Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

# Variational Inference

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 21

## Roadmap

- Big-picture questions
- VI for LDA
- More content questions
- Walkthrough of VI for LDA (HW)

## Big picture content questions

## Big picture content questions

## Big picture content questions

## Big picture content questions

# Big picture content questions

## Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_d \left[ \prod_k \theta_{d,k}^{\alpha_k - 1} \left( \prod_n \prod_i \prod_j (\theta_{d,i}\beta_{i,j})^{w_{d,n}^j} \right) \right] \tag{1}$$

## Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_d \left[ \prod_k \theta_{d,k}^{\alpha_k - 1} \left( \prod_n \prod_i \prod_j (\theta_{d,i} \beta_{i,j})^{w_{d,n}^j} \right) \right] \tag{1}$$

Variational distribution:

$$q(\theta, z) = q(\theta \mid \gamma) q(z \mid \phi) \tag{2}$$

## Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_d \left[ \prod_k \theta_{d,k}^{\alpha_k - 1} \left( \prod_n \prod_i \prod_j (\theta_{d,i} \beta_{i,j})^{w_{d,n}^j} \right) \right] \tag{1}$$

Variational distribution:

$$q(\theta, z) = q(\theta \mid \gamma) q(z \mid \phi) \tag{2}$$

ELBO:

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) = & \mathbb{E}_q [\log p(\theta \mid \alpha)] + \mathbb{E}_q [\log p(z \mid \theta)] + \mathbb{E}_q [\log p(w \mid z, \beta)] \\ & - \mathbb{E}_q [\log q(\theta)] - \mathbb{E}_q [\log q(z)] \end{aligned} \tag{3}$$

## Expectation of $\log$ Dirichlet

- Most expectations are straightforward to compute
- Dirichlet is harder

$$\mathbb{E}_{\text{dir}}\left[p(\theta_i \,|\, \alpha)\right] = \Psi\left(\alpha_i\right) - \Psi\left(\sum_j \alpha_j\right) \tag{4}$$

## Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{5}$$

$$\tag{6}$$

## Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{5}$$

$$= \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i - 1}\right] \tag{6}$$

Log of products becomes sum of logs.

## Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{5}$$

$$= \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i - 1}\right]$$

$$= \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \mathbb{E}_q\left[\sum_i (\alpha_i - 1)\log \theta_i\right] \tag{6}$$

Log of exponent becomes product, expectation of constant is constant

## Expectation 1

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] =& \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i-1}\right\}\right] \qquad (5)\\
=& \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\right\} + \sum_i \log \theta_i^{\alpha_i-1}\right]\\
=& \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \mathbb{E}_q\left[\sum_i (\alpha_i - 1)\log \theta_i\right]\\
=& \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i)\\
& + \sum_i (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right)
\end{aligned}
$$

Expectation of log Dirichlet

## Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] \tag{6}$$

$$\tag{7}$$

## Expectation 2

$$\mathbb{E}_q[\log p(z \mid \theta)] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{6}$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{7}$$

$$\tag{8}$$

Products to sums

## Expectation 2

$$\mathbb{E}_q [\log p(z \mid \theta)] = \mathbb{E}_q \left[ \log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]} \right] \tag{6}$$

$$= \mathbb{E}_q \left[ \sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n == i]} \right] \tag{7}$$

$$= \sum_n \sum_i \mathbb{E}_q \left[ \log \theta_i^{\mathbb{1}[z_n == i]} \right] \tag{8}$$

$$\tag{9}$$

Linearity of expectation

## Expectation 2

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(z \mid \theta)\right] =& \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] && (6)\\
=& \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] && (7)\\
=& \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] && (8)\\
=& \sum_n \sum_i \phi_{ni} \mathbb{E}_q\left[\log \theta_i\right] && (9)\\
& && (10)
\end{aligned}
$$

Independence of variational distribution, exponents become products

## Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{6}$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{7}$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{8}$$

$$= \sum_n \sum_i \phi_{ni} \mathbb{E}_q\left[\log \theta_i\right] \tag{9}$$

$$= \sum_n \sum_i \phi_{ni} \left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{10}$$

Expectation of log Dirichlet

## Complete objective function

$$
\begin{aligned}
L(\gamma,\phi;\alpha,\beta) &= \log\Gamma\left(\textstyle\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij} \\
&- \log\Gamma\left(\textstyle\sum_{j=1}^{k}\gamma_j\right) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k}\gamma_j\right)\right) \\
&- \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni},
\end{aligned}
$$

Note the entropy terms at the end (negative sign)

## Deriving the algorithm

- Compute partial wrt to variable of interest
- Set equal to zero
- Solve for variable

## Update for $\phi$

Derivative of ELBO:

$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda.$$

Solution:

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{11}$$

## Update for $\gamma$

Derivative of ELBO:

$$\frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^{k} \gamma_j\right) \sum_{j=1}^{k} \left(\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j\right).$$

Solution:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}.$$

## Update for $\beta$

Slightly more complicated (requires Lagrange parameter), but solution is obvious:

$$\beta_{ij} \propto \sum_d \sum_n \phi_{dni} w_{dn}^j \tag{12}$$

## Detail content questions

## Detail content questions

## Detail content questions

# Detail content questions

# Detail content questions

## Administrivia

- End of flipped classroom
    - Short content session at start
    - Use the time to meet with teammates
- First deliverable soon!

## Example

- Three topics, same documents as last time

$$\beta = \begin{array}{ccccc} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \end{array} \\ \begin{bmatrix} .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{bmatrix} \tag{13}$$

- Assume uniform $\gamma$: $(2.0, 2.0, 2.0)$
- Compute update for $\phi$

$$\phi_{ni} \propto \beta_{iv} \exp \left( \Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{14}$$

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

## Example

- Three topics, same documents as last time

$$\beta = \begin{bmatrix} \text{cat} & \text{dog} & \text{hamburger} & \text{iron} & \text{pig} \\ .26 & .185 & .185 & .185 & .185 \\ .185 & .185 & .26 & .185 & .185 \\ .185 & .185 & .185 & .26 & .185 \end{bmatrix} \tag{13}$$

- Assume uniform $\gamma$: $(2.0, 2.0, 2.0)$
- Compute update for $\phi$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{14}$$

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

## Update $\phi$ for dog

- $\gamma = (2.000, 2.000, 2.000)$

## Update $\phi$ for dog

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for dog

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for dog

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for dog

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

## Update $\phi$ for pig

- $\gamma = (2.000, 2.000, 2.000)$

## Update $\phi$ for pig

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for pig

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for pig

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for pig

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

## Update $\phi$ for cat

- $\gamma = (2.000, 2.000, 2.000)$

## Update $\phi$ for cat

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.260 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.072$

## Update $\phi$ for cat

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.260 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.072$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for cat

- $\gamma = (2.000, 2.000, 2.000)$
- $\phi(0) \propto 0.260 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.072$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$
- $\phi(0) \propto 0.185 \times \exp\left(\Psi\left(2.000\right) - \Psi\left(2.000 + 2.000 + 2.000\right)\right) = 0.051$

## Update $\phi$ for cat

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

## Update $\gamma$

- Document: dog cat cat pig
- Update equation

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{15}$$

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

## Update $\gamma$

- Document: dog cat cat pig
- Update equation

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{15}$$

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

| | $\phi_0$ | $\phi_1$ | $\phi_2$ |
|---|---|---|---|
| dog | .333 | .333 | .333 |
| cat | .413 | .294 | .294 |
| pig | .333 | .333 | .333 |
| $\alpha$ | 0.1 | 0.1 | 0.1 |
| sum | 1.592 | 1.354 | 1.354 |

- Note: **do not normalize**!

## Update $\gamma$

- Document: dog cat cat pig
- Update equation

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{15}$$

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

| | $\phi_0$ | $\phi_1$ | $\phi_2$ | |
|---|---|---|---|---|
| dog | .333 | .333 | .333 | |
| cat | .413 | .294 | .294 | x2 |
| pig | .333 | .333 | .333 | |
| $\alpha$ | 0.1 | 0.1 | 0.1 | |
| sum | 1.592 | 1.354 | 1.354 | |

- Note: **do not normalize**!

## Update $\beta$

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

## Update $\beta$

- Count up all of the $\phi$ across all documents
- For each topic, divide by total
- Corresponds to maximum likelihood of expected counts
- Unlike Gibbs sampling, no Dirichlet prior