Variational Inference

Machine Learning: Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 19

Variational Inference

- Inferring hidden variables
- Unlike MCMC:
  - Deterministic
  - Easy to gauge convergence
  - Requires dozens of iterations
- Doesn't require conjugacy
- Slightly hairier math

Setup

- $\vec{x} = x_{1:n}$ observations
- $\vec{z} = z_{1:m}$ hidden variables
- α fixed parameters
- Want the posterior distribution

$$p(z \mid x, \alpha) = \frac{p(z, x \mid \alpha)}{\int_z p(z, x \mid \alpha)} \quad (1)$$

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \dots n$:
   2.1 Draw $z_i \sim \text{Mult}(\pi)$
   2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \,|\, x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})} \quad (2)$$

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \dots n$:
   2.1 Draw $z_i \sim \text{Mult}(\pi)$
   2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \,|\, x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})} \tag{2}$$

Consider all means

Motivation

- Can't compute posterior for many interesting models

GMM (finite)

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$
2. For each observation $i = 1 \dots n$:
   - 2.1 Draw $z_i \sim \text{Mult}(\pi)$
   - 2.2 Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_0^2)$

- Posterior is intractable for large n, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \,|\, x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \,|\, z_i, \mu_{1:K})} \tag{2}$$

Consider all assignments

Main Idea

- We create a **variational distribution** over the latent variables

$$q(z_{1:m} \mid \nu) \tag{3}$$

- Find the settings of ν so that q is close to the posterior
- If $q == p$, then this is vanilla EM

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z \mid x)} \right] \tag{4}$$

What does it mean for distributions to be close?

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z \,|\, x)} \right] \quad (4)$$

- Characterizing KL divergence
  - 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

What does it mean for distributions to be close?

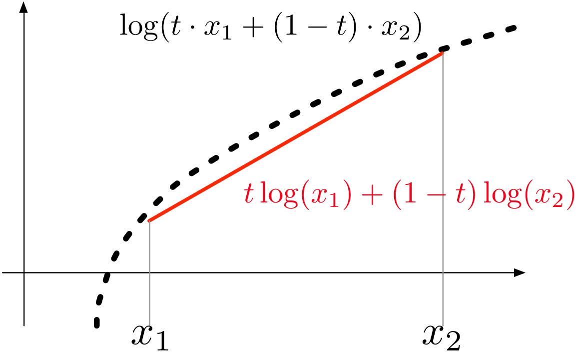
- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\text{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[\log \frac{q(Z)}{p(Z \,|\, x)} \right] \tag{4}$$

- Characterizing KL divergence
  - 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 $\text{KL} = 0$, then distribution are equal

This behavior is often called "mode splitting": we want **a** good solution, not every solution.

Jensen's Inequality: Concave Functions and Expectations



When f is concave

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

If you haven't seen this before, spend fifteen minutes to convince yourself that it's true

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

Add a term that is equal to one

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

Take the numerator to create an expectation

Evidence Lower Bound (ELBO)

- 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

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Evidence Lower Bound (ELBO)

- Apply Jensen's inequality on log probability of data

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

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on on log probability

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

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

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

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

Break quotient into difference

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

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

Apply definition of conditional probability

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

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

Reorganize terms

Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

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

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

Mean field variational inference

- Assume that your variational distribution factorizes

$$q(z_1, \dots, z_m) = \prod_{j=1}^{m} q(z_j) \tag{6}$$

- You may want to group some hidden variables together
- Does not contain the true posterior because hidden variables are dependent

General Blueprint

- Choose q
- Derive ELBO
- Coordinate ascent of each q_i
- Repeat until convergence

Example: Latent Dirichlet Allocation

TOPIC 1

computer, technology, system, service, site, phone, internet, machine

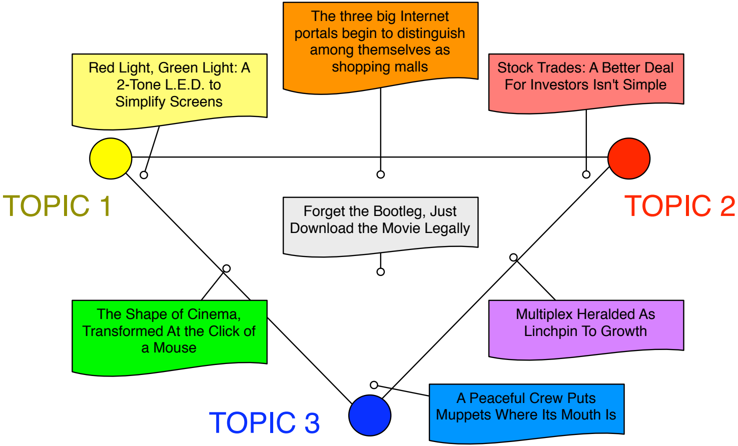
TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

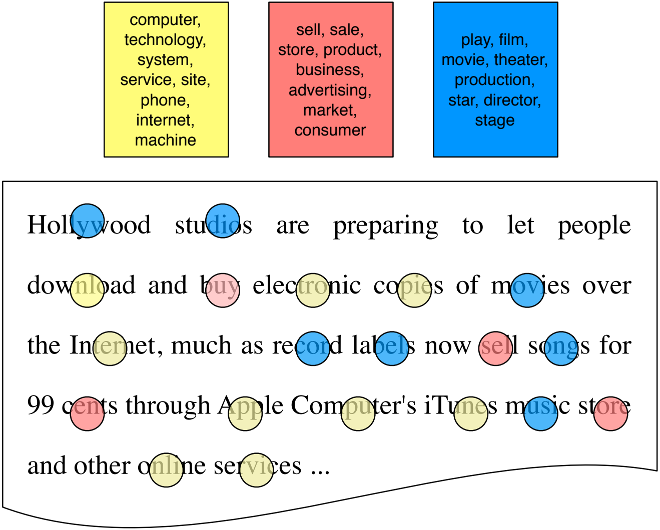
TOPIC 3

play, film, movie, theater, production, star, director, stage

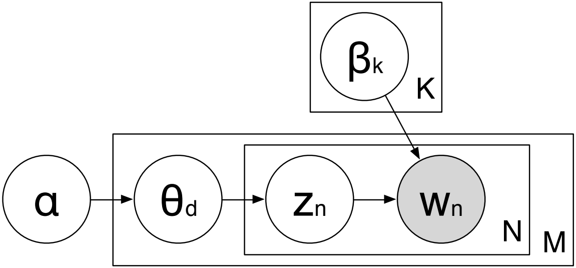
Example: Latent Dirichlet Allocation



Example: Latent Dirichlet Allocation

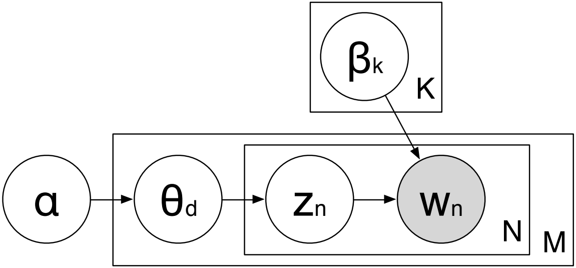


LDA Generative Model



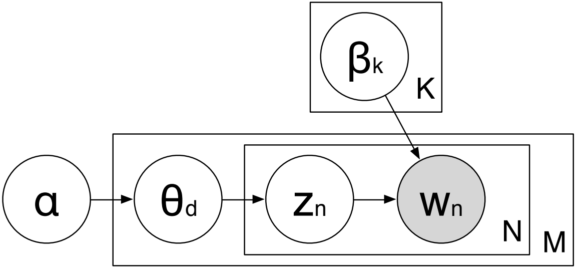
- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k

LDA Generative Model



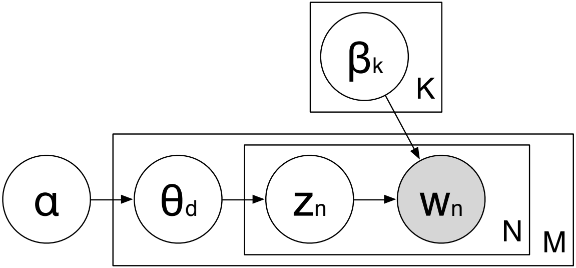
- For each topic $k \in \{1, \ldots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

LDA Generative Model



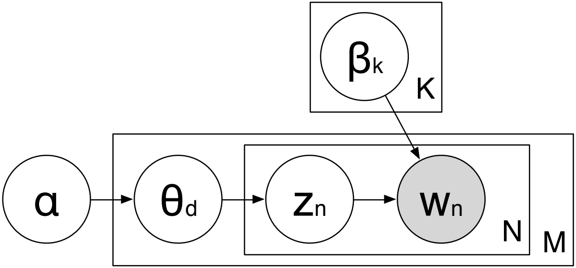
- For each topic $k \in \{1, \dots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

LDA Generative Model



- For each topic $k \in \{1, \dots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

LDA Generative Model



- For each topic $k \in \{1, \dots, K\}$, a multinomial distribution β_k
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \,|\, \alpha, \beta) = \prod_d p(\theta_d \,|\, \alpha) \prod_n p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta, z_{d,n}) \quad (7)$$

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \tag{7}$$

- $p(\theta_d \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_k \theta_{d,k}^{\alpha_k - 1}$ (Dirichlet)

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta, z_{d,n}) \quad (7)$$

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

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \,|\, \alpha, \beta) = \prod_d p(\theta_d \,|\, \alpha) \prod_n p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta, z_{d,n}) \quad (7)$$

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

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \,|\, \alpha, \beta) = \prod_d p(\theta_d \,|\, \alpha) \prod_n p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta, z_{d,n}) \quad (7)$$

Variational distribution:

$$q(\theta, z) = q(\theta \,|\, \gamma) q(z \,|\, \phi) \quad (8)$$

Deriving Variational Inference for LDA

Joint distribution:

$$p(\theta, z, w \,|\, \alpha, \beta) = \prod_d p(\theta_d \,|\, \alpha) \prod_n p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta, z_{d,n}) \quad (7)$$

Variational distribution:

$$q(\theta, z) = q(\theta \,|\, \gamma) q(z \,|\, \phi) \quad (8)$$

ELBO:

$$\begin{aligned} L(\gamma, \phi; \alpha, \beta) =& \mathbb{E}_q \left[\log p(\theta \,|\, \alpha)\right] + \mathbb{E}_q \left[\log p(z \,|\, \theta)\right] + \mathbb{E}_q \left[\log p(w \,|\, z, \beta)\right] \\ & - \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log q(z)\right] \end{aligned} \quad (9)$$

What is the variational distribution?

$$q(\vec{\theta}, \vec{z}) = \prod_d q(\theta_d \mid \gamma_d) \prod_n q(z_{d,n} \mid \phi_{d,n}) \quad (10)$$

- 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

Expectation of log Dirichlet

- Most expectations are straightforward to compute
- Dirichlet is harder

$$\mathbb{E}_{\text{dir}}\left[\log p(\theta_i \,|\, \alpha)\right] = \Psi\left(\alpha_i\right) - \Psi\left(\sum_j \alpha_j\right) \tag{11}$$

Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{12}$$

$$\tag{13}$$

Expectation 1

$$\mathbb{E}_q [\log p(\theta \mid \alpha)] = \mathbb{E}_q \left[\log \left\{ \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \right\} \right] \tag{12}$$

$$= \mathbb{E}_q \left[\log \left\{ \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \right\} + \sum_i \log \theta_i^{\alpha_i - 1} \right] \tag{13}$$

Log of products becomes sum of logs.

Expectation 1

$$\mathbb{E}_q\left[\log p(\theta \mid \alpha)\right] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{12}$$

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

$$= \log \Gamma(\sum_i \alpha_i) - \sum_i \log \Gamma(\alpha_i) + \mathbb{E}_q\left[\sum_i (\alpha_i - 1) \log \theta_i\right] \tag{13}$$

Log of exponent becomes product, expectation of constant is constant

Expectation 1

$$\mathbb{E}_q[\log p(\theta \mid \alpha)] = \mathbb{E}_q\left[\log\left\{\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}\right\}\right] \tag{12}$$

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

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

$$= \log\Gamma(\sum_i \alpha_i) - \sum_i \log\Gamma(\alpha_i)$$

$$+ \sum_i (\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right)\right)$$

Expectation of log Dirichlet

Expectation 2

$$\mathbb{E}_q[\log p(z \mid \theta)] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{13}$$

$$\tag{14}$$

Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (13)$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n == i]}\right] \quad (14)$$

$$(15)$$

Products to sums

Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{13}$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{14}$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{15}$$

$$\tag{16}$$

Linearity of expectation

Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \quad (13)$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \quad (14)$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] \quad (15)$$

$$= \sum_n \sum_i \phi_{ni} \mathbb{E}_q\left[\log \theta_i\right] \quad (16)$$

$$(17)$$

Independence of variational distribution, exponents become products

Expectation 2

$$\mathbb{E}_q\left[\log p(z \mid \theta)\right] = \mathbb{E}_q\left[\log \prod_n \prod_i \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{13}$$

$$= \mathbb{E}_q\left[\sum_n \sum_i \log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{14}$$

$$= \sum_n \sum_i \mathbb{E}_q\left[\log \theta_i^{\mathbb{1}[z_n==i]}\right] \tag{15}$$

$$= \sum_n \sum_i \phi_{ni} \mathbb{E}_q\left[\log \theta_i\right] \tag{16}$$

$$= \sum_n \sum_i \phi_{ni} \left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{17}$$

Expectation of log Dirichlet

Expectation 3

$$\mathbb{E}_q \left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q \left[\log \beta_{z_{d,n}, w_{d,n}}\right] \tag{18}$$

$$\tag{19}$$

Expectation 3

$$\mathbb{E}_q\left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] \tag{18}$$

$$= \mathbb{E}_q\left[\log \prod_{v}^{V} \prod_{i}^{K} \beta_{i,v}^{\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]}\right] \tag{19}$$

$$\tag{20}$$

Expectation 3

$$\mathbb{E}_q \left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q \left[\log \beta_{z_{d,n}, w_{d,n}}\right] \quad (18)$$

$$= \mathbb{E}_q \left[\log \prod_v^V \prod_i^K \beta_{i,v}^{\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]}\right] \quad (19)$$

$$= \sum_v^V \sum_i^K \mathbb{E}_q \left[\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]\right] \log \beta_{i,v} \quad (20)$$

$$(21)$$

Expectation 3

$$\mathbb{E}_q\left[\log p(w \mid z, \beta)\right] = \mathbb{E}_q\left[\log \beta_{z_{d,n}, w_{d,n}}\right] \quad (18)$$

$$= \mathbb{E}_q\left[\log \prod_{v}^{V} \prod_{i}^{K} \beta_{i,v}^{\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]}\right] \quad (19)$$

$$= \sum_{v}^{V} \sum_{i}^{K} \mathbb{E}_q\left[\mathbb{1}\left[v = w_{d,n}, z_{d,n} = i\right]\right] \log \beta_{i,v} \quad (20)$$

$$= \sum_{v}^{V} \sum_{i}^{K} \phi_{n,i} w_{d,n}^{v} \log \beta_{i,v} \quad (21)$$

Entropies

Entropy of Dirichlet

$$\mathbb{H}_q\left[\gamma\right] = -\log\Gamma\left(\sum_j \gamma_j\right) + \sum_i \log\Gamma(\gamma_i) - \sum_i (\gamma_i - 1)\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

Entropies

Entropy of Dirichlet

$$\mathbb{H}_q[\gamma] = -\log\Gamma\left(\sum_j \gamma_j\right) + \sum_i \log\Gamma(\gamma_i) - \sum_i (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

Entropy of Multinomial

$$\mathbb{H}_q[\phi_{d,n}] = -\sum_i \phi_{d,n,i}\log\phi_{d,n,i} \tag{22}$$

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 ϕ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ni}} = \Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right) + \log \beta_{i,v} - \log \phi_{ni} - 1 + \lambda \tag{23}$$

Solution:

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_j \gamma_j\right)\right) \tag{24}$$

Update for γ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \phi_{n,i} - \gamma_i\right) - \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)$$

Update for γ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \phi_{n,i} - \gamma_i\right) - \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)$$

Update for γ

Derivative of ELBO:

$$\frac{\partial \mathcal{L}}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \phi_{n,i} - \gamma_i\right) - \Psi'\left(\sum_j \gamma_j\right) \sum_j \left(\alpha_j + \sum_n \phi_{nj} - \gamma_j\right)$$

Solution:

$$\gamma_i = \alpha_i + \sum_n \phi_{ni} \tag{25}$$

Update for β

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

$$\beta_{ij} \propto \sum_d \sum_n \phi_{dni} w_{dn}^j \tag{26}$$

Overall Algorithm

1. Randomly initialize variational parameters (can't be uniform)
2. For each iteration:
   2.1 For each document, update γ and ϕ
   2.2 For corpus, update β
   2.3 Compute $\mathcal{L}$ for diagnostics
3. Return expectation of variational parameters for solution to latent variables

Relationship with Gibbs Sampling

- Gibbs sampling: sample from the conditional distribution of all other variables
- Variational inference: each factor is set to the exponentiated log of the conditional
- Variational is easier to parallelize, Gibbs faster per step
- Gibbs typically easier to implement

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first ... then match implementation

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first . . . then match implementation
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first . . . then match implementation
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)
- Write the state (checkpointing and debugging)
- Visualize variational parameters

Implementation Tips

- Match derivation exactly at first
- Randomize initialization, but specify seed
- Use simple languages first ... then match implementation
- Try to match variables with paper
- Write unit tests for each atomic update
- Monitor variational bound (with asserts)
- Write the state (checkpointing and debugging)
- Visualize variational parameters
- Cache / memoize gamma / digamma functions

Next class

- Example on toy LDA problem
- Current research in variational inference