10-708 PGM (Spring 2020): Homework 2

Andrew ID:	[your Andrew ID]
Name:	[your first and last name]
Collaborators:	[Andrew IDs of all collaborators, if any]

1 Variational Inference [40 points] (Junxian)

In this problem, we are going to work with approximate posterior inference via variational inference for a given topic model.

The standard Latent Dirichlet Allocation model only models the word co-occurrences, without considering temporal information, i.e. the time when a document is generated. However, a large number of subjects in documents change dramatically over time. It is important to interpret the topics in the context of the timestamps of the documents. To address how topics occur and shift over time, Topics on Time (TOT) model was proposed, by explicitly modeling of time jointly with word co-occurrence patterns (Wang and McCallum, 2006). The model is shown in Figure 1.

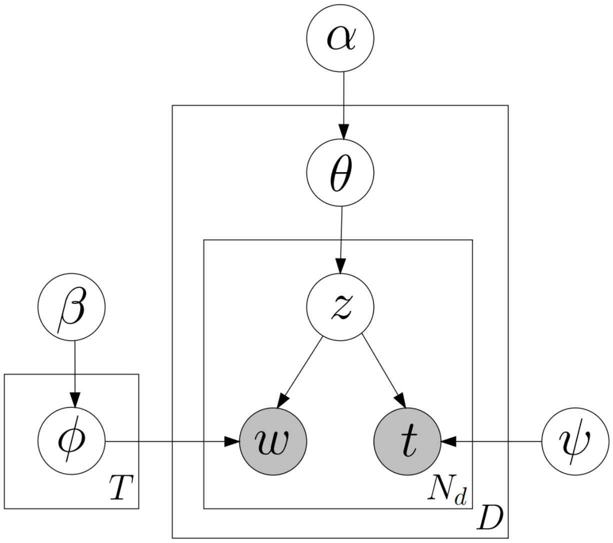

Figure 1: TOT Model

In the model, there are D documents. Each document d contains N_d words $w_{d1}, w_{d2}, ..., w_{dN_d}$. Each word w_{di} has a timestamp $t_{di} \in (0, 1)$, indicating when the document is generated in a relative time scale $(0, 1)$. All words in the same document have the same timestamp. There are K topics (also $T = K$ topics for the notation in the paper and Figure 1) in the document corpora. Each topic follows a multinomial distribution ϕ over the V words in the vocabulary. Each document follows a multinomial distribution θ over the K topics. The prior distribution for ϕ and θ are Dirichlet distributions with parameters β and α respectively. For each topic k, the temporal occurrence follows a Beta distribution $Beta(\psi_{k1}, \psi_{k2})$, where $\psi_k = (\psi_{k1}, \psi_{k2})$ and we use $\boldsymbol{\psi} \in \mathbb{R}_{+}^{K\times 2}$ to denote ψ_k for all topics. Each word w_{di} and its timestamp t_{di} are assumed to be generated from a topic, with a topic label $z_{di} \in \{1, ..., K\}$.

The generative process of this model is described as follows.

1. Draw K multinomials ϕ_k from a Dirichlet prior β, one for each topic k.
2. For each document d,
   - Draw a multinomial θ_d from a Dirichlet prior α;
   - For each word w_{di} in document d,
     (a) Sample a topic z_{di} from multinomial θ_d;
     (b) Sample a word w_{di} from multinomial $\phi_{z_{di}}$;
     (c) Sample a timestamp t_{di} from Beta $\psi_{z_{di}}$.

We use variational EM to approximate the posterior of latent variables and learn model parameters. To do this, a mean field variational distribution needs to be defined, which is parameterized by some parameters called variational parameters. The variational EM algorithm iteratively performs two steps: 1) in the E step, variational parameters are updated; 2) in the M step, model parameters are optimized. Same as in the paper, we consider α and β are predefined fixed hyperparameters with no need to update. Therefore, in M step, only the other model parameters are optimized. The pseudo-code for the proposed algorithm is shown in Algorithm 1.

Algorithm 1 Pseudo-code of variational EM algorithm for TOT model

```
Input: Observations, Topic number $K$, MaxIter, and other optional parameters
Output: Posterior distributions for latent variables, optimized model parameters
Initialize parameters;
Compute and record ELBO with initial parameters
for $k \leftarrow 1$ to $MaxIter$ do
    Update variational variables                                ▷ Stage 1: E-Step
    Update $\boldsymbol{\psi}$ with projected Newton method      ▷ Stage 2: M-Step
    Compute and record ELBO
end for
```

In the TOT model, $\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\phi}$ are latent variables and $\boldsymbol{\psi}$ is the model parameter to be learned. As a start, we use mean-field variational inference and the variational distribution has the form:

$$q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \prod_{k=1}^{K} q(\phi_k|\lambda_k) \prod_{d=1}^{D} \left[q(\theta_d|\gamma_d) \prod_{n=1}^{N_d} q(z_{dn}|\pi_{dn})\right], \tag{1}$$

where $\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\pi}$ are variational parameters that need to be updated in E-step.

Next, we write out the joint distribution of latent and observed variables:

$$p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}|\alpha, \beta, \boldsymbol{\psi}) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} [p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(x_{dn}|z_{dn}, \boldsymbol{\phi}) p(t_{dn}|z_{dn}, \boldsymbol{\psi})] \tag{2}$$

Given Eq. 1 and 2, we can write out ELBO with:

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z})}[\log p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z}) - \log q(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{z})]. \tag{3}$$

Variational EM basically maximizes ELBO w.r.t. variational parameters and model parameters in E- and M-step respectively.

Questions:

1. **[10 points]** Update variational parameters
Derive the update equations of variational parameters, and also specify their distributions. Here you can directly use the conclusion below for the derivation.

$$q_j^* \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{t}, \boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{z})]\},$$

where $\mathbb{E}_{q_{-j}}$ denotes expectation over all latent variables excluding variable j.

Solution

We denote the vocabulary size as V.
$q(\boldsymbol{\theta})$:

$$\begin{aligned} q(\theta_d) &\propto \exp\Big\{\mathbb{E}_{q_{-\theta_d}}\big[\log p(\theta_d|\alpha) + \sum_n \log p(z_{dn}|\theta_d)\big]\Big\} \\ &\propto \exp\Big\{\mathbb{E}_{q_{-\theta_d}}\big[(\alpha-1)\sum_k \log\theta_{dk} + \sum_n\sum_k \mathbf{I}(z_{dn}=k)\log\theta_{dk}\big]\Big\} \\ &\propto \exp\Big\{(\alpha-1)\log\prod_{k=1}^{K}\theta_{dk} + \sum_n\sum_k q(z_{dn}=k)\log\theta_{dk}\Big\} \\ &\propto \prod_{k=1}^{K}\theta_{dk}^{\sum_n q(z_{dn}=k)+\alpha-1}. \end{aligned} \tag{4}$$

Therefore, $q(\theta_d) = Dir(\cdot|\boldsymbol{\gamma})$, where $\gamma_k = \sum_n q(z_{dn}=k)+\alpha$, we further define $\pi_{dnk} = q(z_{dn}=k)$, and $\gamma_k = \sum_n \pi_{dnk} + \alpha$.
$q(\boldsymbol{\phi})$:

$$\begin{aligned} q(\phi_k) &\propto \exp\Big\{\mathbb{E}_{q_{-\phi_k}}\big[\sum_{d,n}\log p(w_{dn}|z_{dn},\boldsymbol{\phi}) + \log p(\phi_k|\beta)\big]\Big\} \\ &\propto \exp\Big\{\mathbb{E}_{q_{-\phi_k}}\big[\sum_{d,n}\sum_{v} q(z_{dn}=k)\mathbf{I}(w_{dn}=v)\log\phi_{kv} + (\beta-1)\log\prod_{v=1}^{V}\phi_{kv}\big]\Big\} \\ &\propto \prod_{v=1}^{V}\phi_{kv}^{\sum_{d,n} q(z_{dn}=k)\mathbf{I}(w_{dn}=v)+\beta-1}. \end{aligned} \tag{5}$$

Therefore, $q(\phi_k) = Dir(\cdot|\lambda_k)$, where $\lambda_{kv} = \beta + \sum_{d,n}\mathbf{I}(w_{dn}=v)\pi_{dnk}$.
$q(\mathbf{z})$:

$$\begin{aligned} q(z_{dn}=k) &\propto \exp\Big\{\mathbb{E}_{q_{-z_{dn}}}\big[\log p(z_{dn}=k|\theta_d) + \log p(w_{dn}|z_{dn}=k,\phi_k) + \log p(t_{dn}|z_{dn}=k,\psi_k)\big]\Big\} \\ &\propto \exp\Big\{\mathbb{E}_{q_{-z_{dn}}}\big[\log\theta_{dk} + \log\phi_{kv} + \log(\frac{t_{dn}^{\psi_{k1}-1}(1-t_{dn})^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})})\big]\Big\} \\ &\propto \frac{t_{dn}^{\psi_{k1}-1}(1-t_{dn})^{\psi_{k2}-1}}{B(\psi_{k1},\psi_{k2})}\exp\big(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di}) + \Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^{V}\lambda_{ki})\big), \end{aligned} \tag{6}$$

where $v = w_{dn}$, $\Psi(\cdot)$ is the digamma function, B represents Beta function.

2. **[10 points]**Update model parameters
Derive the update equations of model parameters, as mentioned before, there is no need to update $\boldsymbol{\alpha}$

and $\boldsymbol{\beta}$. For the updating rule of $\boldsymbol{\psi}$, please be careful that $\boldsymbol{\psi}$ should be constrained as positive. Hint: For a problem with positive solution ($x > 0$), a projected Newton method could be applied:

$$y = (x - H^{-1}g)_+$$

$$x^+ = x + \lambda(y - x)$$

where x is the current variable, y is the projected update, g and H are gradient and Hessian matrix respectively, λ is the step size and x^+ is the updated variable. $(\cdot)_+$ is defined as $s_+ := \max(0, s)$.

Solution

To compute the Hessian matrix and gradient of ψ_k, we first write out the terms in ELBO that contains ψ_k:

$$\begin{aligned}
\mathrm{ELBO}(\psi_k) &= \mathbb{E}_q\Big\{\sum_{d,n}\mathbf{I}(z_{dn}=k)\log p(t_{dn}|z_{dn}=k,\psi_k)\Big\} \\
&= \mathbb{E}_q\Big\{\sum_{d,n}\mathbf{I}(z_{dn}=k)\Big[(\psi_{k1}-1)\log t_{dn} + (\psi_{k2}-1)\log(1-t_{dn}) - \log B(\psi_{k1},\psi_{k2})\Big]\Big\} \\
&= \sum_{d,n}\pi_{dnk}(\psi_{k1}-1)\log t_{dn} + \sum_{d,n}\pi_{dnk}(\psi_{k2}-1)\log(1-t_{dn}) - \sum_{d,n}\pi_{dnk}\log B(\psi_{k1},\psi_{k2}),
\end{aligned} \tag{7}$$

then we derive the first-order and second-order derivative with respect to ψ_k:

$$\frac{\partial \mathrm{ELBO}}{\partial \psi_{k1}} = \sum_{d,n}\Big[\pi_{dnk}\log t_{dn} - \pi_{dnk}[\Psi(\psi_{k1}) - \Psi(\psi_{k1}+\psi_{k2})]\Big], \tag{8}$$

$$\frac{\partial \mathrm{ELBO}}{\partial \psi_{k2}} = \sum_{d,n}\Big[\pi_{dnk}\log(1-t_{dn}) - \pi_{dnk}[\Psi(\psi_{k2}) - \Psi(\psi_{k1}+\psi_{k2})]\Big], \tag{9}$$

$$\frac{\partial \mathrm{ELBO}}{\partial \psi_{k1}\psi_{k2}} = \frac{\partial \mathrm{ELBO}}{\partial \psi_{k2}\psi_{k1}} = \sum_{d,n}\pi_{dnk}\Psi_1(\psi_{k1}+\psi_{k2}), \tag{10}$$

$$\frac{\partial \mathrm{ELBO}}{\partial \psi_{k1}^2} = \sum_{d,n} -\pi_{dnk}\big[\Psi_1(\psi_{k1}) - \Psi_1(\psi_{k1}+\psi_{k2})\big], \tag{11}$$

$$\frac{\partial \mathrm{ELBO}}{\partial \psi_{k2}^2} = \sum_{d,n} -\pi_{dnk}\big[\Psi_1(\psi_{k2}) - \Psi_1(\psi_{k1}+\psi_{k2})\big], \tag{12}$$

where $\Psi_1(\cdot)$ is the first-order derivative of digamma function. Eq. (8), (9), (10), (11), (12) compute the Hessian matrix H_{ψ_k} and gradient g_{ψ_k} of ELBO w.r.t. ψ_k. We use projected Newton method to update ψ_k:

$$\begin{aligned}
\psi_k^* &= (\psi_k^{old} - H_{\psi_k}^{-1}g_{\psi_k})_+ \\
\psi_k^{new} &= \psi_k^{old} + \ell\cdot(\psi_k^* - \psi_k^{old}),
\end{aligned} \tag{13}$$

where ℓ is the step size.

3. **[20 points]** Detailed variational lower bound
Based on the variational distributions, expand Eq. 3 to obtain detailed variational lower bound. The result should be as specific as possible, that is, it can be directly used in the implementation.

Solution

We next expand the terms in Eq. (3) one by one to compute detailed variational lower bound:

$$\mathbb{E}_q\big[\sum_d \log p(\theta_d|\alpha)\big] = \sum_d \Big[\log\Gamma(K\alpha) - K\log\Gamma(\alpha) + \sum_k (\alpha-1)(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di}))\Big], \tag{14}$$

$$\mathbb{E}_q\big[\sum_k \log p(\phi_k|\beta)\big] = \sum_k \Big[\log\Gamma(V\beta) - V\log\Gamma(\beta) + \sum_v (\beta-1)(\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^{V}\lambda_{ki}))\Big], \tag{15}$$

$$\begin{aligned}\sum_{d,n}\mathbb{E}_q\big[\log p(z_{dn}|\theta_d)\big] &= \sum_{d,n}\mathbb{E}_q\big[\sum_k \mathbf{I}(z_{dn}=k)\log\theta_{dk}\big]\\ &= \sum_{d,n,k}\pi_{dnk}\big(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di})\big),\end{aligned} \tag{16}$$

$$\begin{aligned}\sum_{d,n}\mathbb{E}_q\big[\log p(w_{dn}|z_{dn},\boldsymbol{\phi})\big] &= \sum_{d,n}\mathbb{E}_q\big[\sum_{k,v}\mathbf{I}(w_{dn}=v)\mathbf{I}(z_{dn}=k)\log\phi_{kv}\big]\\ &= \sum_{d,n,k,v}\pi_{dnk}\mathbf{I}(w_{dn}=v)\big[\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^{V}\lambda_{ki})\big],\end{aligned} \tag{17}$$

According to Eq. (7), we can directly write out $\sum_{d,n}\mathbb{E}_q\big[\log p(t_{dn}|z_{dn},\boldsymbol{\psi})\big]$:

$$\begin{aligned}\sum_{d,n}\mathbb{E}_q\big[\log p(t_{dn}|z_{dn},\boldsymbol{\psi})\big] = &\sum_{d,n,k}\pi_{dnk}(\psi_{k1}-1)\log t_{dn} + \sum_{d,n,k}\pi_{dnk}(\psi_{k2}-1)\log(1-t_{dn})\\ &- \sum_{d,n,k}\pi_{dnk}\log B(\psi_{k1},\psi_{k2}),\end{aligned} \tag{18}$$

then,

$$\sum_d \mathbb{E}_q[\log q(\theta_d)] = \sum_d\Big[\log\Gamma(\sum_k\gamma_{dk}) - \sum_k\log\Gamma(\gamma_{dk}) + \sum_k(\gamma_{dk}-1)(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di}))\Big], \tag{19}$$

$$\sum_{d,n}\mathbb{E}_q[\log q(z_{dn})] = \sum_{d,n,k}\pi_{dnk}\log\pi_{dnk}, \tag{20}$$

$$\sum_k \mathbb{E}_q[\log q(\phi_k)] = \sum_k\Big[\log\Gamma(\sum_v\lambda_{kv}) - \sum_v\log\Gamma(\lambda_{kv}) + \sum_v(\lambda_{kv}-1)(\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^{V}\lambda_{ki}))\Big], \tag{21}$$

Combine the equations above together, we can write the ELBO as:

$$\begin{aligned}
\text{ELBO} = &\sum_d \Big[\log\Gamma(K\alpha) - K\log\Gamma(\alpha) + \sum_k (\alpha-1)(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^K \gamma_{di}))\Big] \\
&+ \sum_k \Big[\log\Gamma(V\beta) - V\log\Gamma(\beta) + \sum_v (\beta-1)(\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^V \lambda_{ki}))\Big] \\
&+ \sum_{d,n,k} \pi_{dnk}\big(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^K \gamma_{di})\big) \\
&+ \sum_{d,n,k,v} \pi_{dnk}\mathbf{I}(w_{dn}=v)\big[\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^V \lambda_{ki})\big] \\
&+ \sum_{d,n,k} \pi_{dnk}(\psi_{k1}-1)\log t_{dn} + \sum_{d,n,k} \pi_{dnk}(\psi_{k2}-1)\log(1-t_{dn}) - \sum_{d,n,k} \pi_{dnk}\log B(\psi_{k1},\psi_{k2}) \\
&- \sum_d \Big[\log\Gamma(\sum_k \gamma_{dk}) - \sum_k \log\Gamma(\gamma_{dk}) + \sum_k (\gamma_{dk}-1)(\Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^K \gamma_{di}))\Big] \\
&- \sum_{d,n,k} \pi_{dnk}\log\pi_{dnk} \\
&- \sum_k \Big[\log\Gamma(\sum_v \lambda_{kv}) - \sum_v \log\Gamma(\lambda_{kv}) + \sum_v (\lambda_{kv}-1)(\Psi(\lambda_{kv}) - \Psi(\sum_{i=1}^V \lambda_{ki}))\Big]
\end{aligned} \tag{22}$$

Hint: the problem is designed based on the paper (Wang and McCallum, 2006). In the paper, Gibbs sampling was used for posterior inference, and here we are working with variational inference. You may gain better understanding of the model and get some ideas of how to solve the problem by reading the paper.