Variational EM Algorithm And Its Application to Wireless system

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

March 13, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Applied EM to 5G wireless mMTC systems - sparse Bayesian learning
- Today's agenda
  - Discuss limitations of EM and then discuss variational EM which overcomes this limitation
  - Ref: Chap 10.1 of PRML

Limitations of EM algorithm

- EM assumes in E step, tractability in calculating
  - posterior distribution of latent variable $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$
- Variational inference helps when they are not tractable
  - Bypasses the requirement of exactly knowing $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$, by assuming an appropriate $q(\mathbf{Z})$

EM algorithm derivation recap

- Recall that the maximum likelihood is given as

$$\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q \parallel p) \text{ where} \\
\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right) \text{ and } KL(q \parallel p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right)
\end{aligned}$$

- Recall E step calculates $q(\mathbf{Z})$ by maximizing $\mathcal{L}(q,\boldsymbol{\theta}^{\text{old}})$ with respect to $q(\mathbf{Z})$, by fixing $\boldsymbol{\theta}^{old}$
  - Leads to $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$, which is now difficult to calculate
- M step fixes $q(\mathbf{Z})$, and maximizes $L(q,\boldsymbol{\theta})$ wrt $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$

Variational EM (VEM) algorithm (1)

- VEM algorithm assumes
  1. $\mathbf{Z}$ is partitioned into M disjoint groups as $\mathbf{Z}_i$ where $i = 1, \ldots, M$
  2. Posterior distribution $q(\mathbf{Z})$ also factorizes with respect to these partitions as

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i) = \prod_{i=1}^{M} q_i$$

  where q_i is the simplified notation of $q_i(\mathbf{Z}_i)$
- Factorized approximation stems from theoretical physics where it is called mean field theory
  - Assume $\mathbf{Z}$ is independent across these M groups
- E step of VEM calculates q_i by maximizing $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to q_i, by fixing $\boldsymbol{\theta}^{old}$

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}^{old}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})}{q(\mathbf{Z})}\right) = \sum_{\mathbf{Z}} \prod_i q_i(\mathbf{Z}_i) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})}{\prod_i q_i(\mathbf{Z}_i)}\right) \\ &= \sum_{\mathbf{Z}} \prod_i q_i \left(\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old}) - \sum_i \log q_i\right) \end{aligned} \quad (1)$$

- We have to now determine optimal $q_i(\mathbf{Z}_i)$, for $i = 1, \ldots, M$, which will maximize $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$

Variational EM algorithm (2)

- Let's simplify Eq. (1) for $M = 2$, wherein $q(\mathbf{Z}) = q_1(\mathbf{Z}_1)q_2(\mathbf{Z}_2) = q_1 q_2$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}^{old}) &= \sum_{\mathbf{Z}} q_1 q_2 \left\{ \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old}) - (\log q_1 + \log q_2) \right\} \\
&= \sum_{\mathbf{Z}} q_1 q_2 \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old}) - \underbrace{\sum_{\mathbf{Z}_1} \sum_{\mathbf{Z}_2} q_1 q_2 \log q_1}_{T1} - \underbrace{\sum_{\mathbf{Z}_1} \sum_{\mathbf{Z}_2} q_1 q_2 \log q_2}_{T2}
\end{aligned}$$

- T_1 and T_2 can further be simplified as follows

$$\begin{aligned}
T_1 &= \sum_{\mathbf{Z}_1} \sum_{\mathbf{Z}_2} q_1 q_2 \log q_1 = \left(\sum_{\mathbf{Z}_1} q_1 \log q_1 \right) \underbrace{\left(\sum_{\mathbf{Z}_2} q_2 \right)}_{=1} = \sum_{\mathbf{Z}_1} q_1 \log q_1 \\
T_2 &= \sum_{\mathbf{Z}_2} q_2 \log q_2
\end{aligned}$$

- Thus, $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ in terms of $q_1 = q_1(\mathbf{Z}_1)$ reduces to following

$$\mathcal{L}(q, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} q_1 \left(q_2 \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old}) \right) - \sum_{\mathbf{Z}_1} q_1 \log q_1 + \text{constant wrt } q_1$$

Variational EM algorithm (3)

- Equivalently, in terms of $q_j = q_j(\mathbf{Z}_j)$, $\mathcal{L}(q, \boldsymbol{\theta}^{old})$ reduces to

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}^{old}) &= \sum_{\mathbf{Z}_j} q_j \left(\sum_{\mathbf{Z}_{i \neq j}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old}) \prod_{i \neq j} q_i \right) - \sum_{\mathbf{Z}_j} q_j \log q_j + \text{constant} \\ &= \sum_{\mathbf{Z}_j} q_j \log \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}) - \sum_{\mathbf{Z}_j} q_j \log q_j + \text{constant} \qquad (2)\end{aligned}$$

- Here $\log \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}) = \mathbb{E}_{i \neq j}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})\right] + \text{constant}$
  - $\mathbb{E}_{i \neq j}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})\right]$ denotes expectation w.r.t. q distributions over all variables $\mathbf{Z}_i$ for $i \neq j$
  - Constant is because $\log \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old})$ is unscaled distribution
- We keep $q_{i \neq j}$ fixed and maximize $L(q, \boldsymbol{\theta}^{old})$ in (2) w.r.t $q_j(\mathbf{Z}_j)$
- This is done by recognizing the following about Eq. (2)

$$\mathcal{L}(q, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}_j} q_j \log \left(\frac{\tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old})}{q_j} \right) = -KL(q_j \parallel \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}))$$

- RHS of Eq. (2) is a negative KL distance between $q_j(\mathbf{Z}_j)$ and $\tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old})$
  - Maximizing (2) is minimizing KL distance $KL(q_j \parallel \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}))$

Variational EM algorithm (4)

- Minimizing KL distance $KL(q_j \parallel \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}))$ happens when

$$
\begin{aligned}
q_j^*(\mathbf{Z}_j) &= \tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old}) \\
\Rightarrow \log(q_j^*(\mathbf{Z}_j)) &= \log(\tilde{p}(\mathbf{X}, \mathbf{Z}_j|\boldsymbol{\theta}^{old})) = \mathbb{E}_{i \neq j}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})\right] + \text{constant}
\end{aligned}
$$

- Solution says that log of optimal q_j is obtained by
  - Considering the log of complete data likelihood (CDLL)
  - Taking the expectation with respect to all $\{q_i\}$ for $i \neq j$

$$
q_j^*(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})\right]\right)}{\sum_{\mathbf{Z}_j} \exp\left(\mathbb{E}_{i \neq j}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{old})\right]\right)}
$$

- Solution is calculated by cyclically calculating q_j, and replacing each in turn with revised estimate

Summary of variational EM algorithm

- If the current estimate for the parameters is denoted $\boldsymbol{\theta}^{old}$, then variational EM algorithm is
  1. E step: use current parameter values $\boldsymbol{\theta}^{old}$ to find posterior of latent variables $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) = q^*(\mathbf{Z}) = \prod_{i=1}^{M} q_i^*(\mathbf{Z}_i)$
  2. Use $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ to find expectation of CDLL evaluated for some general $\boldsymbol{\theta}$

$$\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

  3. M step: determine the revised parameter estimate $\boldsymbol{\theta}^{new}$ by maximizing expected value of CDLL

$$\boldsymbol{\theta}^{new} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}).$$

- Variational EM resolves tractability in calculating
  - posterior distribution of latent variable $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$