MM Optimization Algorithms

Chathuranga Weeraddana

March 2022

Lecture 5: EM Principle

Landmarks

- roots trace back to
  - H. O. Hartley (1958, EM algorithms)
- a phenomenal contribution from
  - A. P. Dempster et al. (1977, ML ... via the EM algorithm)
  - citations ≈ 66500 (2022-May)

When to Use?

- observations are incomplete
- still you want to compute the ML estimate of parameter θ
- i.e., applied when observations can be viewed as incomplete
  - missing value situations
  - when there are censored or truncated data
  - factor analysis
  - many more

EM Vs MM

- EM = Expectation and Maximization
- an interpretation [1]
  - EM transfers maximization of likelihood $l(\cdot)$ to $Q(\cdot\,|\theta^{(n)})$
    - this transfer is simply the expectation step
  - then $Q(\cdot\,|\theta^{(n)})$ is maximized with respect to θ
  - $Q(\cdot\,|\theta^{(n)})$ is a minorization function [2] of $l(\cdot)$
  - i.e., we have SM (surrogate maximization) principle within EM

[1] See *Optimization Transfer Using Surrogate Objective Functions* by K. Lange et al., 2000.

[2] Up to an irrelevant constant.

- is SM (or MM) [3] is just EM?
  - a problem posed by Xiao-Li Meng [4]
  - given SM construction $\rightarrow$ a corresponding EM construction?
  - is EM class is as rich as SM class?
- Xiao-Li Meng's EM flu $\rightarrow$ no cure so far

[3] There is a slight difference though, see *[Optimization Transfer Using Surrogate Objective Functions]: Rejoinder* by D. R. Hunter and K. Lange, 2000.

[4] See *[Optimization Transfer Using Surrogate Objective Functions]: Discussion* by Xiao-Li Meng, 2000.

Key Idea

- recall..
  - maximization of log-likelihood $l(\cdot)$ is transferred to $Q(\ \cdot\ |\theta^{(n)})$
  - $Q(\ \cdot\ |\theta^{(n)})$ is a minorization function of $l(\cdot)$
  - then $Q(\ \cdot\ |\theta^{(n)})$ is maximized with respect to θ
  - i.e., we have MM principle within EM
- recall that the observations are incomplete

Formulation of the Setting

- denote the complete data by x with likelihood $r_\theta(x)$
- denote the observed data by y with likelihood $s_\theta(y)$
- thus, the conditional density of $x|y$, $k_\theta(x|y)$ is given by

$$k_\theta(x|y) = \frac{r_\theta(x)}{s_\theta(y)} \tag{1}$$

- log-likelihood function of x is $\ln r_\theta(x)$
- log-likelihood function of y (observed data) is $l(\theta) = \ln s_\theta(y)$

- EM literature defines the surrogate $Q(\,\cdot\,|\theta^{(n)})$ as

$$Q(\theta|\theta^{(n)}) = \mathbb{E}\left\{\ln r_\theta(x) \mid y, \theta^{(n)}\right\} \qquad (2)$$
$$= \int_{\mathcal{X}(y)} \ln r_\theta(x) k_{\theta^{(n)}}(x|y)\, dx$$

- heuristic idea:
  - we would like to choose $\theta^\star$ that maximize $\ln r_\theta(x)$
  - but we do not have it because observations are incomplete
  - instead, maximize the expectation of $\ln r_\theta(x)$ given
    - the observations y
    - the current parameter $\theta^{(n)}$

$Q(\,\cdot\,|\theta^{(n)})$ as a Minorization

- it can be shown that [5]

$$\begin{aligned} Q\big(\theta|\theta^{(n)}\big) - l(\theta) &= \mathbb{E}\left\{\ln k_\theta(x|y) \mid y, \theta^{(n)}\right\} \\ &\leq \mathbb{E}\left\{\ln k_{\theta^{(n)}}(x|y) \mid y, \theta^{(n)}\right\} \\ &= Q\big(\theta^{(n)}|\theta^{(n)}\big) - l(\theta^{(n)}) \end{aligned}$$

- thus, $Q(\,\cdot\,|\theta^{(n)})$ is a minorization function of l [6]

[5] See *Additional Reading* section of the courseweb for a sketch of the proof.
[6] Up to an irrelevant constant.

Examples

Cell Probabilities of a Population

- 197 animals distributed multinomially into 4 groups
- observed data $y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$
- cell probabilities are of the form

$$\left(\tfrac{1}{2} + \tfrac{1}{4}\pi, \tfrac{1}{4}(1-\pi), \tfrac{1}{4}(1-\pi), \tfrac{1}{4}\pi\right)$$

  for some π with $0 \leq \pi \leq 1$

- thus the likelihood of observed data is

$$s_\pi(y) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}\left(\tfrac{1}{2}+\tfrac{1}{4}\pi\right)^{y_1}\left(\tfrac{1}{4}-\tfrac{1}{4}\pi\right)^{y_2}\left(\tfrac{1}{4}-\tfrac{1}{4}\pi\right)^{y_3}\left(\tfrac{1}{4}\pi\right)^{y_4}$$

- log-likelihood function l is given by

$$l(\pi) = y_1 \ln(\tfrac{1}{2} + \tfrac{1}{4}\pi) + (y_2 + y_3)\ln(\tfrac{1}{4} - \tfrac{1}{4}\pi) + y_4 \ln \pi + c$$

- maximize $l(\pi)$ subject to $\pi \in [0, 1]$ to determine $\pi^\star$
- in this example
  - observed data $=$ complete data
  - the procedure is straightforward
- what if observed data $\neq$ complete data?

- 197 animals distributed multinomially into 5 groups
- complete data $x = (x_1, x_2, x_3, x_4, x_5)$
- observed data $y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ where
  - $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, and $y_4 = x_5$
  - cell probabilities are of the form

    $$\left(\tfrac{1}{2}, \tfrac{1}{4}\pi, \tfrac{1}{4}(1-\pi), \tfrac{1}{4}(1-\pi), \tfrac{1}{4}\pi\right)$$

    for some π with $0 \leq \pi \leq 1$
- thus the likelihood of complete data is

$$r_\pi(x) = \frac{(\sum_i x_i)!}{x_1!x_2!x_3!x_4!x_5!} \left(\tfrac{1}{2}\right)^{x_1} \left(\tfrac{1}{4}\pi\right)^{x_2} \left(\tfrac{1}{4}-\tfrac{1}{4}\pi\right)^{x_3} \left(\tfrac{1}{4}-\tfrac{1}{4}\pi\right)^{x_4} \left(\tfrac{1}{4}\pi\right)^{x_5}$$

- EM defines the surrogate $Q(\,\cdot\,|\pi^{(n)})$ as

$$Q(\pi|\pi^{(n)}) = \mathbb{E}\left\{\ln r_\pi(x) \mid y, \pi^{(n)}\right\} \tag{3}$$
$$= \int_{\mathcal{X}(y)} \ln r_\pi(x) k_{\pi^{(n)}}(x|y)\, dx$$

- here we have

$$k_{\pi^{(n)}}(x|y) = \frac{y_1!}{x_1!x_2!(\frac{1}{2}+\frac{\pi^{(n)}}{4})^{y_1}}\left(\tfrac{1}{2}\right)^{x_1}\left(\tfrac{1}{4}\pi^{(n)}\right)^{x_2} \tag{4}$$
$$= \frac{250!}{x_1!x_2!(\frac{1}{2}+\frac{\pi^{(n)}}{4})^{250}}\left(\tfrac{1}{2}\right)^{x_1}\left(\tfrac{1}{4}\pi^{(n)}\right)^{x_2}$$

- with some tedious steps it can be shown that [7]

$$Q(\pi|\pi^{(n)}) = \ln\left[(\tfrac{1}{2})^{x_1^{(n)}}(\tfrac{1}{4}\pi)^{x_2^{(n)}}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{18}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{20}(\tfrac{1}{4}\pi)^{34}\right]+\alpha^{(n)}$$

where

$$x_1^{(n)} = \mathbb{E}\{x_1|y,\pi^{(n)}\} = \frac{\frac{1}{2}y_1}{\frac{1}{2}+\frac{1}{4}\pi^{(n)}} = \frac{250}{2+\pi^{(n)}},$$
$$x_2^{(n)} = \mathbb{E}\{x_2|y,\pi^{(n)}\} = \frac{\frac{1}{4}\pi^{(n)}y_1}{\frac{1}{2}+\frac{1}{4}\pi^{(n)}} = \frac{250\pi^{(n)}}{2+\pi^{(n)}}, \qquad (5)$$

and $\alpha^{(n)}$ is an irrelevant constant which does not depend on π

[7] When the underlying distributions are from exponential families, some convenient tricks can be used when computing $Q(\cdot|\theta^{(n)})$. See A. P. Dempster et al. 1977, pp. 2-4.

- maximize $Q(\pi|\pi^{(n)})$ with respect to π to yield

$$\pi^{(n+1)} = \frac{x_2^{(n)} + 34}{x_2^{(n)} + 34 + 38} \tag{6}$$

Algorithm 1 EM for Computing Cell Probabilities

Input: $\pi^{(0)} \in (0,1)$, $n = 0$

while `a stopping criterion true` **do**
    $x_2^{(n)}$ is computed from (5)
    $\pi^{(n+1)}$ is computed from (6) and $n \leftarrow n+1$
end while
return $\pi^{(n)}$

Life of Light Bulbs

- lifetime information of 2 bulbs were observed
- observed data
  - lifetime of the first bulb is y
  - lifetime z of the second bulb is less than t
  - note: z was not observed
- lifetime x of bulbs $\rightarrow$ an exponential density, i.e.,

$$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- z is known $\rightarrow$ ML estimate of λ is computed

- complete data $x = (y, z)$
- observed data y and $z \leq t$
- the likelihood of complete data is

$$r_\lambda(x) = \lambda e^{-\lambda y} \lambda e^{-\lambda z}$$

- EM defines the surrogate $Q(\,\cdot\,|\lambda^{(n)})$ as

$$Q(\lambda|\lambda^{(n)}) = \mathbb{E}\left\{\ln r_\lambda(y,z) \mid y, z \leq t, \lambda^{(n)}\right\} \tag{7}$$

- here we have

$$k_{\lambda^{(n)}}(y, z|y, z \leq t) = \frac{\lambda^{(n)} e^{-\lambda^{(n)} z}}{1 - e^{-\lambda^{(n)} t}}, \quad 0 \leq z \leq t \tag{8}$$

- therefore we have

$$
\begin{aligned}
Q(\lambda|\lambda^{(n)}) &= \mathbb{E}\left\{\ln r_\lambda(y,z) \mid y, z \le t, \lambda^{(n)}\right\} \\
&= \mathbb{E}\left\{\ln[\lambda e^{-\lambda y}\lambda e^{-\lambda z}] \mid y, z \le t, \lambda^{(n)}\right\} \\
&= \ln\lambda - \lambda y + \ln\lambda - \lambda\mathbb{E}\{z|z \le t, \lambda^{(n)}\} \\
&= 2\ln\lambda - \lambda y - \lambda\int_0^t z\,\frac{\lambda^{(n)}e^{-\lambda^{(n)}z}}{1-e^{-\lambda^{(n)}t}}\,dz \\
&= 2\ln\lambda - \lambda y - \lambda\underbrace{\left[\frac{1}{\lambda^{(n)}} - \frac{te^{-\lambda^{(n)}t}}{1-e^{-\lambda^{(n)}t}}\right]}_{w^{(n)}}
\end{aligned}
$$

- maximize $Q(\lambda|\lambda^{(n)})$ with respect to λ to yield

$$\lambda^{(n+1)} = \frac{2}{w^{(n)} + y}$$

- thus an EM algorithm for computing the lifetime of a bulb
  - is readily derived

Mixture of Gaussian

- to be discussed!