ENGR-E 511; ENGR-E 399

Machine Learning for Signal Processing

Module 11:

Probabilistic Topic Modeling

Minje Kim

Department of Intelligent Systems Engineering

Email: minje@indiana.edu

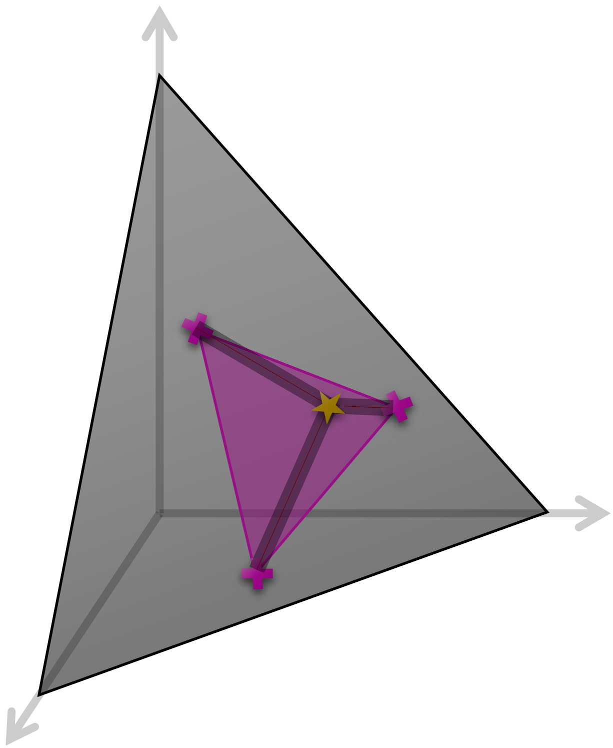
Website: http://minjekim.com

Research Group: http://saige.sice.indiana.edu

Meeting Request: http://doodle.com/minje

INDIANA UNIVERSITY

SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING



Document Generation Process

- A document with a single topic

- Let's invite Prof. K once again
  - We're interested in his journal, saying:

- Since we care a lot about him,
  we want to know his thinking process
- The assumption:
  - There is a probabilistic distribution
    that governs his choice of words when he writes his journal
- The generation of d-th document
  - Sample N words $W_{1:N,d}$
    out of V vocabulary using $\mathrm{Mult}(N, \boldsymbol{\beta}_v)$
- For example
  - ***Today I like my burger I hate spinach***

$$[W_{1,d}, W_{2,d}, W_{3,d}, W_{4,d}, W_{5,d}, W_{6,d}, W_{7,d}, W_{8,d}]^\top$$
$$[v=2, v=1, v=4, v=3, v=6, v=1, v=5, v=7]^\top$$

$\boldsymbol{v}$	$\boldsymbol{\beta}_v$
I	.20
today	.12
my	.10
like	.10
hate	.10
burger	.20
spinach	.18

Document Generation Process

- A document with a single topic

- How do we calculate the parameter β_v from data, if we didn't know it?
  - Maximum Likelihood Estimation (MLE)

$$P(W_{:,d};\boldsymbol{\beta}_v) = \frac{N!}{\prod_{v=1}^{V} n_v!}\prod_{v=1}^{V}\boldsymbol{\beta}_v^{n_v}$$

Note that the order of words doesn't matter

$$\arg\max_{\boldsymbol{\beta}_v} \log P(W_{:,d};\boldsymbol{\beta}_v) = \arg\max_{\boldsymbol{\beta}_v} \sum_{v=1}^{V} n_v \log \boldsymbol{\beta}_v + Const.$$

Log likelihood for convenience

$$\mathcal{LL} = \sum_{v=1}^{V} n_v \log \boldsymbol{\beta}_v + \lambda\left(\sum_{v=1}^{V}\boldsymbol{\beta}_v - 1\right)$$

Objective function
with Lagrange multiplier

$$\frac{\partial \mathcal{LL}}{\partial \boldsymbol{\beta}_v} = \frac{n_v}{\boldsymbol{\beta}_v} + \lambda = 0 \qquad \frac{\partial \mathcal{LL}}{\partial \lambda} = \sum_v \boldsymbol{\beta}_v - 1 = 0$$

$$\Leftrightarrow -\frac{n_v}{\lambda} = \boldsymbol{\beta}_v \qquad \Leftrightarrow \sum_v -\frac{n_v}{\lambda} = 1$$

$$\Leftrightarrow \sum_v -n_v = \lambda$$

$$\Leftrightarrow \frac{n_v}{\sum_v n_v} = \frac{n_v}{N} = \boldsymbol{\beta}_v$$

$\boldsymbol{v}$	$\boldsymbol{\beta_v}$	$\widehat{\boldsymbol{\beta_v}}$
I	.20	.25
today	.12	.125
my	.10	.125
like	.10	.125
hate	.10	.125
burger	.20	.125
spinach	.18	.125

Document Generation Process

- A document with a multiple topics

- So far so good, but how about a more complex case?

  - Well, you're free to assume that a document is generated from a single probabilistic distribution
  - But, what's the meaning of that?
- For this new document, I'd say there are two **topics**
  - Something about eating
  - Something about work
- Probabilistic topic modeling
  - Assumes that a document comprises multiple topics

Document Generation Process

- A document with a multiple topics

- Document generation process (d-th document) using topics:
  - For n-th word in the document (repeat N times)
    - Sample a topic $Z_{n,d} = k$ out of K total topics using $\mathrm{Mult}(\boldsymbol{\Theta}_{1:K,d})$
    - Sample a word $W_{n,d} = v$ out of V total words using $\mathrm{Mult}(\boldsymbol{B}_{1:V,\boldsymbol{Z}_{n,d}})$

v	$X_{1:V,d}$
I	4
today	1
my	3
like	3
hate	1
burger	1
spinach	1
research	1
student	1

$\sim$

$P(W_{n,d} = v)$
.18
.112
.12
.12
.08
.12
.108
.1
.06

$=$

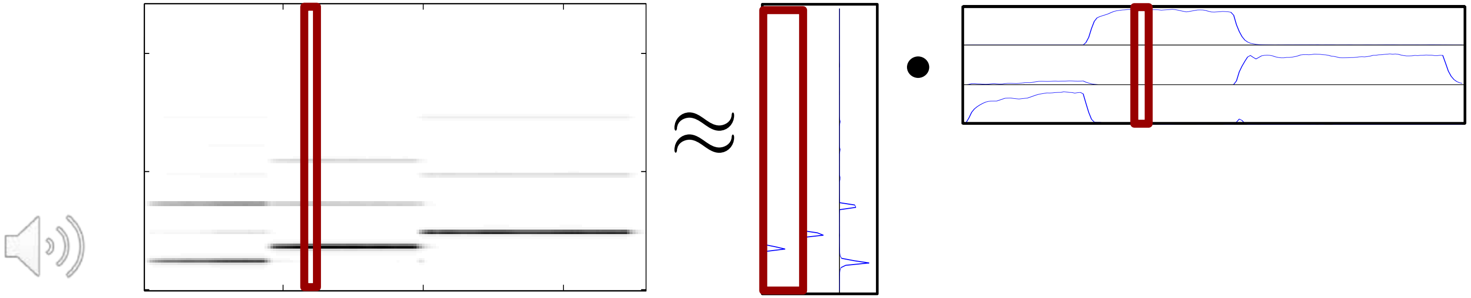
v	$B_{1:V,1}$	$B_{1:V,2}$
I	.20	.15
today	.12	.10
my	.10	.15
like	.10	.15
hate	.10	.05
burger	.20	0
spinach	.18	0
research	0	.25
student	0	.15

$\bullet$

k	…	$\Theta_{k,d-1}$	$\Theta_{k,d}$	$\Theta_{k,d+1}$	…
eating	…	.45	.6	.44	…
working	…	.55	.4	.56	…

  - Eventually $X_{1:V,d} \sim \mathrm{Mult}(N_d, \mathbf{B}_{1:V,1:K}\boldsymbol{\Theta}_{1:K,d})$

Another Take on Topic Modeling
- Sound quanta



- Spectrum generation process (d-th spectrum) using basis vectors:
  - For n-th sound quanta in the ~~document~~ spectrum (repeat N times)
    - Sample a topic $\boldsymbol{Z}_{n,d} = k$ out of K topics using $\mathrm{Mult}(\boldsymbol{\Theta}_{1:K,d})$
    - Sample a sound quanta $W_{n,d} = v$ out of V frequencies using $\mathrm{Mult}(\boldsymbol{B}_{1:V,\boldsymbol{Z}_{n,d}})$
  - If N_d is large enough, the histogram of $W_{1:N_d,d}$ (i.e. $X_{1:V,d}$) will look like a spectrum
- Ring a bell?
  - NMF! (I'm going to revisit this similarity later in this lecture)

Yet Another Take on Topic Modeling

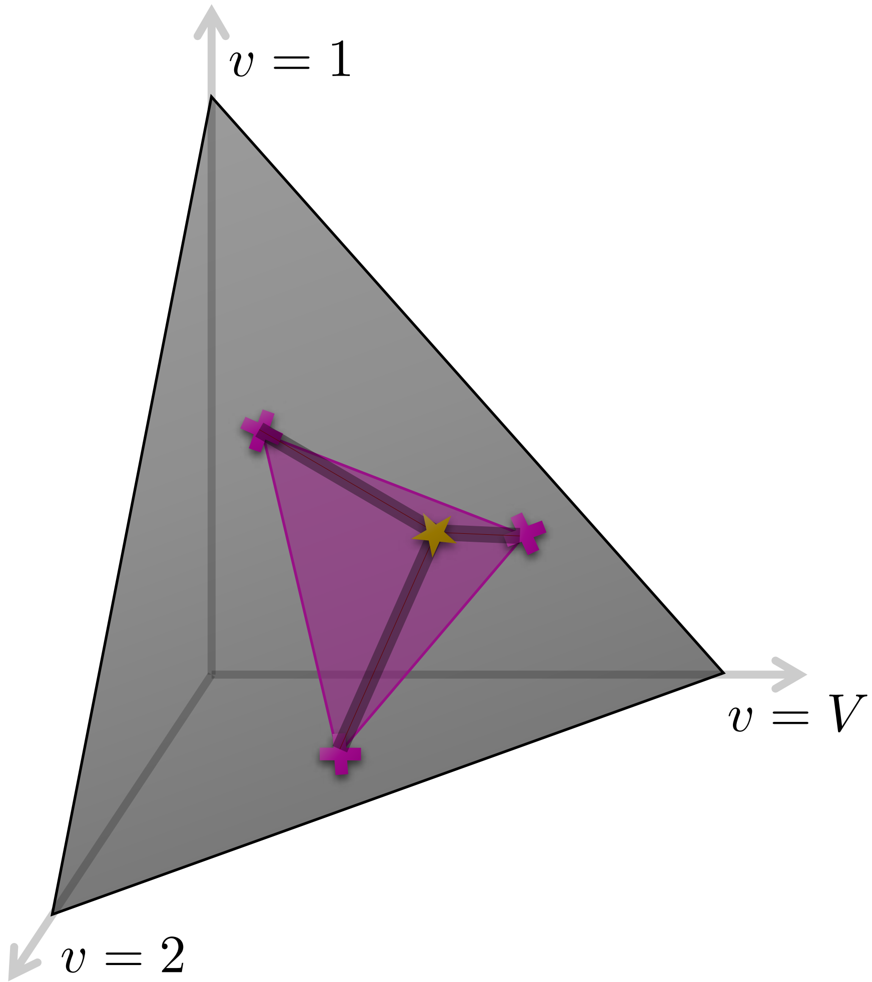
- A geometric interpretation on simplex

$\boldsymbol{B}_{1:V,k}$ One of the $\mathbf{K}$ topics: a distribution over the vocabulary

The convex hull defined by the topics

$\boldsymbol{\Theta}_{1:K,d}$ Contribution of the topics to the $\mathbf{d}$-th document

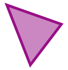
$X_{1:V,d}$ The $\mathbf{d}$-th document

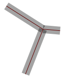


Yet Another Take on Topic Modeling

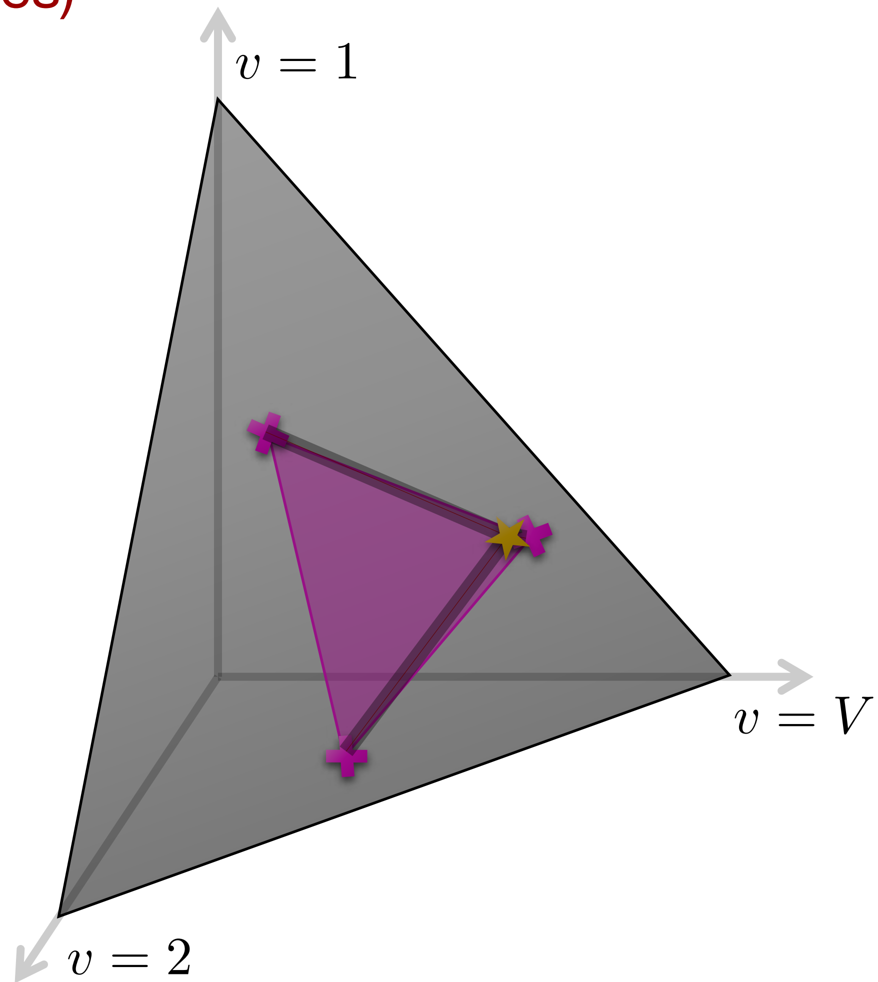
- A geometric interpretation on simplex (three notes)


$\boldsymbol{B}_{1:V,k}$ One of the 3 notes: a distribution over $\mathbf{V}$ Fourier coefficients.


Polyphonic music that can be played by the 3 notes

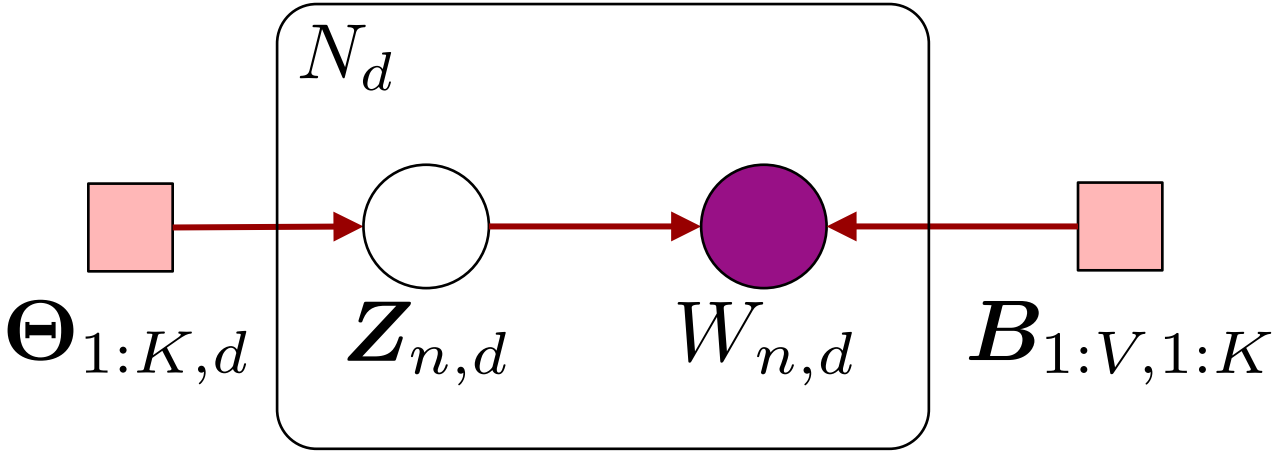

$\boldsymbol{\Theta}_{1:K,d}$ Contribution of the notes to the $\mathbf{d}$-th spectrum

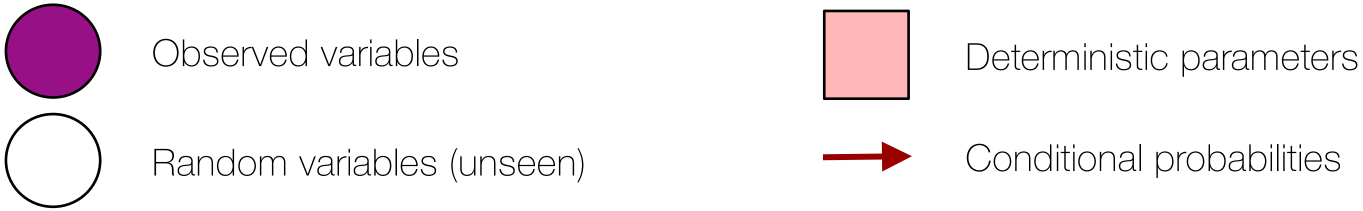

$X_{1:V,d}$ The $\mathbf{d}$-th spectrum



Last-But-Not-Least Take on Topic Modeling

- A graphical model



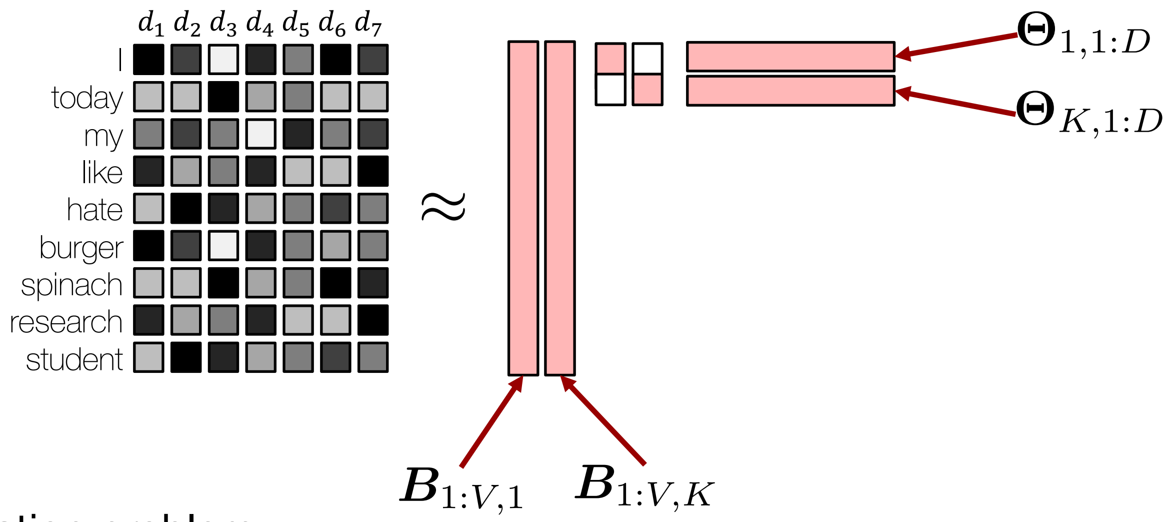


Probabilistic Latent Semantic Indexing

- Or, Probabilistic Latent Semantic Analysis

- Where is this name from?
- Latent Semantic Analysis (LSA)
  - SVD on text data, e.g. Term-Frequency (TF) matrix

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$$



  - This is a matrix factorization problem
- PLSI is a probabilistic version of this matrix factorization problem

Probabilistic Latent Semantic Indexing

- EM on PLSI

- Ready for some math?
- Probability of the word v at the n-th position in the document d

$$P(W_{n,d} = v; \boldsymbol{B}_{v,1:K}, \boldsymbol{\Theta}_{1:K,d}) = \sum_{k=1}^{K} \boldsymbol{B}_{v,k} \boldsymbol{\Theta}_{k,d}$$

- Probability of observing the d-th document with N_d words

The count of v

$$P(W_{1:N_d,d}; \boldsymbol{B}_{1:V,1:K}, \boldsymbol{\Theta}_{1:K,d}) = \prod_{n=1}^{N_d} \sum_{k=1}^{K} \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d} = \prod_{v=1}^{V} \sum_{k=1}^{K} \left(\boldsymbol{B}_{v,k} \boldsymbol{\Theta}_{k,d}\right)^{X_{v,d}}$$

We ignore the word order in this Bag-of-Words model

- Probability of observing the entire collection of D documents

$$P(W; \boldsymbol{B}, \boldsymbol{\Theta}) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} \sum_{k=1}^{K} \left(\boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d}\right)$$

- Objective function

$$\arg\max_{\boldsymbol{B},\boldsymbol{\Theta}} \log P(W; \boldsymbol{B}, \boldsymbol{\Theta}) + \sum_{k=1}^{K} \lambda_k \left(1 - \sum_{v=1}^{V} \boldsymbol{B}_{v,k}\right) + \sum_{d=1}^{D} \psi_d \left(1 - \sum_{k=1}^{K} \boldsymbol{\Theta}_{k,d}\right)$$

Probabilistic Latent Semantic Indexing

- EM on PLSI

- The curse of summation inside logarithm

$$\arg\max_{\boldsymbol{B},\boldsymbol{\Theta}} \sum_{d=1}^{D}\sum_{n=1}^{N_d} \log \sum_{k=1}^{K} \boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d} + \sum_{k=1}^{K}\lambda_k\Big(1-\sum_{v=1}^{V}\boldsymbol{B}_{v,k}\Big) + \sum_{d=1}^{D}\psi_d\Big(1-\sum_{k=1}^{K}\boldsymbol{\Theta}_{k,d}\Big)$$

  - Difficult to differentiate

- Let's look into it

$$\log\sum_{k=1}^{K}\boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d} = \log\sum_{k=1}^{K}\frac{q(k)\boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d}}{q(k)} \overset{\text{Jensen's inequality}}{\geq} \sum_{k=1}^{K} q(k)\log\frac{\boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d}}{q(k)}$$

$$= \sum_{k=1}^{K} q(k)\log\frac{P(W_{n,d},\boldsymbol{Z}_{n,d}=k)}{q(k)} \overset{\text{Bayes' theorem}}{=} \sum_{k=1}^{K} q(k)\log\frac{P(\boldsymbol{Z}_{n,d}=k|W_{n,d})P(W_{n,d})}{q(k)}$$

$$= \sum_{k=1}^{K} q(k)\log\frac{P(\boldsymbol{Z}_{n,d}=k|W_{n,d})}{q(k)} + \sum_{k=1}^{K} q(k)\log P(W_{n,d})$$

$$= -\mathcal{D}_{KL}\Big(q(k)||P(\boldsymbol{Z}_{n,d}=k|\boldsymbol{W}_{n,d})\Big) + \log P(W_{n,d})$$

Probabilistic Latent Semantic Indexing

- EM on PLSI

- So what?

$$\log \sum_{k=1}^{K} \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d} = \log \sum_{k=1}^{K} \frac{q(k) \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d}}{q(k)}$$
$$\geq -\mathcal{D}_{KL}\Big(q(k) || P(\boldsymbol{Z}_{n,d} = k | \boldsymbol{W}_{n,d})\Big) + \log P(W_{n,d})$$

  - If $q(k) = P(\boldsymbol{Z}_{n,d} = k | \boldsymbol{W}_{n,d})$, we can best maximize the effect of introducing the proposal distribution

$$\mathcal{Z}_{n,d,k} = P(\boldsymbol{Z}_{n,d} = k | W_{n,d})$$

- Then, what?

$$\log \sum_{k=1}^{K} \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d} \geq \sum_{k=1}^{K} \mathcal{Z}_{n,d,k} \log \frac{\boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d}}{\mathcal{Z}_{n,d,k}}$$
$$= \sum_{k=1}^{K} \mathcal{Z}_{n,d,k} \log \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d} - \sum_{k=1}^{K} \cancel{\mathcal{Z}_{n,d,k} \log \mathcal{Z}_{n,d,k}}$$

Constant

- The full objective function (for the M-step)

$$\mathcal{LL} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \mathcal{Z}_{n,d,k} \log \boldsymbol{B}_{W_{n,d},k} \boldsymbol{\Theta}_{k,d} + \sum_{k=1}^{K} \lambda_k \Big(1 - \sum_{v=1}^{V} \boldsymbol{B}_{v,k}\Big) + \sum_{d=1}^{D} \psi_d \Big(1 - \sum_{k=1}^{K} \boldsymbol{\Theta}_{k,d}\Big)$$

Probabilistic Latent Semantic Indexing

- EM on PLSI

- In the E-step

$$\mathcal{Z}_{n,d,k} = \frac{P(W_{n,d}|\boldsymbol{Z}_{n,d} = k)P(\boldsymbol{Z}_{n,d} = k)}{\sum_{k=1}^{K} P(W_{n,d}|\boldsymbol{Z}_{n,d} = k)P(\boldsymbol{Z}_{n,d} = k)}$$

$$\mathcal{Z}_{n,d,k} = \frac{\boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d}}{\sum_{k=1}^{K} \boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d}}$$

$$\mathcal{Z}_{v,d,k} = \frac{\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}{\sum_{k=1}^{K} \boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}} = P(\boldsymbol{Z}_{v,d} = k|X_{v,d})$$

Regardless of their positions, same word share the same posterior distribution

$$X_{v,d} = \sum_{n=1}^{N_d} \mathcal{I}(W_{n,d} = v)$$

Probabilistic Latent Semantic Indexing

- EM on PLSI

- M-step

$$\mathcal{LL} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \boldsymbol{\mathcal{Z}}_{n,d,k} \log \boldsymbol{B}_{W_{n,d},k}\boldsymbol{\Theta}_{k,d} + \sum_{k=1}^{K}\lambda_k\Big(1-\sum_{v=1}^{V}\boldsymbol{B}_{v,k}\Big) + \sum_{d=1}^{D}\psi_d\Big(1-\sum_{k=1}^{K}\boldsymbol{\Theta}_{k,d}\Big)$$

I don't care
about the word order!

$$\mathcal{LL} = \sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{k=1}^{K} \boldsymbol{\mathcal{Z}}_{v,d,k} \log \left(\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}\right)^{X_{v,d}} + \sum_{k=1}^{K}\lambda_k\Big(1-\sum_{v=1}^{V}\boldsymbol{B}_{v,k}\Big) + \sum_{d=1}^{D}\psi_d\Big(1-\sum_{k=1}^{K}\boldsymbol{\Theta}_{k,d}\Big)$$

$$\frac{\partial \mathcal{LL}}{\partial \boldsymbol{B}_{v,k}} = \frac{\partial \sum_{d=1}^{D} \boldsymbol{\mathcal{Z}}_{n,d,k} X_{v,d} \log \boldsymbol{B}_{W_{n,d},k}}{\partial \boldsymbol{B}_{v,k}} - \lambda_k = \frac{\sum_{d=1}^{D} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}{\boldsymbol{B}_{v,k}} - \lambda_k = 0$$

$$\Leftrightarrow \sum_{d} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} = \lambda_k \boldsymbol{B}_{v,k}$$

$$\sum_{v}\sum_{d} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} = \lambda_k \sum_{v} \boldsymbol{B}_{v,k} = \lambda_k$$

$$\Leftrightarrow \frac{\sum_{d} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}{\lambda_k} = \boldsymbol{B}_{v,k}$$

Substitute

$$\Leftrightarrow \boldsymbol{B}_{v,k} = \frac{\sum_{d} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}{\sum_{v}\sum_{d} X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}$$

Probabilistic Latent Semantic Indexing

- EM on PLSI

- M-step

Sum along $\mathbf{d}$-axis

$$\boldsymbol{B}_{v,k} = \frac{\sum_d \boxed{X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}}{\sum_v \sum_d X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}$$

Re-weight input using post dist

Normalize along $\mathbf{v}$-axis

Sum along $\mathbf{v}$-axis

$$\boldsymbol{\Theta}_{k,d} = \frac{\sum_v \boxed{X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}}{\sum_k \sum_v X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k}}$$

Re-weight input using post dist

Normalize along $\mathbf{k}$-axis

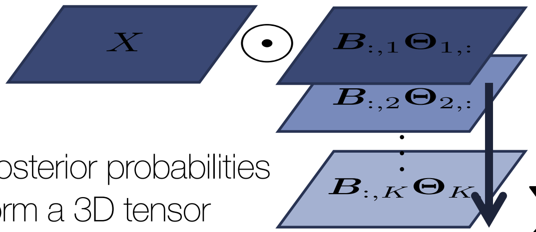
- E-step

$$\boldsymbol{\mathcal{Z}}_{v,d,k} = \frac{\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}{\sum_{k=1}^{K} \boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}$$

Probabilistic Latent Semantic Indexing

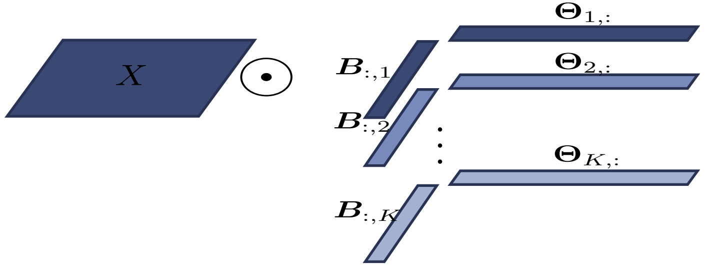
- EM on PLSI: a matrix version

- E-step $\mathcal{Z}_{v,d,k} = \dfrac{\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}{\sum_{k=1}^{K}\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}$
- M-step $\boldsymbol{B}_{v,k} = \dfrac{\sum_d X_{v,d}\mathcal{Z}_{v,d,k}}{\sum_v\sum_d X_{v,d}\mathcal{Z}_{v,d,k}}$ $\qquad \boldsymbol{\Theta}_{k,d} = \dfrac{\sum_v X_{v,d}\mathcal{Z}_{v,d,k}}{\sum_k\sum_v X_{v,d}\mathcal{Z}_{v,d,k}}$

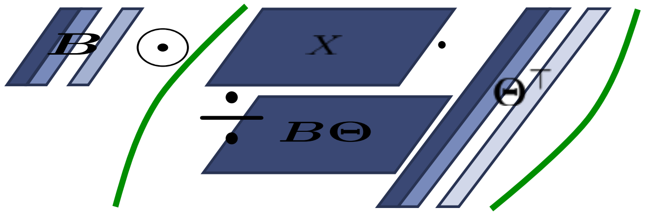


Posterior probabilities form a 3D tensor

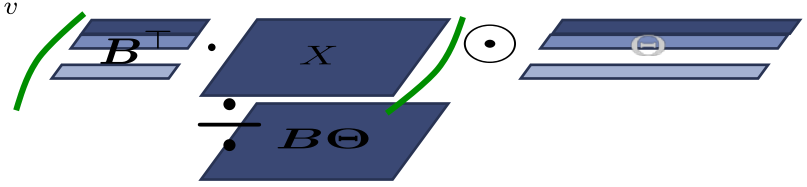
$$\sum_{k=1}^{K}\mathbf{B}_{:,k}\boldsymbol{\Theta}_{k,:} = \mathbf{B}\boldsymbol{\Theta}$$



$$\sum_d X_{v,d}\mathcal{Z}_{v,d,k} =$$



$$\sum_v X_{v,d}\mathcal{Z}_{v,d,k} =$$



Probabilistic Latent Semantic Indexing

- PLSI is equivalent to NMF (with KL divergence)

- We can merge the E-step and M-step

$$\boldsymbol{B} = \boldsymbol{B} \odot \left(\frac{X}{\boldsymbol{B}\boldsymbol{\Theta}} \boldsymbol{\Theta}^\top \right) \qquad \boldsymbol{\Theta} = \boldsymbol{\Theta} \odot \left(\boldsymbol{B}^\top \frac{X}{\boldsymbol{B}\boldsymbol{\Theta}} \right)$$

$$\boldsymbol{B} = \frac{\boldsymbol{B}}{\mathbf{1}^{V \times V} \boldsymbol{B}} \qquad \boldsymbol{\Theta} = \frac{\boldsymbol{\Theta}}{\mathbf{1}^{K \times K} \boldsymbol{\Theta}}$$

- Ring a bell?

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\left\{ \frac{X}{\boldsymbol{W}\boldsymbol{H}} \right\} \boldsymbol{H}^\top}{\mathbf{1}^{V \times D} \boldsymbol{H}^\top}, \qquad \boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^\top \left\{ \frac{X}{\boldsymbol{W}\boldsymbol{H}} \right\}}{\boldsymbol{W}^\top \mathbf{1}^{V \times D}}.$$

- PLSI is equivalent to NMF
  - If NMF is using KL divergence as the error function
  - Except the normalization schemes

The Final Take on Topic Modeling (I Mean It)

- A geometric comparison to NMF

$\boldsymbol{B}_{1:V,k}$ One of the $\mathbf{K}$ topics: a distribution over the vocabulary

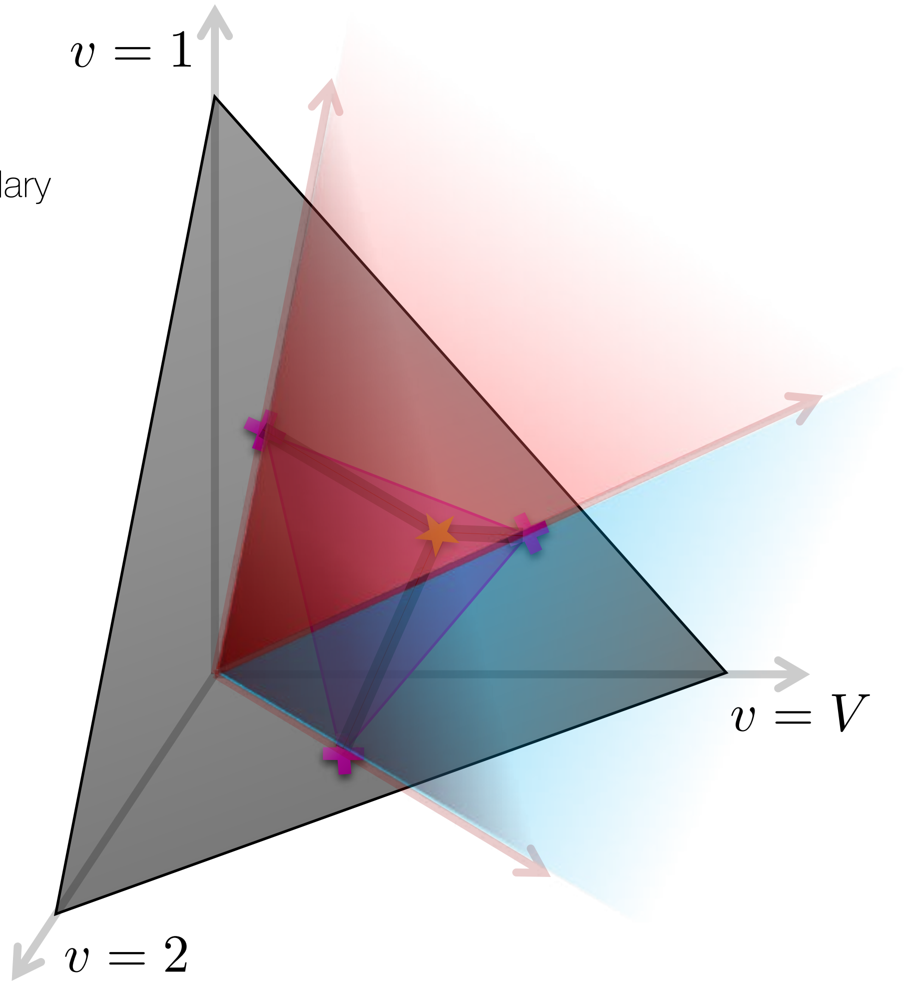
The convex hull defined by the topics

$\boldsymbol{\Theta}_{1:K,d}$ Contribution of the topics to the $\mathbf{d}$-th document

$X_{1:V,d}$ The $\mathbf{d}$-th document

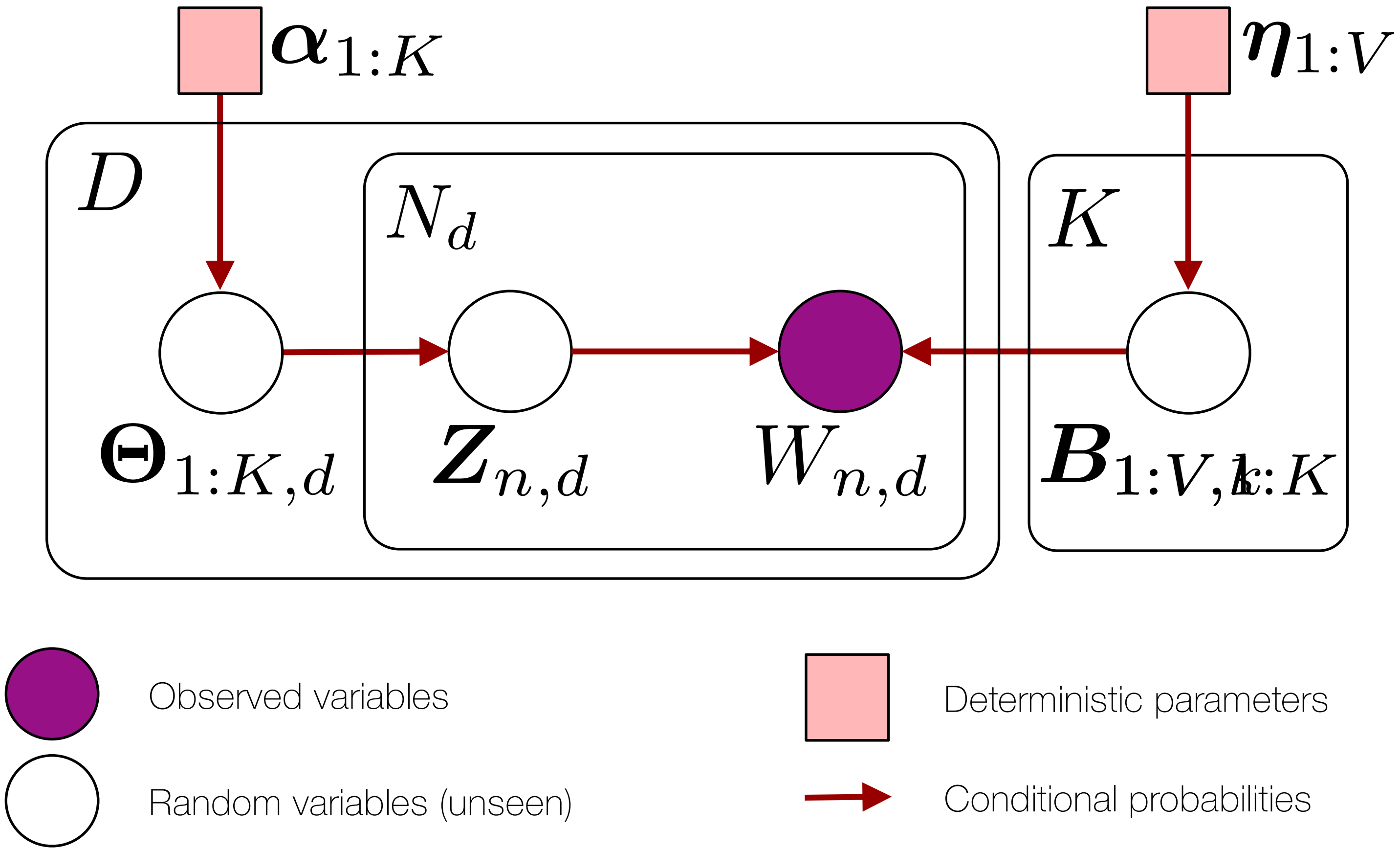
$\boldsymbol{W}$ The NMF basis vectors

$\boldsymbol{WH}$ The convex cone defined by the NMF model



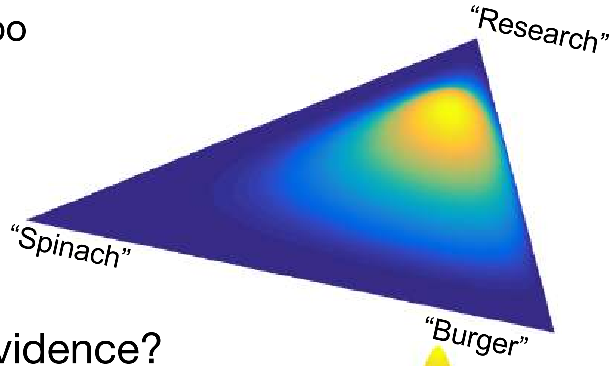
Latent Dirichlet Allocation

- A Bayesian touch on PLSI

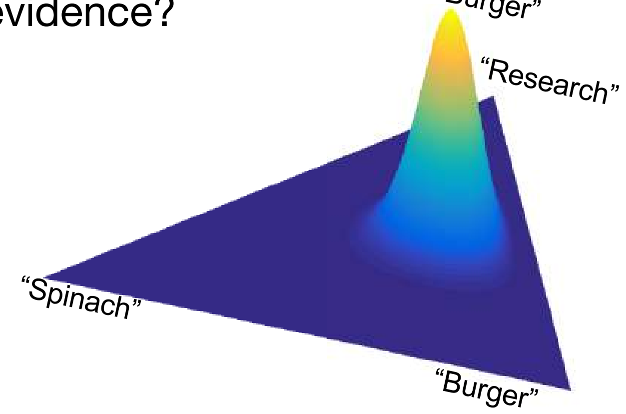


Why Dirichlet?
- Conjugate priors of multinomials

- If the a priori distribution has the similar algebraic form with the likelihood, we call it a **conjugate prior**
  - This ensures the posterior distributions have the same algebraic forms, too
- You observed in Prof. K's journal that he used
  - "Research": 5 times
  - "Burger": 4 times
  - "Spinach": 2 times



- What if you've watched them for a month and accumulate more evidence?
  - "Research": 19 times
  - "Burger": 15 times
  - "Spinach": 7 times



- These are the distribution where he samples from to decide what to write about on that day
- It looks like multinomial, but it's not
  - The RV is not for the counts, but for the parameter $P(\Theta_1 = p_1, \Theta_2 = p_2, \cdots, \Theta_K = p_K)$
  - The counts don't have to be integers (what?)

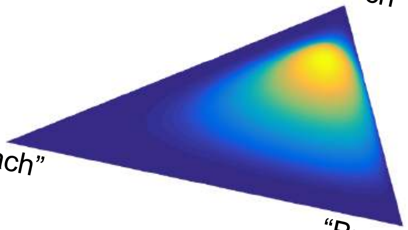
Why Dirichlet?

- Conjugate priors of multinomials

- **Dirichlet distribution**

$$P(\Theta_1 = p_1, \Theta_2 = p_2, \cdots, \Theta_K = p_K | \alpha_1, \alpha_2, \cdots, \alpha_K) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

$$\sum_{k=1}^K p_k = 1, \quad p_k \geq 0 \quad \forall p_k, \quad \alpha_k > 0 \quad \forall \alpha_k$$



  - We call α_k a **hyperparameter** or **pseudo count**
  - The conjugate prior of multinomial distribution
    - Because…

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_K = x_K | \Theta_1 = p_1, \Theta_2 = p_2, \cdots, \Theta_K = p_K)$$

$$\cdot P(\Theta_1 = p_1, \Theta_2 = p_2, \cdots, \Theta_K = p_K)$$

$$= \frac{N!}{\prod_{k=1}^K x_k!} \prod_{k=1}^K p_k^{x_k} \cdot \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

$$= \frac{N!}{\prod_{k=1}^K x_k!} \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{x_k + \alpha_k - 1}$$

$$\propto \prod_{k=1}^K p_k^{x_k + \alpha_k - 1}$$

←MAP is to maximize this w.r.t. **Θ**

Latent Dirichlet Allocation

- The generation process

- For the k-th topic (i.e. the distribution over the vocabulary) out of K
  - Sample a multinomial parameter $\boldsymbol{B}_{1:V,k}$ from $\mathrm{Dir}(\boldsymbol{\eta}_{1:V})$

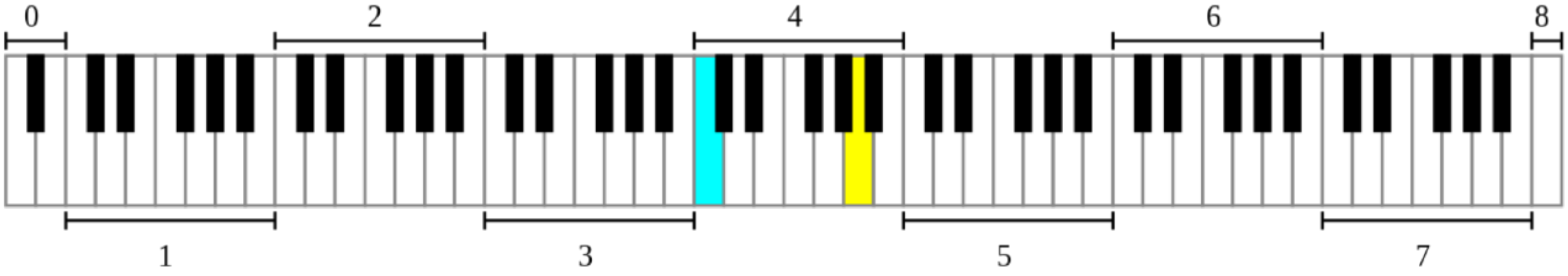
- For the d-th document in the collection of D documents
  - Sample a multinomial parameter $\boldsymbol{\Theta}_{1:K,d}$ from $\mathrm{Dir}(\boldsymbol{\alpha}_{1:K})$
  - For the n-th word among N_d words in the d-th document
    - Sample a topic $\boldsymbol{Z}_{n,d}$ from $\mathrm{Mult}(N_d, \boldsymbol{\Theta}_{1:K,d})$
    - Sample a word $W_{n,d}$ from $\mathrm{Mult}(N_d, \boldsymbol{B}_{1:V,Z_{n,d}})$

- Why are we doing this?
  - By introducing the Dirichlet priors for the parameters,
    we can have some more control over them

The Sparsity of the Topics

- How sparse my music is?

- Do you know how many notes are possible in music?
  - I don't know, perhaps a lot
- In piano, there are 88 keys

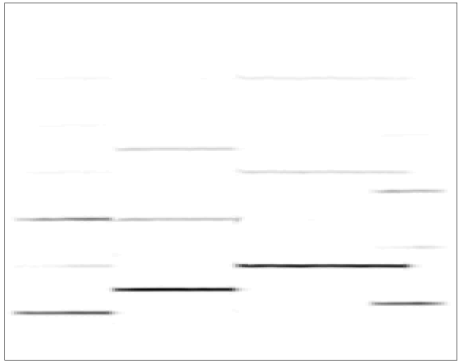


- Do you know how many notes are played at the same time?
  - You can go ahead and count them
  - But, most of the time there are not so many notes at a given time
- When I decompose a music signal, I want to introduce this prior knowledge
  - I mean the sparsity of the notes

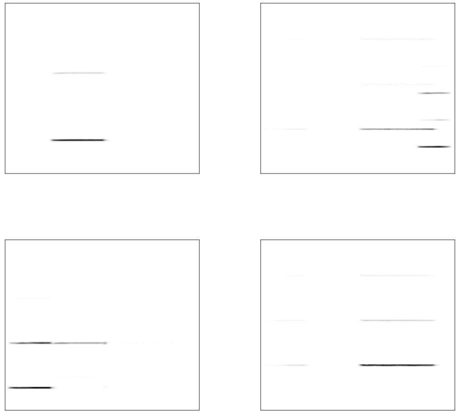
https://upload.wikimedia.org/wikipedia/commons/thumb/2/2e/Piano_Frequencies.svg/1280px-Piano_Frequencies.svg.png

The Sparsity of the Topics

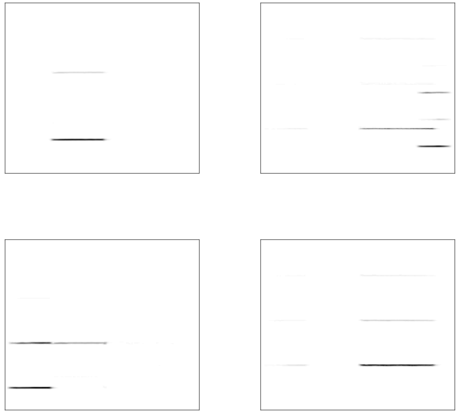
- LDA results on four notes with different hyperparameters



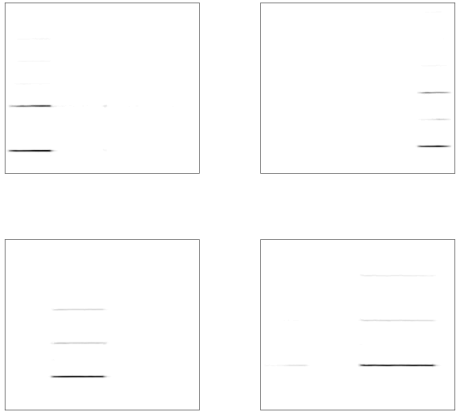
$\boldsymbol{\alpha} = \mathbf{1}^K, \boldsymbol{\eta} = \mathbf{1}^V$



$\boldsymbol{\alpha} = 0.1 \times \mathbf{1}^K, \boldsymbol{\eta} = 0.1 \times \mathbf{1}^V$



$\boldsymbol{\alpha} = 0.01 \times \mathbf{1}^K, \boldsymbol{\eta} = 0.01 \times \mathbf{1}^V$



(Collapsed) Gibbs Sampling for LDA

- Calculating posterior

- Posterior probabilities of LDA is difficult to calculate (as always)

$$P(\boldsymbol{Z}_{1:N_d,1:D}|W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{P(\boldsymbol{Z}_{1:N_d,1:D},W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta})}{\sum_{\boldsymbol{Z}} P(\boldsymbol{Z}_{1:N_d,1:D},W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta})}$$

- Instead, we calculate the individual posterior probability for (n,d)-th word in the collection given all the other observations

$$P(\boldsymbol{Z}_{n,d}=k|\widehat{\boldsymbol{Z}}_{\backslash n,d},W_{1:N_d,1:D})$$

$$\propto \underbrace{\frac{\sum_{n',d'\in\backslash n,d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n',d'}=k)\cdot\mathcal{I}(\boldsymbol{W}_{n',d'}=\boldsymbol{W}_{n,d}) \quad +\boldsymbol{\eta}}{\sum_{v=1}^{V}\sum_{n',d'\in\backslash n,d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n',d'}=k)\cdot\mathcal{I}(\boldsymbol{W}_{n',d'}=\cancel{\boldsymbol{W}_{n,d}}\,v) \; +V\boldsymbol{\eta}}}_{\text{Probability of seeing a word } \mathbf{v} \text{ in the } \mathbf{k}\text{-th topic}}$$

$$\cdot \underbrace{\frac{\sum_{n'\in\backslash n}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n',d}=k) \quad +\boldsymbol{\alpha}}{\sum_{k=1}^{K}\sum_{n'\in\backslash n}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n',d}=k) \; +K\boldsymbol{\alpha}}}_{\text{Probability of choosing the } \mathbf{k}\text{-th topic in the } \mathbf{d}\text{-th document}}$$

(Collapsed) Gibbs Sampling for LDA

- Calculating the parameters

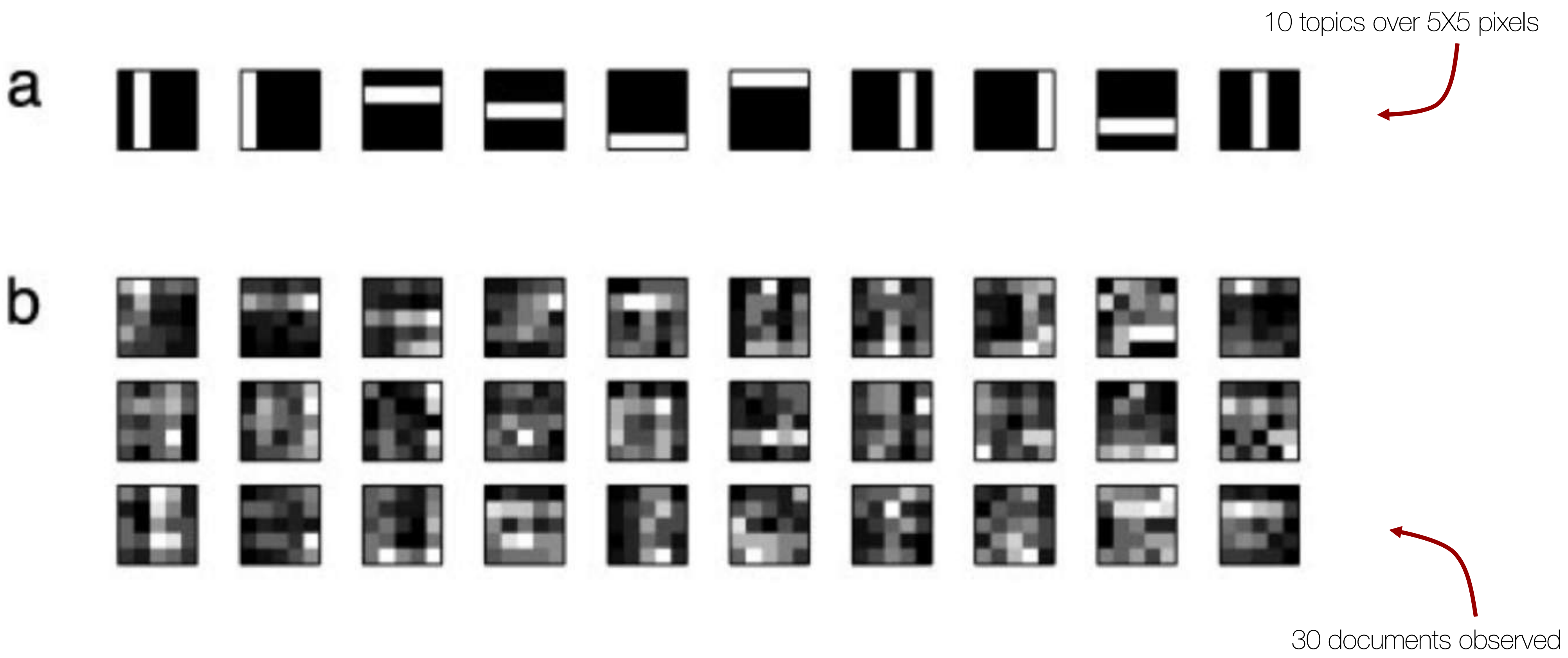
- Calculating the parameters (if needed)

$$\underbrace{\boldsymbol{B}_{v,k} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n,d}=k)\cdot\mathcal{I}(\boldsymbol{W}_{n,d}=v)+\boldsymbol{\eta}}{\sum_{v=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n,d}=k)\cdot\mathcal{I}(\boldsymbol{W}_{n,d}=v)+V\boldsymbol{\eta}}}_{\text{Probability of seeing a word } \mathbf{v} \text{ in the } \mathbf{k}\text{-th topic}}$$

$$\underbrace{\boldsymbol{\Theta}_{k,d} = \frac{\sum_{n=1}^{N_d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n,d}=k)+\boldsymbol{\alpha}}{\sum_{k=1}^{K}\sum_{n=1}^{N_d}\mathcal{I}(\widehat{\boldsymbol{Z}}_{n,d}=k)+K\boldsymbol{\alpha}}}_{\text{Probability of choosing the } \mathbf{k}\text{-th topic in the } \mathbf{d}\text{-th document}}$$

(Collapsed) Gibbs Sampling for LDA

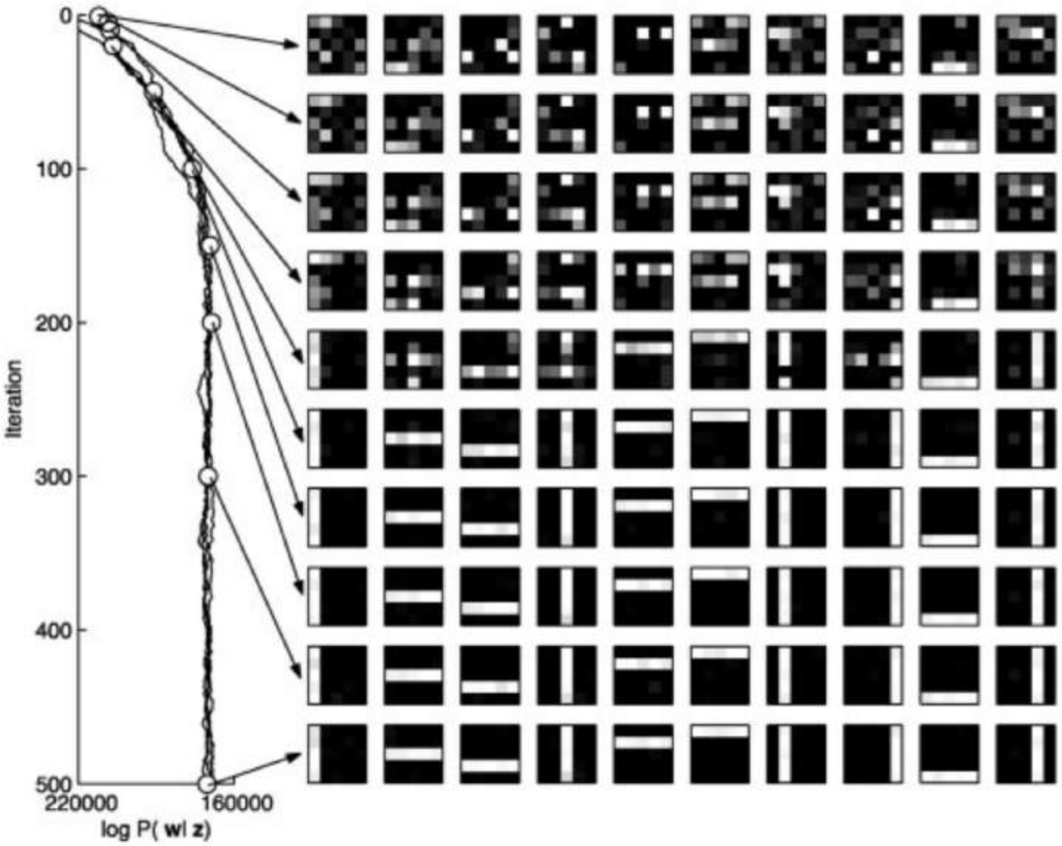
- Another toy example



Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Science, 101, 5228-5235.

(Collapsed) Gibbs Sampling for LDA

- Another toy example



Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Science, 101, 5228-5235

A Tweak for Spectrograms

- Matrices with (pseudo-) counts only

- We have no access to the full word sequence, but their counts
- The posterior probabilities

$$P(Z_{1:N_d,1:D}|W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{P(Z_{1:N_d,1:D},W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta})}{\sum_Z P(Z_{1:N_d,1:D},W_{1:N_d,1:D};\boldsymbol{\alpha},\boldsymbol{\eta})}$$

$$P(\mathcal{Z}_{1:V,1:D}|X_{1:V,1:D};\alpha,\eta) = \frac{P(\mathcal{Z}_{1:V,1:D},X_{1:V,1:D};\alpha,\eta)}{\sum_{\mathcal{Z}} P(\mathcal{Z}_{1:V,1:D}|X_{1:V,1:D};\alpha,\eta)}$$

- Gibbs sampling

$$P(\boldsymbol{\mathcal{Z}}_{v,d}=k|\widehat{\boldsymbol{\mathcal{Z}}}_{\setminus v,d},X_{1:V,1:D})$$

$$\propto \underbrace{\frac{\sum_{d'\in\setminus d}\mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d'}=k)\boldsymbol{X}_{v,d'}+\boldsymbol{\eta}}{\sum_{v=1}^{V}\sum_{d'\in\setminus d}\mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d'}=k)\boldsymbol{X}_{v,d'}+V\boldsymbol{\eta}}}_{\text{Probability of seeing a word } \mathbf{v} \text{ in the } \mathbf{k}\text{-th topic}}$$

$$\cdot \underbrace{\frac{\sum_{v'\in\setminus v}\mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v',d}=k)\mathbf{X}_{v',d}+\boldsymbol{\alpha}}{\sum_{k=1}^{K}\sum_{v'\in\setminus v}\mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v',d}=k)\mathbf{X}_{v',d}+K\boldsymbol{\alpha}}}_{\text{Probability of choosing the } \mathbf{k}\text{-th topic in the } \mathbf{d}\text{-th document}}$$

A Tweak for Spectrograms

- Matrices with (pseudo-) counts only

- Parameters

$$\underbrace{\boldsymbol{B}_{v,k} = \frac{\sum_{d=1}^{D} \mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d} = k) X_{v,d} + \boldsymbol{\eta}}{\sum_{v=1}^{V} \sum_{d=1}^{D} \mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d} = k) X_{v,d} + V\boldsymbol{\eta}}}_{\text{Probability of seeing a word } \mathbf{v} \text{ in the } \mathbf{k}\text{-th topic}}$$

$$\underbrace{\boldsymbol{\Theta}_{k,d} = \frac{\sum_{v=1}^{V} \mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d} = k) X_{v,d} + \boldsymbol{\alpha}}{\sum_{k=1}^{K} \sum_{v=1}^{V} \mathcal{I}(\widehat{\boldsymbol{\mathcal{Z}}}_{v,d} = k) X_{v,d} + K\boldsymbol{\alpha}}}_{\text{Probability of choosing the } \mathbf{k}\text{-th topic in the } \mathbf{d}\text{-th document}}$$

- Ring a bell?
  - PLSI update rules (M-step)
    - Before the reformulation for speed-up

$$\boldsymbol{B}_{v,k} = \frac{\sum_d X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k}}{\sum_v \sum_d X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k}} \qquad \boldsymbol{\Theta}_{k,d} = \frac{\sum_v X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k}}{\sum_v \sum_d X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k}}$$

A Tweak for Spectrograms

- An EM version for the MAP estimation

- I wouldn't get into the details for this, but an EM version looks similar

$$\boldsymbol{B}_{v,k} = \frac{\sum_d X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + \boldsymbol{\eta}}{\sum_v \sum_d X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + V\boldsymbol{\eta}} \qquad \boldsymbol{\Theta}_{k,d} = \frac{\sum_v X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + \boldsymbol{\alpha}}{\sum_k \sum_v X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + K\boldsymbol{\alpha}}$$

$$\boldsymbol{\mathcal{Z}}_{v,d,k} = \frac{\boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}{\sum_{k=1}^{K} \boldsymbol{B}_{v,k}\boldsymbol{\Theta}_{k,d}}$$

- So far we've considered the hyperparameters are scalar
  - What if not?

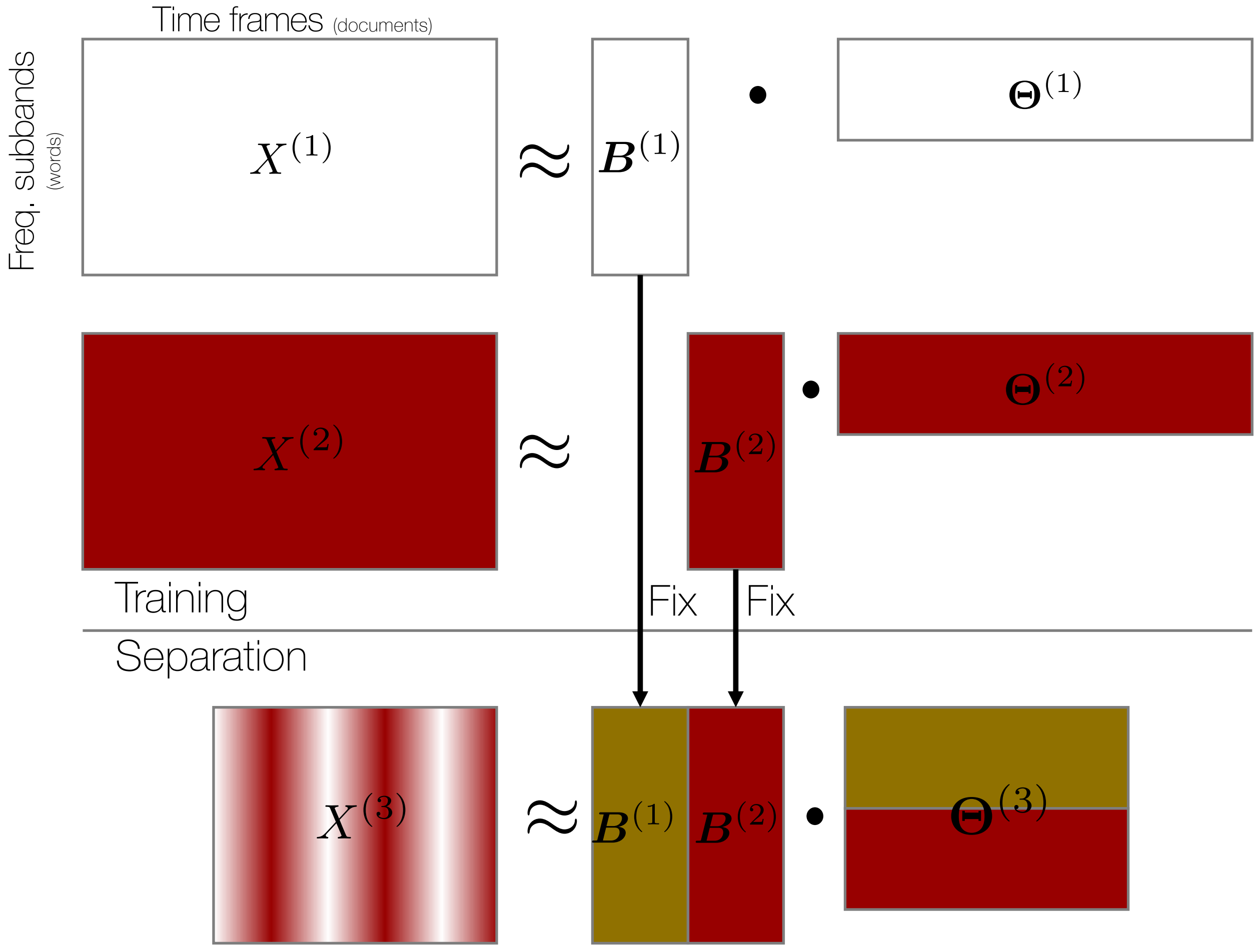
$$\boldsymbol{B}_{v,k} = \frac{\sum_d X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + \boldsymbol{\mathcal{H}}_{1:V,1:K}}{\sum_v \sum_d X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + \sum_v \boldsymbol{\mathcal{H}}_{1:V,1:K}}$$

Each basis vector has its own prior

$$\boldsymbol{\Theta}_{k,d} = \frac{\sum_v X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + \boldsymbol{\alpha}}{\sum_k \sum_v X_{v,d}\boldsymbol{\mathcal{Z}}_{v,d,k} + K\boldsymbol{\alpha}}$$

PLSI as a Source Separation Model

- You already know how to do this



LDA as a Source Separation Model

- You already know how to do this, too

- Reduced matrix computation version of LDA
- First, train your $\boldsymbol{\mathcal{H}}^{(1)}$ and $\boldsymbol{\mathcal{H}}^{(2)}$ using PLSI $\boldsymbol{\mathcal{H}} = [\boldsymbol{\mathcal{H}}^{(1)}, \boldsymbol{\mathcal{H}}^{(2)}]$
  - From your training data
- Do LDA updates either using Gibbs sampling

$$\boldsymbol{B}_{v,k} = \frac{\sum_d X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k} + \xi \boldsymbol{\mathcal{H}}_{1:V,1:K}}{\sum_v \sum_d X_{v,d} \boldsymbol{\mathcal{Z}}_{v,d,k} + \xi \sum_v \boldsymbol{\mathcal{H}}_{1:V,1:K}}$$

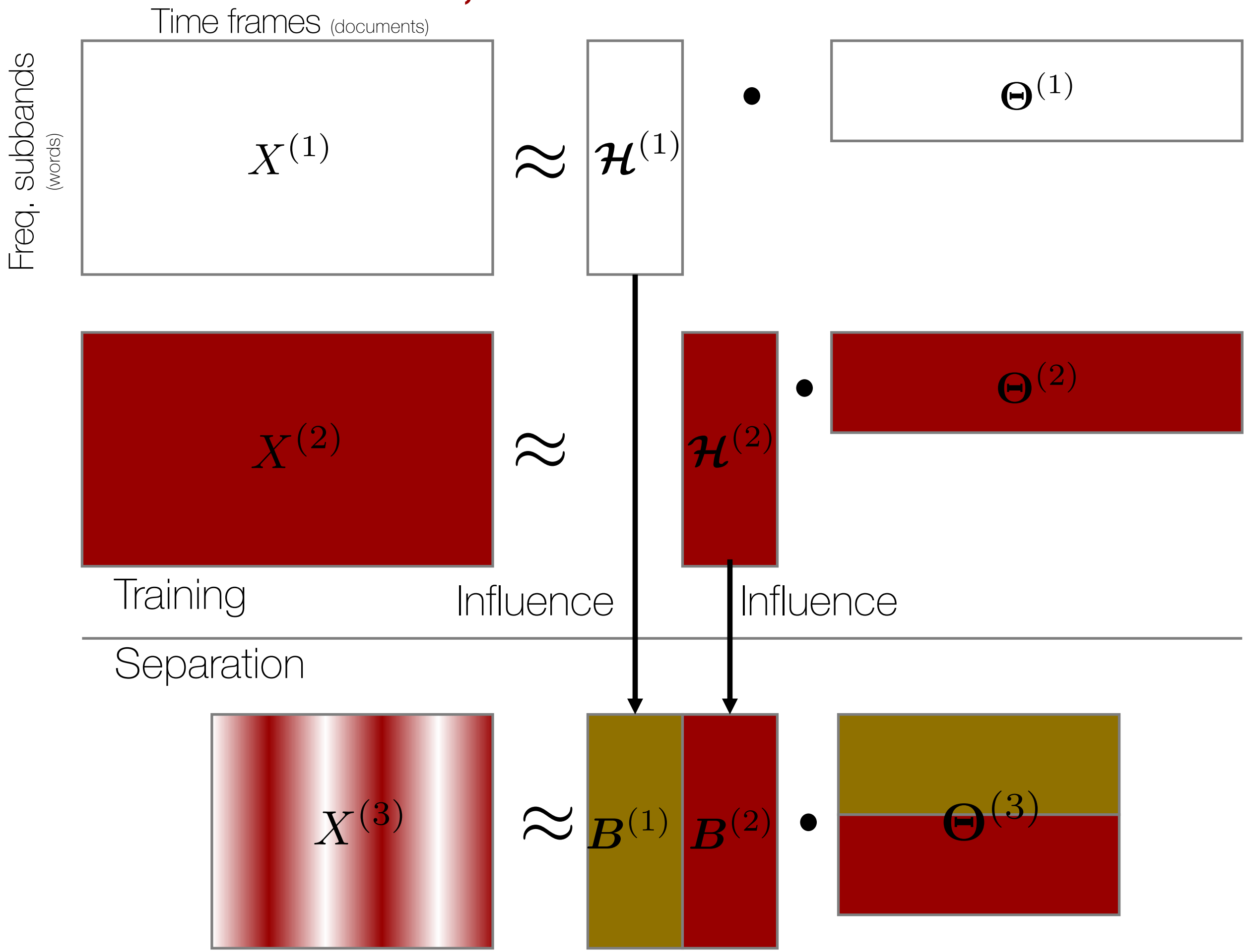
- Or EM

$$\boldsymbol{B} = \boldsymbol{B} \odot \left(\frac{X}{\boldsymbol{B}\boldsymbol{\Theta}} \boldsymbol{\Theta}^\top \right) + \xi \boldsymbol{\mathcal{H}} \qquad \boldsymbol{B} = \frac{\boldsymbol{B}}{\boldsymbol{1}^{V \times V} \boldsymbol{B}}$$

- If ξ is very very large
  - The other terms don't contribute
  - Same as PLSI with fixed basis vectors!

LDA as a Source Separation Model

- You already know how to do this, too



Speech Denoising Using Topic Modeling
- PLSI vs. LDA

- LDA can adapt to the variation

Training speech
(same person)

Training noise

Test mixture

PLSI: 11.38 (SNR)

LDA: 10.79 dB (SNR)

Training speech
(different person)

PLSI: 8.78 (SNR)

LDA: 10.97 dB (SNR)

Sentiment Analysis on Tweets

- Positive vs negative

- Which table looks topics from positive tweets (or negative)?

Topic 3	Topic 8
'apple'	'apple'
'thanks'	'ios5'
'apps'	'twitter'
'new'	'ipad'
'today'	'phone'
'only'	'screen'
'dear'	'upgrade'
'update'	'mac'
'phone'	**'love'**
'next'	**'better'**
'using'	**'little'**
'well'	**'top'**
'ups'	'come'
'post'	**'awesome'**
'yesterday'	'pull'
'awesome'	'bottom'
'sure'	'available'
'thx'	'asked'
'hard'	'messages'
'change'	**'missing'**
'impressed'	**'stuck'**
'calls'	'tweeting'

Topic 19	Topic 16
'apple'	'apple'
'iphone'	**'why'**
'love'	'really'
'new'	'itunes'
'still'	**'f**k'**
'ipad'	'know'
'restore'	'newsstand'
'fail'	**'wont'**
'nice'	'music'
'nothing'	'put'
'bestbuy'	'time'
'fixed'	**'well'**
'charge'	'appstore'
'lost'	**'wtf'**
'want'	'umber'
'wish'	'shows'
'hate'	'folder'
'turned'	'key'
'annoyed'	'design'
'down'	'changes'
'gone'	'component'
'repair'	**'hate'**

Reading

- PLSI
  - Original papers:
    http://dl.acm.org/citation.cfm?id=312649
    http://dl.acm.org/citation.cfm?id=2073829
  - Derivation: https://arxiv.org/pdf/1212.3900.pdf
- LDA
  - Original LDA: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
  - Gibbs sampling: http://www.pnas.org/content/101/suppl_1/5228.full.pdf
  - https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- Kevin Murphy, “Machine Learning: a Probabilistic Perspective”,
  - http://site.ebrary.com/lib/iub/detail.action?docID=10597102
  - Chapter 27.3

Thank You!