JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Training Topic Models on Streaming Text Data

Sophie Burkhardt and Zahra Ahmadi

Johannes Gutenberg-Universität, Mainz

October 04, 2019

Outline

- Latent Dirichlet Allocation (LDA) - *Sophie*
  - Gibbs Sampling
  - Variational Bayes
- **Practical Exercise: Use gensim library to train an LDA model**
- Online Learning and Concept Drifting Data Streams - *Zahra*
- Online LDA - *Sophie*
- **Practical Exercise: Implement and Train Online LDA Model to Analyze Data over Time**

Acknowledgements

- Some Slides adapted from
  - ChengXiang Zhai
  - Ido Abramovich
  - Ramesh Nallapati

Topic Models

- Discover topics in large amounts of text data (e.g. news, social media, scientific papers etc.)
- Provide clustering of documents
- Get semantic description of discovered topics

Topic Models

- Why use topic modeling?
  - Quick way of finding major themes in large text datasets
- Different topic modeling techniques:
  - Hierarchical topic models:
    - PAM (Pachinko Allocation Model): Finds Super-Topics and Sub-Topics
    - Nested CRP: Discovers topic hierarchies with arbitrary depth
  - Dynamic Topic Model: Topics change over time
  - Labeled Topic Models: Use annotations as labels

LDA topics

"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

LDA's view of a document

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.

"Arts" **"Budgets"** **"Children"** **"Education"**

How are Topic Models relevant for Non-Text Data…?

- Discover patterns in biological data
  - Clustering
  - Data classification
  - Feature extraction
- Gene sequence data
- Protein sequence data
- Predicting protein functions
- Patterns in images

A summary of the analogies between document-topic-word and a biological object in the relevant studies (see ""Document-word-topic" in biological data" section)

Reference	Words	Topics	Documents	Biological dataset
Rogers et al. (2005), Masada et al. (2009), Perina et al. (2010), Bicego et al. (2010a, b, 2012), Lee et al. (2014)	Genes	Functional groups	Samples	Expression microarray data
Masseroli et al. (2012), Pinoli et al. (2013, 2014), Youngs et al. (2014)	Ontological terms	Latent relationship	Proteins	Protein annotations
Chen et al. (2010, 2012a, b), La Rosa et al. (2015), Zhang et al. (2015)	K-mers of DNA sequences	Taxonomic category/components of the whole genome	DNA sequences	Genomic sequences
Caldas et al. (2009)	Gene sets	Biological process	Experiments	Gene expression dataset
Coelho et al. (2010)	Object classes	Fundamental patterns	Images	Fluorescence images
Konietzny et al. (2011)	A fixed-sized vocabulary of words based on the gene annotations	Functional modules of protein families	Genome annotations	A set of genome annotations
Bisgin et al. (2013)	Endpoint measurements	Diagnostic topics	Drugs	Expression of the HCS endpoints
Chen et al. (2011), Randhave and Sonkamble (2014)	Functional elements (NCBI taxonomic level indicators, indicator of gene orthologous groups and KEGG pathway indicators)	Functional groups	Samples	Genome set

Liu et al. (2016) An overview of topic modeling and its current applications in bioinformatics SpringerPlus vol. 5

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:

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

An infinite sequence of random is *infinitely exchangeable* if every finite subsequence is exchangeable

bag-of-words Assumption

Word order is ignored

“bag-of-words” – exchangeability

Theorem (De Finetti, 1935) – 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:

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

For some random variable θ

Notation and terminology

A *word* is an item from a vocabulary indexed by {1,…,V}. We represent words using unit-basis vectors. The *v*th word is represented by a V-vector *w* such that $w^v = 1$ and $w^v = 1$ for $u \neq v$. A *document* is a sequence of N words denoted by $d = (w_1, w_2, \dots, w_n)$, where w_n is the *n*th word in the sequence.

A *corpus* is a collection of M documents denoted by

$D = \{d_1, d_2, \dots, d_M\}$

LDA – generative process

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n

$$[\beta]_{k \times V} \quad \beta_{ij} = p(w^j = 1 | z^i = 1)$$

Dirichlet distribution

A *k*-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex, and has the following probability density on this simplex:

$$(1)\ p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

The LDA equations

$$(2)\ p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\,p(w_n|z_n,\beta)$$

$$(3)\ p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)\,p(w_n|z_n,\beta)\right)d^k\theta$$

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\,p(w_{dn}|z_{dn},\beta)\right)d^k\theta_d$$

LDA and exchangeability

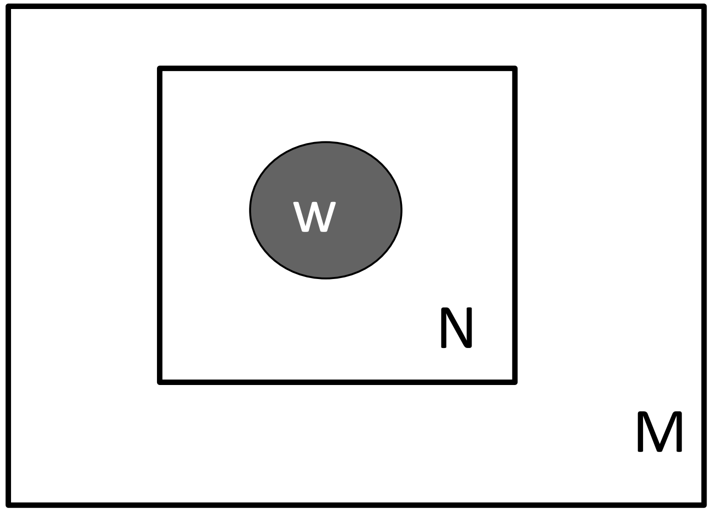
We assume that words are generated by topics and that those topics are infinitely exchangeable within a document.

By de Finetti's theorem:

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

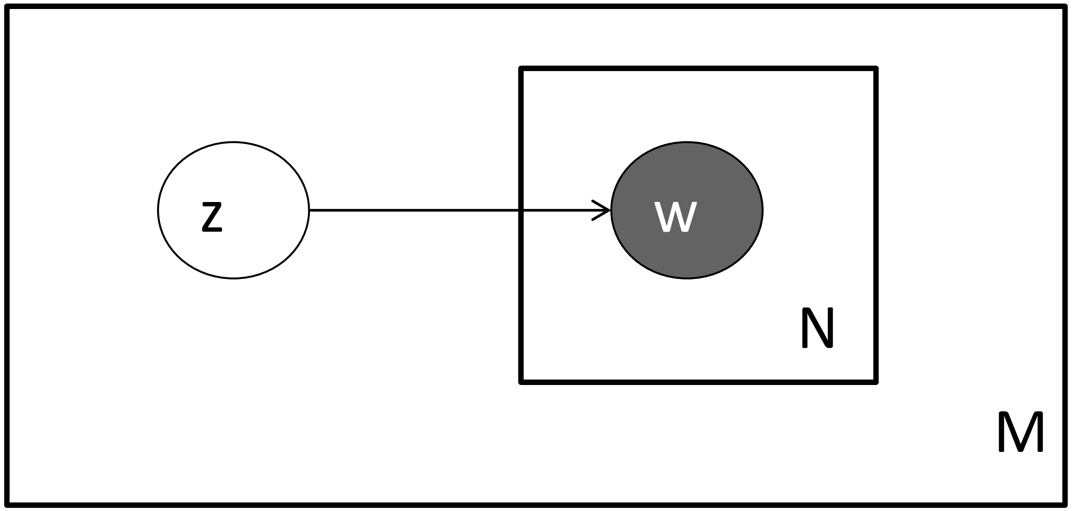
By marginalizing out the topic variables, we get eq. 3 in the previous slide.

Unigram model



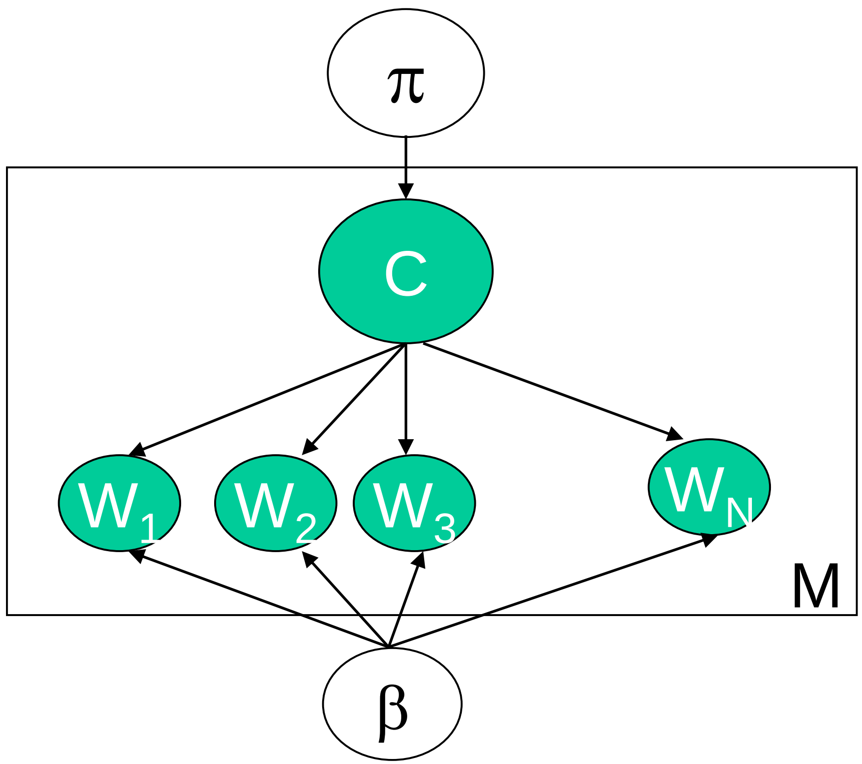
$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

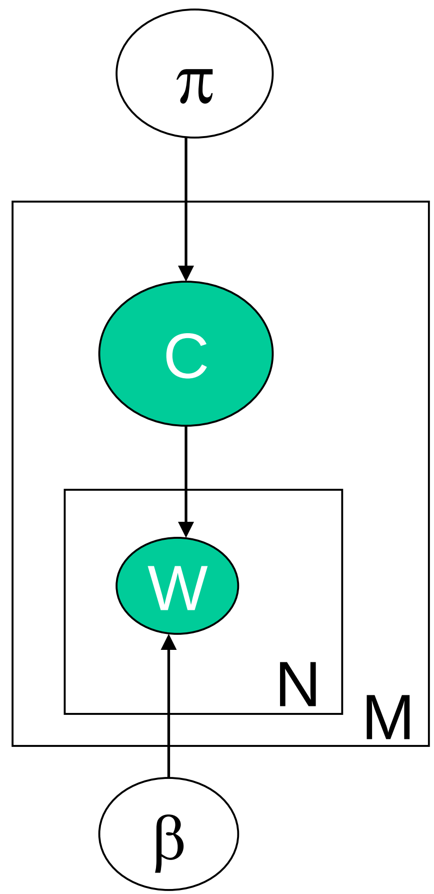
Mixture of unigrams



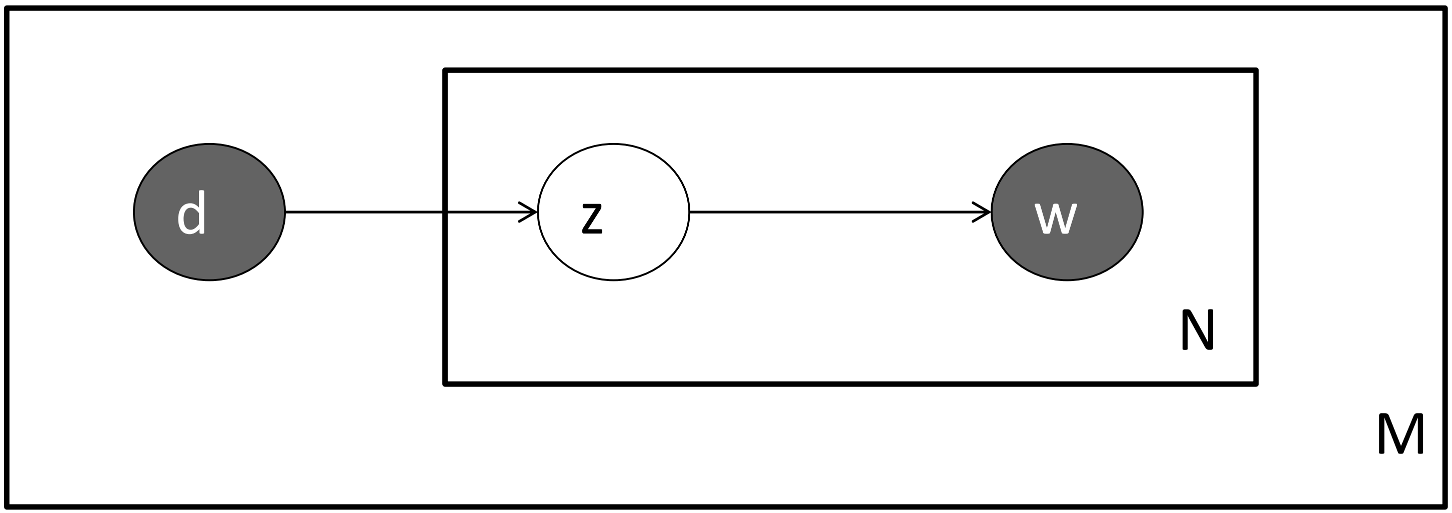
$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$

Naïve Bayes Model: Compact representation





Probabilistic LSI

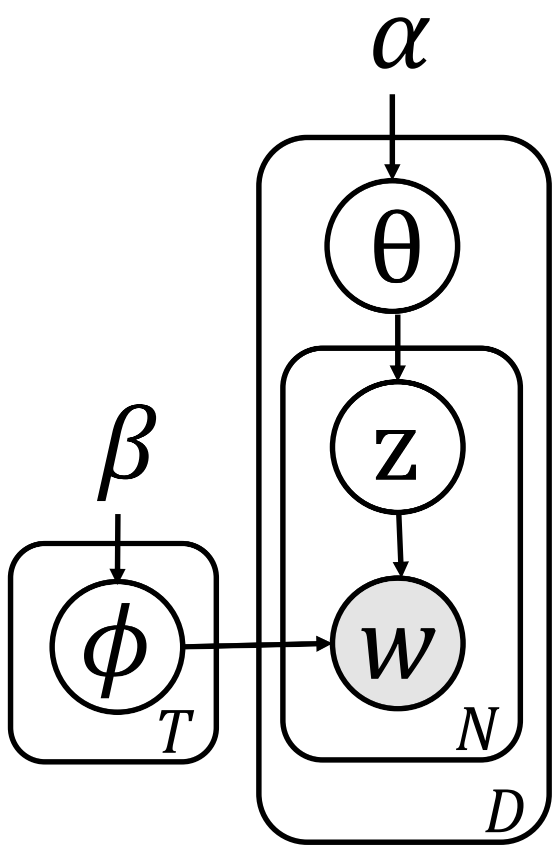


$$p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d)$$

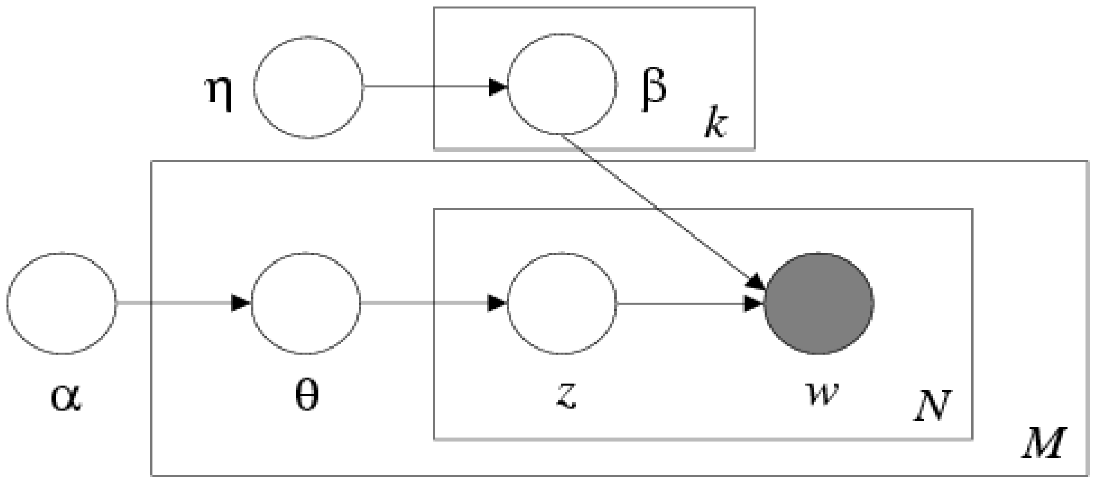
LDA

- Documents are mixtures of topics
- Matrix decomposition: $\gamma_n = \theta_n \times \phi$ where γ_n is the distribution over words for the nth document

$$
\begin{aligned}
\theta &\sim Dirichlet(\alpha) \\
\phi &\sim Dirichlet(\beta) \\
z &\sim Discrete(\theta) \\
w &\sim Discrete(\phi_z)
\end{aligned}
$$



Smoothed LDA



Introduces Dirichlet smoothing on β to avoid the “zero frequency problem”

More Bayesian approach

Inference and parameter learning similar to unsmoothed LDA

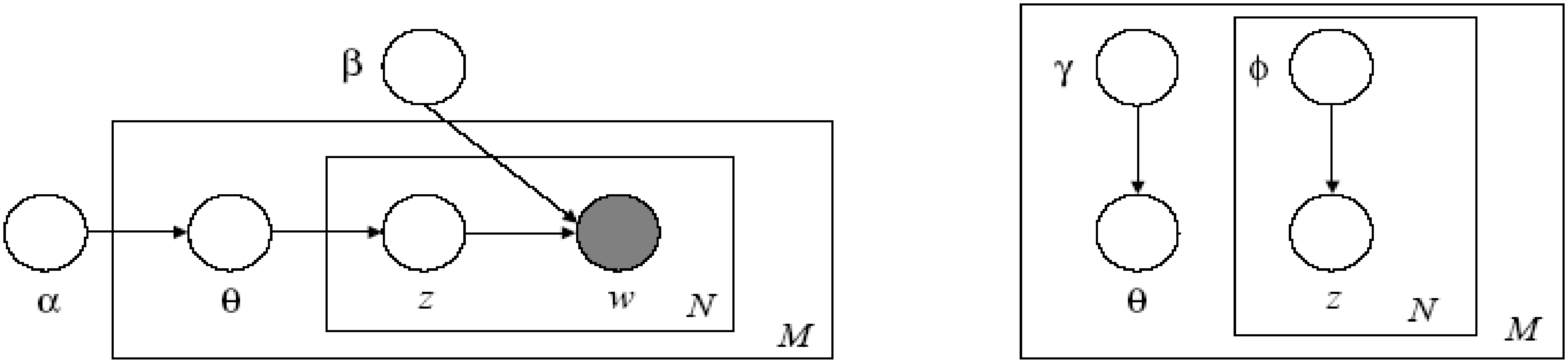
Inference

We want to compute the posterior dist. Of the hidden variables given a document:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}$$

Unfortunately, this is intractable to compute in general.

Variational inference



$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \phi) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

Parameter estimation

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \beta)$$

Variational EM

- (E Step) For each document, find the optimizing values of the variational parameters (γ, φ) with α, β fixed.
- (M Step) Maximize variational distribution w.r.t. α, β for the γ and φ values found in the E step.

Online Training

- Can be trained step by step using minibatches
- Hoffman, Blei, Bach, „Online Learning for Latent Dirichlet Allocation“ NIPS(2010).
- Variational updates:
- $\phi^t = (1-\rho)\phi^{t-1} + \rho\hat{\phi}^t$
- $\hat{\phi}^t$ is the estimate of the distribution based on the current minibatch

Gibbs sampling

Applicable when joint distribution is hard to evaluate but conditional distribution is known

Sequence of samples comprises a Markov Chain

Stationary distribution of the chain is the joint distribution

1. Initialise $x_{0,1:n}$.
2. For $i = 0$ to $N - 1$
   - Sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
   - Sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

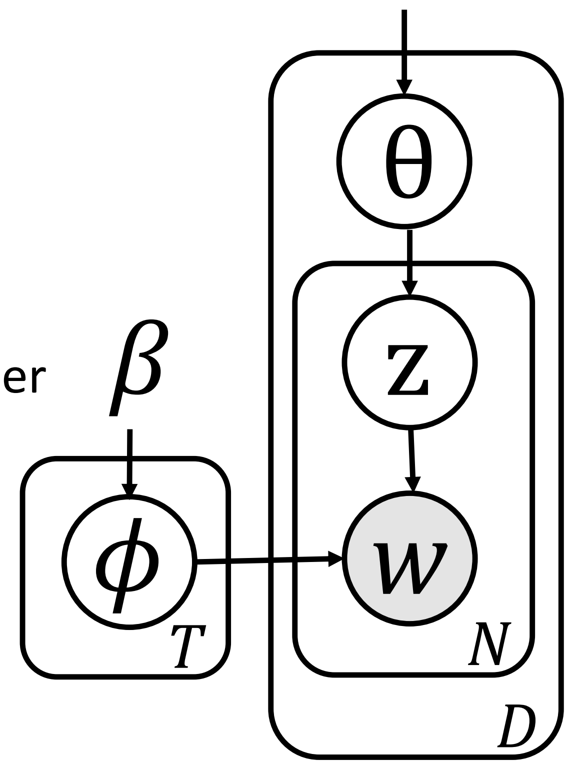
     $\vdots$
   - Sample $x_j^{(i+1)} \sim p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

     $\vdots$
   - Sample $x_n^{(i+1)} \sim p(x_n | x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

Collapsed Gibbs Sampling

$$P(z = t|\text{rest}) \propto \frac{n_{wt}+\beta}{n_t+\sum\beta}(n_{td}+\alpha)$$

- t: topic, w: word
- n_{wt} : number of times topic *t* and word *w* occur together
- n_{td} : number of times topic *t* occurs in document *d*
- Leads to topic models that have
  - few topics per document,
  - few words per topic



Document modeling

Unlabeled data – our goal is density estimation.

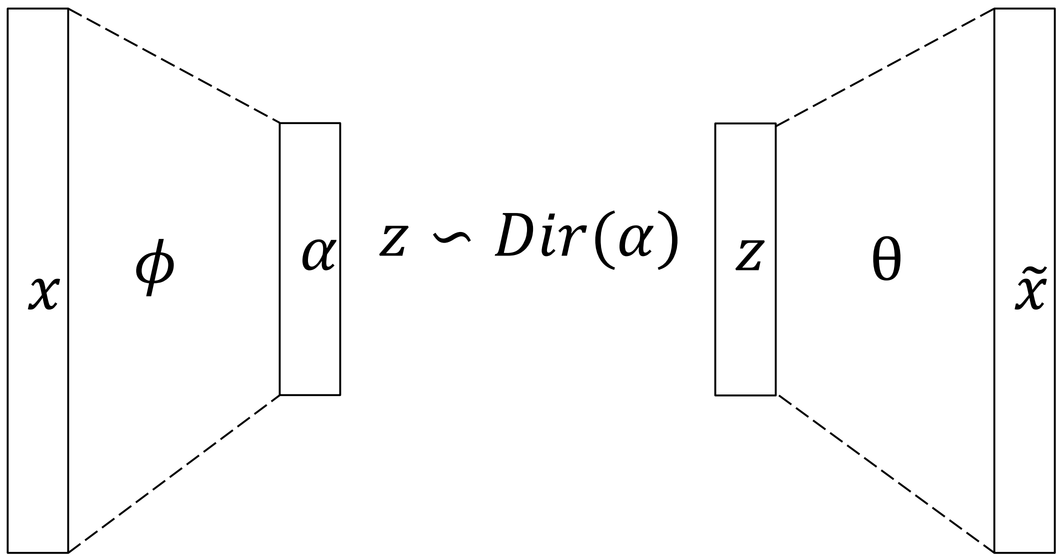
Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

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

.

Outlook

- Bayesian models increasingly trained using deep neural networks
  - Variational Autoencoders (prior on the latent variables)
  - Bayes by Backprop (prior on the weights)
- Several approaches for topic models using autoencoders exist
- Efficient hierarchical models still an open problem
- Consideration of word order and word structure possible -> many directions of research



Summary

- Based on the exchangeability assumption
- Can be viewed as a dimensionality reduction technique
- Exact inference is intractable, we can approximate instead
  - Variational Inference
  - Gibbs Sampling

Outline

- Latent Dirichlet Allocation (LDA) - *Sophie*
  - Gibbs Sampling
  - Variational Bayes
- **Practical Exercise: Use gensim library to train an LDA model**
- **Online Learning and Concept Drifting Data Streams - *Zahra***
- **Online LDA - *Sophie***
- **Practical Exercise: Implement and Train Online LDA Model to Analyze Data over Time**

Outline

- Latent Dirichlet Allocation (LDA) - *Sophie*
  - Gibbs Sampling
  - Variational Bayes
- **Practical Exercise: Use gensim library to train an LDA model**
- Online Learning and Concept Drifting Data Streams - *Zahra*
- **Online LDA - *Sophie***
- **Practical Exercise: Implement and Train Online LDA Model to Analyze Data over Time**

Refugee Crisis Dataset

- Arrival wave started in August 2015
- Many related events and polarized debates

- Data: German media between 01/2016 and 05/2017, filtered according to relevance to refugee crisis
  - Journalistic media (no social media)
- 208,683 articles
- 71,633 features

JG|U

Research Questions

- How can we use this unique dataset?
  - Understand public opinion (what the world thinks)
    - Main concerns of different groups
  - How is public opinion influenced by certain events?
    - How does it evolve over time?
  - Discover biases in the media
    - What is reported disproportionately often?
    - Are important topics ignored?

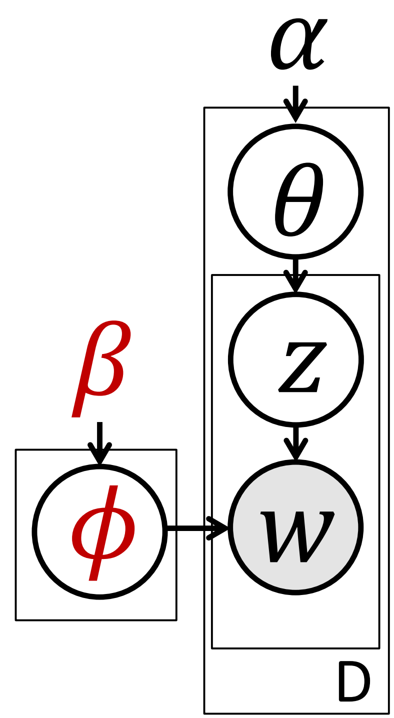
Topic Modeling

- Problems of most techniques:
  - Complex models take lots of time to produce questionable results
- Solution:
  - Focus on simple but effective methods based on standard LDA

Latent Dirichlet Allocation: Collapsed Gibbs Sampling

$$\mathrm{P}(\mathrm{z} = \mathrm{k}|\mathrm{rest}) \propto \frac{n_{kw}+\beta_{kw}}{n_k+\sum_w \beta_{kw}}(n_{kd}+\alpha)$$

- k: topic, w: word
- n_{kw} : number of times topic *k* and word *w* occur together
- n_{kd} : number of times topic *k* occurs in document *d*
- Leads to topic models that have
  - few topics per document,
  - few words per topic
- Topic-word distribution: $\phi \sim Dir(\beta)$



Online Topic Models

- Split data into different time slices $D = \{D^1, \ldots, D^{t-1}, D^t\}$
- Goal:
  - Learn topic word distributions ϕ_{kw}^t for each time slot t.

$$\phi^1 \longrightarrow \quad \phi^2 \longrightarrow \quad \phi^3 \longrightarrow$$

On-line LDA (AlSumait et al.)

- AlSumait et al. On-line LDA, ICDM (2008)
- Parameters β are a weighted mixture of $\phi^1, \ldots, \phi^{t-1}$
- $\beta_k^t = \sum_{t'=1}^{t-1} \omega^{t'} \phi_k^{t'}$
- Problem:
  - Have to keep all matrices ϕ^t for all time slots in memory
- Keep only last time slot ➡ information lost from previous time slots

Online Variational Bayes (Hoffman et al.)

- Inspiration for our method
- Hoffman, Blei, Bach, „Online Learning for Latent Dirichlet Allocation“ NIPS(2010).
- Variational updates:
- $\phi^t = (1-\rho)\phi^{t-1} + \rho\hat{\phi}^t$
- Problem:
  - ϕ^t is heavily influenced by the previous time slot ϕ^{t-1}
  - We want a model to find the topics specific to one time slot only
  - This model converges to one global distribution

- $\rho \equiv (\tau_0 + t)^{-\kappa}$, e.g. $\kappa \in (0.5,1], \tau_0 \geq 0$ and t is the iteration number

Our Online Topic Model

- Variational updates for parameter β:
- $\beta^t = (1-\rho)\beta^{t-1} + \rho\phi^{t-1}$

- Advantage:
  - Several iterations over documents for one time slot will learn a distribution ϕ^t specific to time slot t

Comparison to Online Variational Bayes

- Different purpose
- Online variational Bayes:
  - increasingly small updates to incrementally improve one global model
- Our method:
  - Learn a chain of separate models that are linked through their priors
  - Individual models may be trained with variational Bayes or Gibbs sampling

Online Variational Bayes

$$\begin{tikzcd} \beta^1 \arrow[d] & \beta^2 \arrow[d] & \\ \phi^1 \arrow[r] & \phi^2 \arrow[r] & {} \end{tikzcd}$$

Our Method

$$\begin{tikzcd} \beta^1 \arrow[d] \arrow[r] & \beta^2 \arrow[d] \arrow[r] & {} \\ \phi^1 \arrow[ur] & \phi^2 \arrow[ur] & \end{tikzcd}$$

Results (1)

- Number of topics: 100
- One time slot: 10,000 documents
- 100 iterations per time slot

- Wide range of topics related to e.g.: financial system, border security, brexit, EU, EU commission, different political parties, countries (Turkey, Greece, Syria, Libya etc.), conflict between EU and hungary, deportation to Afghanistan

Results (2)

- Events that are reflected in one topic about the AfD (political party):
  - Bundestag elections, 24 September 2017
  - Erwin Sellering (SPD, Mecklenburg Vorpommern) resigns 30 May 2017
  - Helmut Seifen (AfD) was elected in NRW landtag election, 14 May 2017
  - Landtag election Saarland, 26 March 2017
  - AfD announces they want to leave Paris climate agreement, 9 March 2017
  - AfD party convention, 30 April 2016, 22 April 2017

2016-01-18
afd men pictures
germany members
pegida petry
fugitives leave …

→

2016-07-05
afd **party_convention**
april elected named
remain perceived
germany best racism
relationships nationalism
level effort color …

2017-01-11
afd party
bundestag_election petry
percent polls april poll
party_convention frauke
union lucke government
cdu grünen linken …

→

2017-02-06
afd party parties germany
april cdu **bundestag_election**
election left right_populist
alternative saar
landtag_election
election_campaign …

→

2017-03-08
afd german members government majority keeps participates stop planned us president exit **climate_change paris** …

→

2017-03-31
afd refugee_politics **seifen helmut** election_campaign worker mobilization leave party parties citizens currently strong records elections refugee_numbers times terror …

2017-04-16
afd cdu parties bundestag politics party coalition answer spd grünen linke wagenknecht linnemann ask vote vacuum linken get contribute union …

→

2017-05-01
vorpommern mecklenburg afd strongest force parliament berlin union germany party according_to cdu parties brussels problems social_democratic landtag_election …

2016-01-18
job_market integration benefits draws hartz fast currently receives job refugees clear federal agency berlin federal_government job_center nahles

2017-02-11
integration club job_market german refugees racism projects schools football young companies pupils accomplish foundation state responsible offers

2017-02-23
integration labor_market nahles refugees ii beginning residential_status german tuesday andrea meanwhile would spd benefits

2017-03-24
trump csu frankfurt union merkel refugees integration berlin iran choose put police darmstadt live usa riots german afghanistan cdu

Summary Online Topic Modeling

- Online topic modeling method
  - Topics related over time
  - Time slot specific topics

- Open problems:
  - Separate facts from opinion
  - Separate different view points on certain topics
    - E.g. articles sympathetic with AfD or opposed to it

More work on topic models

- For work on online nonparametric topic models (number of topics not fixed) check out the code on github: https://github.com/sophieburkhardt/HybridHDP
- For topic models trained with variational autoencoders: https://github.com/sophieburkhardt/dirichlet-vae-topic-models
- Work on supervised topic models for multi-label classification: https://github.com/sophieburkhardt/Multi-Label-Topic-Modeling

Outline

- Latent Dirichlet Allocation (LDA) - *Sophie*
  - Gibbs Sampling
  - Variational Bayes
- **Practical Exercise: Use gensim library to train an LDA model**
- Online Learning and Concept Drifting Data Streams - *Zahra*
- Online LDA - *Sophie*
- **Practical Exercise: Implement and Train Online LDA Model to Analyze Data over Time**