Machine Learning (6.867) – Topic models

David Sontag

Massachusetts Institute of Technology

Lecture 17, Nov. 9, 2017

Course announcements

- HW3 due this Tuesday, 11/14
- No recitation this Friday – but there is an exercise for the week
- Readings on course information sheet have been updated (*TBD*s filled in)
- Readings for today's lecture:
  - **Applications of Topic Models**, Boyd-Graber, Hu, Mimno, in Foundations and Trends in Information Retrieval, 2017 (Sections 1 & 9)
  - Available for free within MIT

Today's lecture: outline

- **Warm up: topic mixture models**
  - General case of EM algorithm
  - Example derivation for topic mixture models
- Latent Dirichlet allocation
- Extensions of the basic approach
  - Polylingual topic models
  - Author-topic model
- Using topic models – inference and learning
  - Approximate inference
  - SGD with black-box variational inference

Expectation maximization

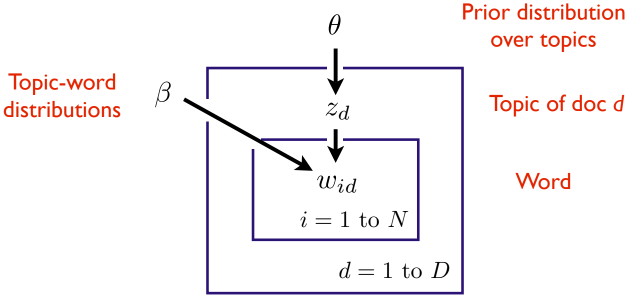
Algorithm is as follows:

1. Write down the **complete log-likelihood** $\log p(\mathbf{x}, \mathbf{z}; \theta)$ in such a way that it is linear in $\mathbf{z}$
2. Initialize θ_0, e.g. at random or using a good first guess
3. Repeat until convergence:

$$\theta_{t+1} = \arg\max_{\theta} E_{p(\mathbf{z}|\mathbf{x};\theta_t)}[\log p(\mathbf{x}, \mathbf{Z}; \theta)]$$

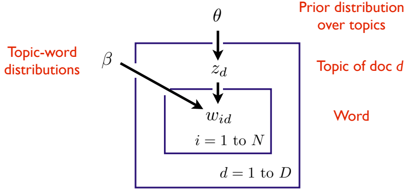
- Notice that $\log p(\mathbf{x}, \mathbf{Z}; \theta)$ is a random function because $\mathbf{Z}$ is unknown
- By linearity of expectation, objective decomposes into expectation terms and data terms
- "E" step corresponds to computing the objective (i.e., the **expectations**)
- "M" step corresponds to **maximizing** the objective

Application to mixture models



- Example of "plate" notation for graphical models
  - Variables within a plate are replicated in a conditionally independent manner
- This model is a type of (discrete) **mixture model**
  - Called *multinomial* naive Bayes (a word can appear multiple times)
  - Document is generated from a single topic
- Notation: we will use both K and T to denote number of topics

EM for mixture models



- The complete likelihood is $p(\mathbf{w}, \mathbf{Z}; \theta, \beta) = \prod_{d=1}^{D} p(\mathbf{w}_d, Z_d; \theta, \beta)$, where

$$p(\mathbf{w}_d, Z_d; \theta, \beta) = \theta_{Z_d} \prod_{i=1}^{N} \beta_{Z_d, w_{id}}$$

- Trick #1: re-write this as

$$p(\mathbf{w}_d, Z_d; \theta, \beta) = \prod_{k=1}^{K} \theta_k^{1[Z_d=k]} \prod_{i=1}^{N} \prod_{k=1}^{K} \beta_{k, w_{id}}^{1[Z_d=k]}$$

EM for mixture models

- Thus, the complete log-likelihood is:

$$\log p(\mathbf{w}, \mathbf{Z}; \theta, \beta) = \sum_{d=1}^{D} \left(\sum_{k=1}^{K} 1[Z_d = k] \log \theta_k + \sum_{i=1}^{N} \sum_{k=1}^{K} 1[Z_d = k] \log \beta_{k, w_{id}} \right)$$

- In the "E" step, we take the expectation of the complete log-likelihood with respect to $p(\mathbf{z} \mid \mathbf{w}; \theta^t, \beta^t)$, applying linearity of expectation, i.e.

$$E_{p(\mathbf{z} \mid \mathbf{w}; \theta^t, \beta^t)}[\log p(\mathbf{w}, \mathbf{z}; \theta, \beta)] =$$

$$\sum_{d=1}^{D} \left(\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \theta_k + \sum_{i=1}^{N} \sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \beta_{k, w_{id}} \right)$$

- In the "M" step, we maximize this with respect to θ and β

EM for mixture models

- Just as with complete data, this maximization can be done in closed form
- First, re-write expected complete log-likelihood from

$$\sum_{d=1}^{D}\left(\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \theta_k + \sum_{i=1}^{N}\sum_{k=1}^{K} p(Z_d = k \mid \mathbf{w}; \theta^t, \beta^t) \log \beta_{k,w_{id}}\right)$$

to

$$\sum_{k=1}^{K} \log \theta_k \sum_{d=1}^{D} p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t) + \sum_{k=1}^{K}\sum_{w=1}^{W} \log \beta_{k,w} \sum_{d=1}^{D} N_{dw}\, p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t)$$

- We then have that

$$\theta_k^{t+1} = \frac{\sum_{d=1}^{D} p(Z_d = k \mid \mathbf{w}_d; \theta^t, \beta^t)}{\sum_{\hat{k}=1}^{K}\sum_{d=1}^{D} p(Z_d = \hat{k} \mid \mathbf{w}_d; \theta^t, \beta^t)}$$

Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
…	…	…

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

   where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to N, sample the **topic** z_i of the i'th word

$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

   where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed collection of distributions on words)

Generative model for a document in LDA

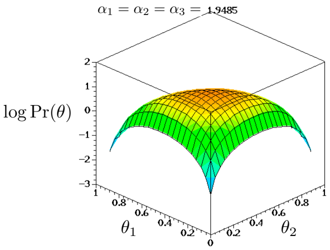
1. Sample the document's **topic distribution** θ (aka topic vector)

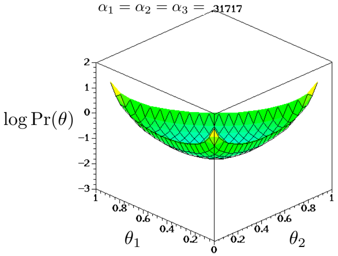
$$\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^{T}$ are hyperparameters.The Dirichlet density, defined over $\Delta = \{\vec{\theta} \in \mathbb{R}^T : \forall t \; \theta_t \geq 0, \sum_{t=1}^{T} \theta_t = 1\}$, is:

$$p(\theta_1, \ldots, \theta_T) \propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$

For example, for T=3 ($\theta_3 = 1 - \theta_1 - \theta_2$):





Generative model for a document in LDA

3. … and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^{T}$ are the *topics* (a fixed collection of distributions on words)

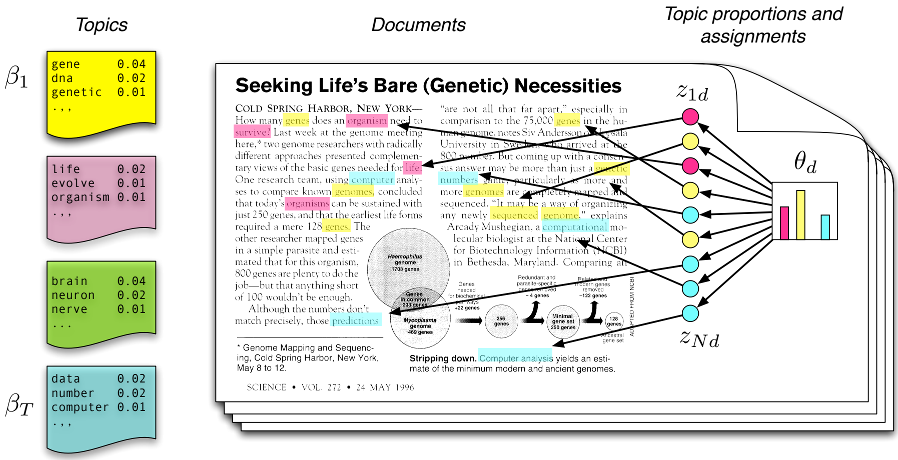
Documents

Topics

politics .0100	religion .0500	sports .0105
president .0095	hindu .0092	baseball .0100
obama .0090	judiasm .0080	soccer .0055
washington .0085	ethics .0075	basketball .0050
religion .0060	buddhism .0016	football .0045
...	...	...

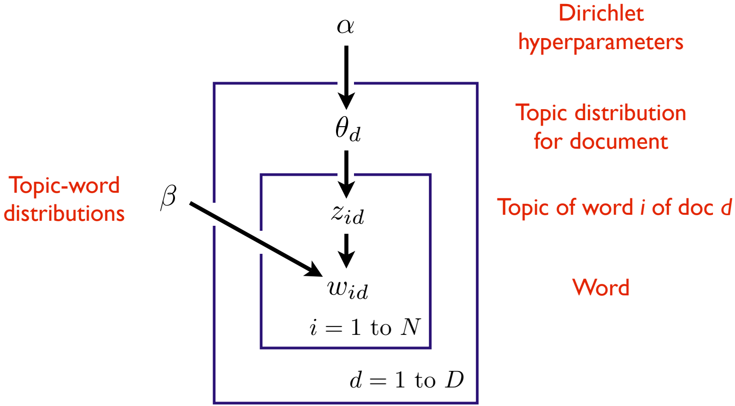
$$\beta_t = \{\, p(w \mid z = t) \,\}$$

Example of using LDA



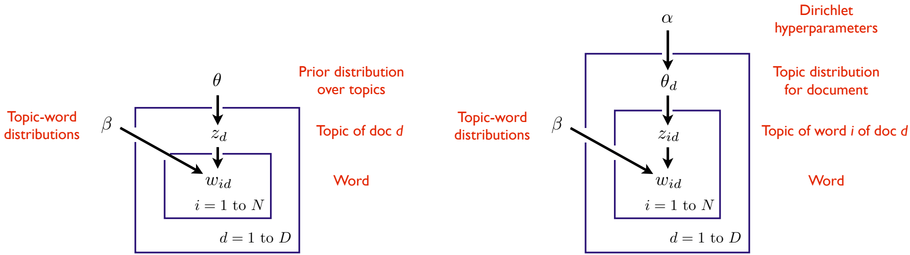
(Blei, *Introduction to Probabilistic Topic Models*, 2011)

"Plate" notation for LDA model



Variables within a plate are replicated in a conditionally independent manner

Comparison of mixture and admixture models



- Model on left is a **mixture model**
  - Called *multinomial* naive Bayes (a word can appear multiple times)
  - Document is generated from a single topic
- Model on right (LDA) is an **admixture model**
  - Document is generated from a distribution over topics

Demo

Explore topic models of:

- Politics over time
- State-of-the-union addresses
- Literary studies (explanation)
- Wikipedia

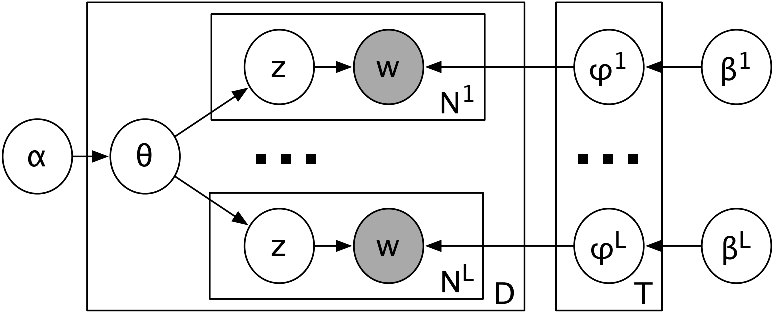
Today's lecture: outline

- Warm up: topic mixture models
  - General case of EM algorithm
  - Example derivation for topic mixture models
- Latent Dirichlet allocation
- **Extensions of the basic approach**
  - Polylingual topic models
  - Author-topic model
- Using topic models – inference and learning
  - Approximate inference
  - SGD with black-box variational inference

Polylingual topic models (Mimno et al., EMNLP '09)

- Goal: topic models that are aligned across languages
- Training data: corpora with multiple documents in each language
  - EuroParl corpus of parliamentary proceedings (11 western languages; exact translations)
  - Wikipedia articles (12 languages; not exact translations)
- How to do this?

Polylingual topic models (Mimno et al., EMNLP '09)



Learned topics

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

Learned topics

DA	børn familie udnyttelse børns børnene seksuel
DE	kinder kindern familie ausbeutung familien eltern
EL	παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής
EN	**children family child sexual families exploitation**
ES	niños familia hijos sexual infantil menores
FI	lasten lapsia lapset perheen lapsen lapsiin
FR	enfants famille enfant parents exploitation familles
IT	bambini famiglia figli minori sessuale sfruttamento
NL	kinderen kind gezin seksuele ouders familie
PT	crianças família filhos sexual criança infantil
SV	barn barnen familjen sexuellt familj utnyttjande

Discussion

- How would you use this?
- How could you extend this?

Author-topic model (Rosen-Zvi et al., UAI '04)

- Goal: topic models that take into consideration who the authors are
- Training data: corpora with label for who wrote each document
  - Papers from NIPS conference from 1987 to 1999
  - Twitter posts from US politicians
- Why do this?
- How to do this?

Author-topic model (Rosen-Zvi et al., UAI ’04)

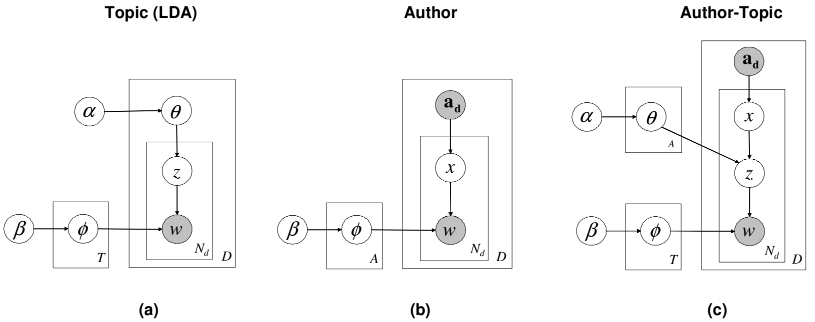


Figure 1: Generative models for documents. (a) Latent Dirichlet Allocation (LDA; Blei et al., 2003), a topic model. (b) An author model. (c) The author-topic model.

x denotes the author of a single word

Most likely author for a topic

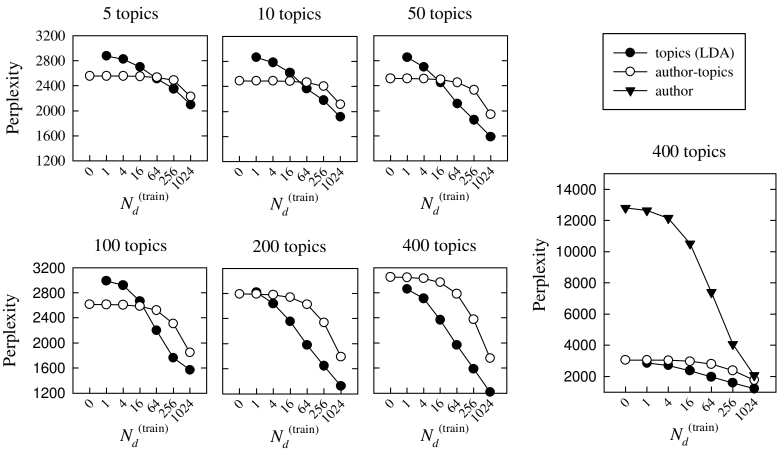
TOPIC 31	
WORD	**PROB.**
SPEECH	0.0823
RECOGNITION	0.0497
HMM	0.0234
SPEAKER	0.0226
CONTEXT	0.0224
WORD	0.0166
SYSTEM	0.0151
ACOUSTIC	0.0134
PHONEME	0.0131
CONTINUOUS	0.0129
AUTHOR	**PROB.**
Waibel_A	0.0936
Makhoul_J	0.0238
De-Mori_R	0.0225
Bourlard_H	0.0216
Cole_R	0.0200
Rigoll_G	0.0191
Hochberg_M	0.0176
Franco_H	0.0163
Abrash_V	0.0157
Movellan_J	0.0149

TOPIC 61	
WORD	**PROB.**
BAYESIAN	0.0450
GAUSSIAN	0.0364
POSTERIOR	0.0355
PRIOR	0.0345
DISTRIBUTION	0.0259
PARAMETERS	0.0199
EVIDENCE	0.0127
SAMPLING	0.0117
COVARIANCE	0.0117
LOG	0.0112
AUTHOR	**PROB.**
Bishop_C	0.0563
Williams_C	0.0497
Barber_D	0.0368
MacKay_D	0.0323
Tipping_M	0.0216
Rasmussen_C	0.0215
Opper_M	0.0204
Attias_H	0.0155
Sollich_P	0.0143
Schottky_B	0.0128

TOPIC 71	
WORD	**PROB.**
MODEL	0.4963
MODELS	0.1445
MODELING	0.0218
PARAMETERS	0.0205
BASED	0.0116
PROPOSED	0.0103
OBSERVED	0.0100
SIMILAR	0.0083
ACCOUNT	0.0069
PARAMETER	0.0068
AUTHOR	**PROB.**
Omohundro_S	0.0088
Zemel_R	0.0084
Ghahramani_Z	0.0076
Jordan_M	0.0075
Sejnowski_T	0.0071
Atkeson_C	0.0070
Bower_J	0.0066
Bengio_Y	0.0062
Revow_M	0.0059
Williams_C	0.0054

TOPIC 100	
WORD	**PROB.**
HINTON	0.0329
VISIBLE	0.0124
PROCEDURE	0.0120
DAYAN	0.0114
UNIVERSITY	0.0114
SINGLE	0.0111
GENERATIVE	0.0109
COST	0.0106
WEIGHTS	0.0105
PARAMETERS	0.0096
AUTHOR	**PROB.**
Hinton_G	0.2202
Zemel_R	0.0545
Dayan_P	0.0340
Becker_S	0.0266
Jordan_M	0.0190
Mozer_M	0.0150
Williams_C	0.0099
de-Sa_V	0.0087
Schraudolph_N	0.0078
Schmidhuber_J	0.0056

Perplexity as a function of number of observed words

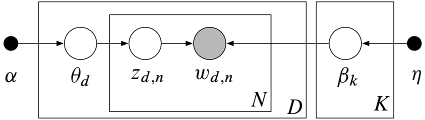


$$\text{perplexity}(\mathbf{w}_{test,d} \mid \mathbf{w}_{train,d}, \mathbf{a}_d) = \exp\left[-\frac{\ln p(\mathbf{w}_{test,d} | \mathbf{w}_{train,d}, \mathbf{a}_d)}{N_{test,d}}\right]$$

Today's lecture: outline

- Warm up: topic mixture models
  - General case of EM algorithm
  - Example derivation for topic mixture models
- Latent Dirichlet allocation
- Extensions of the basic approach
  - Polylingual topic models
  - Author-topic model
- **Using topic models – inference and learning**
  - Approximate inference
  - SGD with black-box variational inference

Posterior Inference

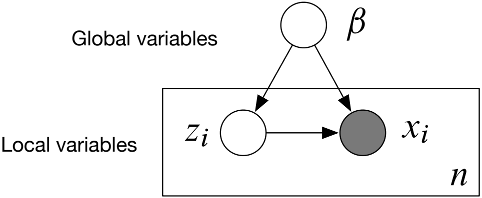


- The posterior of the latent variables given the documents is

$$p(\beta, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) = \frac{p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{\int_\beta \int_{\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\beta, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$
- We use approximate inference
  1. Monte-carlo methods (e.g. Gibbs sampling)
  2. Variational algorithms (e.g. mean-field)

A Generic Class of Models

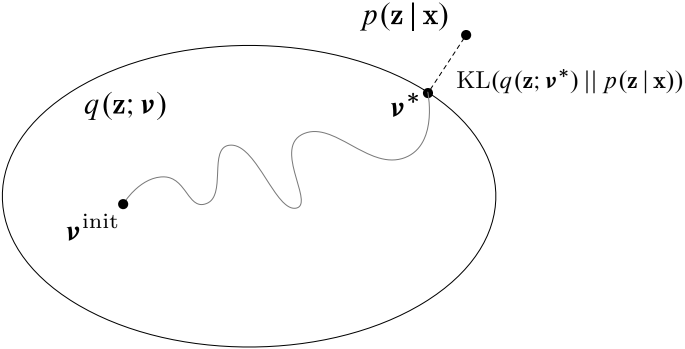


$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta)$$

- The observations are $\mathbf{x} = x_{1:n}$.
- The **local** variables are $\mathbf{z} = z_{1:n}$.
- The **global** variables are β.
- The ith data point x_i only depends on z_i and β.

Compute $p(\beta, \mathbf{z} \mid \mathbf{x})$.

Variational Inference



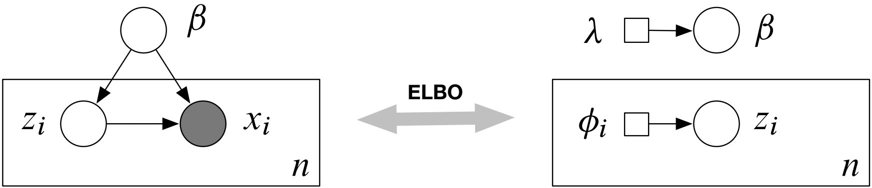
Minimize KL between $q(\beta, \mathbf{z}; \boldsymbol{\nu})$ and the posterior $p(\beta, \mathbf{z} \mid \mathbf{x})$.

The Evidence Lower Bound

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_q \left[\log p(\beta, \mathbf{z}, \mathbf{x})\right] - \mathbb{E}_q \left[\log q(\beta, \mathbf{z}; \boldsymbol{\nu})\right]$$

- KL is intractable; VI optimizes the **evidence lower bound** (ELBO) instead.
  - It is a lower bound on $\log p(\mathbf{x})$.
  - Maximizing the ELBO is equivalent to minimizing the KL.
- The ELBO trades off two terms.
  - The first term prefers $q(\cdot)$ to place its mass on the MAP estimate.
  - The second term encourages $q(\cdot)$ to be diffuse.
- Caveat: The ELBO is not convex.

Mean-field Variational Inference



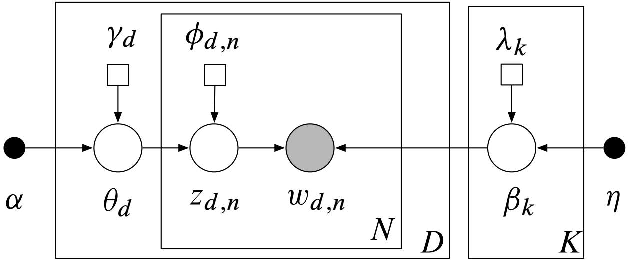
- We need to specify the form of $q(\beta, \mathbf{z})$.
- The **mean-field family** is fully factorized,

$$q(\beta, \mathbf{z}; \lambda, \phi) = q(\beta; \lambda) \prod_{i=1}^{n} q(z_i; \phi_i).$$

- To **learn**, we do stochastic gradient ascent on the evidence lower bound (ELBO),

$$\mathcal{L}(\lambda, \phi) = \mathbb{E}_q \left[\log p(\beta, \mathbf{z}, \mathbf{x})\right] - \mathbb{E}_q \left[\log q(\beta, \mathbf{z})\right].$$

Mean-field Variational Inference for LDA



- The local variables are the per-document variables θ_d and $\mathbf{z}_d$.
- The global variables are the topics $\beta_1, \ldots, \beta_K$.
- The variational distribution is

$$q(\beta, \boldsymbol{\theta}, \mathbf{z}) = \prod_{k=1}^{K} q(\beta_k; \lambda_k) \prod_{d=1}^{D} q(\theta_d; \gamma_d) \prod_{n=1}^{N} q(z_{d,n}; \phi_{d,n})$$