
MM Algorithm

2012-9-25

Yijuan Hu

The MM algorithm is not an algorithm, but a **strategy** for constructing optimization algorithms.

An MM algorithm operates by creating a **surrogate function** that minorizes or majorizes the objective function. When the surrogate function is optimized, the objective function is driven uphill or downhill as needed.

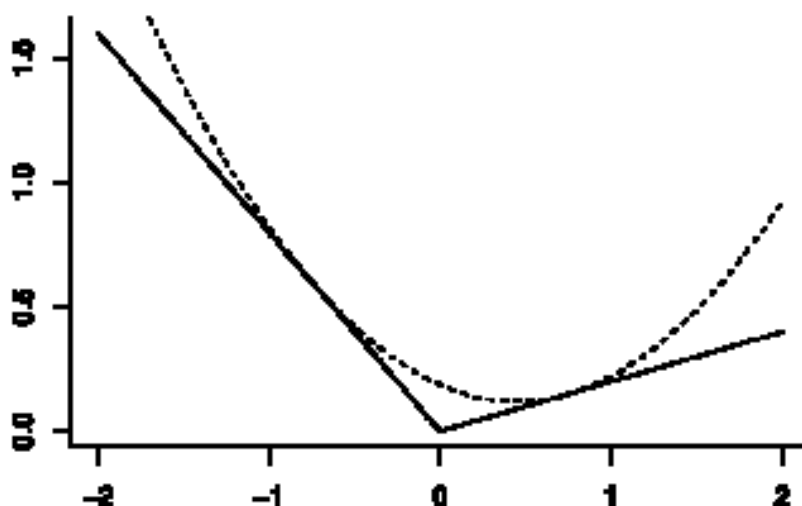
In minimization MM stands for **majorize/minimize**, and in maximization MM stands for **minorize/maximize**.

The **EM algorithm** from statistics is a special case.

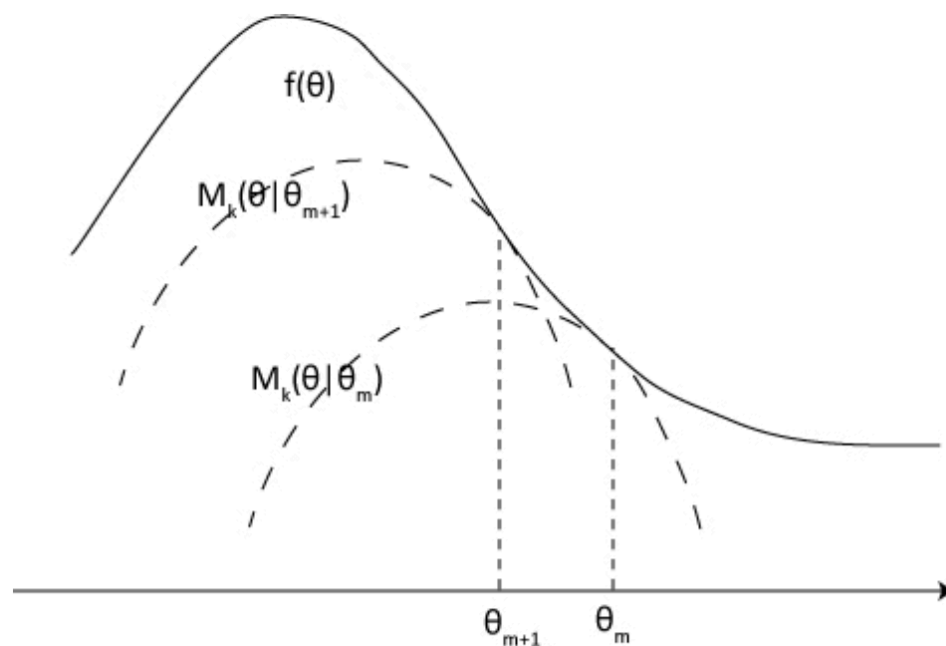
- We first focus on the **minimization** problem, in which MM = majorize/minimize.
- A function $g(\theta|\theta^{(k)})$ is said to **majorize** the function $f(\theta)$ at $\theta^{(k)}$ provided

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

- We choose a majorizing function $g(\theta|\theta^{(k)})$ and **minimize** it (rather than minimizing $f(\theta)$). Denote $\theta^{(k+1)} = \arg \min_{\theta} g(\theta|\theta^{(k)})$. Iterate until $\theta^{(k)}$ converges.
- **Descent property:** $f(\theta^{(k+1)}) \leq g(\theta^{(k+1)}|\theta^{(k)}) \leq g(\theta^{(k)}|\theta^{(k)}) = f(\theta^{(k)})$.



- In a **maximization** problem, MM = minorize/maximize.
- To maximize $f(\theta)$, we **minorize** it by a surrogate function $g(\theta|\theta^{(k)})$ and maximize $g(\theta|\theta^{(k)})$ to produce the next iterate $\theta^{(k+1)}$.
- A function $g(\theta|\theta^{(k)})$ is said to minorize the function $f(\theta)$ at $\theta^{(k)}$ provided that $-g(\theta|\theta^{(k)})$ majorizes $-f(\theta)$.



One of the key criteria in judging majorizing or minorizing functions is their **ease of optimization**

Successful MM algorithms in high-dimensional parameter spaces often rely on surrogate functions in which the individual parameter components are **separated**, i.e., for $\theta = (\theta_1, \dots, \theta_p)$,

$$g(\theta \mid \theta^{(k)}) = \sum_{j=1}^p q_j(\theta_j),$$

where $q_j(\cdot)$ are univariate functions.

Because the p univariate functions may be **optimized one by one**, this makes the surrogate function easier to optimize at each iteration.

- **Numerical stability:** warranted by the descent property
- **Simplicity:** substitute a simple optimization problem for a difficult optimization problem.
 - It can turn a non-differentiable problem into a smooth problem (Example 2).
 - It can separate the parameters of a problem (Example 3).
 - It can linearize an optimization problem (Example 3).
 - It can deal gracefully with equality and inequality constraints (Example 4).
 - It can generate an algorithm that avoids large matrix inversion (5).
- Iteration is the price we pay for simplifying the original problem.

- **(EM)** The E-step creates a surrogate function by identifying a complete-data log-likelihood function and evaluating it with respect to the observed data. The M-step maximizes the surrogate function. Every EM algorithm is an example of an MM algorithm.
- **(EM)** demands creativity in identifying the **missing data (complete data)** and technical skill in calculating an often complicated conditional expectation and then maximizing it analytically.
- **(MM)** pay attentions to the **convexity** of the objective function and **inequalities**.
- **(MM)** easier to understand and sometimes easier to apply than EM algorithms.

Inequalities to construct majorizing/minorizing function — 7/?? —

- **Property of convex function:** $\kappa(\theta)$ is called convex if for any $\theta_1, \theta_2, \lambda \in [0, 1]$

$$\kappa(\lambda\theta_1 + (1 - \lambda)\theta_2) \leq \lambda\kappa(\theta_1) + (1 - \lambda)\kappa(\theta_2)$$

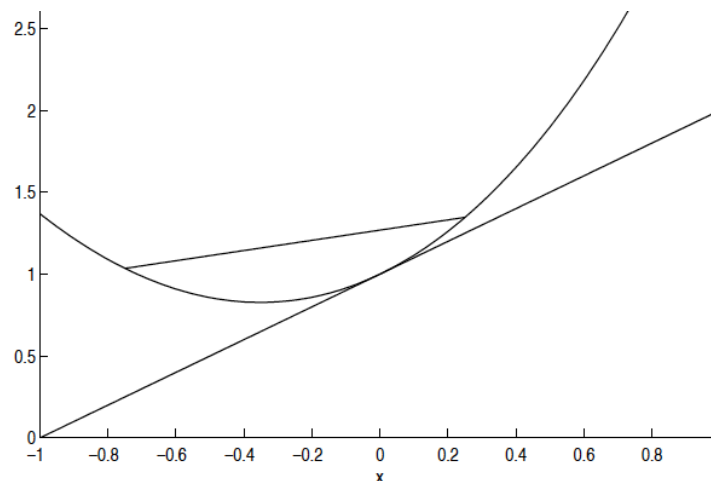
- **Jensen's Inequality:** For a convex function $\kappa(x)$ and any random variable X ,

$$\kappa[E(X)] \leq E[\kappa(X)]$$

- **Supporting hyperplanes:** If $\kappa(\cdot)$ is convex and differentiable, then

$$\kappa(\theta) \geq \kappa(\theta^{(k)}) + [\nabla\kappa(\theta^{(k)})]'(\theta - \theta^{(k)}),$$

and equality holds when $\theta = \theta^{(k)}$.



- **Quadratic upper bound:** Suppose a convex function $\kappa(\theta)$ is twice differentiable and have bounded curvature, so we can find a positive definite matrix M such that $M - \nabla^2 \kappa(\theta)$ is nonnegative definite. Then we can majorize $\kappa(\theta)$ by a quadratic function with sufficient high curvature and tangent to $\kappa(\theta)$ at $\theta^{(k)}$, i.e.,

$$\kappa(\theta) \leq \kappa(\theta^{(k)}) + [\nabla \kappa(\theta^{(k)})]' (\theta - \theta^{(k)}) + \frac{1}{2} (\theta - \theta^{(k)})' M (\theta - \theta^{(k)})$$

- **Arithmetic-Geometric Mean Inequality:** For nonnegative x_1, \dots, x_m ,

$$\sqrt[m]{\prod_{i=1}^m x_i} \leq \frac{1}{m} \sum_{i=1}^m x_i,$$

and the equality holds iff $x_1 = x_2 = \dots = x_m$.

- **Cauchy-Schwartz Inequality:** For p -vectors x and y ,

$$x'y \leq \|x\| \cdot \|y\|,$$

where $\|x\| = \sqrt{\sum_{i=1}^p x_i^2}$ is the norm of the vector.

- By **Jensen's inequality** and the convexity of the function $-\log(y)$, we have for probability densities $a(y)$ and $b(y)$ that

$$-\log \left\{ \mathbb{E} \left[\frac{a(Y)}{b(Y)} \right] \right\} \leq \mathbb{E} \left[-\log \frac{a(Y)}{b(Y)} \right].$$

- If Y has the density $b(y)$, then $\mathbb{E}[a(Y)/b(Y)] = 1$, so the left-hand side above vanishes and we obtain

$$\mathbb{E}[\log a(Y)] \leq \mathbb{E}[\log b(Y)],$$

which is sometimes known as the **information inequality** (Kullback-Leibler information).

- It is this inequality that guarantees that a minorizing function is constructed in the E-step of any EM algorithm, making every EM algorithm an MM algorithm.

- We have the **decomposition**

$$h^{(k)}(\theta) \equiv \mathbb{E}\{\log f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)|Y_{\text{obs}}, \theta^{(k)}\} = \mathbb{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta)|Y_{\text{obs}}, \theta^{(k)}\} + \log g(Y_{\text{obs}}|\theta)$$

- By the **information inequality**,

$$\mathbb{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta)|Y_{\text{obs}}, \theta^{(k)}\} \leq \mathbb{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})|Y_{\text{obs}}, \theta^{(k)}\}$$

- We obtain the **surrogate function** that minorizes the objective function

$$\log g(Y_{\text{obs}}|\theta) \geq h^{(k)}(\theta) - \mathbb{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})|Y_{\text{obs}}, \theta^{(k)}\} \quad (1)$$

- **Note:** The second term of (1) does not depend on θ .

- Consider the sequence of numbers y_1, \dots, y_n . The sample median θ minimizes the **non-differentiable objective function**

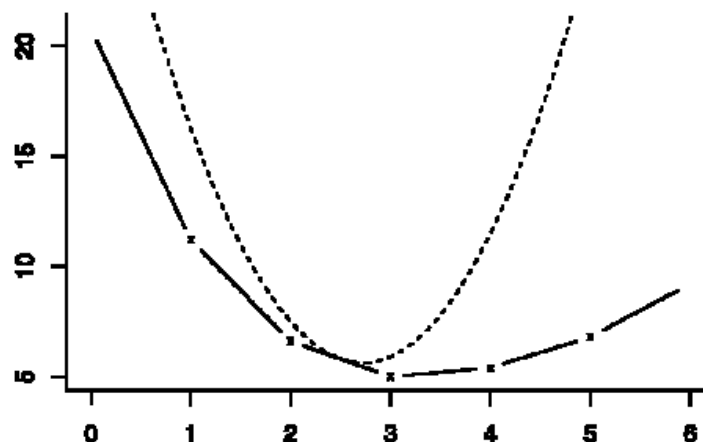
$$f(\theta) = \sum_{i=1}^n |y_i - \theta|.$$

- The **quadratic function**

$$h_i(\theta|\theta^{(k)}) = \frac{(y_i - \theta)^2}{2|y_i - \theta^{(k)}|} + \frac{1}{2}|y_i - \theta^{(k)}|$$

majorizes $|y_i - \theta|$ at the point $\theta^{(k)}$.

- Hence, $g(\theta|\theta^{(k)}) = \sum_{i=1}^n h_i(\theta|\theta^{(k)})$ majorizes $f(\theta)$.
- Note:** The majorization relation between functions is **closed** under the formation of sums, nonnegative products, limits, and composition with an increasing function.



- The **minimum** of the sum of the quadratic

$$g(\theta|\theta^{(k)}) = \frac{1}{2} \sum_{i=1}^n \left[\frac{(y_i - \theta)^2}{|y_i - \theta^{(k)}|} + |y_i - \theta^{(k)}| \right]$$

occurs at

$$\theta^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} y_i}{\sum_{i=1}^n w_i^{(k)}}$$

for $w_i^{(k)} = |y_i - \theta^{(k)}|^{-1}$.

- This algorithm works except when a weight $w_i^{(k)} = \infty$. It generalizes to **sample quantiles**, **least L1 regression** and **quantile regression**.

- Consider a sports league with n teams. Assign team i the skill level θ_i , where $\theta_1 = 1$ for identifiability reason. Bradley and Terry proposed the model

$$\Pr(i \text{ beats } j) = \frac{\theta_i}{\theta_i + \theta_j}.$$

- If b_{ij} is the number of times i beats j , then the likelihood of the data is

$$L(\theta) = \prod_{i \neq j} \left(\frac{\theta_i}{\theta_i + \theta_j} \right)^{b_{ij}}.$$

We estimate θ by **maximizing** $f(\theta) = \ln L(\theta)$ and then rank the teams on the basis of the estimates.

- The form of the log-likelihood

$$f(\theta) = \sum_{i \neq j} b_{ij} [\ln \theta_i - \ln(\theta_i + \theta_j)]$$

suggests that we **linearize** on the term $-\ln(\theta_i + \theta_j)$ to **separate parameters**.

- By the **supporting hyperplane property** and the convexity of $-\ln(\cdot)$,

$$-\ln y \geq -\ln x - x^{-1}(y - x).$$

- The above inequality produces the **surrogate**

$$g(\theta|\theta^{(k)}) = \sum_{i \neq j} b_{ij} \left[\ln \theta_i - \ln(\theta_i^{(k)} + \theta_j^{(k)}) - \frac{\theta_i + \theta_j}{\theta_i^{(k)} + \theta_j^{(k)}} + 1 \right].$$

- Because the parameters are **separated**, we can easily find the optimal point

$$\theta_i^{(k+1)} = \frac{\sum_{i \neq j} b_{ij}}{\sum_{i \neq j} (b_{ij} + b_{ji}) / (\theta_i^{(k)} + \theta_j^{(k)})}.$$

- Consider the problem of **minimizing** $f(\theta)$ subject to the **constraints** $v_j(\theta) \geq 0$ for $1 \leq j \leq q$, where each $v_j(\theta)$ is a concave, differentiable function.

- By the **supporting hyperplane property** and the convexity of $-v_j(\theta)$,

$$v_j(\theta^{(k)}) - v_j(\theta) \geq \left[\nabla v_j(\theta^{(k)}) \right]' (\theta^{(k)} - \theta). \quad (2)$$

- Again, by the **supporting hyperplane property** and the convexity of $-\ln(\cdot)$, we have $-\ln y + \ln x \geq -x^{-1}(y - x)$ and then

$$v_j(\theta^{(k)}) \left[-\ln v_j(\theta) + \ln v_j(\theta^{(k)}) \right] \geq v_j(\theta^{(k)}) - v_j(\theta). \quad (3)$$

- By (2) and (3),

$$v_j(\theta^{(k)}) \left[-\ln v_j(\theta) + \ln v_j(\theta^{(k)}) \right] + \left[\nabla v_j(\theta^{(k)}) \right]' (\theta - \theta^{(k)}) \geq 0,$$

and the equality holds when $\theta = \theta^{(k)}$.

- Summing over j and multiplying by a positive tuning parameter ω , we construct the **surrogate function** that majorizes $f(\theta)$,

$$g(\theta|\theta^{(k)}) = f(\theta) + \omega \sum_{j=1}^q \left[v_j(\theta^{(k)}) \ln \frac{v_j(\theta^{(k)})}{v_j(\theta)} + [\nabla v_j(\theta^{(k)})]' (\theta - \theta^{(k)}) \right] \geq f(\theta)$$

- **Note:**

- **Majorization disposes of the inequality constraints.**
- The presence of $\ln v_j(\theta)$ ensures $v_j(\theta) \geq 0$.
- An initial point $\theta^{(0)}$ must be selected with all inequality constraints strictly satisfied. All iterates stay within the interior region but allows strict inequalities to become equalities in the limit.
- The minimization step of the MM algorithm can be carried out approximately by **Newton's method**.
- Where there are linear equality constraints $A\theta = b$ in addition to the inequality constraints $v_j(\theta) \geq 0$, these should be enforced by introducing **Lagrange multipliers** during the minimization of $g(\theta|\theta^{(k)})$.

- We have an $n \times 1$ vector Y of binary responses and an $n \times p$ matrix X of predictors. The logistic regression model assumes that

$$\pi_i(\theta) \equiv \Pr(Y_i = 1) = \frac{\exp(\theta' x_i)}{1 + \exp(\theta' x_i)}.$$

Then the **log likelihood** is

$$l(\theta) \equiv \sum_{i=1}^n Y_i \theta' x_i - \sum_{i=1}^n \log \{1 + \exp(\theta' x_i)\}.$$

- The **Hessian** can be obtained by direct differentiation:

$$\nabla^2 l(\theta) = - \sum_{i=1}^n \pi_i(\theta) [1 - \pi_i(\theta)] x_i x_i'. \quad (4)$$

- Since $\pi_i(\theta) [1 - \pi_i(\theta)]$ is bounded above by $1/4$, we may define the negative definite matrix $M = -1/4 X'X$ such that $\nabla^2 l(\theta) - M$ is nonnegative definite. Thus,

$$g(\theta|\theta^{(k)}) = l(\theta^{(k)}) + [\nabla l(\theta^{(k)})]' (\theta - \theta^{(k)}) + \frac{1}{2}(\theta - \theta^{(k)})' M (\theta - \theta^{(k)})$$

is a **quadratic lower bound** of $l(\theta)$.

- The MM algorithm proceeds by **maximizing** $g(\theta|\theta^{(k)})$, giving

$$\begin{aligned}\theta^{(k+1)} &= \theta^{(k)} - M^{-1} \nabla l(\theta^{(k)}) \\ &= \theta^{(k)} + 4(X'X)^{-1} X' [Y - \pi(\theta^{(k)})].\end{aligned}$$

- Computational advantage of the MM algorithm over Newton-Raphson
 - **MM**: invert $X'X$ **only once**.
 - **NR**: invert the Hessian (??) **for every iteration**.

- **Convergence rate**

- NR: a quadratic rate $\lim \|\theta^{(k+1)} - \hat{\theta}\| / \|\theta^{(k+1)} - \hat{\theta}\|^2 = c$ (constant)
- MM: a linear rate $\lim \|\theta^{(k+1)} - \hat{\theta}\| / \|\theta^{(k+1)} - \hat{\theta}\| = c < 1$

- **Complexity of each iteration**

- NR: require evaluation and inversion of Hessian, $O(p^3)$
- MM: separates parameters, $O(p)$ or $O(p^2)$

- **Stability of the algorithm**

- NR: behave poorly if started too far from an optimum point
- MM: guaranteed to increase/decrease the objective function at every iteration

In conclusion, well-designed MM algorithms tend to require more iterations but simpler iterations than Newton-Raphson; thus MM sometimes enjoy an advantage in computation speed. Numerical stability also enters the balance sheet.

- Quantile regression (Hunter and Lange, 2000)
- Survival analysis (Hunter and Lange, 2002)
- Paired and multiple comparisons (Hunter 2004)
- Variable selection (Hunter and Li, 2002)
- DNA sequence analysis (Sabatti and Lange, 2002)