Variational Inference for LDA

Zhao Zhou

The Hong Kong University of Science and Technology
Outline

1. Generative Process of LDA
2. Exponential Family
3. Newton Method
4. Variational Inference
   - E-Step
   - M-Step
5. Conclusion
Notations and terminology

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by \{1, \ldots, V\}.

- A *document* is a sequence of \(N\) words denoted by \(\mathbf{w} = (w_1, w_2, \ldots, w_N)\), where \(w_n\) is the \(n\)th word in the sequence.

- A *corpus* is a collection of \(M\) documents denoted by \(D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}\).
Latent Dirichlet allocation

LDA assumes the following generative process for each document \( w \) in a corpus \( D \):

- Choose \( N \sim \text{Poisson}(\xi) \).
- Choose \( \theta \sim \text{Dir}(\alpha) \).
- For each of the \( N \) words \( w_n \):
  - Choose a topic \( z_n \sim \text{Mult}(\theta) \).
  - Choose a word \( w_n \) from \( p(w_n|z_n, \beta) \).

Note that:

- The dimension of the Dirichlet distribution (topic variable) is known and fixed.
- The word probabilities are parameterized by a \( k \times V \) matrix \( \beta \) where \( \beta_{ij} = p(w^j = 1|z^i = 1) \).
- The randomness of \( N \) is ignored in subsequent slides.
Given the parameters $\alpha$ and $\beta$, the joint distribution of a topic mixture $\theta$, a set of topics $z$, and a set of $N$ words $w$ is:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

The marginal distribution of a document is:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d\theta.$$
An exponential family distribution has the form

\[ p(x|\eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\} \]

The different parts of this equation are

- The natural parameter \( \eta \)
- The sufficient statistic \( t(x) \)
- The underlying measure \( h(x) \)
- The log normalizer \( a(\eta) \)

\[ a(\eta) = \log \int h(x) \exp\{\eta^T t(x)\} \]
The derivatives of the log normalizer gives the moments of the sufficient statistics

\[
\frac{d}{d\eta} a(\eta) = \frac{d}{d\eta} \left( \log \int \exp\{\eta^T t(x)\} h(x) dx \right)
= \int t(x) \exp\{\eta^T t(x)\} h(x) dx 
= \frac{\int \exp\{\eta^T t(x)\} h(x) dx}{\int \exp\{\eta^T t(x)\} h(x) dx}
= \int t(x) \exp\{\eta^T t(x) - a(\eta)\} h(x) dx
= E[t(X)]
\]
Computing $E[\log(\theta|\alpha)]$

- The Dirichlet distribution $p(\theta|\alpha)$:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\sum_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i-1}$$

$$= \exp\left\{(\sum_{i=1}^{K} (\alpha_i - 1) \log \theta_i) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)\right\}$$

- Sufficient statistics: $\log \theta_i$.

- Log normalizer: $\sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i)$
Computing $E[\log(\theta|\alpha)]$

- The expectation $E[\log(\theta|\alpha)]$ is:

$$E[\log \theta_i|\alpha] = a(\alpha)' = (\sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i))'$$

$$= \psi(\alpha_i) - \psi(\sum_{j=1}^{K} \alpha_j).$$

where $\psi$ is the digamma function, the first derivative of the log Gamma function.
Unconstrained minimization

- Suppose $f$ convex, twice continuously differentiable.
- Assume optimal value $p^* = \inf_x f(x)$ is attained.
- Interpreted as iterative methods for solving optimality condition

$$\nabla f(x^*) = 0.$$
Newton Step

- Newton step:
  \[ \Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x). \]

- Interpretations:
  - \( x + \Delta x_{nt} \) minimizes second order approximation
    \[ \hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v \]
  - \( x + \Delta x_{nt} \) solves linearized optimality condition.
    \[ \nabla f(x + v) \approx \nabla \hat{f}(x + v) = \nabla f(x) + \nabla^2 f(x) v = 0. \]
Newton decrement

- A measure of the proximity of $x$ to $x^*$

\[ \lambda(x) = \left( \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \right)^{1/2} \]

- equal to the norm of the Newton step in the quadratic Hessian norm

\[ \lambda(x) = (\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt})^{1/2} \]
Backtracking Line Search

- **Exact Line search:**

\[
 t = \arg \min_{t > 0} f(x + t\Delta x)
\]

- **Backtracking line search (with parameters \(\alpha \in (0, 1/2), \beta \in (0, 1)\)):**
  - Starting at \(t = 1\), repeat \(t = \beta t\) until
  
  \[
  f(x + t\Delta x) < f(x) + \alpha t\nabla f(x)^T \Delta x.
  \]
Newton Method

Repeat

• Compute the Newton step and decrement.

\[ \Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x) \]

\[ \lambda^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \]

• Stopping criterion. quit if \( \frac{\lambda^2}{2} \leq \epsilon \).

• Line search. Choose step size \( t \) by backtracking line search.

• Update \( x = x + t \nabla x_{nt} \).
Inference

- The posterior distribution of hidden variable:

\[
p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}
\]

- This distribution is intractable to compute since

\[
p(w|\alpha, \beta) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \int (\prod_{i=1}^k \theta_i^{\alpha_i-1}) (\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_{ijn}}) d\theta
\]

due to the coupling between $\theta$ and $\beta$. 
The variational distribution on latent variables:

\[ q(\theta, z | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^{N} q(z_n | \phi_n). \]

An optimization problem that determines the values of \( \gamma \) and \( \phi \) with respect to KL-Divergence \( D \):

\[ (\gamma^*, \phi^*) = \arg \min_{\gamma, \phi} D(q(\theta, z | \gamma, \phi) \| p(\theta, z | w, \alpha, \beta)) \]
Now, we denote $q(\theta, z | \gamma, \phi)$ by $q$.

The KL-Divergence between $q$ and $p(\theta, z | w, \alpha, \beta)$ is

$$D(q||p) = E_q[\log q] - E_q[\log p(\theta, z | w, \alpha, \beta)]$$

$$= E_q[\log q] - E_q[\log p(\theta, z, w | \alpha, \beta)] + \log p(w | \alpha, \beta)$$

Using Jensen’s inequality, we bound $p(w | \alpha, \beta)$ by

$$\log p(w | \alpha, \beta) = \log \int \sum_z p(\theta, z, w | \alpha, \beta) d\theta$$

$$= \log \int \sum_z p(\theta, z, w | \alpha, \beta) \frac{q(\theta, z)}{q(\theta, z)} d\theta$$

$$\geq \int \sum_z q(\theta, z) \log \frac{p(\theta, z, w | \alpha, \beta)}{q(\theta, z)} d\theta$$

$$= E_q[\log p(w, z, w | \alpha, \beta)] - E_q[\log q(\theta, z)].$$
KL-Divergence

- We denote $E_q[\log p(w, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)]$ by $L(\gamma, \phi; \alpha, \beta)$.
- Then we have

$$\log p(w|\alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)).$$

- Maximizing the lower bound $L(\gamma, \phi; \alpha, \beta)$ with respect to $\gamma$ and $\phi$ is equivalent to minimizing the KL-Divergence between the variational posterior probability and the true posterior probability.
Variational Inference

- Expand $L(\gamma, \phi; \alpha, \beta)$ using the factorizations of $p$ and $q$:

$$L(\gamma, \phi; \alpha, \beta) = E_q[\log p(w, z, w|\alpha, \beta)] - E_q[\log q(\theta, z)]$$

$$= E_q[\log p(\theta|\alpha)] + E_q[\log p(z|\theta)] + E_q[\log p(w|z, \beta)] - E_q[\log q(\theta)] - E_q[\log q(z)]$$

- Compute the five terms, respectively.
Computing $E_q[\log p(\theta|\alpha)]$

- $E_q[\log p(\theta|\alpha)]$ is given by

$$E_q[\log p(\theta|\alpha)] = \sum_{i=1}^{K} (\alpha_i - 1) E_q[\log \theta_i]$$

$$+ \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i).$$

- $\theta$ is generated by $\text{Dir}(\theta|\gamma)$: $E_q[\log \theta_i] = \psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)$.

- Then we have:

$$E_q[\log p(\theta|\alpha)] = \sum_{i=1}^{K} (\alpha_i - 1) \psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)$$

$$+ \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i).$$
Computing $E_q[\log p(z|\theta)]$

$E_q[\log p(z|\theta)]$ is given by

$$E_q[\log p(z|\theta)] = E_q\left[\sum_{n=1}^{N} \sum_{i=1}^{K} z_{ni} \log \theta_i \right]$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} E_q[z_{ni}] E_q[\log \theta_i]$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni} (\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j))$$

where $z$ is generated from $Mult(z|\phi)$ and $\theta$ is generated from $Dir(\theta|\gamma)$. 
Computing $E_q[\log p(w|z, \beta)]$

$E_q[\log p(w|z, \beta)]$ is given by

$$E_q[\log p(w|z, \beta)] = E_q[\sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} z_{ni} w_n^j \log \beta_{ij}]$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} E_q[z_{ni}] w_n^j \log \beta_{ij}$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij}$$
Computing $E_q[\log q(\theta|\gamma)]$

$E_q[\log q(\theta|\gamma)]$ is given by

$$
E_q[\log p(\theta|\gamma)] = \sum_{i=1}^{k} (\gamma_i - 1)E_q[\log \theta_i] + \log \Gamma(\sum_{i=1}^{k} \gamma_i) - \sum_{i=1}^{k} \log \Gamma(\gamma_i)
$$

Then, we have

$$
E_q[\log p(\theta|\gamma)] = \log \Gamma(\sum_{i=1}^{k} \gamma_i) - \sum_{i=1}^{k} \log \Gamma(\gamma_i)
$$

$$
+ \sum_{i=1}^{k} (\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j))
$$
Computing $E_q[\log q(z|\phi)]$

$E_q[\log q(z|\phi)]$ is given by

\[
E_q[\log q(z|\phi)] = E_q\left[\sum_{n=1}^{N} \sum_{i=1}^{k} z_{ni} \log \phi_{ni}\right]
\]

\[
= \sum_{n=1}^{N} \sum_{i=1}^{k} E_q[z_{ni}] \log \phi_{ni}
\]

\[
= \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}
\]
Finally, $L(\gamma, \phi; \alpha, \beta)$ is

\[
(\gamma, \phi; \alpha, \beta) = \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)
\]
\[
+ \sum_{i=1}^{K} (\alpha_i - 1)(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right))
\]
\[
+ \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} w_{nj} \log \beta_{ij}
\]
\[
- \log \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) - \sum_{i=1}^{K} \log \Gamma(\gamma_i) + \sum_{i=1}^{K} (\gamma_i - 1)(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{K} \gamma_j\right))
\]
\[
- \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni}.
\]
Variational Multinomial

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to $\phi_{ni}$:

$$L_{\phi_{ni}} = \phi_{ni}(\psi(\gamma_{i}) - \psi(\sum_{j=1}^{K} \gamma_{j})) + \phi_{ni} \log \beta_{iv}$$

$$- \phi_{ni} \log \phi_{ni} + \lambda(\sum_{j=1}^{K} \phi_{ni} - 1).$$
Taking derivatives with respect to $\phi_{ni}$:

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

Setting this derivative to zero yields

$$\phi_{ni} \propto \beta_{iv} \exp(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)).$$
Variational Dirichlet

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to $\gamma_i$:

$$L_\gamma = \sum_{i=1}^{K} (\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j))(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i)$$

$$- \log \Gamma(\sum_{j=1}^{K} \gamma_j) + \sum_{i=1}^{K} \log \Gamma(\gamma_i)$$
Variational Dirichlet

- Taking the derivative with respect to $\gamma_i$

$$\frac{\partial L}{\partial \gamma_i} = \psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \psi''(\sum_{j=1}^{K} \gamma_j) \sum_{j=1}^{K} (\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j)$$

- Setting this equation to zero yields:

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

1. initialize $\phi_{ni}^0 = \frac{1}{K}$ for all $i$ and $n$.
2. initialize $\gamma_i = \alpha_i + \frac{N}{K}$ for all $i$
3. repeat
4. for $n = 1$ to $N$
5. for $i = 1$ to $K$
   1. $\phi_{ni}^{t+1} = \beta_{iw_n} \exp(\psi(\gamma_i^t))$.
   2. normalize $\phi_{n}^{t+1}$ to sum 1.
6. $\gamma^{t+1} = \alpha + \sum_{n=1}^{N} \phi_{n}^{t+1}$
7. until convergence
Parameter Estimation

- In the variational E-step, maximize the lower bound $L(\gamma, \phi; \alpha, \beta)$ with respect to the variational parameters $\gamma$ and $\phi$.
- In the M-step, maximize the bound with respect to the model parameters $\alpha$ and $\beta$. 
Conditional Multinomials

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to $\beta$:

$$L_\beta = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{dni} w^j_{dn} \log \beta_{ij} + \sum_{i=1}^{K} \lambda_i (\sum_{j=1}^{V} \beta_{ij} - 1).$$

- Taking the derivative with respect to $\beta_{ij}$ and setting it to zero:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w^j_{dn}.$$
Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to $\alpha$:

$$L_\alpha = \sum_{d=1}^{M} \left( \log \Gamma \left( \sum_{j=1}^{K} \alpha_j \right) - \sum_{i=1}^{K} \log \Gamma (\alpha_i) \right)$$

$$+ \sum_{i=1}^{K} \left( (\alpha_i - 1) \left( \psi(\gamma_{di}) - \psi \left( \sum_{j=1}^{K} \gamma_{dj} \right) \right) \right)$$

Taking the derivative with respect to $\alpha_i$

$$\frac{\partial L}{\partial \alpha_i} = M \left( \psi \left( \sum_{j=1}^{K} \alpha_j \right) - \psi (\alpha_i) \right) + \sum_{d=1}^{M} \left( \psi(\gamma_{di}) - \psi \left( \sum_{j=1}^{K} \gamma_{dj} \right) \right).$$

It is difficult to compute $\alpha_i$ by setting the derivative to zero.
Newton Method

- Compute the Hessian Matrix by

\[ \frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j} = M(\psi'(\sum_{j=1}^{K} \alpha_j) - \delta(i, j)\psi'(\alpha_i)). \]

- Input this Hessian Matrix and the derivative to Newton Method.
Conclusion

- Variational Inference is used for approximating intractable integrals arising in Bayesian network.
- Variational Inference can be seen as an extension of the EM algorithm which computes the entire posterior distribution of latent variables.
- Usually, the derived "best" variational distribution is the same family as the corresponding prior distribution over the variable.
- A good template proof for variational inference on other topic models.