MM Algorithms

Haiqin Yang

Department of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

Contents are from

1. Hunter, D. R., and Lange, K. (2004), "A Tutorial on MM Algorithms," The American Statistician, 58, 30-37.
2. Zhou, H. and Lange, K. (2010), "MM algorithms for some discrete multivariate distributions," Journal of Computational and Graphical Statistics, 19(3):645-665.

Outline

Definition of Majorization and Property

- A function $g(\theta|\theta^n)$ is said to majorize the function $f(\theta)$ at θ^n provided

$$\begin{aligned} f(\theta^n) &= g(\theta^n|\theta^n) \\ f(\theta) &\leq g(\theta|\theta^n) \quad \text{for all } \theta. \end{aligned}$$

- A descent property: $f(\theta^{n+1}) \leq f(\theta^n)$
  1. Minimizing $g(\theta)$ given θ^n yields

$$g(\theta^{n+1}|\theta^n) \leq g(\theta^n|\theta^n)$$

Definition of Majorization and Property

- A function $g(\theta|\theta^n)$ is said to majorize the function $f(\theta)$ at θ^n provided

$$\begin{aligned} f(\theta^n) &= g(\theta^n|\theta^n) \\ f(\theta) &\leq g(\theta|\theta^n) \quad \text{for all } \theta. \end{aligned}$$

- A descent property: $f(\theta^{n+1}) \leq f(\theta^n)$
  1. Minimizing $g(\theta)$ given θ^n yields
  $$g(\theta^{n+1}|\theta^n) \leq g(\theta^n|\theta^n)$$
  2. $f(\theta) \leq g(\theta|\theta^n)$ derives
  $$f(\theta^{n+1}) - g(\theta^{n+1}|\theta^n) \leq 0 = f(\theta^n) - g(\theta^n|\theta^n)$$

Definition of Majorization and Property

- A function $g(\theta|\theta^n)$ is said to majorize the function $f(\theta)$ at θ^n provided

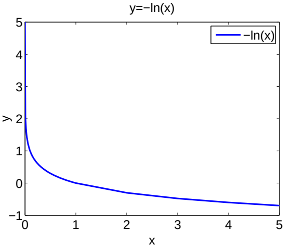
$$\begin{aligned} f(\theta^n) &= g(\theta^n|\theta^n) \\ f(\theta) &\leq g(\theta|\theta^n) \quad \text{for all } \theta. \end{aligned}$$

- A descent property: $f(\theta^{n+1}) \leq f(\theta^n)$
  1. Minimizing $g(\theta)$ given θ^n yields
  $$g(\theta^{n+1}|\theta^n) \leq g(\theta^n|\theta^n)$$
  2. $f(\theta) \leq g(\theta|\theta^n)$ derives
  $$f(\theta^{n+1}) - g(\theta^{n+1}|\theta^n) \leq 0 = f(\theta^n) - g(\theta^n|\theta^n)$$
  3. Summing up together, one has
  $$\begin{aligned} f(\theta^{n+1}) &= g(\theta^{n+1}|\theta^n) + f(\theta^{n+1}) - g(\theta^{n+1}|\theta^n) \\ &\leq g(\theta^n|\theta^n) + f(\theta^n) - g(\theta^n|\theta^n) \\ &= f(\theta^n) \end{aligned}$$

Outline

Jensen' Inequality

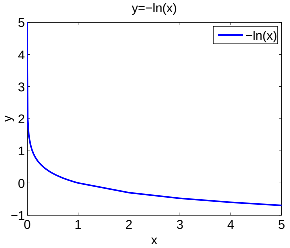
- $-\ln(x)$ is convex



Jensen' Inequality

- $-\ln(x)$ is convex
- For probability densities $a(x)$ and $b(x)$, we have

$$-\ln\left\{E_b\left[\frac{a(x)}{b(x)}\right]\right\} \leq -E_b\left[\ln\frac{a(x)}{b(x)}\right]$$



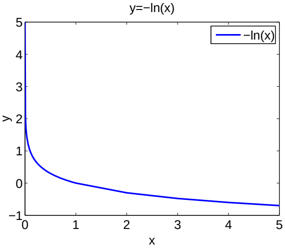
Jensen' Inequality

- $-\ln(x)$ is convex
- For probability densities $a(x)$ and $b(x)$, we have

$$-\ln\left\{E_b\left[\frac{a(x)}{b(x)}\right]\right\} \leq -E_b\left[\ln\frac{a(x)}{b(x)}\right]$$

- If x has the density $b(x)$, then

$$E_b[a(x)/b(x)] = 1.$$



Jensen' Inequality

- $-\ln(x)$ is convex
- For probability densities $a(x)$ and $b(x)$, we have

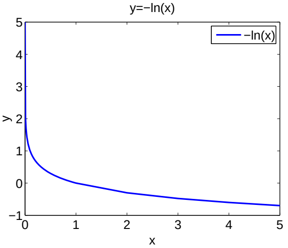
$$-\ln\left\{E_b\left[\frac{a(x)}{b(x)}\right]\right\} \leq -E_b\left[\ln\frac{a(x)}{b(x)}\right]$$

- If x has the density $b(x)$, then

$$E_b[a(x)/b(x)] = 1.$$

- Hence, the left hand side above vanishes and we obtain

$$E_b[\ln a(x)] \leq E_b[lnb(x)]$$



Outline

Procedure

- An EM algorithm operates by identifying a theoretical complete data space.
- It consists of the *expectation* step and the *maximization* step.
- In the E step, the conditional expectation of the complete data log-likelihood is calculated wrt. the observed data. The *surrogate* function created by the E step is a minorizing function.
- In the M step, this minorizing function is maximized wrt. the parameters of the underlying model.
- Every EM algorithm is an example of an MM algorithm.

QA

Thanks for your attention!