Application of EM Algorithm to Wireless system

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

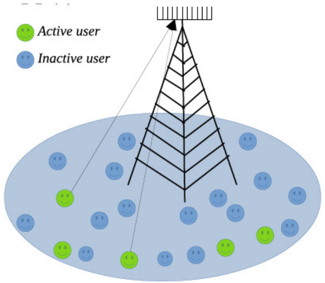
March 11, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Discussed an alternative view of EM algorithm
  - Proved that EM maximizes log likelihood while maximizing the lower bound
- Today's agenda
  - Apply EM to 5G wireless mMTC systems - sparse Bayesian learning

5G mMTC systems model (recap)

- Consider a mMTC system with M single-antenna mMTC devices and N-antenna base-station (BS)



- Only few mMTC active devices transmit data which BS need to process
- BS does not know which devices are active. All active M mMTC devices transmit simultaneously
- Total number of mMTC devices $M \gg N$ and number of active mMTC devices $K < N \ll M$
- Received signal assuming all devices are active $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$
- Tx signal $\mathbf{x} = [x_1, \cdots, x_M]^T$, rx signal $\mathbf{y} = [y_1, \cdots, y_N]^T$, and noise $\mathbf{n} = [n_1, \cdots, n_N]^T$
- Sparse transmit vector $\mathbf{x}$ contains only $K \ll M$ non-zero values $\mathbf{x} = [1, 0, , 1, , 0, 0, 0 \cdots, 0]^T$

EM algorithm for SBL in 5G mMTC systems (1)

- If the current estimate of parameters is denoted $\boldsymbol{\theta}^{old}$, then EM algorithm is
  1. E step: use current parameter values $\boldsymbol{\theta}^{old}$ to find posterior of latent variables $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$
  2. Use $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$ to find expectation of log 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 this function
$$\boldsymbol{\theta}^{new} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}).$$
- 5G mMTC data model is $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$ with unknown noise variance β^{-1}
- We assume
  - $\mathbf{X} = \mathbf{y}$ and $\mathbf{H}$ as observed variable
  - $\mathbf{x}$ as latent variable $\mathbf{z}$ with prior $p\left(\mathbf{x}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\mathbf{0}, (\text{diag}(\boldsymbol{\alpha}))^{-1}\right)$ and $\text{diag}(\boldsymbol{\alpha}) = \text{diag}(\alpha_1, \cdots, \alpha_M)$
    - Gaussian prior promotes sparsity in $\mathbf{x}$ – Note $\boldsymbol{\alpha}$ has M dimensions
  - Parameter $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \beta\}$
- EM will calculate
  - posterior distribution $p(\mathbf{x}|\mathbf{y}, \mathbf{H}, \boldsymbol{\alpha}, \beta)$ in Step 1
  - expected value (using above posterior) of log of CDLL $p(\mathbf{y}, \mathbf{H}, \mathbf{x}|\boldsymbol{\alpha}, \beta)$ in Step 2
  - Point estimate of the parameters by maximizing above expectation

EM algorithm Step 1 – calculation of posterior distribution

- For, 5G mMTC data model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$
- EM in E Step will first calculate posterior distribution $p(\mathbf{x}|\mathbf{y}, \mathbf{H}, \boldsymbol{\alpha}, \beta)$
- Recall from lecture 10, the posterior density of $\mathbf{x}$ is Gaussian with $p\left(\mathbf{x}|\mathbf{y}, \mathbf{H}, \beta\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right)$
  - posterior covariance matrix and mean

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \left(\beta \mathbf{H}^T \mathbf{H} + \mathrm{diag}(\boldsymbol{\alpha})\right)^{-1} \text{ and } \boldsymbol{\mu}_{\mathbf{x}} = \beta \boldsymbol{\Sigma}_{\mathbf{x}} \mathbf{H}^T \mathbf{y}$$

  - We also have

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T \Rightarrow \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T$$

- EM in E Step will then first calculate expectation of log of CDLL

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{H}, \mathbf{x}|\boldsymbol{\alpha}, \beta) &= \log\{p(\mathbf{y}|\mathbf{H}, \mathbf{x}, \boldsymbol{\alpha}, \beta)p(\mathbf{H}, \mathbf{x}|\boldsymbol{\alpha}, \beta)\} = \log\{p(\mathbf{y}|\mathbf{H}, \mathbf{x}, \boldsymbol{\alpha}, \beta)p(\mathbf{x}|\boldsymbol{\alpha})\} \\ &= \log p(\mathbf{y}|\mathbf{H}, \mathbf{x}, \beta) + \log p(\mathbf{x}|\boldsymbol{\alpha}) \end{aligned}$$

- For model $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, we have $p(\mathbf{y}|\mathbf{H}, \mathbf{x}, \beta) = \mathcal{N}(\mathbf{H}\mathbf{x}, \beta^{-1}\mathbf{I})$ and prior $\mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha})^{-1})$
  - Note $\mathbf{y}$ is N dimensional Gaussian and prior on $\mathbf{x}$ is M dimensional Gaussian

EM algorithm Step 1 – calculation of expectation wrt posterior

- EM computes $\mathbb{E}_{\mathbf{x}}$ of log of CDLL under pos. dis. $p\left(\mathbf{x}|\mathbf{y},\mathbf{H},\beta\right)$ and maximizes it to calculate $\boldsymbol{\alpha}$, β

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}} \log p(\mathbf{y},\mathbf{H},\mathbf{x}|\boldsymbol{\alpha},\beta) &= \mathbb{E}_{\mathbf{x}}[\log p(\mathbf{y}|\mathbf{H},\mathbf{x},\beta) + \log p(\mathbf{x}|\boldsymbol{\alpha})] \\
&= \mathbb{E}_{\mathbf{x}}[\log \mathcal{N}(\mathbf{Hx},\beta^{-1}\mathbf{I}) + \log \mathcal{N}(\mathbf{0},\mathrm{diag}(\boldsymbol{\alpha})^{-1})] \\
&= \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[N\log\beta + \sum_{i=1}^{M}\log\alpha_i - \beta||\mathbf{y}-\mathbf{Hx}||^2 - \mathbf{x}^T\mathrm{diag}(\boldsymbol{\alpha})\mathbf{x}\right] \\
&= \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[N\log\beta + \sum_{i=1}^{M}\log\alpha_i - \beta||\mathbf{y}-\mathbf{Hx}||^2 - \mathrm{Tr}\left(\mathrm{diag}(\boldsymbol{\alpha})\mathbf{x}\mathbf{x}^T\right)\right] \\
&= \frac{1}{2}N\log\beta + \sum_{i=1}^{M}\log\alpha_i - \beta\underbrace{\mathbb{E}_{\mathbf{x}}\left[||\mathbf{y}-\mathbf{Hx}||^2\right]}_{T_1} - \underbrace{\mathbb{E}_{\mathbf{x}}\left[\mathrm{Tr}(\mathrm{diag}(\boldsymbol{\alpha})\mathbf{x}\mathbf{x}^T)\right]}_{T_2} \quad (1)
\end{aligned}
$$

Calc. of hyper-parameter α and noise precision β using EM (1)

- Simplifying T_1:

$$
\begin{aligned}
T_1 &= \mathbb{E}_{\mathbf{x}}[||\mathbf{y}-\mathbf{Hx}||^2] = \mathbb{E}_{\mathbf{x}}\left[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{Hx} + \mathbf{x}^T\mathbf{H}^T\mathbf{Hx}\right] \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\mathbb{E}_{\mathbf{x}}\{\mathbf{x}\} + \mathrm{Tr}\left[\mathbf{H}\mathbb{E}_{\mathbf{x}}\{\mathbf{xx}^T\}\mathbf{H}^T\right] \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}} + \mathrm{Tr}\left[(\mathbf{H}(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T + \boldsymbol{\Sigma}_{\mathbf{x}})\mathbf{H}^T)\right] \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}} + \mathrm{Tr}\left(\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T\mathbf{H}^T\right) + \mathrm{Tr}\left(\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^T\right) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}} + \mathrm{Tr}\left((\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}})^T(\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}})\right) + \mathrm{Tr}\left(\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^T\right) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}} + (\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}})^T(\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}) + \mathrm{Tr}\left(\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^T\right) \\
&= ||\mathbf{y}-\mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}||^2 + \mathrm{Tr}\left[\mathbf{H}^T\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\right] \qquad (2)
\end{aligned}
$$

- Simplifying T_2:

$$
\begin{aligned}
T_2 &= \mathbb{E}_{\mathbf{x}}\left[\mathrm{Tr}\left(\mathrm{diag}(\boldsymbol{\alpha})\mathbf{xx}^T\right)\right] = \mathrm{Tr}\left[\mathrm{diag}(\boldsymbol{\alpha})\mathbb{E}_{\mathbf{x}}\left(\mathbf{xx}^T\right)\right] = \mathrm{Tr}\left[\mathrm{diag}(\boldsymbol{\alpha})\left(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^T + \boldsymbol{\Sigma}_{\mathbf{x}}\right)\right] \\
&= \mathrm{Tr}\left[\boldsymbol{\mu}_{\mathbf{x}}^T\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\mu}_{\mathbf{x}} + \mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\Sigma}_{\mathbf{x}}\right] = \sum_{i=1}^{M}\alpha_i\left((\boldsymbol{\mu}_{\mathbf{x}}(i))^2 + \boldsymbol{\Sigma}_{\mathbf{x}}(i,i)\right) \qquad (3)
\end{aligned}
$$

Calc. of hyper-parameter α and noise precision β using EM (2)

- On substituting (2) and (3) in (1),we get

$$\mathbb{E}_{\mathbf{x}}[\log p(\mathbf{y}, \mathbf{H}, \mathbf{x}|\boldsymbol{\alpha}, \beta)] = \frac{1}{2}[N \log \beta + \sum_{i=1}^{M} \log \alpha_i - \beta||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}||^2 - \beta \mathrm{Tr}(\mathbf{H}^T \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{x}}) - \sum_{i=1}^{M} \alpha_i((\boldsymbol{\mu}_{\mathbf{x}}(i))^2 + \boldsymbol{\Sigma}_{\mathbf{x}}(i,i))] \quad (4)$$

- Maximizes it w.r.t to α_i by differentiating (4) and setting the result equal to zero:

$$\frac{1}{2\alpha_i^{new}} - \frac{1}{2}[\boldsymbol{\Sigma}_{\mathbf{x}}(i,i) + (\boldsymbol{\mu}_{\mathbf{x}}(i))^2] = 0 \Rightarrow \alpha_i^{new} = \frac{1}{\boldsymbol{\Sigma}_{\mathbf{x}}(i,i) + (\boldsymbol{\mu}_{\mathbf{x}}(i))^2}$$

- Maximizes it w.r.t to β by differentiating (4) w.r.t to β and setting it equal to zero:

$$\frac{N}{2\beta^{new}} - \frac{1}{2}||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}||^2 - \frac{1}{2}\mathrm{Tr}[\mathbf{H}^T \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{x}}] = 0 \quad (5)$$

$$(\beta^{new})^{-1} = \frac{1}{N}\left(||\mathbf{y} - \mathbf{H}\boldsymbol{\mu}_{\mathbf{x}}||^2 + \mathrm{Tr}[\mathbf{H}^T \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{x}}]\right)$$

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 $q(\mathbf{Z})$