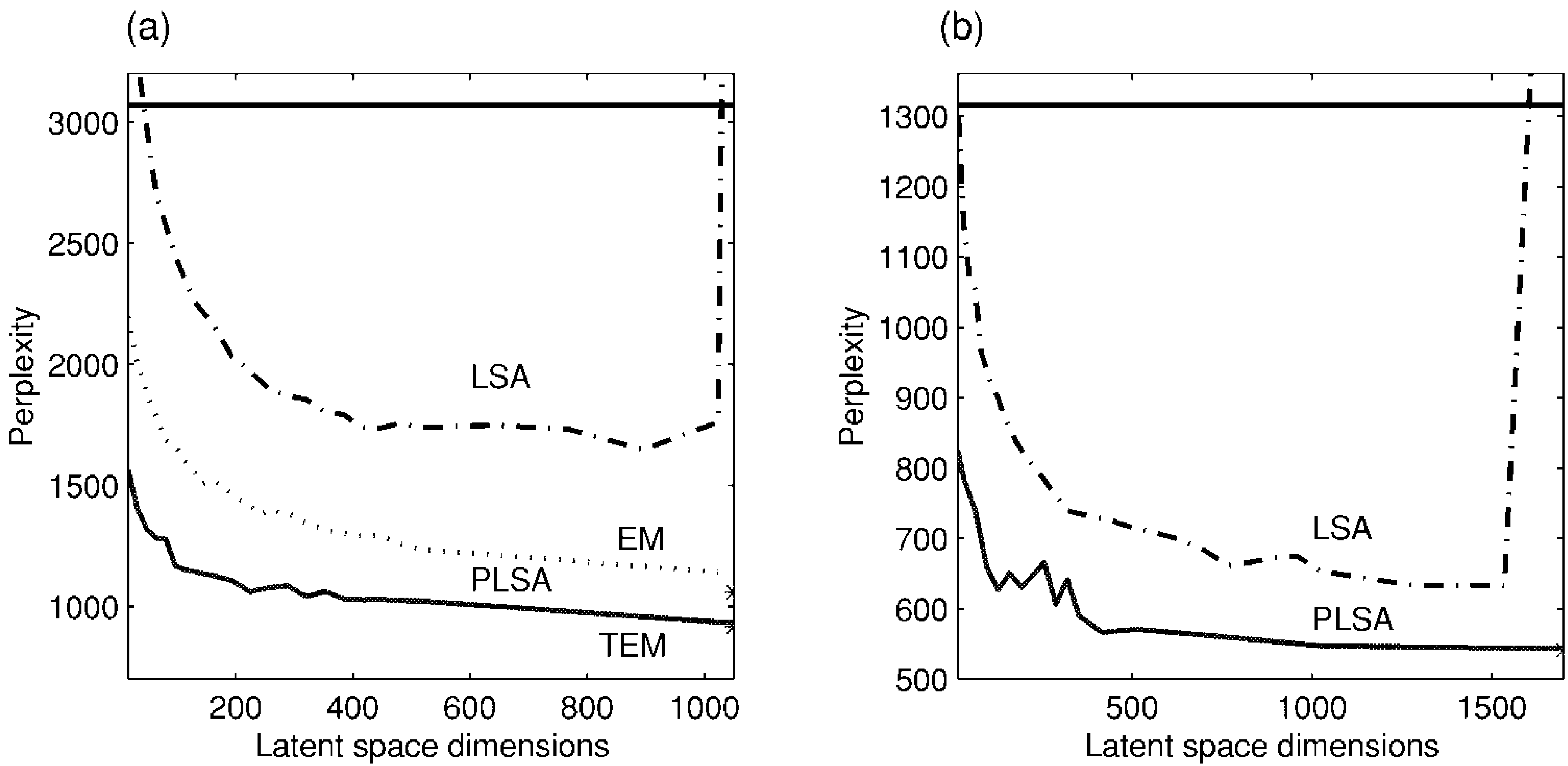


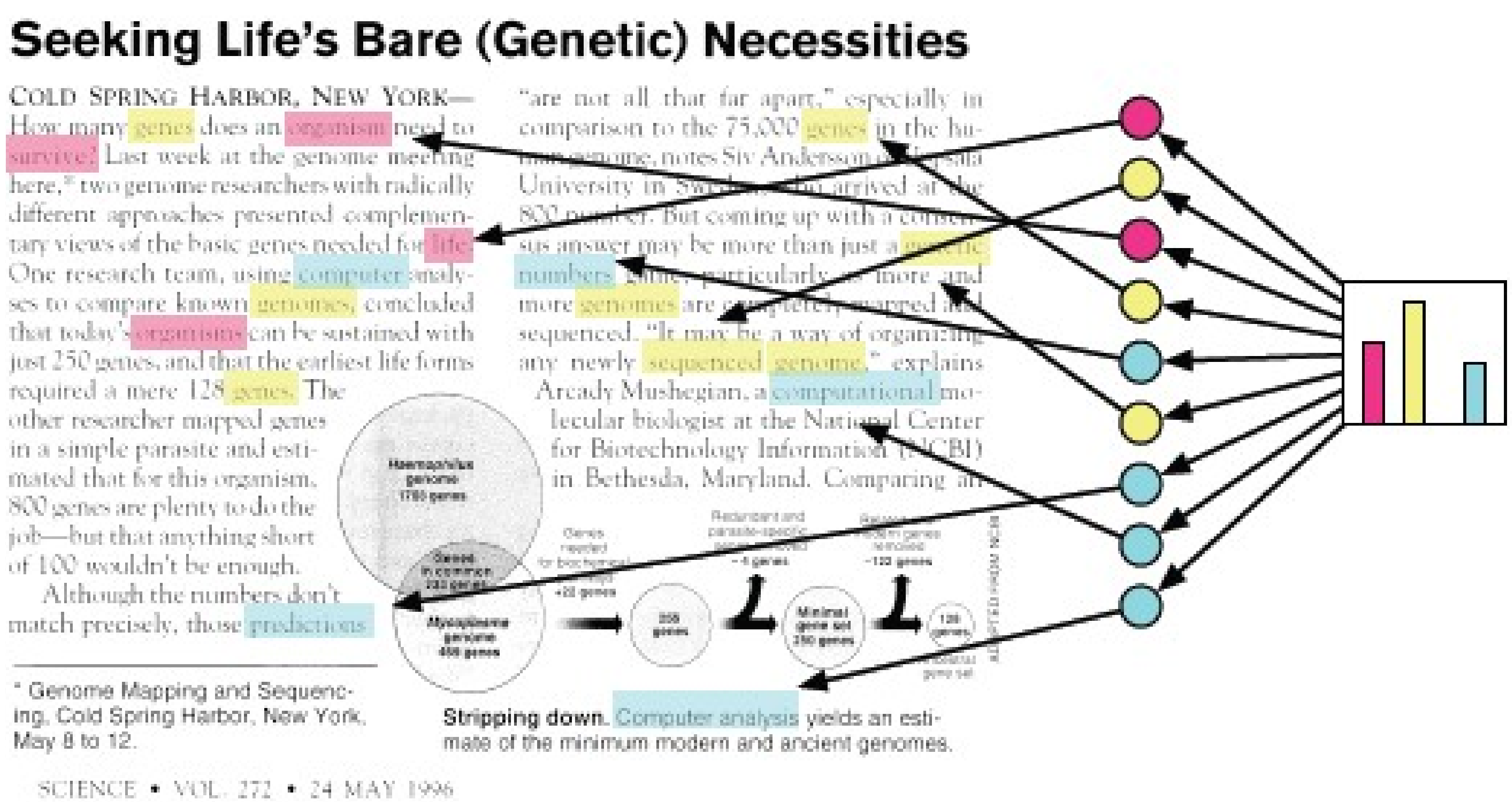
Figure 6. Perplexity results as a function of the latent space dimensionality for (a) the MED data (rank 1033) and (b) the LOB data (rank 1674). Plotted results are for LSA (dashed-dotted curve) and PLSA (trained by TEM = solid curve, trained by early stopping EM = dotted curve). The upper baseline is the unigram model corresponding to marginal independence. The star at the right end of the PLSA denotes the perplexity of the largest trained aspect models ($K = 2048$).

pLSI summary

- Probabilistic variant of LSI
  - Equivalent to NMF with particular normalization
- Better experimental results than LSI
- Good on ‘closed’ corpora
  - But tied on the fixed corpus
    - No generative model!
- Computationally expensive
  - Indexing and querying
- Number of latent concepts has to be selected
  - BIC, AIC, asses with held-out data with different k

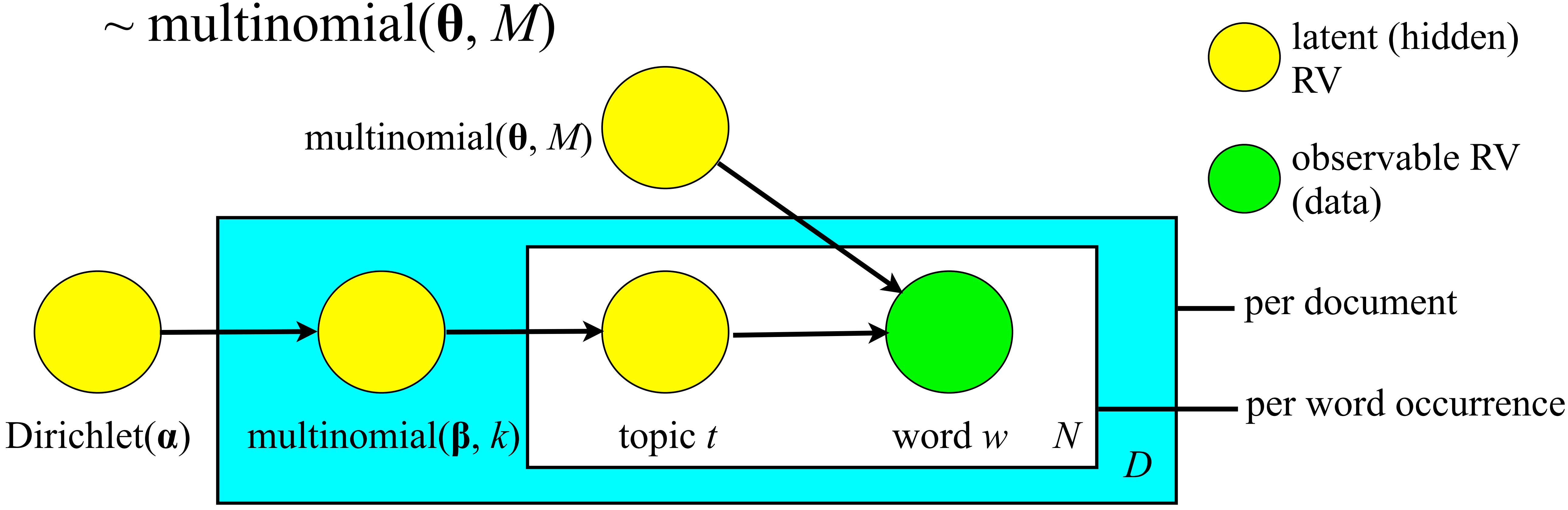
Latent Dirichlet allocation (LDA)

- Multiple-cause mixture model (MCMM)
- Documents contain multiple topics
  - Topics are expressed by specific word distributions
- LDA provides a *generative model* for this
  - *Dirichlet topic mixtures*

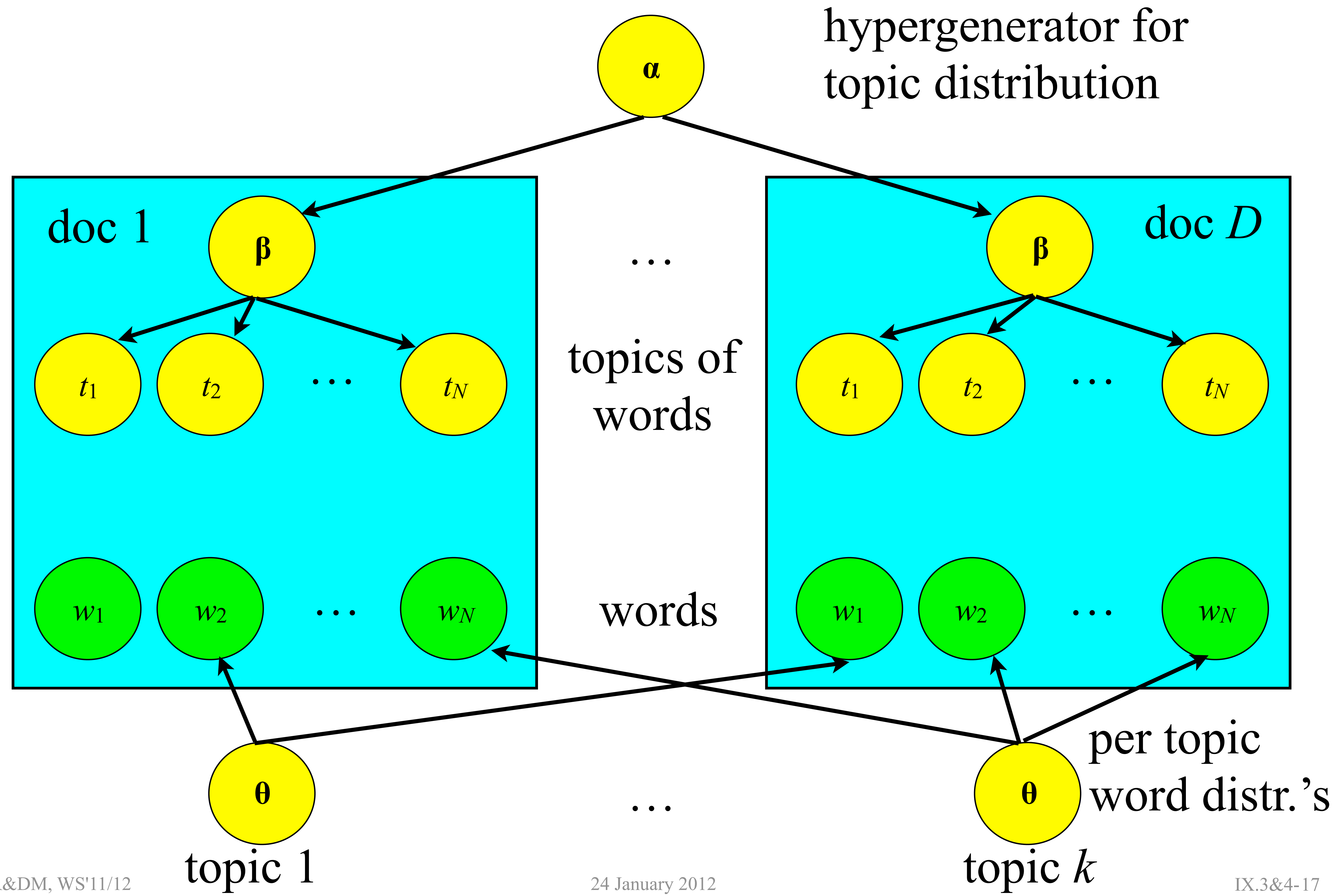


LDA generative model

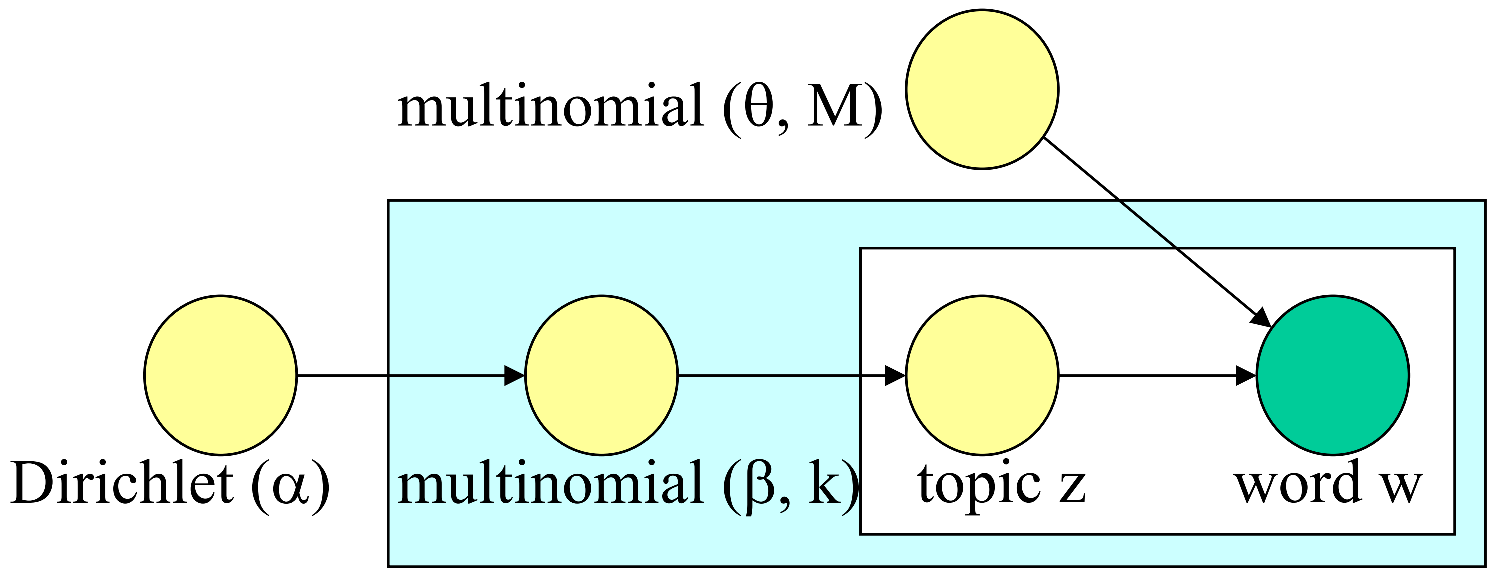
- For each document d
  - Choose doc length N (# word occurrences) ~ Poisson(λ)
  - Choose topic-probability parameters $\boldsymbol{\beta}$ ~ Dirichlet($\boldsymbol{\alpha}$)
  - For each N word occurrences in d (at position n)
    - Choose one of k topics t_n ~ multinomial($\boldsymbol{\beta}$, k)
    - Choose one of M words w_n from per-topic distribution ~ multinomial($\boldsymbol{\theta}$, M)



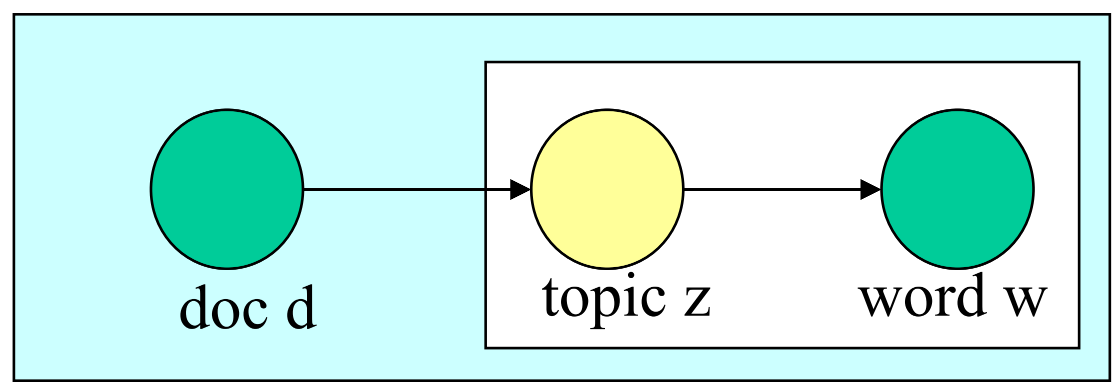
LDA: instance-level model



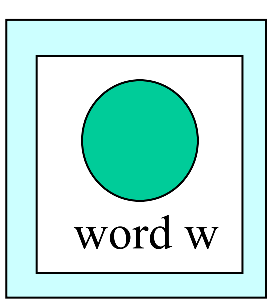
Comparison to other latent-topic models



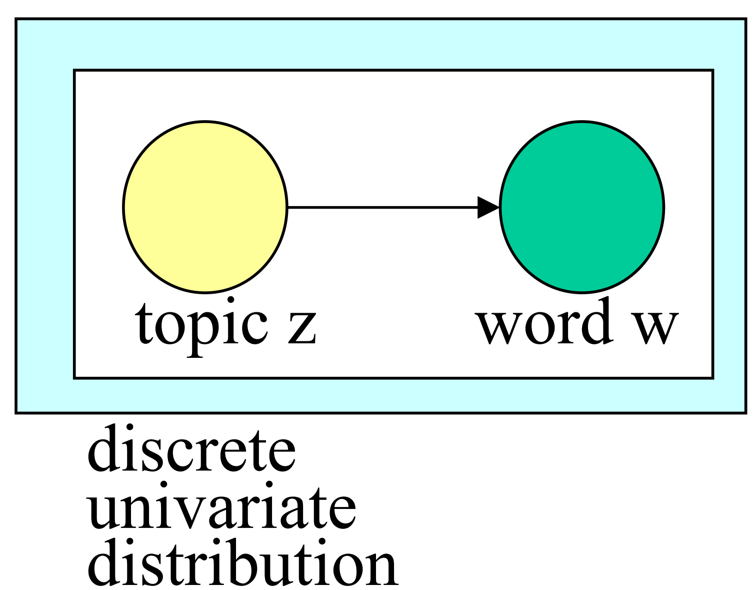
LDA



pLSI



simple unigram model



single-cause mixture of unigrams

Computing LDA

Pdf of Dirichlet: $$f(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \beta_1^{\alpha_1 - 1} \cdots \beta_k^{\alpha_k - 1}$$

Probability of document d given **α** and **β**:

$$\Pr[d \mid \boldsymbol{\alpha}, \theta] = \int f(\beta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{t_n=1}^{k} \beta_{t_n} \theta_{t_n, w_n} \right) d\beta$$

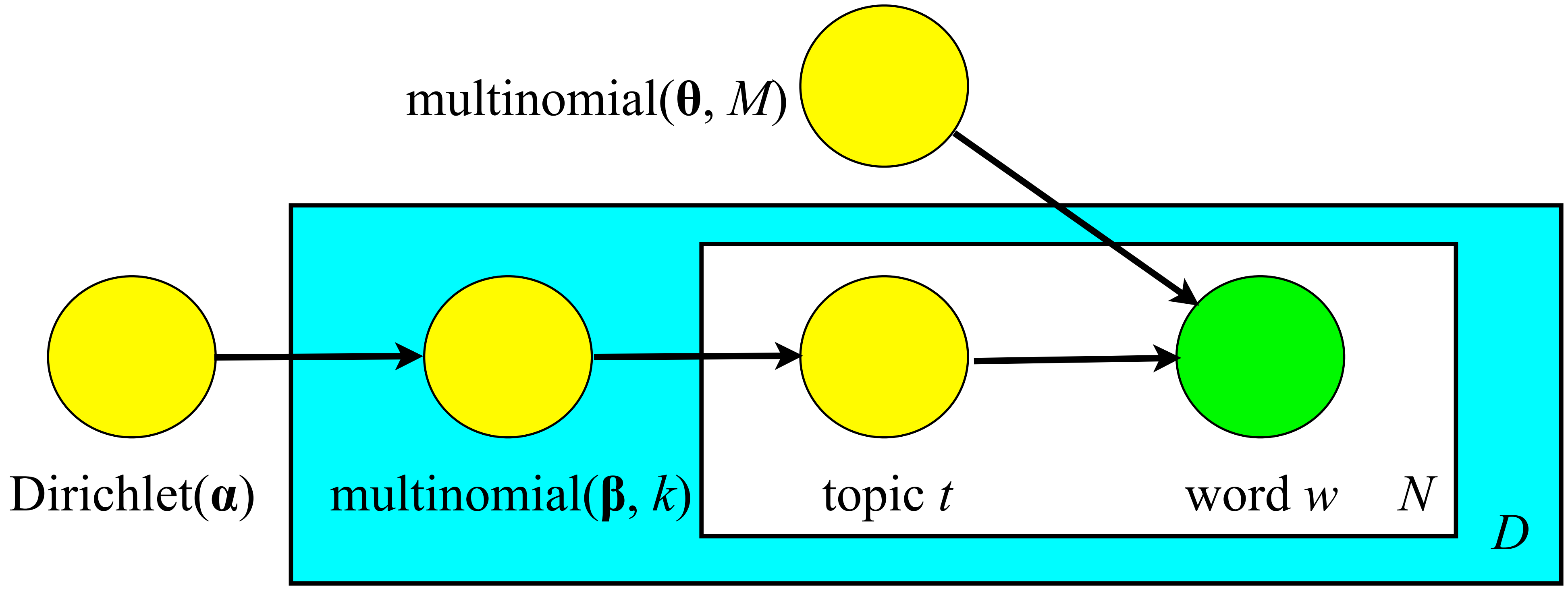
$$= \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \beta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{t_n=1}^{k} \beta_{t_n} \theta_{t_n, w_n} \right) d\beta$$

⇒ Posterior probability is *intractable*!

Variational inference

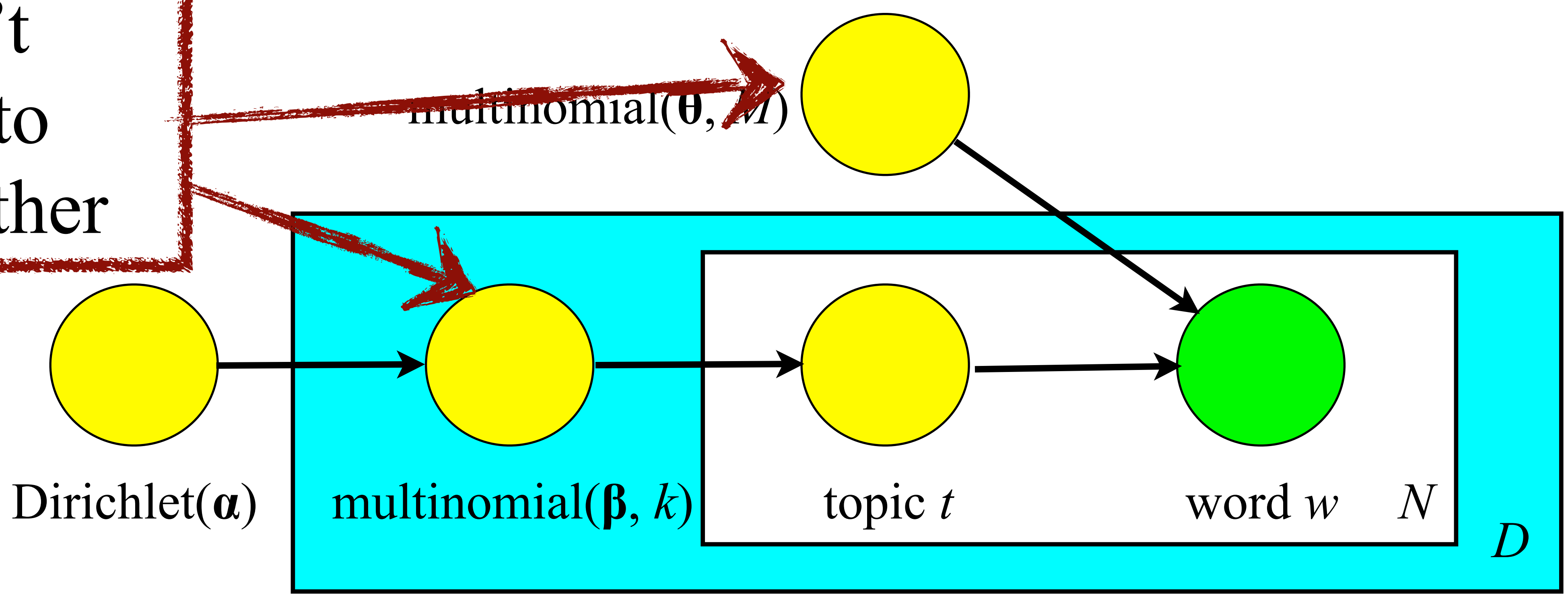
- Consider a family of tractable lower-bound functions
- In E-step, find optimal parameters for these lower-bound functions
- In M-step, use the fixed lower-bound distribution to find parameters to maximize the log-likelihood
  - In M-step we update parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$
  - Full details in [Blei, Ng, Jordan: *Latent Dirichlet Allocation*, J. Mach. Learn. Res., 3, 2003]

Lower-bound distributions



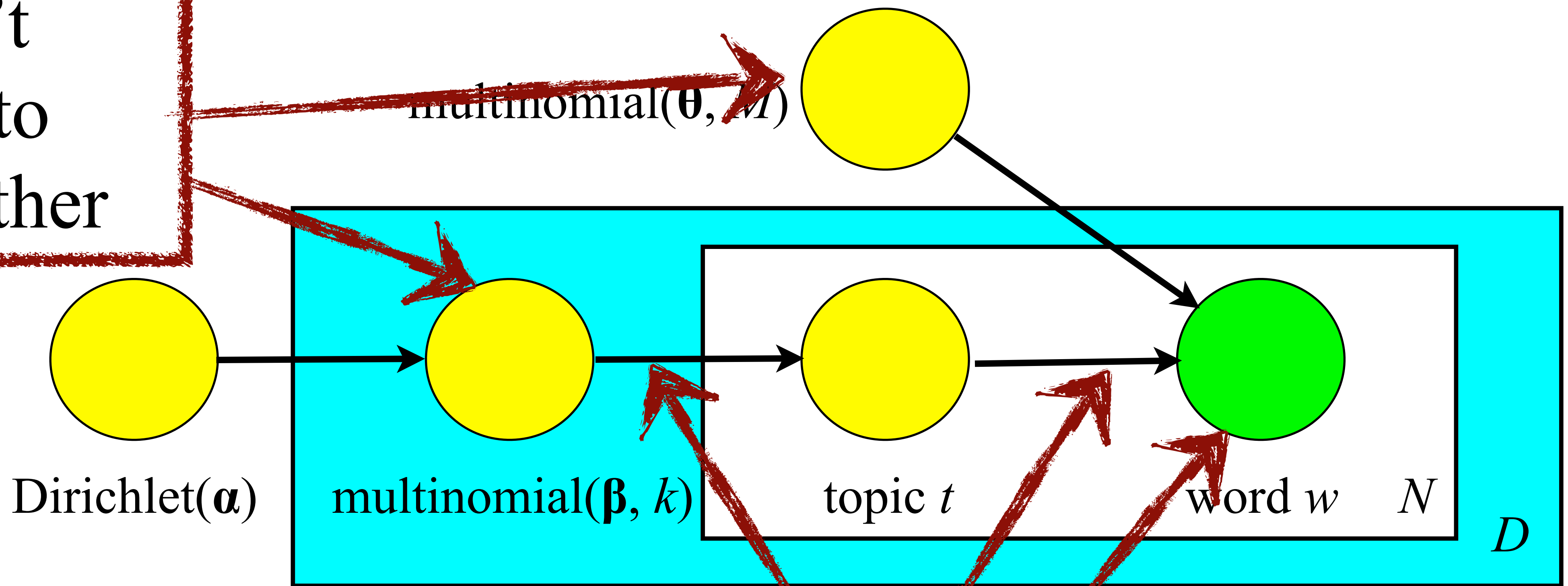
Lower-bound distributions

These shouldn't have anything to do with each other



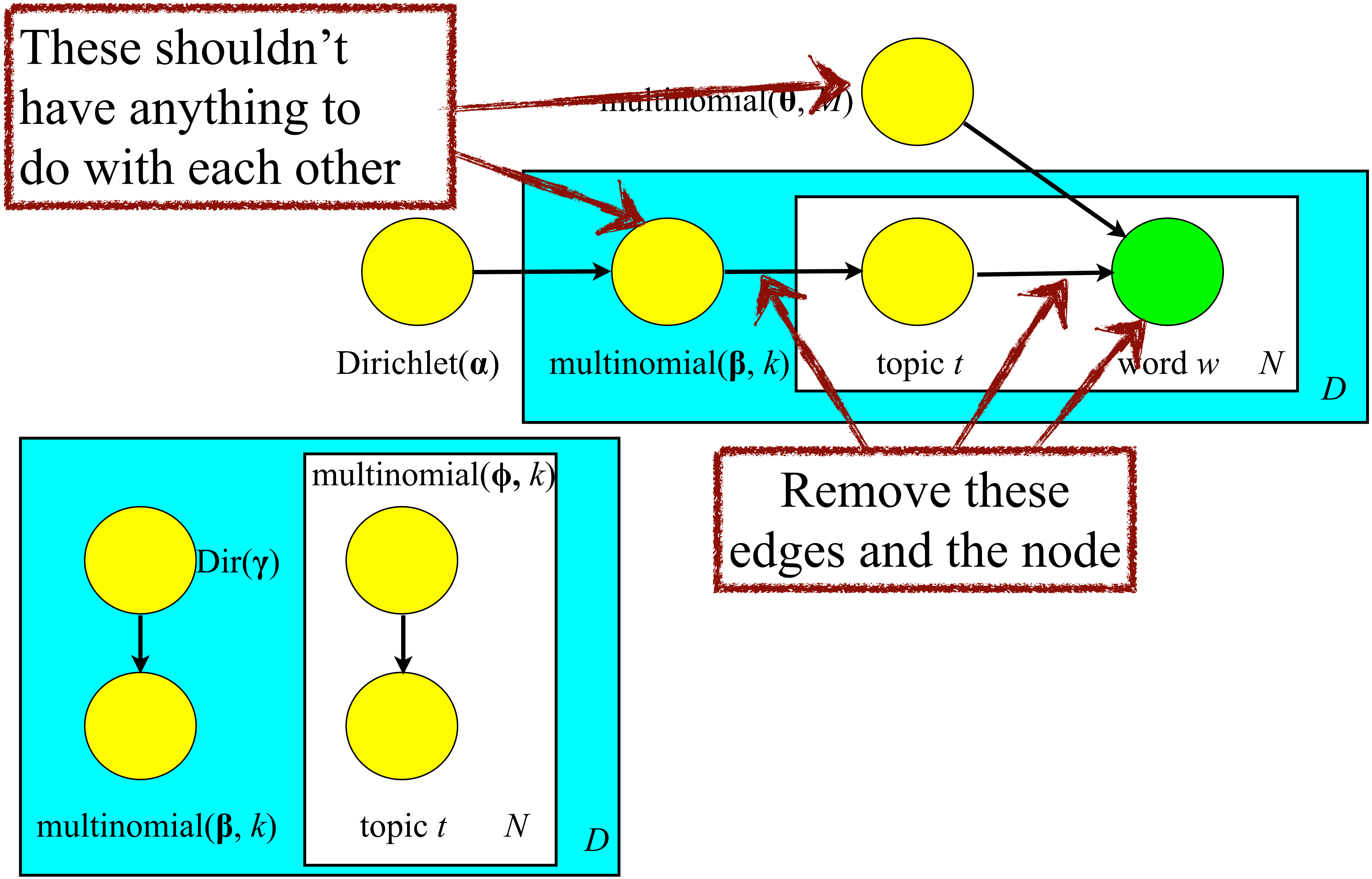
Lower-bound distributions

These shouldn't have anything to do with each other



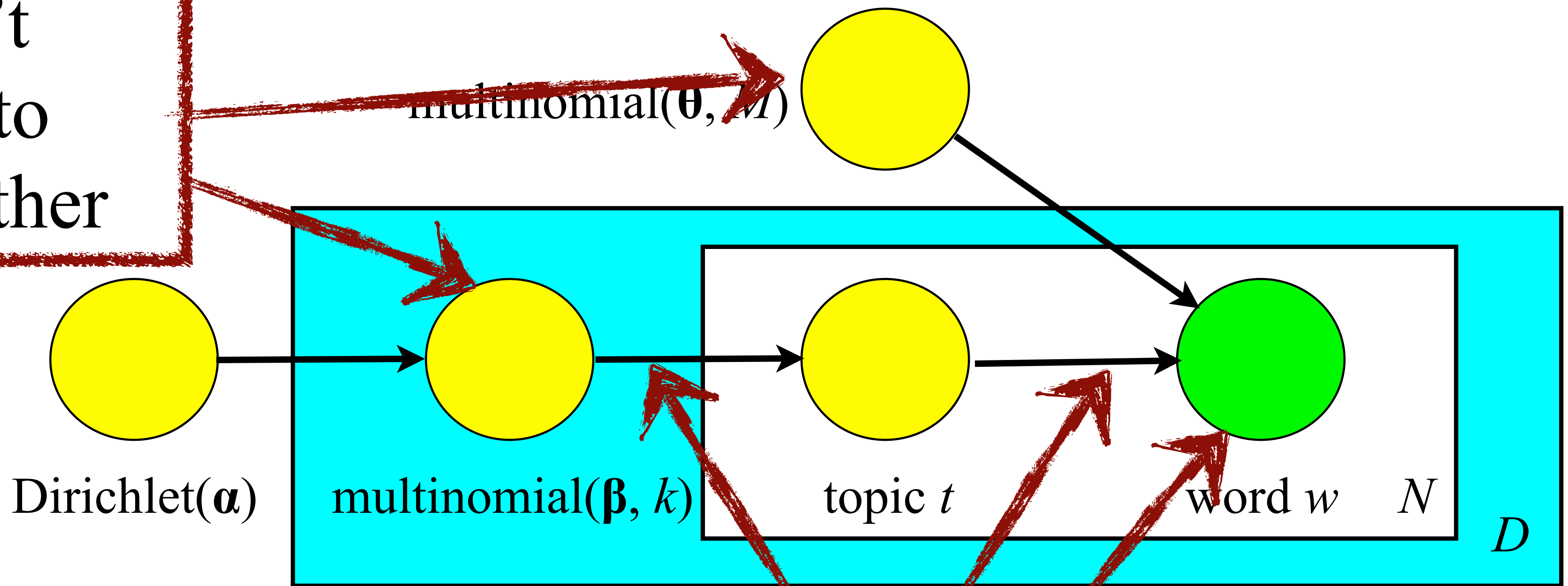
Remove these edges and the node

Lower-bound distributions

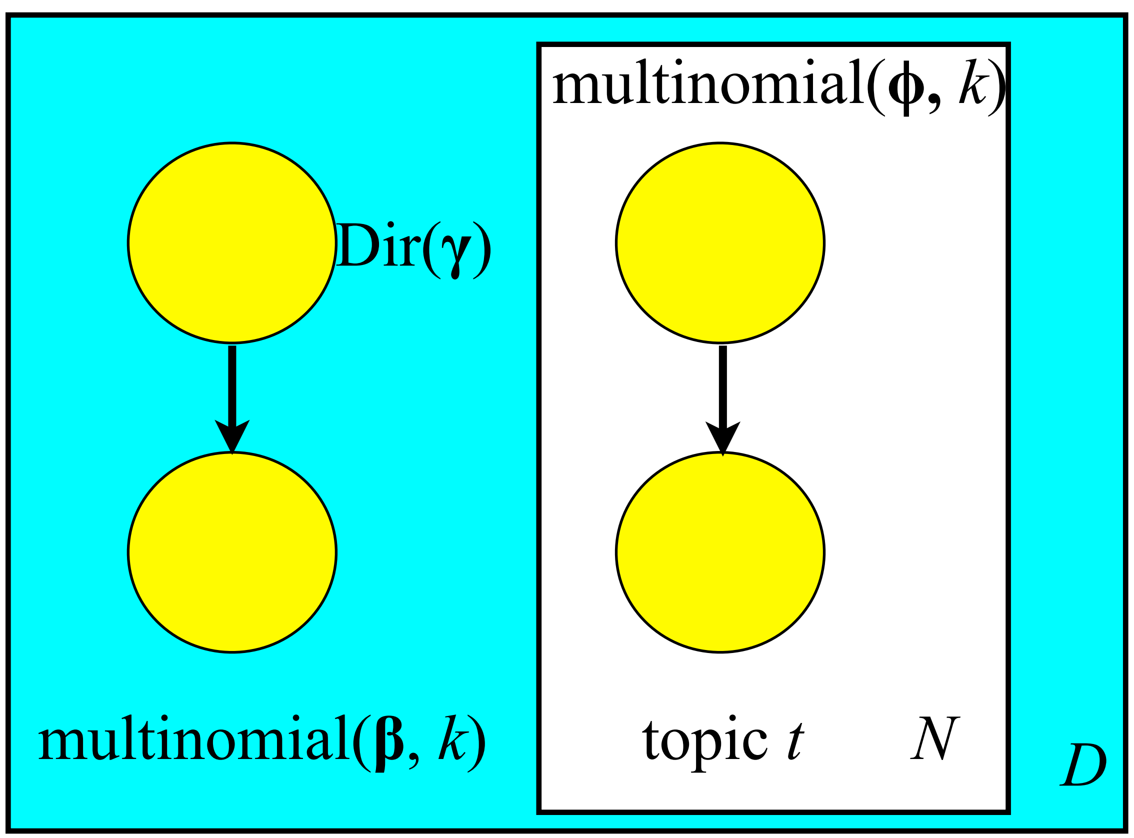


Lower-bound distributions

These shouldn't have anything to do with each other



Remove these edges and the node

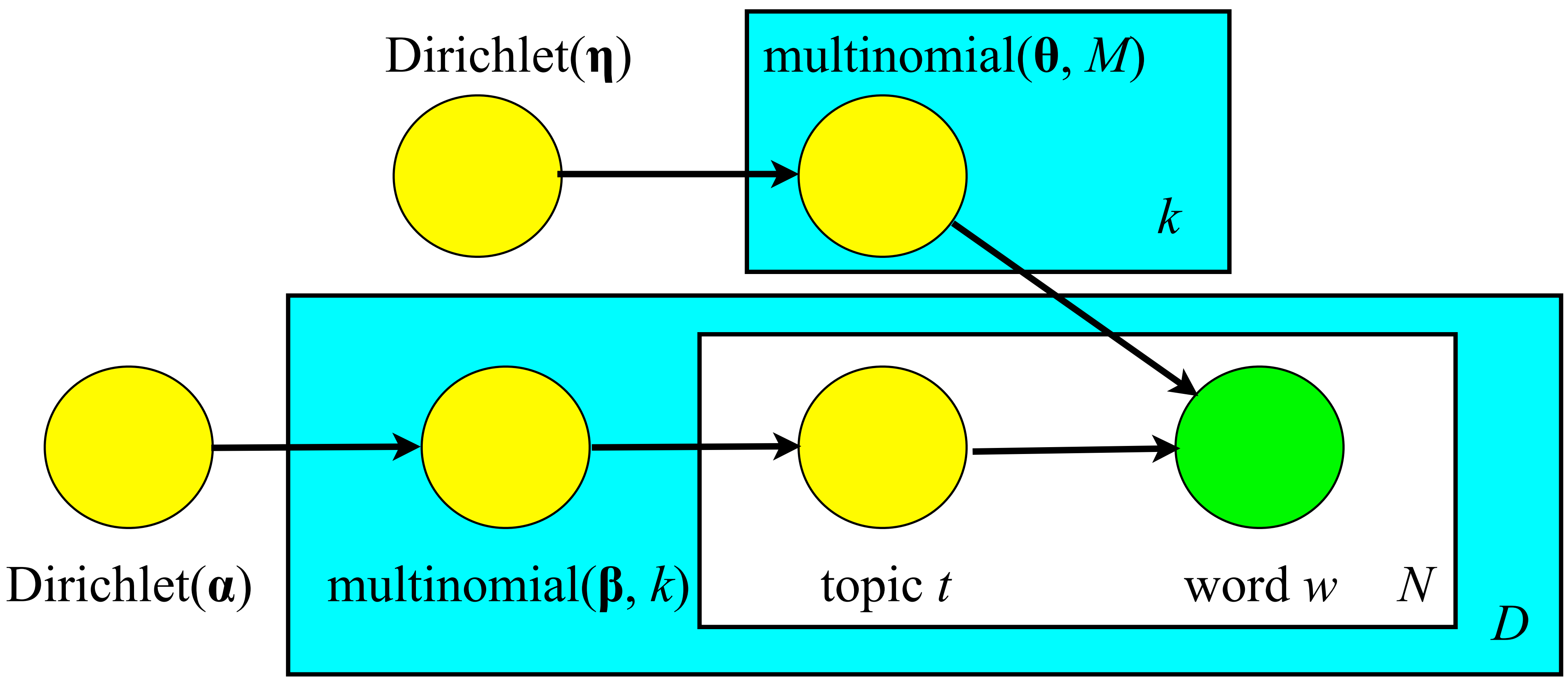


Find **γ*** and **ϕ*** to minimize KL divergence between variational $q(\ldots)$ and data log-likelihood $p(\ldots)$:

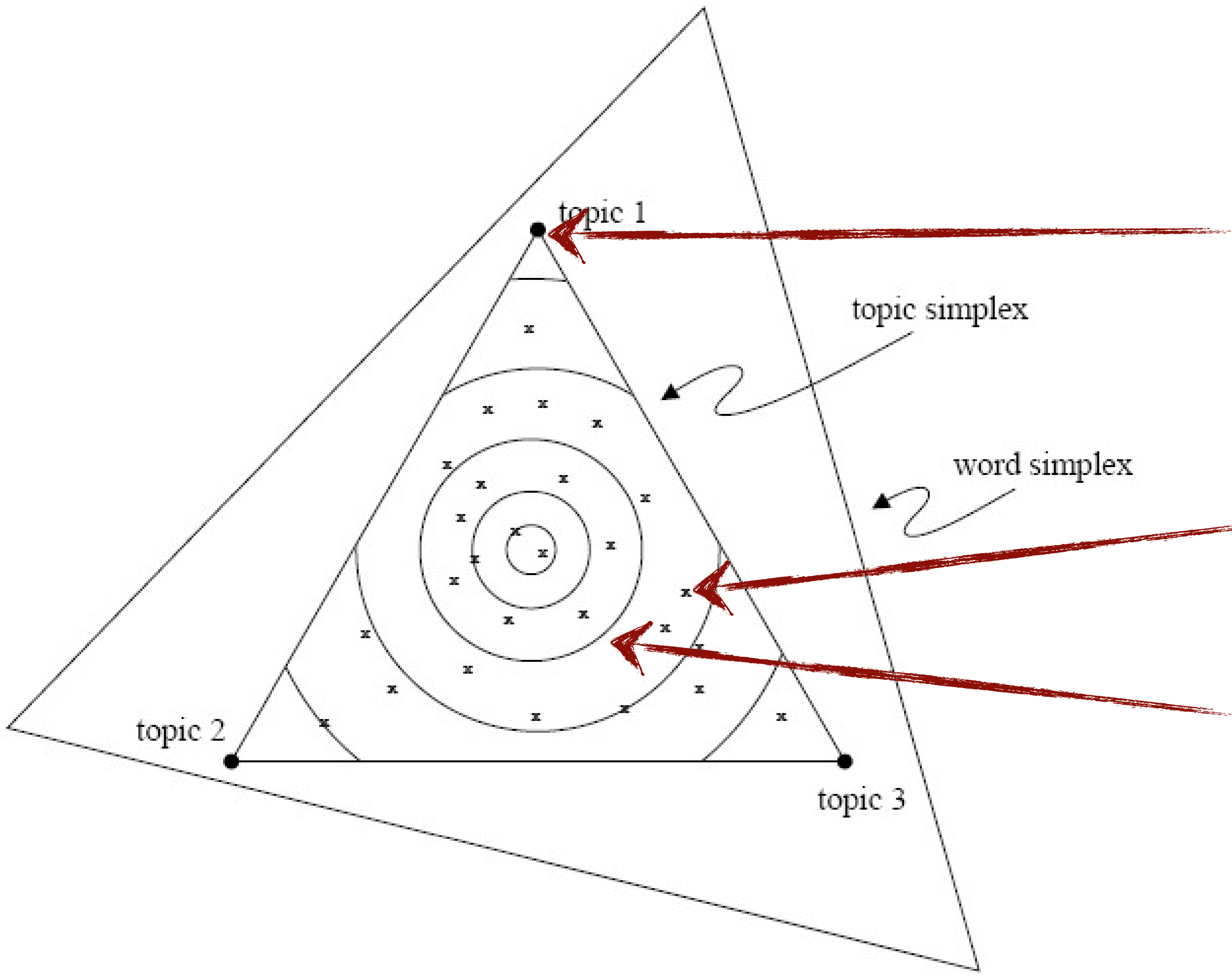
$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \underset{(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\arg\min}\, \mathrm{D}\big(\mathrm{q}(\boldsymbol{\beta}, \mathbf{t} \mid \gamma, \phi) \,\|\, \mathrm{p}(\beta, \mathbf{t} \mid \mathrm{d}, \boldsymbol{\alpha}, \boldsymbol{\theta})\big)$$

Extended LDA

- A new document arrives that has never-before seen word
  - The word gets 0 probability ⇒ document gets 0 prob.
- Answer: *smoothing*
  - Assign each word non-zero probability



Geometry of latent-topic models



single-cause latent-topic model places docs on corners of topic simplex

pLSI places docs at discrete points in topic simplex

LDA imposes smooth distribution in topic simplex and can place docs at arbitrary points in the simplex

Source: Blei et al., 2003

LDA experimental results: example

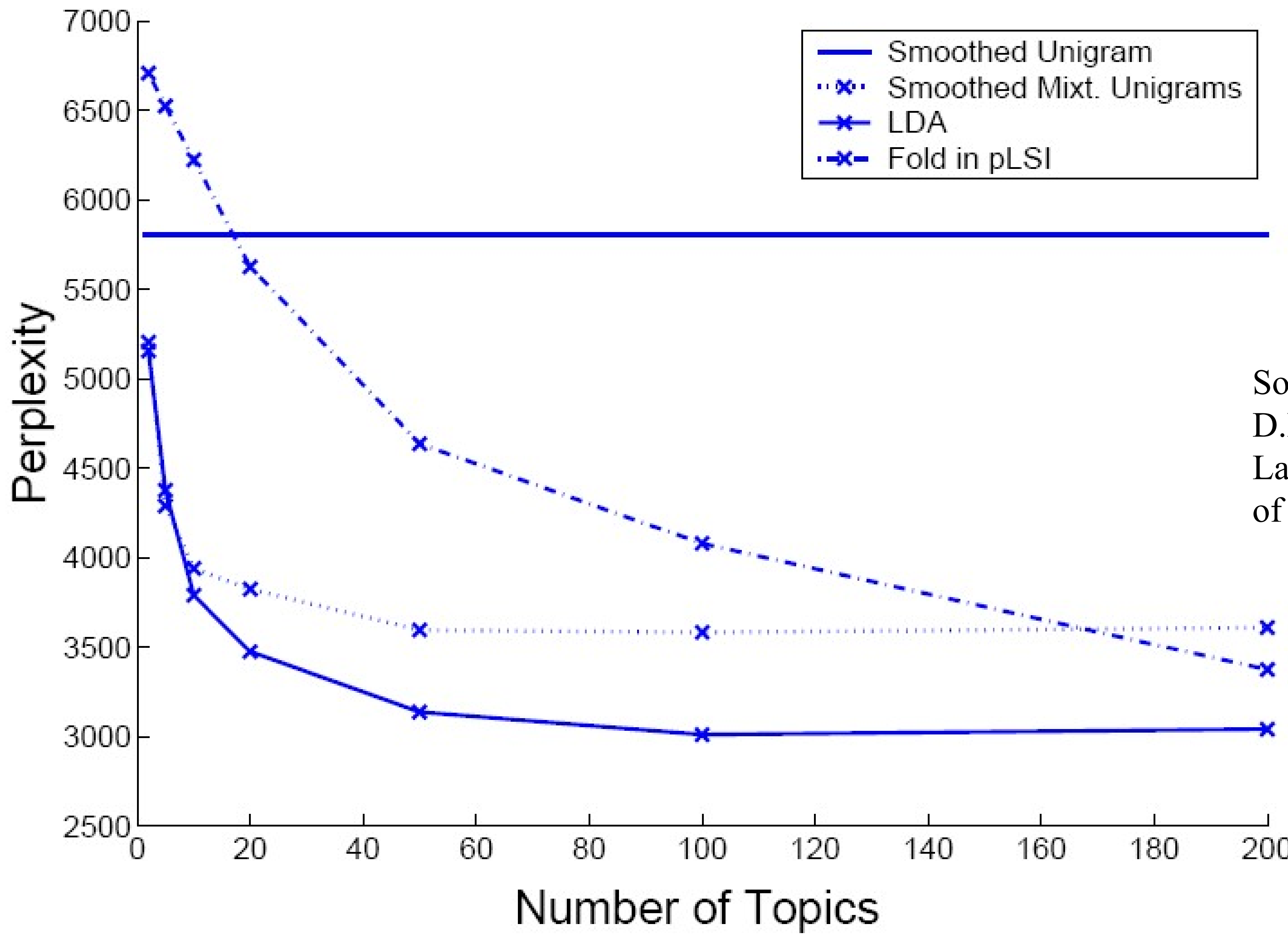
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	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.

Source: Blei et al., 2003

LDA experimental results: perplexity

on corpus of 16333 AP newswire articles



Source:
D.M. Blei, A.Y. Ng, M.I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research 2003

Summary of LDA

- Adds a generative model to pLSI
- Generally thought to be better than pLSI
  - Some recent work enhances pLSI and makes it better
  - pLSI = LDA with uniform Dirichlet prior
- Expensive computations and expensive query processing

IX.4 Dimensionality reduction

1. Curse of dimensionality
2. Matrix factorization to help – Feature extraction
3. Johnson–Lindenstrauss lemma
4. Feature selection

Zaki & Meira, Ch. 6 & 8

Curse of dimensionality

- Many data mining algorithms need to work in high-dimensional data
- But life gets harder as dimensionality increases
  - The volume grows too fast

Curse of dimensionality

- Many data mining algorithms need to work in high-dimensional data
- But life gets harder as dimensionality increases
  - The volume grows too fast
    - 100 points evenly-spaced points in unit interval have max distance between adjacent points of 0.01

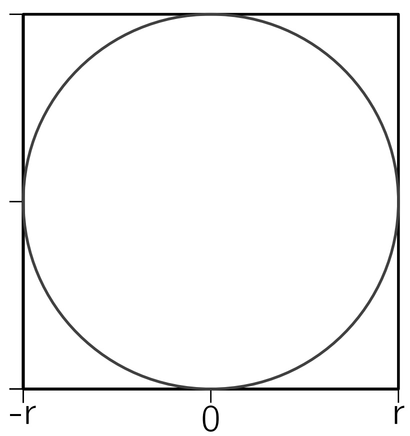
Curse of dimensionality

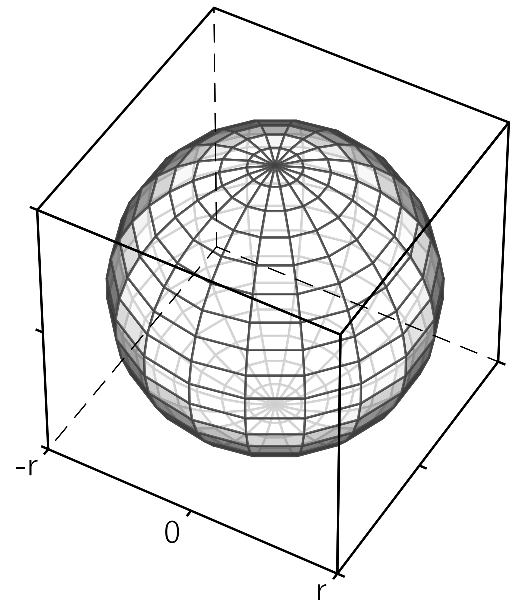
- Many data mining algorithms need to work in high-dimensional data
- But life gets harder as dimensionality increases
  - The volume grows too fast
    - 100 points evenly-spaced points in unit interval have max distance between adjacent points of 0.01
    - To get that distance for adjacent points in 10-dimensional unit hypercube requires 10^{20} points
    - Factor of 10^{18} increase

Hypersphere and hypercube

- Hypercube is d-dimensional cube with edge length $2r$
  - Volume: $\mathrm{vol}(H_d(2r)) = (2r)^d$
- Hypersphere is the d-dimensional ball of radius r
  - $\mathrm{vol}(S_1(r)) = 2r$
  - $\mathrm{vol}(S_2(r)) = \pi r^2$
  - $\mathrm{vol}(S_3(r)) = 4/3\ \pi r^3$
  - $\mathrm{vol}(S_d(r)) = K_d r^d$, where $K_d = \dfrac{\pi^{d/2}}{\Gamma(d/2+1)}$
    - $\Gamma(d/2 + 1) = (d/2)!$ for even d

Hypersphere within hypercube

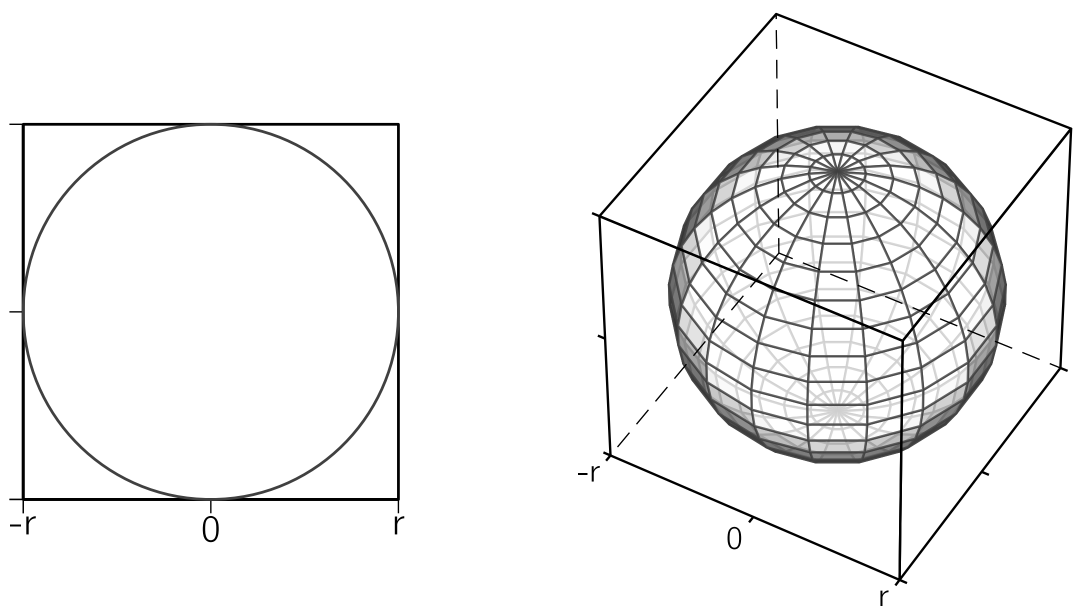




Fraction of volume hypersphere has of surrounding hypercube:

higher dimensions

Hypersphere within hypercube

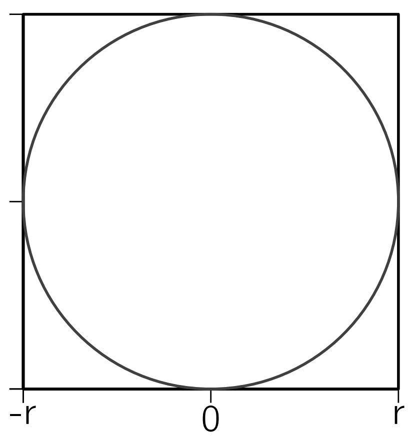


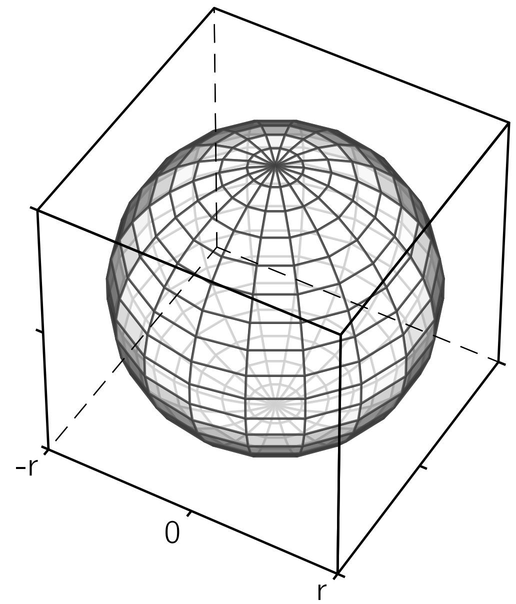
Fraction of volume hypersphere has of surrounding hypercube:

$$\lim_{d \to \infty} \frac{\mathrm{vol}(S_d(r))}{\mathrm{vol}(H_d(2r))} = \lim_{d \to \infty} \frac{\pi^{d/2}}{2^d \Gamma(d/2+1)} \to 0$$

higher dimensions

Hypersphere within hypercube





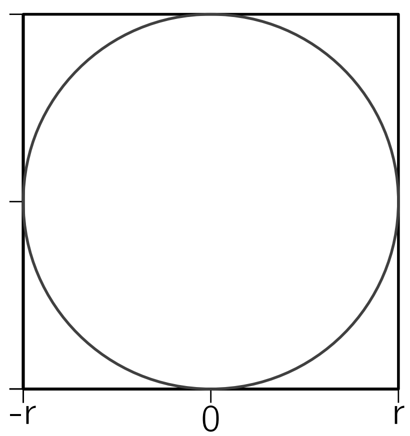
Mass is in the corners!

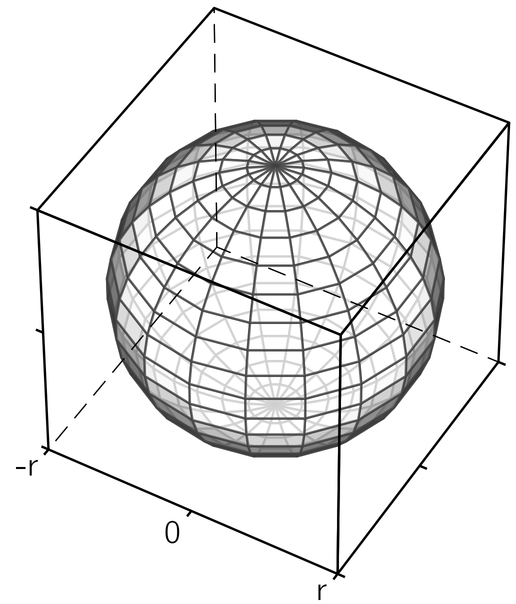
Fraction of volume hypersphere has of surrounding hypercube:

$$\lim_{d\to\infty} \frac{\mathrm{vol}(S_d(r))}{\mathrm{vol}(H_d(2r))} = \lim_{d\to\infty} \frac{\pi^{d/2}}{2^d\Gamma(d/2+1)} \to 0$$

higher dimensions

Hypersphere within hypercube

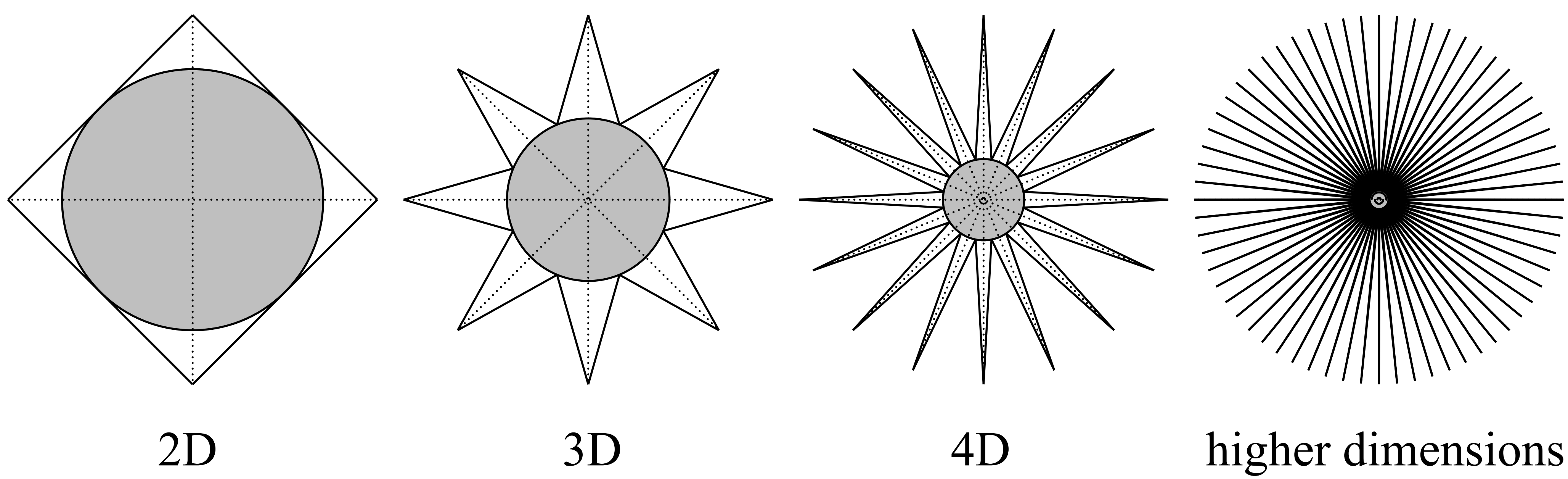




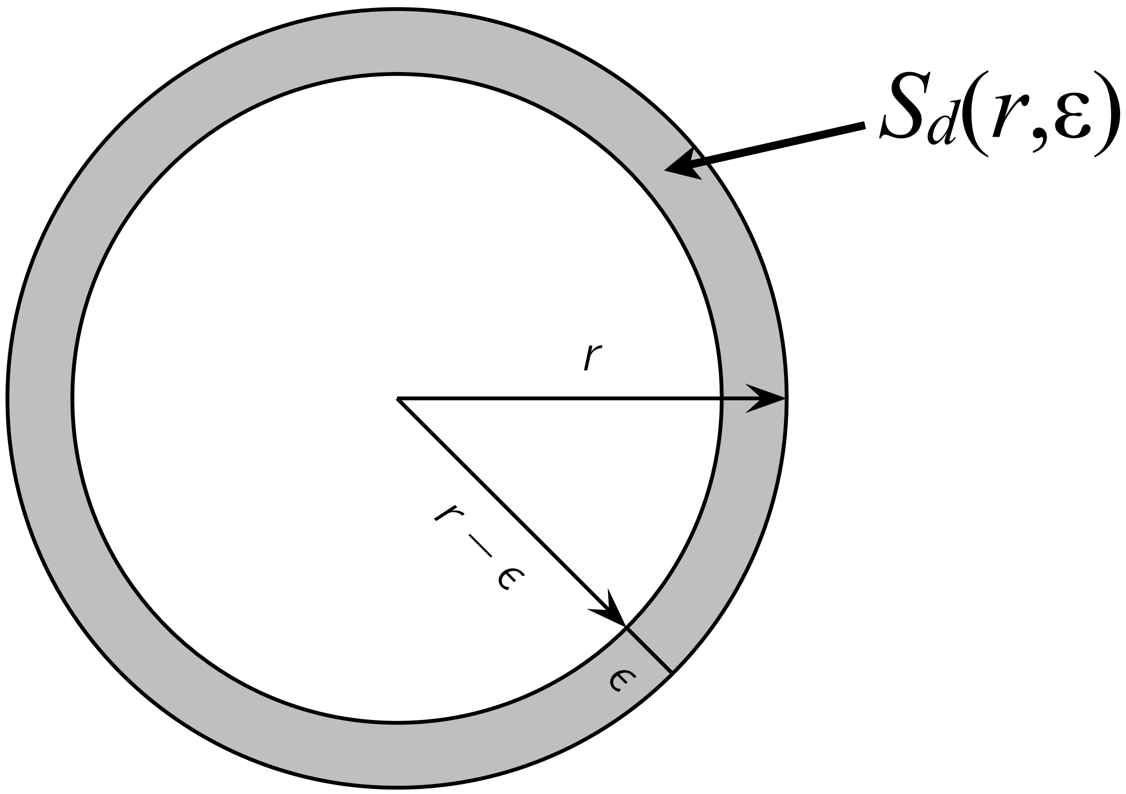
Mass is in the corners!

Fraction of volume hypersphere has of surrounding hypercube:

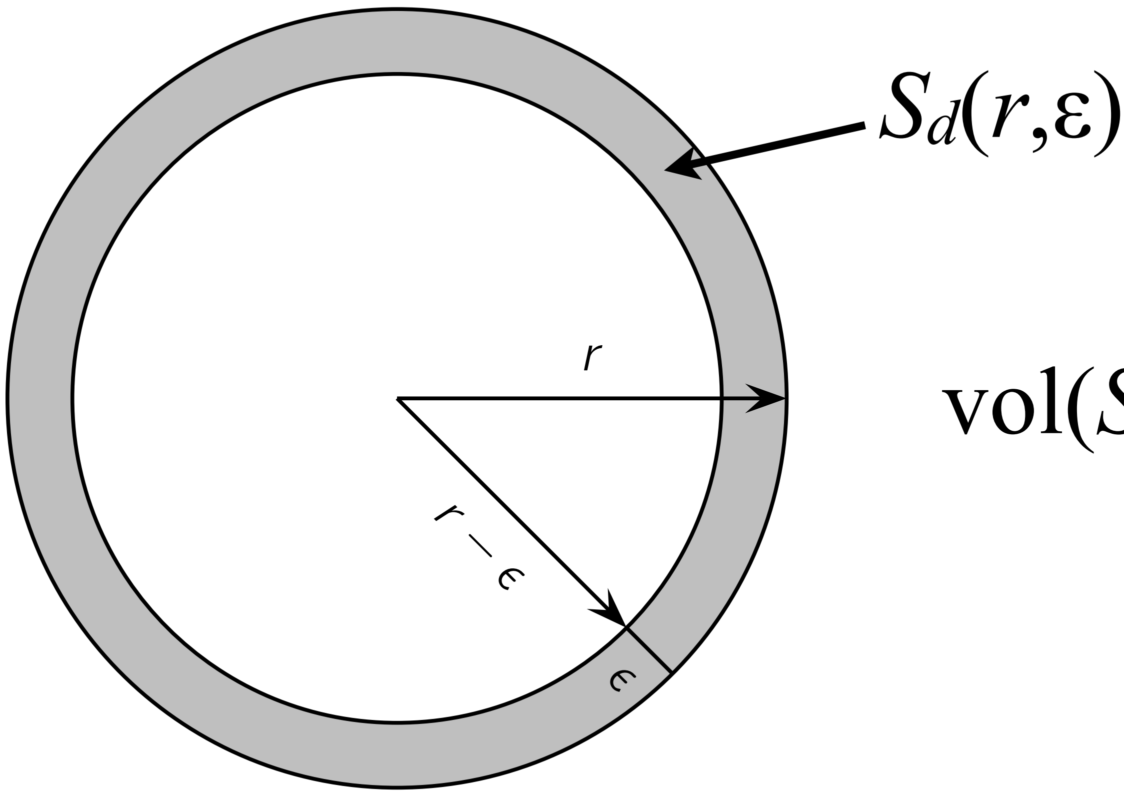
$$\lim_{d\to\infty} \frac{\mathrm{vol}(S_d(r))}{\mathrm{vol}(H_d(2r))} = \lim_{d\to\infty} \frac{\pi^{d/2}}{2^d \Gamma(d/2+1)} \to 0$$



Volume of thin shell of hypersphere

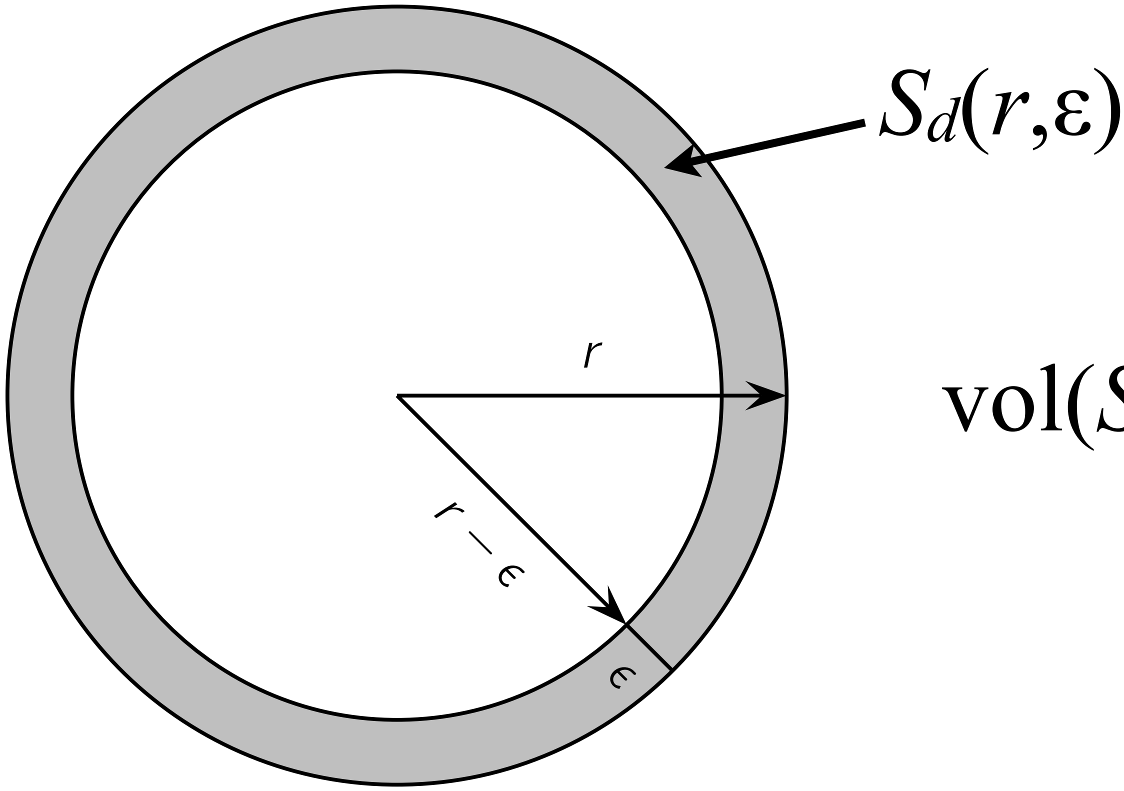


Volume of thin shell of hypersphere



$$\begin{aligned}\mathrm{vol}(S_d(r,\varepsilon)) &= \mathrm{vol}(S_d(r)) - \mathrm{vol}(S_d(r-\varepsilon)) \\ &= K_d r^d - K_d(r-\varepsilon)^d\end{aligned}$$

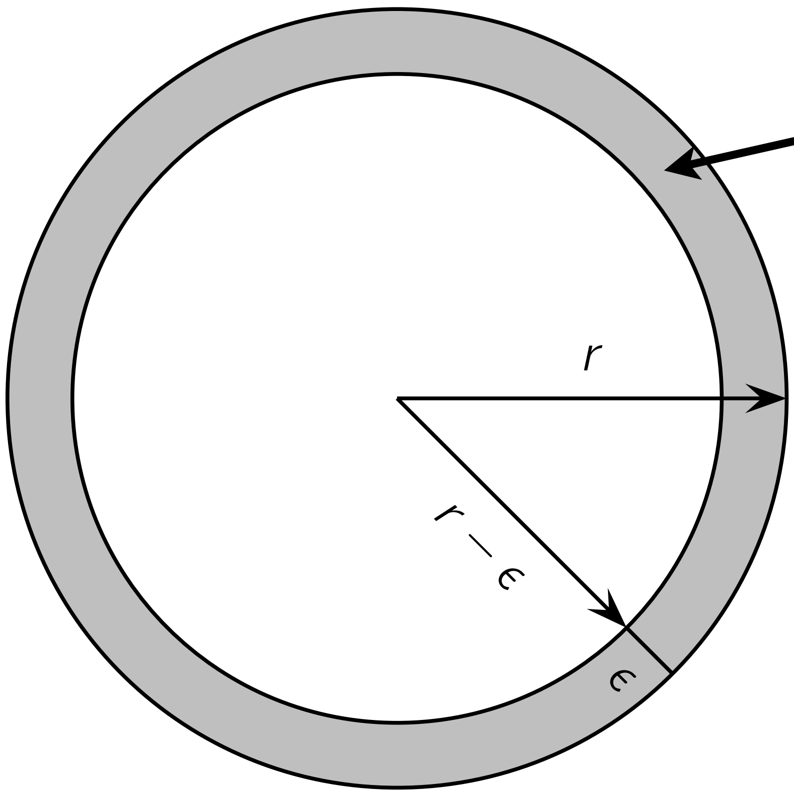
Volume of thin shell of hypersphere



$$\mathrm{vol}(S_d(r,\varepsilon)) = \mathrm{vol}(S_d(r)) - \mathrm{vol}(S_d(r-\varepsilon))$$
$$= K_d r^d - K_d(r-\varepsilon)^d$$

Fraction of volume in the shell: $\dfrac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = 1 - \left(1 - \dfrac{\epsilon}{r}\right)^d$

Volume of thin shell of hypersphere

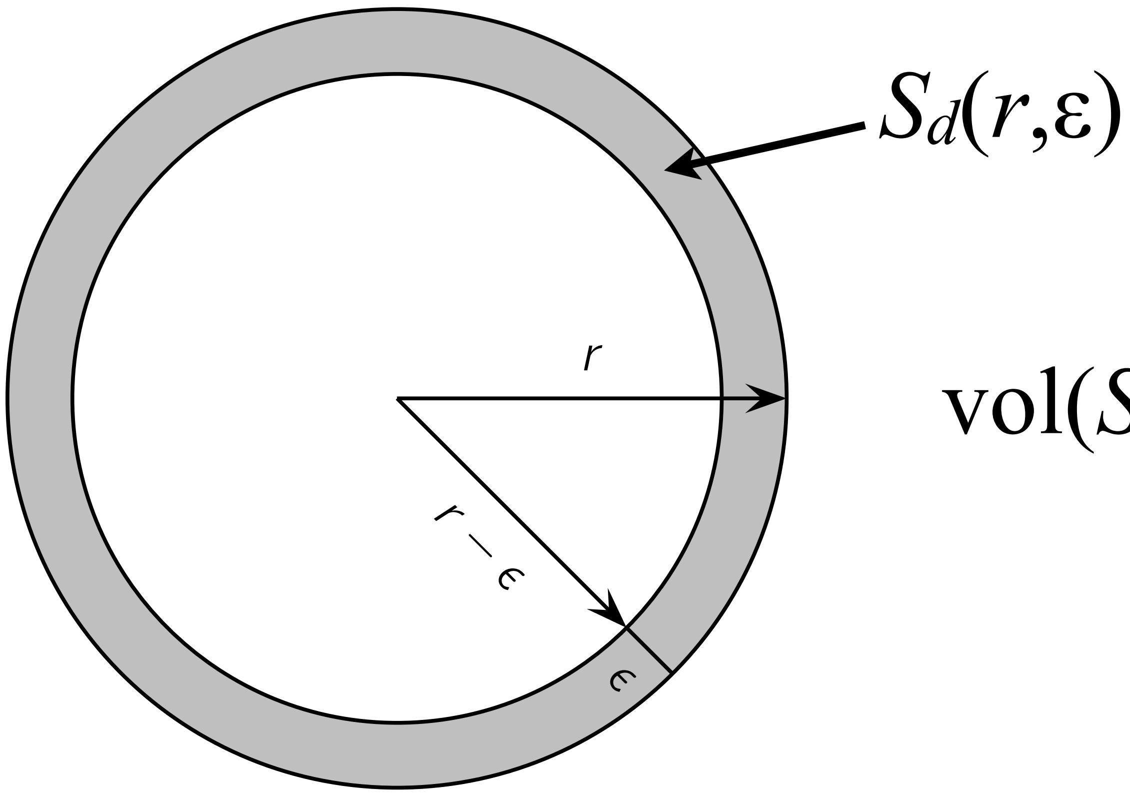


$$\mathrm{vol}(S_d(r,\varepsilon)) = \mathrm{vol}(S_d(r)) - \mathrm{vol}(S_d(r-\varepsilon))$$
$$= K_d r^d - K_d(r-\varepsilon)^d$$

Fraction of volume in the shell: $\dfrac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = 1 - \left(1 - \dfrac{\epsilon}{r}\right)^d$

$$\lim_{d\to\infty} \frac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d \to 1$$

Volume of thin shell of hypersphere



$$\mathrm{vol}(S_d(r,\varepsilon)) = \mathrm{vol}(S_d(r)) - \mathrm{vol}(S_d(r-\varepsilon))$$
$$= K_d r^d - K_d(r-\varepsilon)^d$$

Fraction of volume in the shell: $\dfrac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = 1 - \left(1 - \dfrac{\epsilon}{r}\right)^d$

$$\lim_{d\to\infty} \frac{\mathrm{vol}(S_d(r,\epsilon))}{\mathrm{vol}(S_d(r))} = \lim_{d\to\infty} 1 - \left(1 - \frac{\epsilon}{r}\right)^d \to 1$$

Mass is in the shell!

Feature extraction

- Aim: reduce the number of features by replacing them with new ones
- Tools: PCA (and other matrix factorizations)
  - Typical matrix factorizations give linear transformation
    - Projection of data to small-dimensional subspace
  - Using so-called *kernel trick* we can have non-linear transformations
    - See Zaki & Meira for more on kernel trick

Johnson–Lindenstrauss lemma

- Finding the decomposition can be expensive
- Decompositions give only *global* guarantees
  - Any pair of points can have very different distances
- Can we guarantee *local* similarity?

Johnson–Lindenstrauss lemma. Given $\varepsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = O(\varepsilon^{-2} \log n)$. For every set X of n points in $\mathbb{R}^d$ there exists $F\colon \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\boldsymbol{x}_i, \boldsymbol{x}_j \in X$

$$(1-\varepsilon)\,\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leqslant \|F(\mathbf{x}_i) - F(\mathbf{x}_j)\|^2 \leqslant (1+\varepsilon)\,\|\mathbf{x}_i - \mathbf{x}_j\|^2$$

How to find the projections?

- We need to find an k-by-d matrix $\boldsymbol{R} = (r_{ij})$ such that function $\boldsymbol{x} \mapsto \boldsymbol{Rx}$ satisfies JL
- Remarkably, if we select $r_{ij} \sim N(0,1)$, $\boldsymbol{R}$ satisfies JL with high probability
  - That is, JL holds for *all* points of X with high probability
- Achlioptas has show that we can also select $\Pr[r_{ij} = 1] = 1/2$ and $\Pr[r_{ij} = -1] = 1/2$ or $\Pr[r_{ij} = 1] = 1/6$, $\Pr[r_{ij} = 0] = 2/3$, $\Pr[r_{ij} = -1] = 1/6$
  - Sparse matrix

Feature selection

- Sometimes we want to retain the original features
  - Interpretability
  - Sparsity
  - …
- We can select the most important features and work only on them
  - Greedy algorithm: start with one feature and add new ones based on how much they improve
    - Improvement can be hard to compute
  - One can also use CX matrix decomposition
    - Matrix C selects the features