2020
26
22

University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chenhao Tan

Lecture 21: Variational Inference

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Learning Objectives

- Understand evaluations
- Understand the intuition behind variational inference

Outline

- Evaluations of topic models
- Variational inference

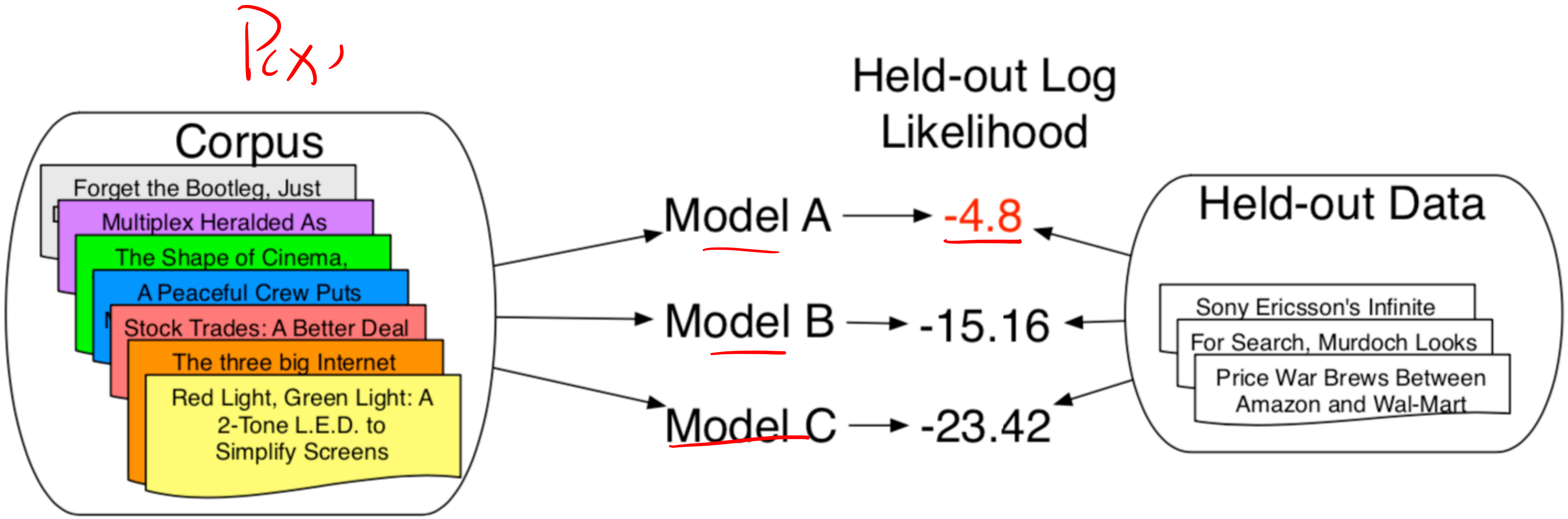
Outline

- Evaluations of topic models
- Variational inference

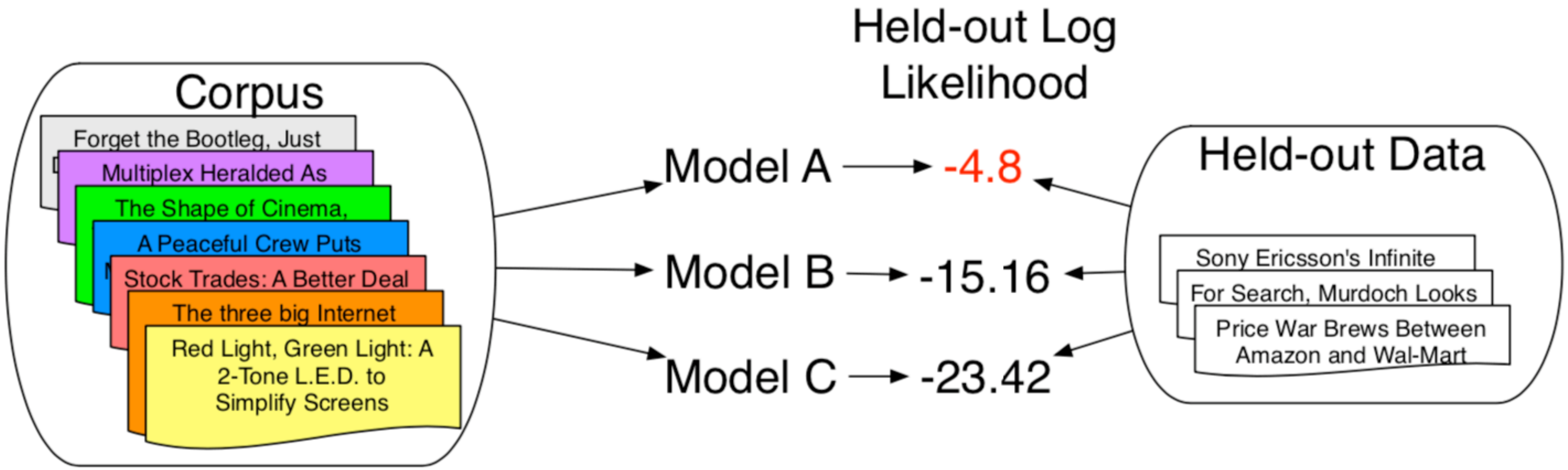
Evaluating Topic Models

- Held-out log likelihood
- Word intrusion Jordan Boyd-Graber

Held-out log likelihood



Held-out log likelihood



$$p(w \mid \alpha, \beta) = \int_{\theta,z} p(\theta, z, w \mid \alpha, \beta)$$

$$= \prod_d \int_{\theta_d} p(\theta_d \mid \alpha) \prod_n \sum_{z_{d,n}} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Word Intrusion

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Word Intrusion

1. Take the highest probability words from a topic

Original Topic
dog, cat, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic
dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder
dog, cat, apple, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic
dog, cat, horse, pig, cow

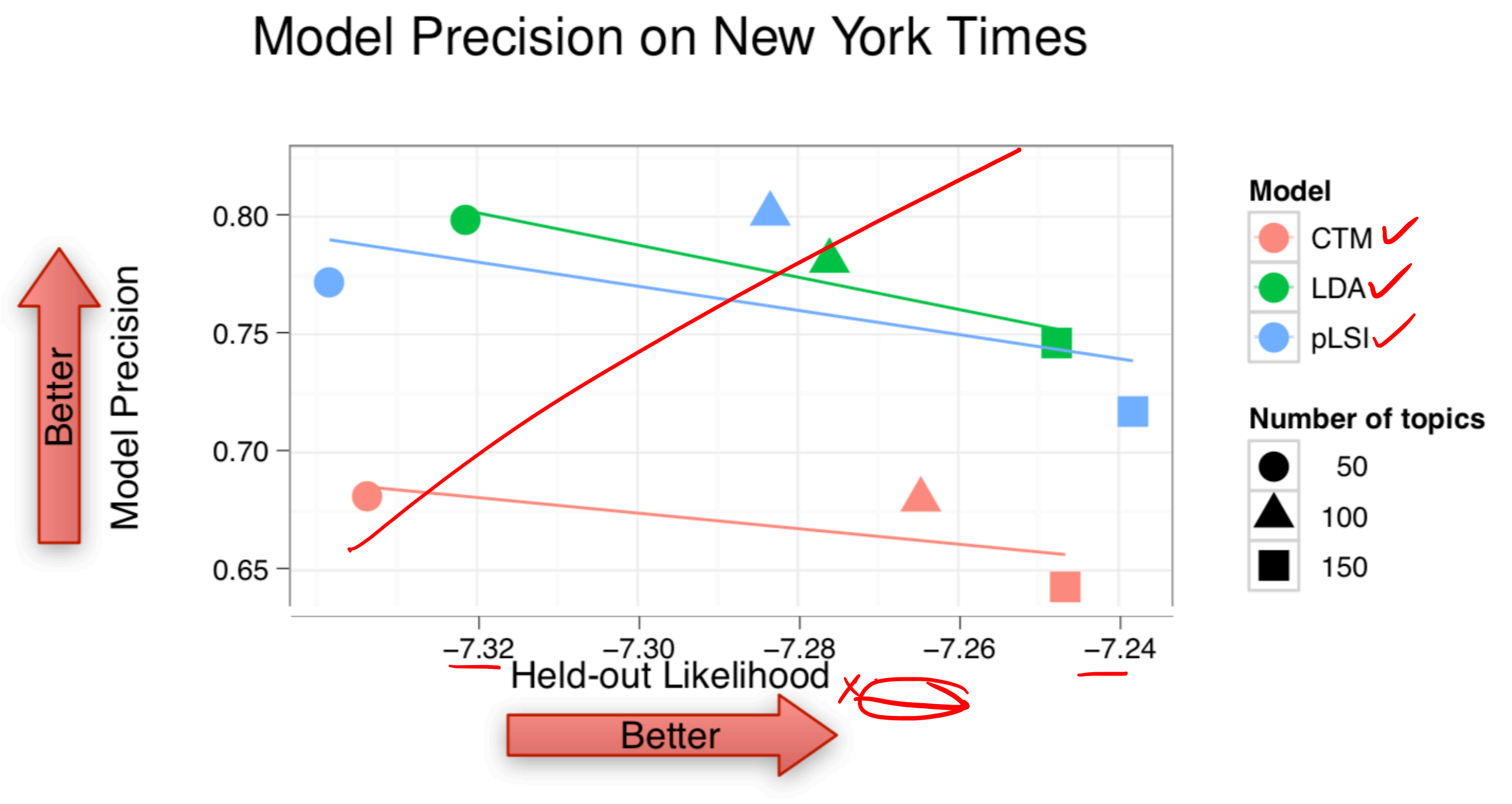
2. Take a high-probability word from another topic and add it

Topic with Intruder
dog, cat, apple, horse, pig, cow

3. We ask users to find the word that doesn't belong

Hypothesis
If the topics are interpretable, users will consistently choose true intruder

Interpretability and likelihood



within a model, higher likelihood $\neq$ higher interpretability

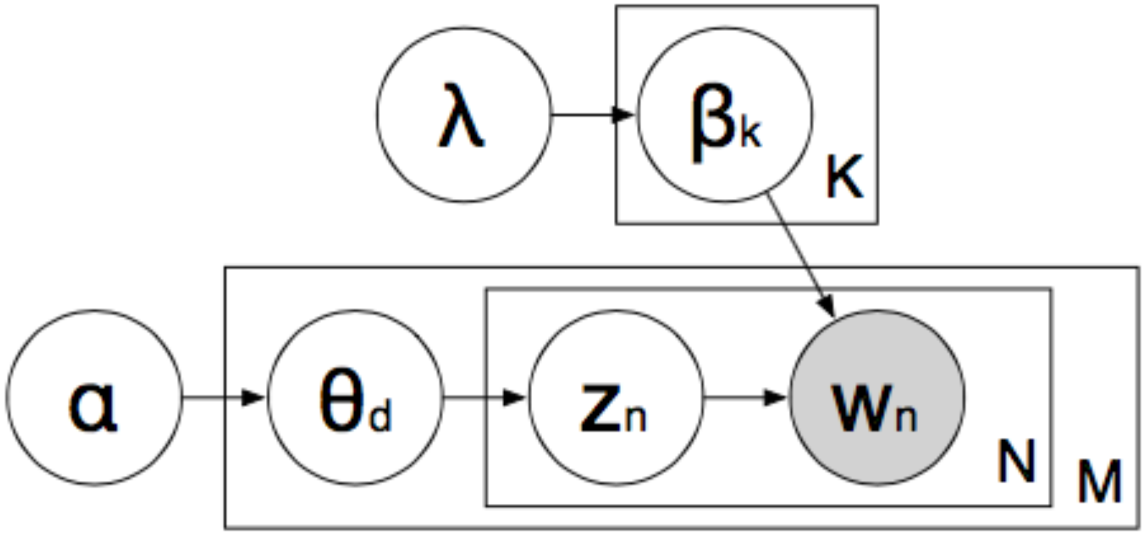
Evaluation takeaway

- Measure what you care about
- If you care about prediction, likelihood is great
- If you care about a particular task, measure that

Outline

- Evaluations of topic models
- Variational inference

Joint distribution



$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Joint distribution

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)
- $p(z_{d,n} \mid \theta_d) = \underline{\theta_{d,z_{d,n}}}$ (Draw from Multinomial)
- $p(w_{d,n} \mid \beta, z_{d,n}) = \underline{\beta_{z_{d,n},w_{d,n}}}$ (Draw from Multinomial)

Posterior distribution

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Posterior distribution:

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

$$p(w \mid \alpha, \beta) = \int_{\theta, z} p(\theta, z, w \mid \alpha, \beta)$$

$$= \prod_d \int_{\theta_d} p(\theta_d \mid \alpha) \prod_n \sum_{z_{d,n}} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Variational inference

ELBO

Instead of estimating the posterior distribution directly, we use a distribution of simpler forms, $q(\theta, z)$ to approximate $P(\theta, z \mid w, \alpha, \beta)$. We try to minimize the difference between p and q

$q(\theta, z) = q(\theta) q(z)$

$$\min \mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(\theta, z)}{p(\theta, z \mid w)} \right] \Rightarrow \max \text{ ELBO}$$

- If q and p are high, we're happy
- If q is high but p isn't, we pay a price
- If q is low, we don't care
- If $\mathrm{KL} = 0$, then distribution are equal

q is usually referred to as the variational distribution.

KL divergence and evidence lower bound

$$KL(q \| p) = E_q \log \frac{q(\theta, z)}{p(\theta, z \mid w)}$$

$$p(\theta, z \mid w) = \frac{p(\theta, z, w)}{p(w)}$$

$$= E_q \log \frac{q(\theta, z)\, p(w)}{p(\theta, z, w)}$$

$$= E_q \log q(\theta, z) - E_q \log p(\theta, z, w) + \underline{E_q \log p(w)}$$

$$E_q \log p(w) = \sum_{\theta, z} q(\theta, z) \log p(w) = 1 \cdot \log p(w)$$

$$= E_q \log q(\theta, z) - E_q \log p(\theta, z, w) + \log p(w) \geq 0$$

$$\log p(w) \geq \underline{E_q \log p(\theta, z, w) - E_q \log q(\theta, z)}$$

KL divergence and evidence lower bound

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)}$$

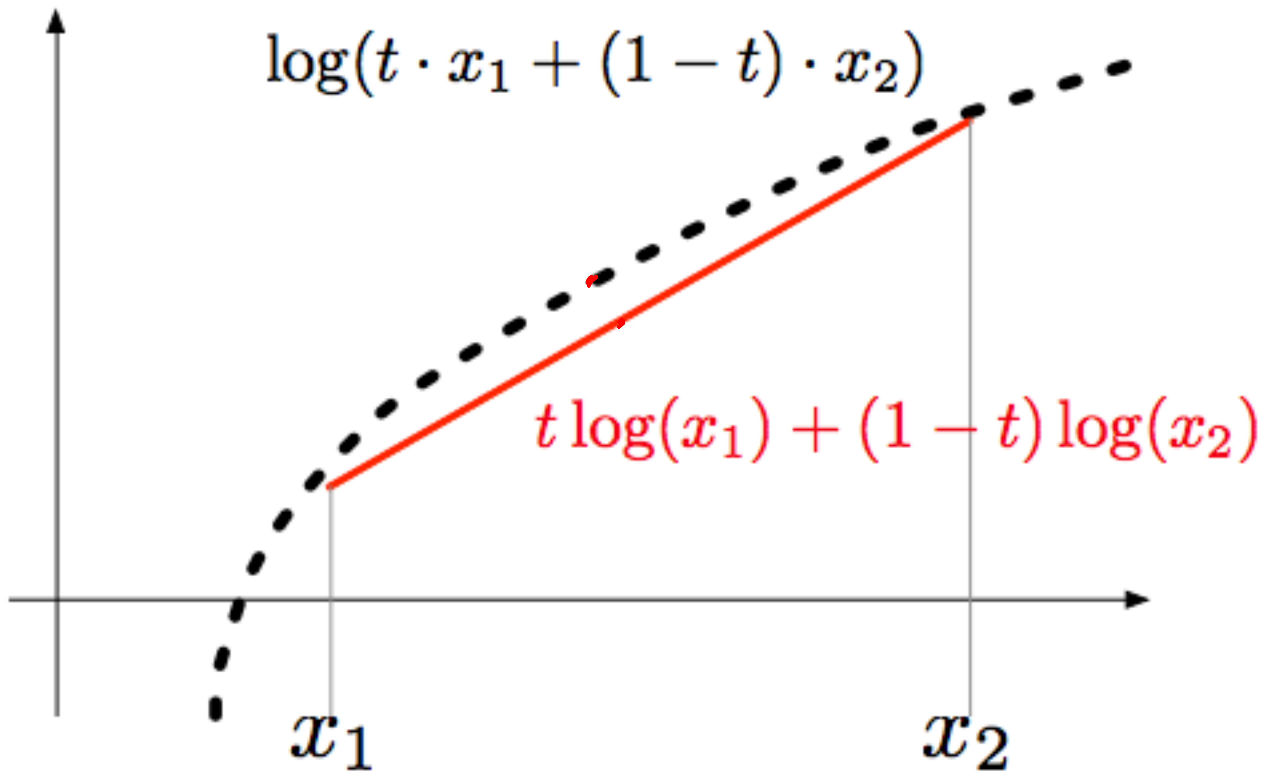
- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) \,||\, p(z \mid x)) &= \mathbb{E}_q \left[\log \frac{q(z)}{p(z \mid x)} \right] \\
&= \mathbb{E}_q \left[\log q(z) \right] - \mathbb{E}_q \left[\log p(z \mid x) \right] \\
&= \mathbb{E}_q \left[\log q(z) \right] - \mathbb{E}_q \left[\log p(z, x) \right] + \log p(x) \\
&= -\left(\mathbb{E}_q \left[\log p(z, x) \right] - \mathbb{E}_q \left[\log q(z) \right] \right) + \log p(x)
\end{aligned}$$

- Negative of ELBO (plus constant); minimizing KL divergence is the same as maximizing ELBO

A different way to get ELBO

- Jensen's inequality



When f is concave

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

Evidence Lower Bound

- Apply Jensen's inequality on log probability of data

$$\underline{\log p(x)} = \log\left[\int_z p(x,z)\right]$$

$$= \log \int_z p(x,z) \frac{q(z)}{q(z)} dz$$

$$= \log \int_z q(z) \cdot \frac{p(x,z)}{q(z)} dz$$

$$\geq \int_z q(z) \log \frac{p(x,z)}{q(z)} dz$$

$$= \int_z q(z) \log p(x,z)\, dz - \int_z q(z) \log q(z)\, dz$$

$$= E_q \log p(x,z) - E_q \log q(z)$$

Evidence Lower Bound

- Apply Jensen's inequality on log probability of data

$$
\begin{aligned}
\log p(x) &= \log\left[\int_z p(x,z)\right] \\
&= \log\left[\int_z p(x,z)\frac{q(z)}{q(z)}\right] \\
&= \log\left[\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right] \\
&\geq \mathbb{E}_q\left[\log p(x,z)\right] - \mathbb{E}_q\left[\log q(z)\right]
\end{aligned}
$$

Apply Jensen's equality and use log difference

Variational inference

- Propose variational distribution q $P_{\theta}(x)$
- Find ELBO (evidence lower bound) using q
- Set derivatives to 0 and update variables

Variational distribution for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Posterior distribution:

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

$q(\theta, z)$

$$p(w \mid \alpha, \beta) = \int_{\theta, z} p(\theta, z, w \mid \alpha, \beta)$$

$$= \prod_d \int_{\theta_d} p(\theta_d \mid \alpha) \prod_n \sum_{z_{d,n}} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n})$$

Variational distribution for LDA

$$q(\theta, z) = \prod_{d} q(\theta_d|\gamma_d) \prod_{z} q(z|\phi)$$

$$\frac{\Gamma(\sum \gamma_d)}{\prod_i \Gamma(\gamma_{di})} \prod_i \theta_{di}^{\gamma_{di}-1} \qquad q(z|\phi) = \phi_z$$

- Variational document distribution over topics γ_d
  - Vector of length K for each document
  - Non-negative
  - Doesn't sum to 1.0
- Variational token distribution over topic assignments $\phi_{d,n}$
  - Vector of length K for every token
  - Non-negative, sums to 1.0

Overall Algorithm

1. Randomly initialize variational parameters (can't be uniform)
2. For each iteration:

   variational document-topic

   2.1 For each document, update γ and ϕ

   2.2 For corpus, update $\beta \to$ topic-word

   2.3 Compute $\mathcal{L}$ for diagnostics
3. Return expectation of variational parameters for solution to latent variables

Updates to Maximize ELBO

For each document d, γ means γ_d here, ϕ_n means $\phi_{d,n}$ here:

V=Wn

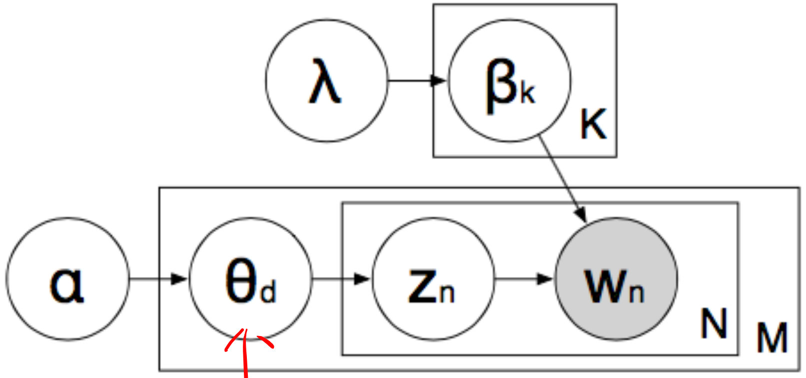
M×N×K $$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right)$$

K×V

M×K $$\gamma_i = \alpha_i + \sum_n \phi_{ni}$$

For the entire corpus,

$$\beta_{ij} \propto \sum_d \sum_n \phi_{dni} w_{dn}^j$$



Example

- Three topics

$$\beta = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{bmatrix} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \\ .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{bmatrix} \begin{array}{c} \\ =1 \\ \\ \\ \end{array} \tag{4}$$

- Assume uniform γ: $(2.0, 2.0, 2.0)$
- Compute update for ϕ

$$\phi_{ni} \propto \beta_{iv} \exp \left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{5}$$

- For the first word (dog) in the document: **dog cat cat pig**

Update ϕ for dog $n=0$

$$\beta = \begin{array}{ccccc} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \\ \end{array}$$
$$\left[\begin{array}{ccccc} .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{array}\right]$$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right)$$

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(1) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(2) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- After normalization: {0.333, 0.333, 0.333}

Update ϕ for pig $n=1$

$$\beta = \begin{array}{c} \begin{array}{ccccc} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \end{array} \\ \left[\begin{array}{ccccc} .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{array}\right] \end{array}$$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right)$$

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(1) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(2) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- After normalization: {0.333, 0.333, 0.333}

Update ϕ for cat

$n=2$

$$\beta = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{ccccc} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \\ \end{array}$$

$$\begin{bmatrix} .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{bmatrix}$$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right)$$

$\beta_{0,cat}$

$\beta_{1,cat}$

$\beta_{2,cat}$

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.260 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.072$
- $\phi(1) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(2) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- After normalization: {0.413, 0.294, 0.294}

Update γ

- Document: dog cat cat pig
- Update equation

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{6}$$

- Assume $\alpha = (.1, .1, .1)$

	ϕ_0	ϕ_1	ϕ_2	
dog	.333	.333	.333	
cat	.413	.294	.294	x2
pig	.333	.333	.333	
α	0.1	0.1	0.1	
sum	1.592	1.354	1.354	

- Note: **do not normalize**!

Update β

- Count up all of the ϕ across all documents
- For each topic, divide by total
- Corresponds to maximum likelihood of expected counts

Recap

- Topic models: a neat way to model discrete count data
- Variational inference converts intractable optimization to maximizing ELBO