Machine Learning 2

8. Latent Dirichlet Allocation (LDA)

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Outline

1. The LDA Model
2. Learning LDA via Gibbs Sampling
3. Learning LDA via Collapsed Gibbs Sampling
4. Learning LDA via Variational Inference
5. Supervised LDA

Syllabus

		A. Advanced Supervised Learning
Tue. 9.12.	(1)	A.1 Generalized Linear Models
Wed. 10.12.	(2)	A.2 Gaussian Processes
Tue. 16.12.	(3)	A.3 Advanced Support Vector Machines
Wed. 17.12.	(4)	A.4 Neural Networks
Tue. 6.1.	(5)	A.5 Ensembles
Wed. 7.1.	(6)	A.5b Ensembles (ctd.)
Tue. 13.1.	(7)	A.6 Sparse Linear Models — L1 regularization
Wed. 14.1.	(8)	A.6b Sparse Linear Models — L1 regularization (ctd.)
Tue. 20.1.	(9)	A.7. Sparse Linear Models — Further Methods
		B. Complex Predictors
Wed. 21.1.	(10)	B.1 Latent Dirichlet Allocation (LDA)
Tue. 27.1.	(11)	B.1b Latent Dirichlet Allocation (LDA; ctd.)
Wed. 28.1.	(12)	—
Tue. 3.2.	(13)	
Wed. 4.2.	(14)	

Outline

Documents / Finite Discrete Sequences

- instances $x_n \in \mathcal{A}^*$ are **discrete sequences**
  - $\mathcal{A} := \{1, \ldots, A\}$ called **dictionary** / **alphabet** ($A \in \mathbb{N}$), where $a \in A$ denotes the a-th **word** / **symbol** / **token**.
  - $\mathcal{A}^* := \bigcup_{\ell=1}^{L} \mathcal{A}^\ell$ called **documents** / **finite $\mathcal{A}$-sequences**.
  - $M_n := |x_n| := \ell$ called **length** (for $x_n \in \mathcal{A}^\ell$).
  - $x_{n,m}$ called **m-th word of x_n**.

- if there are no sequential effects (order does not matter), documents can be described by their **word frequencies** (**bag of words**):

$$\Phi_{n,a} := |\{\ell \in \{1, \ldots, |x_n|\} \mid x_{n,\ell} = a\}$$

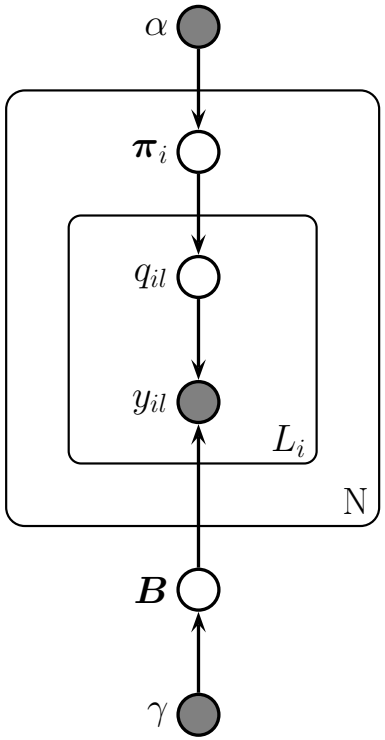
The LDA Model

$$
\begin{aligned}
p(x_{n,m} \mid z_{n,m} = k, \phi) &:= \mathrm{Cat}(x_{n,m} \mid \phi_k), \quad && n = 1, \ldots, N, m = 1, \ldots, M_n \\
p(z_{n,m} \mid \pi_n) &:= \mathrm{Cat}(z_{n,m} \mid \pi_n), \quad && n = 1, \ldots, N, m = 1, \ldots, M_n \\
p(\phi_k \mid \beta) &:= \mathrm{Dir}(\phi_k \mid \beta 1_A), \quad && k = 1, \ldots, K \\
p(\pi_n \mid \gamma) &:= \mathrm{Dir}(\pi_n \mid \gamma 1_K), \quad && n = 1, \ldots, N
\end{aligned}
$$

- $z_{n,m} \in \{1, \ldots, K\}$: topic the m-th word of document n belongs to.
- $\phi_k \in \Delta^A$: word probabilities of topic k.
- $\pi_n \in \Delta^K$: topic probabilities of document n.
- $\beta, \gamma \in \mathbb{R}^+$: priors of ϕ and π.

The LDA Model

$$
\begin{aligned}
p(x_{n,m} \mid z_{n,m} = k, \phi) &:= \mathrm{Cat}(x_{n,m} \mid \phi_k), \\
p(z_{n,m} \mid \pi_n) &:= \mathrm{Cat}(z_{n,m} \mid \pi_n), \\
p(\phi_k \mid \beta) &:= \mathrm{Dir}(\phi_k \mid \beta\, 1_A), \\
p(\pi_n \mid \gamma) &:= \mathrm{Dir}(\pi_n \mid \gamma\, 1_K),
\end{aligned}
$$



- $z_{n,m} \in \{1, \ldots, K\}$: topic the m-th word of document n belongs to.
- $\phi_k \in \Delta^A$: word probabilities of topic k.
- $\pi_n \in \Delta^K$: topic probabilities of document n.
- $\beta, \gamma \in \mathbb{R}^+$: priors of ϕ and π.

[Mur12, fig. 27.2]

Example

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

[Mur12, fig. 27.4]

Example

Document #29795

Bix beiderbecke, at age^{060} fifteen^{207}, sat^{174} on the slope^{071} of a bluff^{055} overlooking^{027} the mississippi^{137} river^{137}. He was listening^{077} to music^{077} coming^{009} from a passing^{043} riverboat. The music^{077} had already captured^{006} his heart^{157} as well as his ear^{119}. It was jazz^{077}. Bix beiderbecke had already had music^{077} lessons^{077}. He showed^{002} promise^{134} on the piano^{077}, and his parents^{035} hoped^{268} he might consider^{118} becoming a concert^{077} pianist^{077}. But bix was interested^{268} in another kind^{050} of music^{077}. He wanted^{268} to play^{077} the cornet. And he wanted^{268} to play^{077} jazz^{077}...

Document #1883

There is a simple^{050} reason^{106} why there are so few periods^{078} of really great theater^{082} in our whole western^{046} world. Too many things^{300} have to come right at the very same time. The dramatists must have the right actors^{082}, the actors^{082} must have the right playhouses, the playhouses must have the right audiences^{082}. We must remember^{288} that plays^{082} exist^{143} to be performed^{077}, not merely^{050} to be read^{254}. (even when you read^{254} a play^{082} to yourself, try^{288} to perform^{062} it, to put^{174} it on a stage^{078}, as you go along.) as soon^{028} as a play^{082} has to be performed^{082}, then some kind^{126} of theatrical^{082}...

Document #21359

Jim^{296} has a game^{166} book^{254}. Jim^{296} reads^{254} the book^{254}. Jim^{296} sees^{081} a game^{166} for one. Jim^{296} plays^{166} the game^{166}. Jim^{296} likes^{081} the game^{166} for one. The game^{166} book^{254} helps^{081} jim^{296}. Don^{180} comes^{040} into the house^{038}. Don^{180} and jim^{296} read^{254} the game^{166} book^{254}. The boys^{020} see a game^{166} for two. The two boys^{020} play^{166} the game^{166}. The boys^{020} play^{166} the game^{166} for two. The boys^{020} like the game^{166}. Meg^{282} comes^{040} into the house^{282}. Meg^{282} and don^{180} and jim^{296} read^{254} the book^{254}. They see a game^{166} for three. Meg^{282} and don^{180} and jim^{296} play^{166} the game^{166}. They play^{166}...

[Mur12, fig. 27.5]

Outline

Learning via Parameter Sampling

The loglikelihood

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)$$

describes the **distribution of the parameters given the data**.
If we can **sample parameters** from this distribution

$$\theta_1, \theta_2, \ldots, \theta_S \sim p(\theta \mid \mathcal{D})$$

we can

- estimate **expected parameter values** and **their variances** from this parameter sample:

$$\hat{\theta} := E(\theta \mid \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\theta_s, \qquad V(\theta \mid \mathcal{D}) \approx \frac{1}{S-1}\sum_{s=1}^{S}(\theta_s - E(\theta \mid \mathcal{D}))^2$$

- **predict targets** for new instances x via model averaging:

$$p(y \mid x, \theta_{1:S}) = \frac{1}{S}\sum_{s=1}^{S} p(y \mid x, \theta_s)$$

Sampling

- for most closed-form distributions $p(x)$ there exist efficient sampling methods
  - categorical, normal, ...
- but most loglikelihoods are not closed-form distributions.
  - but for example products thereof.

Gibbs Sampling

- task: sample from $p(x_1, \ldots, x_N)$
- problem:
  - assume sampling from the joint distribution $p(x_1, \ldots, x_N)$ is difficult.
  - assume sampling from marginals $p(x_n)$ or partial conditionals $p(x_n \mid \text{some } x_{n'})$ is also difficult.
  - assume sampling from all **full conditionals** $p(x_n \mid x_{\_n})$ is easy.

Gibbs Sampling

- task: sample from $p(x_1, \ldots, x_N)$
- problem:
  - assume sampling from the joint distribution $p(x_1, \ldots, x_N)$ is difficult.
  - assume sampling from marginals $p(x_n)$ or partial conditionals $p(x_n \mid \text{some } x_{n'})$ is also difficult.
  - assume sampling from all **full conditionals** $p(x_n \mid x_{-n})$ is easy.

Gibbs sampling: given last sample x^s, sample x^{s+1} one variable at a time:

$$
\begin{aligned}
x_1^{s+1} &\sim p(x_1 \mid x_{2:N} = x_{2:N}^s) \\
x_2^{s+1} &\sim p(x_2 \mid x_{1:1} = x_{1:1}^{s+1}, x_{3:N} = x_{3:N}^s) \\
&\vdots \\
x_n^{s+1} &\sim p(x_n \mid x_{1:n-1} = x_{1:n-1}^{s+1}, x_{n+1:N} = x_{n+1:N}^s) \\
&\vdots \\
x_N^{s+1} &\sim p(x_N \mid x_{1:N-1} = x_{1:N-1}^{s+1})
\end{aligned}
$$

Gibbs Sampling

- the distribution created by the Gibbs sampler eventually will converge to $p(x_1, \ldots, x_N)$
- start Gibbs sampling with an arbitrary x^0
  - but ensure that $p(x^0) > 0$!
  - also consider restarts.
- throw away the first examples (**burn in**; 100-10,000).
  - only after a while the chain has converged to the stationary distribution $p(x_1, \ldots, x_N)$.
  - typical are 100-10,000 examples
- sometimes some variables can be marginalized out, improving the performance of the Gibbs sampler (**collapsed Gibbs sampling**, **Rao-Blackwellisation**)

Gibbs Sampling for LDA

$$
\begin{aligned}
p(x_{n,m} \mid z_{n,m} = k, \phi) &:= \mathrm{Cat}(x_{n,m} \mid \phi_k) &&= \phi_{k,x_{n,m}} \\
p(z_{n,m} \mid \pi_n) &:= \mathrm{Cat}(z_{n,m} \mid \pi_n) &&= \pi_{n,z_{n,m}} \\
p(\phi_k \mid \beta) &:= \mathrm{Dir}(\phi_k \mid \beta\, 1_A) &&\propto \prod_{a=1}^{A} \phi_{k,a}^{\beta_a - 1} \\
p(\pi_n \mid \gamma) &:= \mathrm{Dir}(\pi_n \mid \gamma\, 1_K) &&\propto \prod_{k=1}^{K} \pi_{n,k}^{\gamma_k - 1}
\end{aligned}
$$

Full conditionals:

$$
p(z_{n,m} = k \mid \phi, \pi_n) \propto p(x_{n,m} \mid z_{n,m} = k, \phi) p(z_{n,m} = k \mid \pi_n) = \phi_{k,x_{n,m}} \pi_{n,k}
$$

Gibbs Sampling for LDA

$$p(x_{n,m} \mid z_{n,m} = k, \phi) := \mathrm{Cat}(x_{n,m} \mid \phi_k) = \phi_{k,x_{n,m}}$$

$$p(z_{n,m} \mid \pi_n) := \mathrm{Cat}(z_{n,m} \mid \pi_n) = \pi_{n,z_{n,m}}$$

$$p(\phi_k \mid \beta) := \mathrm{Dir}(\phi_k \mid \beta 1_A) \propto \prod_{a=1}^{A} \phi_{k,a}^{\beta_a - 1}$$

$$p(\pi_n \mid \gamma) := \mathrm{Dir}(\pi_n \mid \gamma 1_K) \propto \prod_{k=1}^{K} \pi_{n,k}^{\gamma_k - 1}$$

Full conditionals:

$$p(\pi_n \mid z_n, \phi) \propto p(\pi_n \mid \gamma) \prod_{m=1}^{M_n} p(z_{n,m} = k \mid \pi_n)$$

$$\propto \prod_{k=1}^{K} \pi_{n,k}^{\gamma_k - 1} \prod_{m=1}^{M_n} \prod_{k=1}^{K} \pi_{n,k}^{\delta(z_{n,m}=k)}$$

$$= \mathrm{Dir}((\gamma_k + \sum_{m=1}^{M} \delta(z_{n,m} = k))_{k=1:K})$$

Gibbs Sampling for LDA

$$p(x_{n,m} \mid z_{n,m} = k, \phi) := \mathrm{Cat}(x_{n,m} \mid \phi_k) = \phi_{k,x_{n,m}}$$

$$p(z_{n,m} \mid \pi_n) := \mathrm{Cat}(z_{n,m} \mid \pi_n) = \pi_{n,z_{n,m}}$$

$$p(\phi_k \mid \beta) := \mathrm{Dir}(\phi_k \mid \beta\, 1_A) \propto \prod_{a=1}^{A} \phi_{k,a}^{\beta_a - 1}$$

$$p(\pi_n \mid \gamma) := \mathrm{Dir}(\pi_n \mid \gamma\, 1_K) \propto \prod_{k=1}^{K} \pi_{n,k}^{\gamma_k - 1}$$

Full conditionals:

$$p(\phi_k \mid z, \pi) \propto \prod_{n=1}^{N} \prod_{m=1}^{M_n} p(z_{n,m} = k \mid \pi_n)\, p(\phi_k \mid \beta)$$

$$= \mathrm{Dir}((\beta_a + \sum_{n=1}^{N} \sum_{m=1}^{M} \delta(x_{n,m} = a, z_{n,m} = k))_{a=1:A})$$

Outline

Counts

$$c_{n,a,k} := \sum_{m=1}^{M_n} \delta(x_{n,m} = a, z_{n,m} = k)$$

$$c_{n,k} := \sum_{a=1}^{A} c_{n,a,k}$$

$$c_{a,k} := \sum_{n=1}^{N} c_{n,a,k}$$

$$c_{k} := \sum_{a=1}^{A} \sum_{n=1}^{N} c_{n,a,k}$$

Marginals over π and ϕ

$$p(z \mid \beta) = \prod_{n=1}^{M_n} \int (\prod_{m=1}^{M_n} \mathrm{Cat}(z_{n,m} \mid \pi_n)) \mathrm{Dir}(\pi_n \mid \gamma 1_K) d\pi_n$$

$$= \left(\frac{\Gamma(K\gamma)}{\Gamma(\gamma)^K} \right)^N \prod_{n=1}^{N} \frac{\prod_{k=1}^{K} \Gamma(c_{n,k} + \gamma)}{\Gamma(M_n + K\gamma)}$$

Marginals over π and ϕ

$$p(z \mid \beta) = \prod_{n=1}^{M_n} \int (\prod_{m=1}^{M_n} \mathrm{Cat}(z_{n,m} \mid \pi_n)) \mathrm{Dir}(\pi_n \mid \gamma 1_K) d\pi_n$$

$$= \left(\frac{\Gamma(K\gamma)}{\Gamma(\gamma)^K}\right)^N \prod_{n=1}^{N} \frac{\prod_{k=1}^{K} \Gamma(c_{n,k} + \gamma)}{\Gamma(M_n + K\gamma)}$$

$$p(x \mid z, \beta) = \prod_{k=1}^{K} \int (\prod_{(n,m):z_{n,m}=k} \mathrm{Cat}(x_{n,m} \mid \phi_k)) \mathrm{Dir}(\phi_k \mid \beta 1_K) d\phi_k$$

$$= \left(\frac{\Gamma(A\beta)}{\Gamma(\beta)^A}\right)^K \prod_{k=1}^{K} \frac{\prod_{a=1}^{A} \beta(c_{a,k} + \beta)}{\Gamma(c_k + A\beta)}$$

Marginals over π and ϕ

$$p(z_{n,m} \mid z_{-(n,m)}, x, \beta, \gamma) = \frac{p(z \mid x, \beta, \gamma)\, p(x \mid z, \beta)}{p(z_{-(n,m)} \mid x_{-(n,m)}, \beta, \gamma)\, p(x_{-(n,m)} \mid z_{-(n,m)}, \beta)}$$

Marginals over π and ϕ

$$p(z_{n,m} \mid z_{-(n,m)}, x, \beta, \gamma) = \frac{p(z \mid x, \beta, \gamma)\, p(x \mid z, \beta)}{p(z_{-(n,m)} \mid x_{-(n,m)}, \beta, \gamma)\, p(x_{-(n,m)} \mid z_{-(n,m)}, \beta)}$$

Now let $c^{-}_{n,a,k}$ be the counts for the leave-one-out sample $x_{(n,m)}, z_{-(n,m)}$ (all but m-th word of document n).

$$c^{-}_{n,a,k} = \begin{cases} c_{n,a,k} - 1, & \text{for } x_{n,m} = a, z_{n,m} = k \\ c_{n,a,k}, & \text{else} \end{cases}$$

- all terms other than for $x_{n,m} = a, z_{n,m} = k$ cancel out.
- terms for $x_{n,m} = a, z_{n,m} = k$ can be simplified via $\Gamma(x+1)/\Gamma(x) = x$

Marginals over π and ϕ

$$p(z_{n,m} \mid z_{-(n,m)}, x, \beta, \gamma) = \frac{p(z \mid x, \beta, \gamma)\, p(x \mid z, \beta)}{p(z_{-(n,m)} \mid x_{-(n,m)}, \beta, \gamma)\, p(x_{-(n,m)} \mid z_{-(n,m)}, \beta)}$$

Now let $c^{-}_{n,a,k}$ be the counts for the leave-one-out sample $x_{(n,m)}, z_{-(n,m)}$ (all but m-th word of document n).

$$c^{-}_{n,a,k} = \begin{cases} c_{n,a,k} - 1, & \text{for } x_{n,m} = a, z_{n,m} = k \\ c_{n,a,k}, & \text{else} \end{cases}$$

- all terms other than for $x_{n,m} = a, z_{n,m} = k$ cancel out.
- terms for $x_{n,m} = a, z_{n,m} = k$ can be simplified via $\Gamma(x+1)/\Gamma(x) = x$

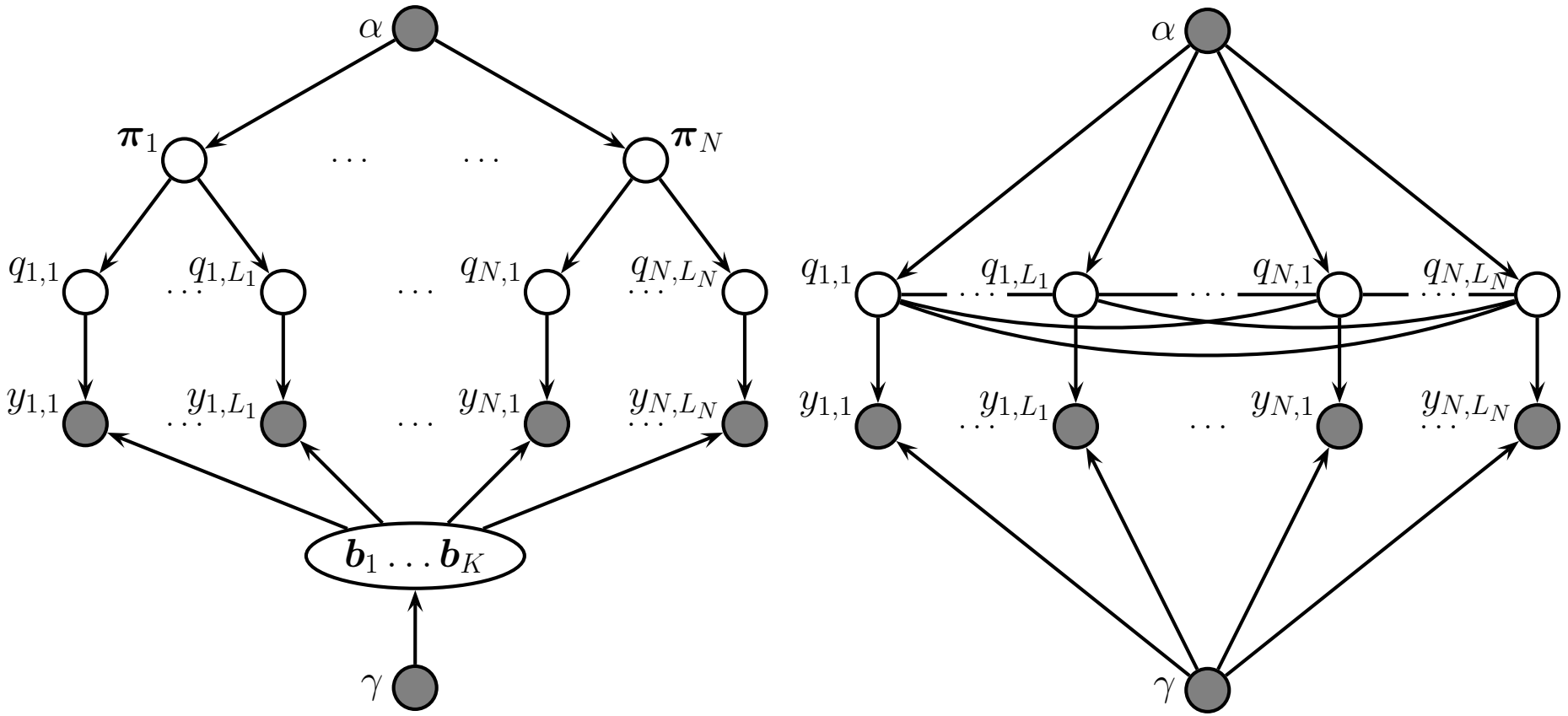
$$p(z_{n,m} = k \mid z_{-(n,m)}, x, \beta, \gamma) = \frac{c^{-}_{x_{n,m},k} + \gamma}{c^{-}_{k} + A\gamma} \frac{c^{-}_{n,k} + \beta}{M_n + K\beta}$$

Collapsed LDA Implementation

- assign all $z_{n,m}$ randomly
- compute $c_{n,a,k}$
- for $s := 1, \ldots, S$:
  - for $n := 1, \ldots, N, \quad m := 1, \ldots, M_n$:

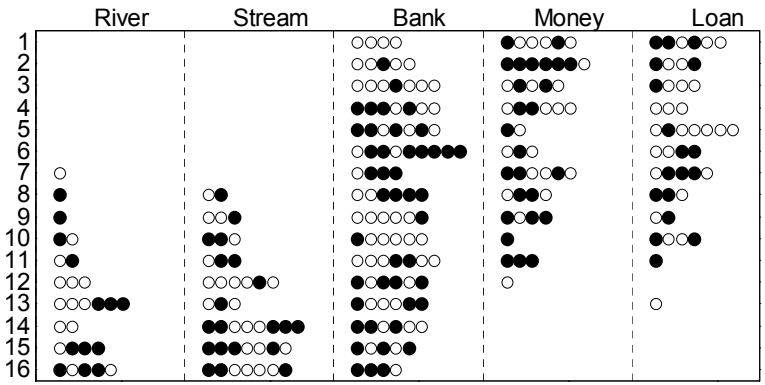
$$
\begin{aligned}
c_{x_{n,m},z_{n,m}} &:= c_{x_{n,m},z_{n,m}} - 1 \\
c_{n,z_{n,m}} &:= c_{n,z_{n,m}} - 1 \\
c_{z_{n,m}} &:= c_{z_{n,m}} - 1 \\
z_{n,m} &:= k \sim \frac{c^{-}_{x_{n,m},k} + \gamma}{c^{-}_{k} + A\gamma} \frac{c^{-}_{n,k} + \beta}{M_n + K\beta} \\
c_{x_{n,m},z_{n,m}} &:= c_{x_{n,m},z_{n,m}} + 1 \\
c_{n,z_{n,m}} &:= c_{n,z_{n,m}} + 1 \\
c_{z_{n,m}} &:= c_{z_{n,m}} + 1
\end{aligned}
$$

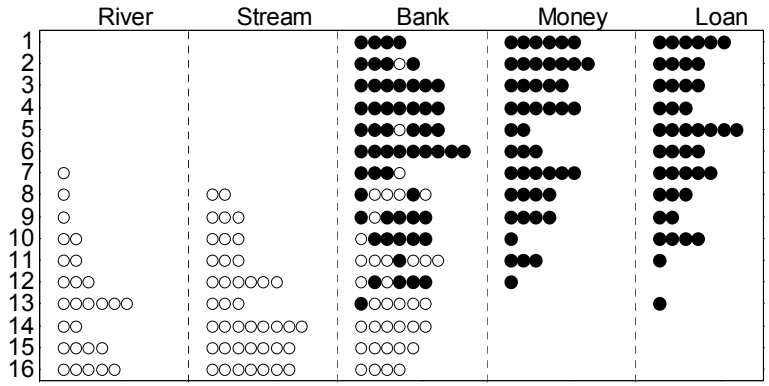
LDA vs Collapsed LDA



[Mur12, fig. 27.7]

Collapsed LDA / Example





[Mur12, fig. 27.8]

Outline

Variational Inference via Mean Field Approximation

To solve the inference problem

$$\text{compute } p(x_1, \ldots, x_N)$$

for intractable p, approximate p with a fully factorized density q

$$p(x_1, \ldots, x_N) \approx q(x_1, \ldots, x_N \mid \theta) := \prod_{n=1}^{N} q_n(x_n \mid \theta_n)$$

A good approximation should minimize the KL divergence of p and q:

$$(\theta_1, \ldots, \theta_N) := \arg\min_{\theta_1, \ldots, \theta_N} \mathrm{KL}(q||p)$$

$$\mathrm{KL}(q||p) := E_x\left(\; q(x) \log \frac{q(x)}{p(x)} \; \right)$$

which can be solved via coordinate descent:

$$\log q_n(x_n \mid \theta_n) = E_{x_{-1} \sim q_{-n}}(\tilde{p}(x_1, \ldots, x_N)) + \text{const}$$

where $\tilde{p}$ can be an unnormalized version of p.

Learning LDA via Mean Field Approximation

Mean field approximation

$$q(\pi_n \mid \tilde{\pi}_n) := \mathrm{Dir}(\pi_n \mid \tilde{\pi}_n)$$
$$q(z_{n,m} \mid \tilde{z}_{n.m}) := \mathrm{Cat}(z_{n,m} \mid \tilde{z}_{n,m})$$

in the E-step of EM leads to

E-step:

$$\tilde{z}_{n,m,k} = \Psi(\tilde{\pi}_{n,k}) - \Psi(\sum_{k'} \tilde{\pi}_{n,k'})$$

$$\tilde{\pi}_{n,k} = \gamma + \sum_{m} \tilde{z}_{n,m,k}$$

M-step:

$$\hat{\phi}_{a,k} = \beta + \sum_{n} \sum_{m} \tilde{z}_{n,m,k} \delta(x_{n,m} = a)$$

Note: $E_{\pi_{n,k} \sim \mathrm{Dir}(\tilde{\pi}_{n,k})}(\log \pi_{n,k}) = \Psi(\tilde{\pi}_{n,k}) - \Psi(\sum_{k'} \tilde{\pi}_{n,k'})$ with Ψ the digamma function.

Outline

...

Further Readings

- L1 regularization: [Mur12, chapter 13.3–5], [HTFF05, chapter 3.4, 3.8, 4.4.4], [Bis06, chapter 3.1.4].
  - LAR, LARS: [HTFF05, chapter 3.4.4], [Mur12, chapter 13.4.2],
- Non-convex regularizers: [Mur12, chapter 13.6].
- Automatic Relevance Determination (ARD): [Mur12, chapter 13.7], [HTFF05, chapter 11.9.1], [Bis06, chapter 7.2.2].
- Sparse Coding: [Mur12, chapter 13.8].

References

Christopher M. Bishop.
Pattern recognition and machine learning, volume 1.
springer New York, 2006.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin.
The elements of statistical learning: data mining, inference and prediction, volume 27.
Springer, 2005.

Kevin P. Murphy.
Machine learning: a probabilistic perspective.
The MIT Press, 2012.