Expectation-Propagation for Latent Dirichlet Allocation

Tomonari MASADA @ Nagasaki University

November 22, 2018

1

This manuscript contains a derivation of the update formulae of expectation-propagation (EP) presented in the following paper:

> Thomas Minka and John Lafferty.
> Expectation-Propagation for the Generative Aspect Model.
> in Proc. of the 18th Conference on Uncertainty in Artificial Intelligence, pp. 352-359, 2002.[1]

2

A joint distribution for document d in LDA can be written as below.

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta}_d|\boldsymbol{\alpha}) &= p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})p(\mathbf{x}|\boldsymbol{\theta}_d) \\ &= p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{i=1}^{n_d} p(x_{di}|\boldsymbol{\theta}_d) = p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{i=1}^{n_d} \Big(\sum_k \theta_{dk}\phi_{kx_{di}} \Big) \\ &= p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_w \Big(\sum_k \theta_{dk}\phi_{kw} \Big)^{n_{dw}} , \end{aligned} \tag{1}$$

where $p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})$ is a Dirichlet prior distribution.

We approximate $p(w|\boldsymbol{\theta}_d) = \sum_k \theta_{dk}\phi_{kw}$ by $t_w(\boldsymbol{\theta}_d) = s_w \prod_k \theta_{dk}^{\beta_{wk}}$.

Then we obtain an approximated joint distribution as follows:

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta}_d|\boldsymbol{\alpha}) \approx p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_w t_w(\boldsymbol{\theta}_d)^{n_{dw}} &= p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_w \Big(s_w \prod_k \theta_{dk}^{\beta_{wk}} \Big)^{n_{dw}} \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_w s_w^{n_{dw}} \prod_k \theta_{dk}^{\alpha_k - 1 + \sum_w \beta_{wk} n_{dw}} \end{aligned} \tag{2}$$

Let $\gamma_{dk} = \alpha_k + \sum_w \beta_{wk} n_{dw}$. That is,

$$p(\mathbf{x}_d, \boldsymbol{\theta}_d|\boldsymbol{\alpha}) \approx \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_w s_w^{n_{dw}} \prod_k \theta_{dk}^{\gamma_{dk}-1} . \tag{3}$$

An approximated posterior $q(\boldsymbol{\theta}_d)$ can be obtained as follows:

$$\begin{aligned} q(\boldsymbol{\theta}_d) &= \frac{\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_w s_w^{n_{dw}} \prod_k \theta_{dk}^{\gamma_{dk}-1}}{\int \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_w s_w^{n_{dw}} \prod_k \theta_{dk}^{\gamma_{dk}-1} d\boldsymbol{\theta}_d} \\ &= \frac{\prod_k \theta_{dk}^{\gamma_{dk}-1}}{\int \prod_k \theta_{dk}^{\gamma_{dk}-1} d\boldsymbol{\theta}_d} \\ &= \frac{\Gamma(\sum_k \gamma_{dk})}{\prod_k \Gamma(\gamma_{dk})} \prod_k \theta_{dk}^{\gamma_{dk}-1} . \end{aligned} \tag{4}$$

[1] http://research.microsoft.com/en-us/um/people/minka/papers/aspect/

3

For each word token in document d, we remove its contribution to $\prod_w t_w(\boldsymbol{\theta}_d)^{n_{dw}} = \prod_w \left(s_w \prod_k \theta_{dk}^{\beta_{wk}}\right)^{n_{dw}}$.

When the token is a token of word w, this corresponds to a division by $s_w \prod_k \theta_{dk}^{\beta_{wk}}$.

Note that we here consider a removal of word token, not a removal of word type.

Then we obtain an 'old' posterior after this division as follows:

$$q^{\backslash w}(\boldsymbol{\theta}_d) = \frac{\Gamma\big(\sum_k(\gamma_{dk} - \beta_{wk})\big)}{\prod_k \Gamma(\gamma_{dk} - \beta_{wk})} \prod_k \theta_{dk}^{\gamma_{dk} - \beta_{wk} - 1} = \frac{\Gamma(\sum_k \gamma_{dk}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} , \tag{5}$$

where $\gamma_{dk}^{\backslash w} \equiv \gamma_{dk} - \beta_{wk}$.

By combining $p(w|\boldsymbol{\theta}_d) = \sum_k \theta_{dk}\Phi_{kw}$ with $q^{\backslash w}(\boldsymbol{\theta}_d)$, we obtain a something similar to joint distribution as follows:

$$(\sum_k \theta_{dk}\phi_{kw}) \cdot \frac{\Gamma(\sum_k \gamma_{dk}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} .$$

Therefore, by integrating $\boldsymbol{\theta}_d$ out, we obtain a something similar to evidence as follows:

$$\int \sum_k \theta_{dk}\phi_{kw} \frac{\Gamma(\sum_k \gamma_{dk}^{\backslash w})}{\prod_{k'} \Gamma(\gamma_{dk'}^{\backslash w})} \prod_{k'} \theta_{dk'}^{\gamma_{dk'}^{\backslash w} - 1} d\boldsymbol{\theta}_d = \frac{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}}{\sum_k \gamma_{dk}^{\backslash w}} .$$

Consequently, we obtain a new posterior as follows:

$$\begin{aligned}
\tilde{q}(\boldsymbol{\theta}_d) &= \frac{\sum_k \gamma_{dk}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big(\sum_k \theta_{dk}\phi_{kw}\Big) \frac{\Gamma(\sum_k \gamma_{dk}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} \\
&= \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big(\sum_k \theta_{dk}\phi_{kw}\Big) \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} ,
\end{aligned} \tag{6}$$

where $\hat{\gamma_d}^{\backslash w} \equiv \sum_k \gamma_{dk}^{\backslash w}$.

4

Based on Eq. (6), we calculate $E_{\tilde{q}}[\theta_{dk}]$ and $E_{\tilde{q}}[\theta_{dk}^2]$.

$$\begin{aligned}
E_{\tilde{q}}[\theta_{dk}] &= \int \theta_{dk}\tilde{q}(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \\
&= \int \theta_{dk} \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big(\sum_k \theta_{dk}\phi_{kw}\Big) \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d \\
&= \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \int \theta_{dk} \Big(\sum_k \theta_{dk}\phi_{kw}\Big) \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d \\
&= \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \Big\{ \int \theta_{dk}^2 \phi_{kw} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d + \sum_{k' \neq k} \int \theta_{dk}\theta_{dk'}\phi_{k'w} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d \Big\} \\
&= \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \Big\{ \phi_{kw} \int \theta_{dk}^2 \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d + \sum_{k' \neq k} \phi_{k'w} \int \theta_{dk}\theta_{dk'} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w} - 1} d\boldsymbol{\theta}_d \Big\} \\
&= \frac{\hat{\gamma_d}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\Gamma(\hat{\gamma_d}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \Big\{ \phi_{kw} \int \theta_{dk}^{\gamma_{dk}^{\backslash w} + 1} \prod_{l \neq k} \theta_{dl}^{\gamma_{dl}^{\backslash w} - 1} d\boldsymbol{\theta}_d + \sum_{k' \neq k} \phi_{k'w} \int \theta_{dk}^{\gamma_{dk}^{\backslash w}} \theta_{dk'}^{\gamma_{dk'}^{\backslash w}} \prod_{l \neq k \wedge l \neq k'} \theta_{dl}^{\gamma_{dl}^{\backslash w} - 1} d\boldsymbol{\theta}_d \Big\}
\end{aligned}$$

It should be noted that $\frac{\prod_k \Gamma(\gamma_{dk})}{\Gamma(\sum_k \gamma_{dk})} = \int \prod_k \theta_{dk}^{\gamma_{dk}-1} d\boldsymbol{\theta}_d$ holds for any $\boldsymbol{\gamma}_d = (\gamma_{d1}, \ldots, \gamma_{dK})$. This is how we obtain the normalizing constant of Dirichlet distribution. Therefore, $\int \theta_{dk}^{\gamma_{dk}^{\backslash w}+1} \prod_{l \neq k} \theta_{dl}^{\gamma_{dl}^{\backslash w}-1} d\boldsymbol{\theta}_d =$

$\frac{\Gamma(\gamma_{dk}^{\backslash w}+2)\prod_{l\neq k}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)}$. Further, $\int \theta_{dk}^{\gamma_{dk}^{\backslash w}}\theta_{dk'}^{\gamma_{dk'}^{\backslash w}}\prod_{l\neq k\wedge l\neq k'}\theta_{dl}^{\gamma_{dl}^{\backslash w}-1}d\boldsymbol{\theta}_d = \frac{\Gamma(\gamma_{dk}^{\backslash w}+1)\Gamma(\gamma_{dk'}^{\backslash w}+1)\prod_{l\neq k\wedge l\neq k'}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)}$. With these results, we can continue the formula rearrangement as follows:

$$
\begin{aligned}
\therefore E_{\tilde{q}}[\theta_{dk}] &= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\Gamma(\hat{\gamma}_d{}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \Big\{ \phi_{kw} \frac{\Gamma(\gamma_{dk}^{\backslash w}+2)\prod_{l\neq k}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)} \\
&\qquad + \sum_{k'\neq k} \phi_{k'w} \frac{\Gamma(\gamma_{dk}^{\backslash w}+1)\Gamma(\gamma_{dk'}^{\backslash w}+1)\prod_{l\neq k\wedge l\neq k'}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)} \Big\} \\
&= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big\{ \phi_{kw} \frac{\Gamma(\hat{\gamma}_d{}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \cdot \frac{\Gamma(\gamma_{dk}^{\backslash w}+2)\prod_{l\neq k}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)} \\
&\qquad + \sum_{k'\neq k} \phi_{k'w} \frac{\Gamma(\hat{\gamma}_d{}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \cdot \frac{\Gamma(\gamma_{dk}^{\backslash w}+1)\Gamma(\gamma_{dk'}^{\backslash w}+1)\prod_{l\neq k\wedge l\neq k'}\Gamma(\gamma_{dl}^{\backslash w})}{\Gamma(\hat{\gamma}_d{}^{\backslash w}+2)} \Big\} \\
&= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big\{ \phi_{kw} \frac{\gamma_{dk}^{\backslash w}(\gamma_{dk}^{\backslash w}+1)}{\hat{\gamma}_d{}^{\backslash w}(\hat{\gamma}_d{}^{\backslash w}+1)} + \sum_{k'\neq k} \phi_{k'w} \frac{\gamma_{dk}^{\backslash w}\gamma_{dk'}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}(\hat{\gamma}_d{}^{\backslash w}+1)} \Big\} \\
&= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\gamma_{dk}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}} \cdot \frac{\phi_{kw} + \sum_k \phi_{kw}\gamma_{dk}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}+1}
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
E_{\tilde{q}}[\theta_{dk}^2] &= \int \theta_{dk}^2 \tilde{q}(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \\
&= \int \theta_{dk}^2 \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big(\sum_k \theta_{dk}\phi_{kw} \Big) \frac{\Gamma(\hat{\gamma}_d{}^{\backslash w})}{\prod_k \Gamma(\gamma_{dk}^{\backslash w})} \prod_k \theta_{dk}^{\gamma_{dk}^{\backslash w}-1} d\boldsymbol{\theta}_d \\
&= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \Big\{ \phi_{kw} \frac{\gamma_{dk}^{\backslash w}(\gamma_{dk}^{\backslash w}+1)(\gamma_{dk}^{\backslash w}+2)}{\hat{\gamma}_d{}^{\backslash w}(\hat{\gamma}_d{}^{\backslash w}+1)(\hat{\gamma}_d{}^{\backslash w}+2)} + \sum_{k'\neq k} \phi_{k'w} \frac{\gamma_{dk}^{\backslash w}(\gamma_{dk}^{\backslash w}+1)\gamma_{dk'}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}(\hat{\gamma}_d{}^{\backslash w}+1)(\hat{\gamma}_d{}^{\backslash w}+2)} \Big\} \\
&= \frac{\hat{\gamma}_d{}^{\backslash w}}{\sum_k \phi_{kw}\gamma_{dk}^{\backslash w}} \cdot \frac{\gamma_{dk}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}} \cdot \frac{\gamma_{dk}^{\backslash w}+1}{\hat{\gamma}_d{}^{\backslash w}+1} \cdot \frac{2\phi_{kw} + \sum_k \phi_{kw}\gamma_{dk}^{\backslash w}}{\hat{\gamma}_d{}^{\backslash w}+2} \; .
\end{aligned} \tag{8}
$$

We determine a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\gamma}'_d)$ so that the mean $\frac{\gamma'_{dk}}{\hat{\gamma}'_d}$ and the *average* second moment[2] $\frac{1}{K}\sum_k \frac{\gamma'_{dk}(\gamma'_{dk}+1)}{\hat{\gamma}'_d(\hat{\gamma}'_d+1)}$ are equal to those of $\tilde{q}(\boldsymbol{\theta}_d)$, where $\hat{\gamma}'_d = \sum_k \gamma'_{dk}$. This procedure is called *moment matching*.

To be precise, we obtain $\boldsymbol{\gamma}'_d$ by solving the following equations:

$$\frac{\gamma'_{dk}}{\hat{\gamma}'_d} = E_{\tilde{q}}[\theta_{dk}] \quad \text{for each } k\text{, and} \tag{9}$$

$$\frac{1}{K}\sum_k \frac{\gamma'_{dk}(\gamma'_{dk}+1)}{\hat{\gamma}'_d(\hat{\gamma}'_d+1)} = \frac{1}{K}\sum_k E_{\tilde{q}}[\theta_{dk}^2] \; . \tag{10}$$

[2]http://www.ucl.ac.uk/statistics/research/pdfs/135.zip

We can obtain γ'_{dk} as below.

$$\begin{aligned}
\gamma'_{dk} &= E_{\tilde{q}}[\theta_{dk}]\hat{\gamma}'_d \\
\sum_k \frac{E_{\tilde{q}}[\theta_{dk}]\hat{\gamma}'_d(E_{\tilde{q}}[\theta_{dk}]\hat{\gamma}'_d + 1)}{\hat{\gamma}'_d(\hat{\gamma}'_d + 1)} &= \sum_k E_{\tilde{q}}[\theta^2_{dk}] \\
\sum_k E_{\tilde{q}}[\theta_{dk}](E_{\tilde{q}}[\theta_{dk}]\hat{\gamma}'_d + 1) &= \sum_k E_{\tilde{q}}[\theta^2_{dk}](\hat{\gamma}'_d + 1) \\
\sum_k E_{\tilde{q}}[\theta_{dk}]^2\hat{\gamma}'_d + \sum_k E_{\tilde{q}}[\theta_{dk}] &= \sum_k E_{\tilde{q}}[\theta^2_{dk}]\hat{\gamma}'_d + \sum_k E_{\tilde{q}}[\theta^2_{dk}] \\
\sum_k \big(E_{\tilde{q}}[\theta_{dk}]^2 - E_{\tilde{q}}[\theta^2_{dk}]\big)\hat{\gamma}'_d &= \sum_k \big(E_{\tilde{q}}[\theta^2_{dk}] - E_{\tilde{q}}[\theta_{dk}]\big) \\
\therefore \hat{\gamma}'_d &= \frac{\sum_k(E_{\tilde{q}}[\theta^2_{dk}] - E_{\tilde{q}}[\theta_{dk}])}{\sum_k(E_{\tilde{q}}[\theta_{dk}]^2 - E_{\tilde{q}}[\theta^2_{dk}])} \qquad (11) \\
\therefore \gamma'_{dk} &= \frac{\sum_k(E_{\tilde{q}}[\theta^2_{dk}] - E_{\tilde{q}}[\theta_{dk}])}{\sum_k(E_{\tilde{q}}[\theta_{dk}]^2 - E_{\tilde{q}}[\theta^2_{dk}])} \cdot E_{\tilde{q}}[\theta_{dk}] \qquad (12)
\end{aligned}$$

5

We now have a Dirichlet distribution that approxiates the posterior $\tilde{q}(\boldsymbol{\theta}_d)$ in Eq. (6), i.e.,

$$\tilde{q}(\boldsymbol{\theta}_d) = \frac{\sum_k \gamma^{\setminus w}_{dk}}{\sum_k \phi_{kw}\gamma^{\setminus w}_{dk}} \Big(\sum_k \theta_{dk}\phi_{kw} \Big) \frac{\Gamma(\sum_k \gamma^{\setminus w}_{dk})}{\prod_k \Gamma(\gamma^{\setminus w}_{dk})} \prod_k \theta_{dk}^{\gamma^{\setminus w}_{dk} - 1} .$$

To obtain a density function of Dirichlet distribution from Eq. (6), we replace the term $(\sum_k \theta_{dk}\phi_{kw})$ by $t'_w(\boldsymbol{\theta}_d) = s'_w \prod_k \theta_{dk}^{\beta'_{wk}}$. We set the resulting function equal to the density function of $\mathrm{Dir}(\boldsymbol{\gamma}'_d)$ and obtain the following equation:

$$\frac{\sum_k \gamma^{\setminus w}_{dk}}{\sum_k \phi_{kw}\gamma^{\setminus w}_{dk}} s'_w \prod_k \theta_{dk}^{\beta'_{wk}} \frac{\Gamma(\sum_k \gamma^{\setminus w}_{dk})}{\prod_k \Gamma(\gamma^{\setminus w}_{dk})} \prod_k \theta_{dk}^{\gamma^{\setminus w}_{dk} - 1} = \frac{\Gamma(\sum_k \gamma'_{dk})}{\prod_k \Gamma(\gamma'_{dk})} \prod_k \theta_{dk}^{\gamma'_{dk} - 1} . \qquad (13)$$

Consequently, s'_w and β'_{wk} are determined as follows:

$$\beta'_{wk} = \gamma'_{dk} - \gamma^{\setminus w}_{dk} , \qquad (14)$$

$$s'_w = \frac{\sum_k \phi_{kw}\gamma^{\setminus w}_{dk}}{\sum_k \gamma^{\setminus w}_{dk}} \frac{\Gamma(\sum_k \gamma'_{dk})}{\prod_k \Gamma(\gamma'_{dk})} \frac{\prod_k \Gamma(\gamma^{\setminus w}_{dk})}{\Gamma(\sum_k \gamma^{\setminus w}_{dk})} . \qquad (15)$$