The MM Algorithm

Kenneth Lange

Departments of Biomathematics,
Human Genetics, and Statistics

UCLA

April, 2007

Overview of the MM Algorithm

1. The MM algorithm is not an algorithm, but a prescription for constructing optimization algorithms.

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

3. 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.

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

History of the MM Algorithm

1. Ortega and Rheinboldt (1970) enunciate the principle in the context of line search methods.

2. de Leeuw and Heiser (1977) present an MM algorithm for multidimensional scaling contemporary with the classic Dempster et al. (1977) paper on EM algorithms.

3. Subsequent appearances: robust regression (Huber, 1981), quadratic lower bound principle (Böhning & Lindsay,1988), medical imaging (Lange et al, 1987, De Pierro, 1995; Lange & Fessler, 1995), quantile regression (Hunter & Lange, 2000), survival analysis (Hunter & Lange, 2002), paired and multiple comparisons (Hunter, 2004), variable selection (Hunter & Li, 2002), DNA sequence analysis (Sabatti & Lange, 2002), and discriminant analysis (Lange & Wu, 2006).

Rationale for the MM Principle

1. It can generate an algorithm that avoids matrix inversion.

2. It can separate the parameters of a problem.

3. It can linearize an optimization problem.

4. It can deal gracefully with equality and inequality constraints.

5. It can turn a non-differentiable problem into a smooth problem.

Majorization and Definition of the Algorithm

1. A function $g(\theta \mid \theta^n)$ is said to majorize the function $f(\theta)$ at θ^n provided

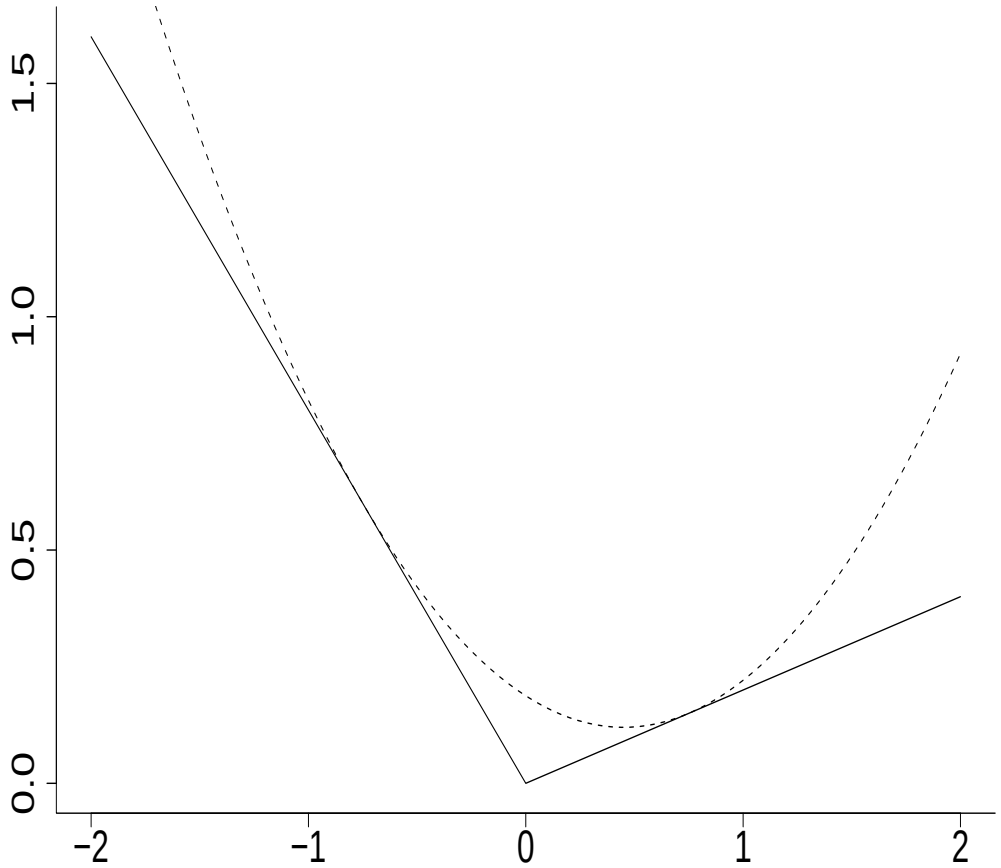
$$\begin{aligned} f(\theta^n) &= g(\theta^n \mid \theta^n) \\ f(\theta) &\leq g(\theta \mid \theta^n) \qquad \text{for all } \theta. \end{aligned}$$

   The majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function.

2. A function $g(\theta \mid \theta^n)$ is said to minorize the function $f(\theta)$ at θ^n provided $-g(\theta \mid \theta^n)$ majorizes $-f(\theta)$.

3. In minimization, we choose a majorizing function $g(\theta \mid \theta^n)$ and minimize it. This produces the next point θ^{n+1} in the algorithm.

A Quadratic Majorizer



Descent Property

1. An MM minimization algorithm satisfies the descent property $f(\theta^{n+1}) \leq f(\theta^n)$ with strict inequality unless both

$$\begin{aligned} g(\theta^{n+1} \mid \theta^n) &= g(\theta^n \mid \theta^n) \\ f(\theta^{n+1}) &= g(\theta^{n+1} \mid \theta^n). \end{aligned}$$

2. The descent property follows from the definitions and

$$\begin{aligned} f(\theta^{n+1}) &= g(\theta^{n+1} \mid \theta^n) + f(\theta^{n+1}) - g(\theta^{n+1} \mid \theta^n) \\ &\leq g(\theta^n \mid \theta^n) + f(\theta^n) - g(\theta^n \mid \theta^n) \\ &= f(\theta^n). \end{aligned}$$

3. The descent property makes the MM algorithm very stable.

Generic Methods of Majorization and Minorization

1. Jensen's inequality — EM algorithms

2. Chord above the graph property of a convex function — image reconstruction

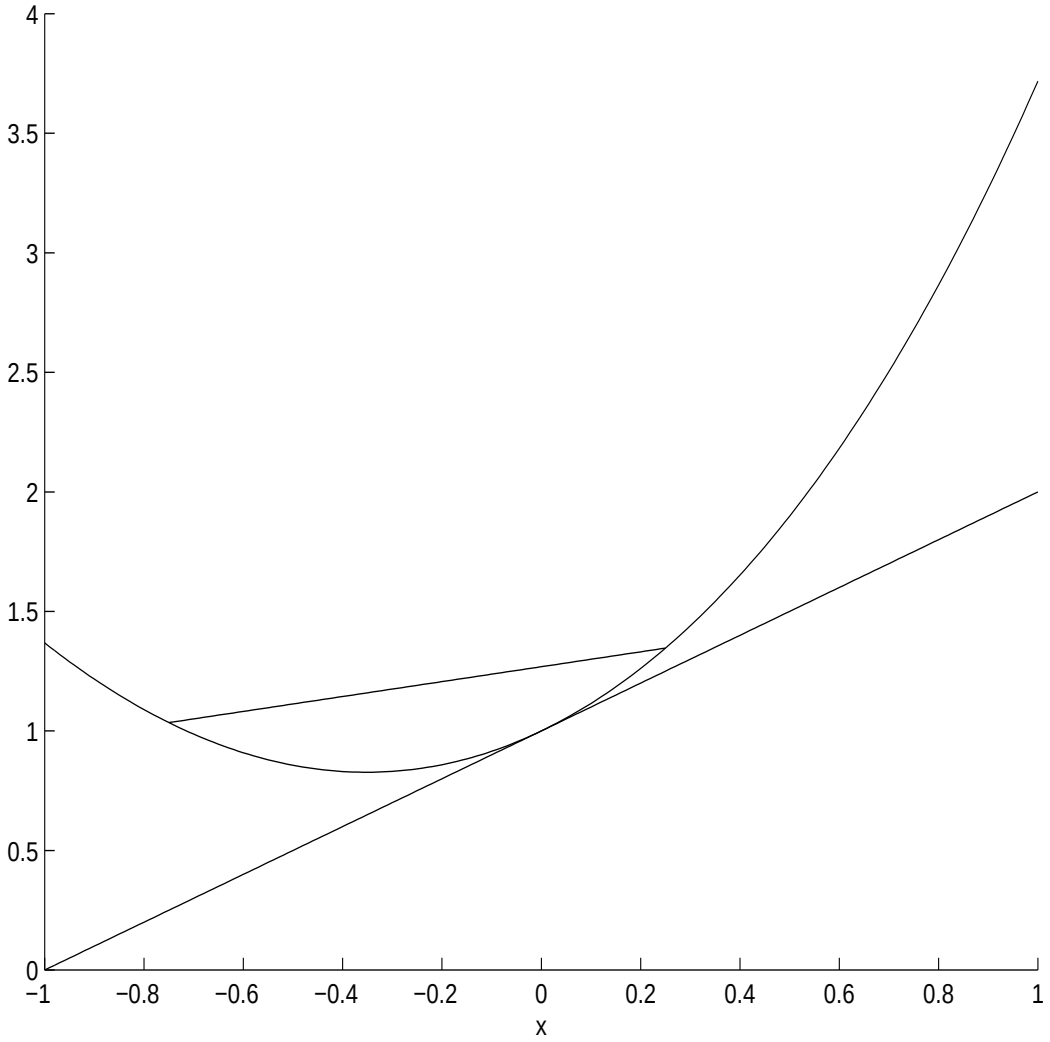
3. Supporting hyperplane property of a convex function

4. Quadratic upper bound principle — Böhning and Lindsay

5. Arithmetic-geometric mean inequality

6. The Cauchy-Schwartz inequality — multidimensional scaling

Chord and Supporting Hyperplane Properties



Example 1: Finding a Sample Median

1. Consider the sequence of numbers $y_1, \ldots, y_m$. The sample median θ minimizes the non-differentiable criterion

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

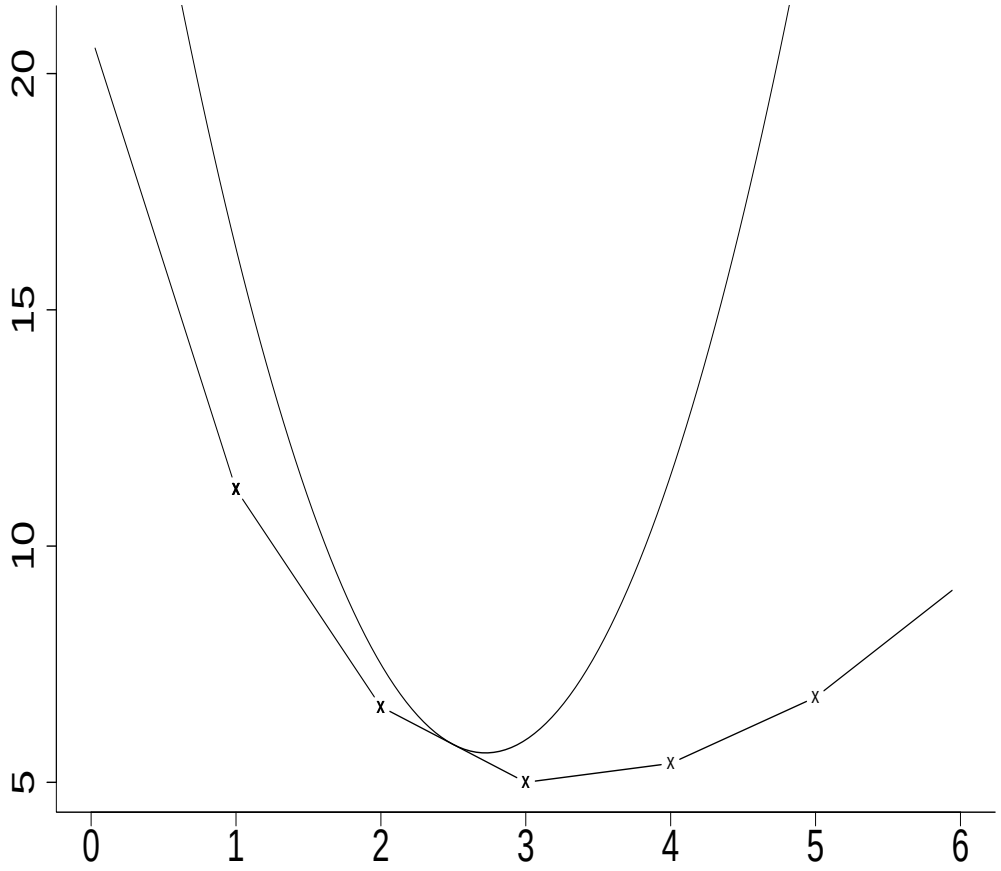
2. The quadratic function

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

majorizes $|y_i - \theta|$ at the point θ^n.

3. Hence, $g(\theta \mid \theta^n) = \sum_{i=1}^{n} h_i(\theta \mid \theta^n)$ majorizes $f(\theta)$.

A Sum of Quadratic Majorizers



Example 1: MM Algorithm

1. The minimum of the quadratic

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

   occurs at

$$\theta^{n+1} \;=\; \frac{\sum_{i=1}^{n} w_i^n y_i}{\sum_{i=1}^{n} w_i^n}$$

   for $w_i^n = |y_i - \theta^n|^{-1}$.

2. The algorithm works except when a weight $w_i^n = \infty$. It generalizes to sample quantiles, to least L_1 regression, and to quantile regression.

Example 2: Bradley-Terry Ranking

1. Consider a sports league with m teams. Assign team i the skill level θ_i. Bradley and Terry proposed the model

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

2. To ensure that the skill levels are identifiable, set $\theta_1 = 1$.

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

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

Example 2: Loglikelihood

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

2. The form of the loglikelihood
$$f(\theta) \;=\; \sum_{i,j} b_{ij} \left[\ln \theta_i - \ln(\theta_i + \theta_j)\right]$$
suggests that we work on the term $-\ln(\theta_i + \theta_j)$ if we want to separate parameters.

3. Hence we apply the supporting hyperplane property
$$h(y) \;\geq\; h(x) + \nabla h(x)^t (y - x),$$
of a convex function $h(x)$. For the choice $h(x) = -\ln x$, this minorization amounts to $-\ln y \geq -\ln x - \frac{1}{x}(y - x)$.

Example 2: Minorization

1. The minorization $-\ln y \geq -\ln x - (y-x)x^{-1}$ produces the surrogate

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

2. Because the parameters are separated, the optimal point

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

   is easy to find.

3. Under natural assumptions, these MM iterates converge to the unique maximum likelihood point.

Example 3: Random Graph Model

1. Random graphs provide interesting models of connectivity in genetics and internet node ranking.

2. In a simplified version of the Chatterjee and Diaconis model, we assign a propensity $p_i \in [0, 1]$ to each node i.

3. An edge between nodes i and j then forms independently with probability $p_i p_j$. In other words, for a handshake to occur, both parties must agree.

4. The most obvious statistical question in the model is how to estimate the p_i from data. Once this is done, we can rank nodes by their estimated propensities.

Example 3: Loglikelihood

1. If E denotes the edge set of the graph, then the loglikelihood can be written as

$$L(p) = \sum_{\{i,j\}\in E} \left[\ln p_i + \ln p_j\right] + \sum_{\{i,j\}\notin E} \ln(1 - p_i p_j). \quad (1)$$

   Here $\{i, j\}$ denotes a generic unordered pair.

2. The logarithms $\ln(1 - p_i p_j)$ are the bothersome terms in the loglikelihood.

3. We will minorize each of these by exploiting the concavity of the function $\ln(1 - x)$.

Example 3: Two Successive Minorizations

1. Using the concavity of $\ln(1-x)$ gives the minorization

$$\ln(1-p_ip_j) \;\geq\; \ln(1-p_{ni}p_{nj}) - \frac{1}{1-p_{ni}p_{nj}}(p_ip_j - p_{ni}p_{nj})$$

   and eliminates the logarithm.

2. This minorization is not quite good enough to separate parameters, however. Separation can be achieved by invoking the second minorizing inequality

$$-p_ip_j \;\geq\; -\frac{1}{2}(\frac{p_{nj}}{p_{ni}}p_i^2 + \frac{p_{ni}}{p_{nj}}p_j^2).$$

   Note again that equality holds when all $p_i = p_{ni}$.

Example 3: MM Algorithm

1. It follows that $L(p)$ is minorized by the function

$$g(p \mid p_n) = \sum_{\{i,j\} \in E} \left[\ln p_i + \ln p_j \right] + \sum_{\{i,j\} \notin E} \left[\ln(1 - p_{ni}p_{nj}) - \frac{1}{1 - p_{ni}p_{nj}} \frac{1}{2} \left(\frac{p_{nj}}{p_{ni}} p_i^2 + \frac{p_{ni}}{p_{nj}} p_j^2 \right) \right].$$

2. If we set $\frac{\partial}{\partial p_i} g(p \mid p_n) = 0$ and let d_i denote the degree of node i, then the MM update is

$$p_{n+1,i} = \sqrt{p_{ni}} \left[\frac{d_i}{\sum_{\{i,j\} \notin E} \frac{p_{nj}}{1 - p_{ni}p_{nj}}} \right]^{1/2}. \qquad (2)$$

Example 3: Comments on the MM Algorithm

1. If by chance $p_{n+1,i} > 1$, then it is prudent to set $p_{n+1,i} = 1-\epsilon$ for $\epsilon > 0$ very small.

2. When $d_i = 0$, the MM algorithm makes the sensible choice $p_{n+1,i} = 0$.

3. Step doubling produces an algorithm close to the Chatterjee and Diaconis algorithm for estimating the propensities.

4. This derivation extends to random directed graphs and generates MM updates for ingoing and outgoing propensities.

Example 4: Transmission Tomography

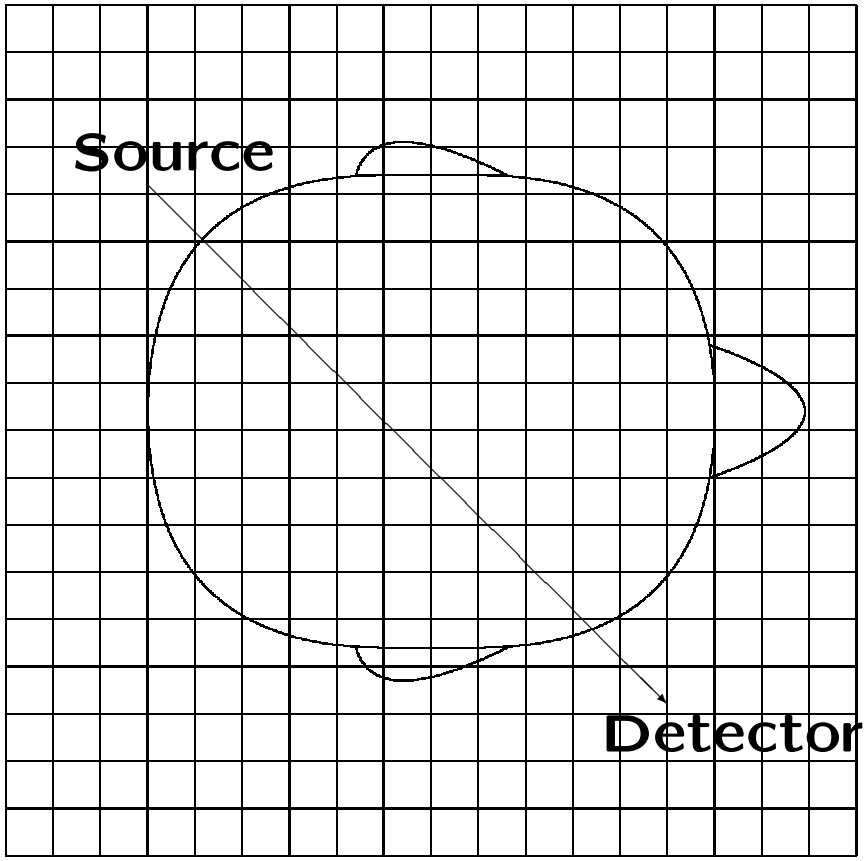
1. In transmission tomography, high energy photons are beamed from an external source through the body to an external detector.

2. The plane region of an X-ray slice is divided into small pixels, and pixel j is assigned attenuation coefficient θ_j.

3. The number of photons beamed along projection i (line of flight) is Poisson distributed with mean d_i. Transmitted counts y_i for different projections are independent.

4. A photon entering pixel j along projection i successfully traverses the pixel with Poisson probability $e^{-l_{ij}\theta_j}$, where l_{ij} is the intersection length of the projection and pixel.

Example 4: Cartoon of Transmission Tomography



Example 4: Loglikelihood

1. The probability that a photon transmitted along projection i is detected is given by the exponentiated line integral $e^{-\langle l_i, \theta \rangle}$, where $\langle l_i, \theta \rangle = \sum_j l_{ij} \theta_j$.

2. The loglikelihood under the model for transmission tomography is

$$f(\theta) = \sum_i f_i(\langle l_i, \theta \rangle) = \sum_i \left[-d_i e^{-\langle l_i, \theta \rangle} + y_i \ln d_i - y_i \langle l_i, \theta \rangle - \ln y_i! \right].$$

3. Note that the function $f_i(s) = -d_i e^{-s} - ys + c_i$ is strictly concave.

Example 4: Minorization

1. Now consider the composition of a concave function $f_i(s)$ and a linear function $\langle l_i, \theta \rangle$, and set

$$\alpha_{ij}^n = \frac{l_{ij}\theta_j^n}{\langle l_i, \theta^n \rangle}.$$

   Because $f_i(s)$ is concave,

$$f_i(\langle l_i, \theta \rangle) = f_i\left(\sum_j \alpha_{ij}^n \frac{\theta_j}{\theta_j^n}\langle l_i, \theta^n \rangle\right) \geq \sum_j \alpha_{ij}^n f_i\left(\frac{\theta_j}{\theta_j^n}\langle l_i, \theta^n \rangle\right).$$

   with equality when $\theta = \theta^n$.

2. This term-by-term minorization for each projection yields an overall minorization $g(\theta \mid \theta^n)$ that separates the parameters.

Example 4: MM Algorithm

1. Maximization of the surrogate function can be accomplished by applying Newton's method parameter by parameter.

2. This treatment omits smoothing terms. These also can be minorized to give a surrogate function with separated parameters.

3. The Poisson model accounts for random variation and allows simultaneous smoothing.

4. The images produced by the Poisson model and the MM algorithm are superior to images produced by standard Fourier methods using the Radon transform. Thus, patient X-ray doses can be lowered without compromising image quality.

Example 5: Machine Learning Discriminant Analysis

1. To discriminate among $k+1$ categories, choose $k+1$ vectors in R^k to form the vertices of a regular tetrahedron. Each training case i is assigned to a vertex via its indicator u_i.

2. For p observed cases, define the regularized risk function

$$R(A,b) \;=\; \sum_{i=1}^{p} \|u_i - Ax_i - b\|_\epsilon + \lambda \sum_{j=1}^{k} \|a_j\|^2,$$

where a_j^t is the jth row of a $k \times q$ matrix A of regression coefficients, b is a $k \times 1$ column vector of intercepts, and $\|v\|_\epsilon = \max\{\|v\| - \epsilon, 0\}$ denotes ϵ-insensitive distance on R^k for a fixed $\epsilon > 0$. Estimate A and b by minimizing $R(A,b)$.

Example 5: Rationale for Risk Function

1. Choosing the vertices of a regular tetrahedron makes all vertices equidistant.

2. Euclidean distance is less sensitive to large residuals than Euclidean distance squared.

3. In predicting membership, it does not make much difference how close the linear predictor is to the artificial indicator u_i. Hence, ϵ-insensitive distance.

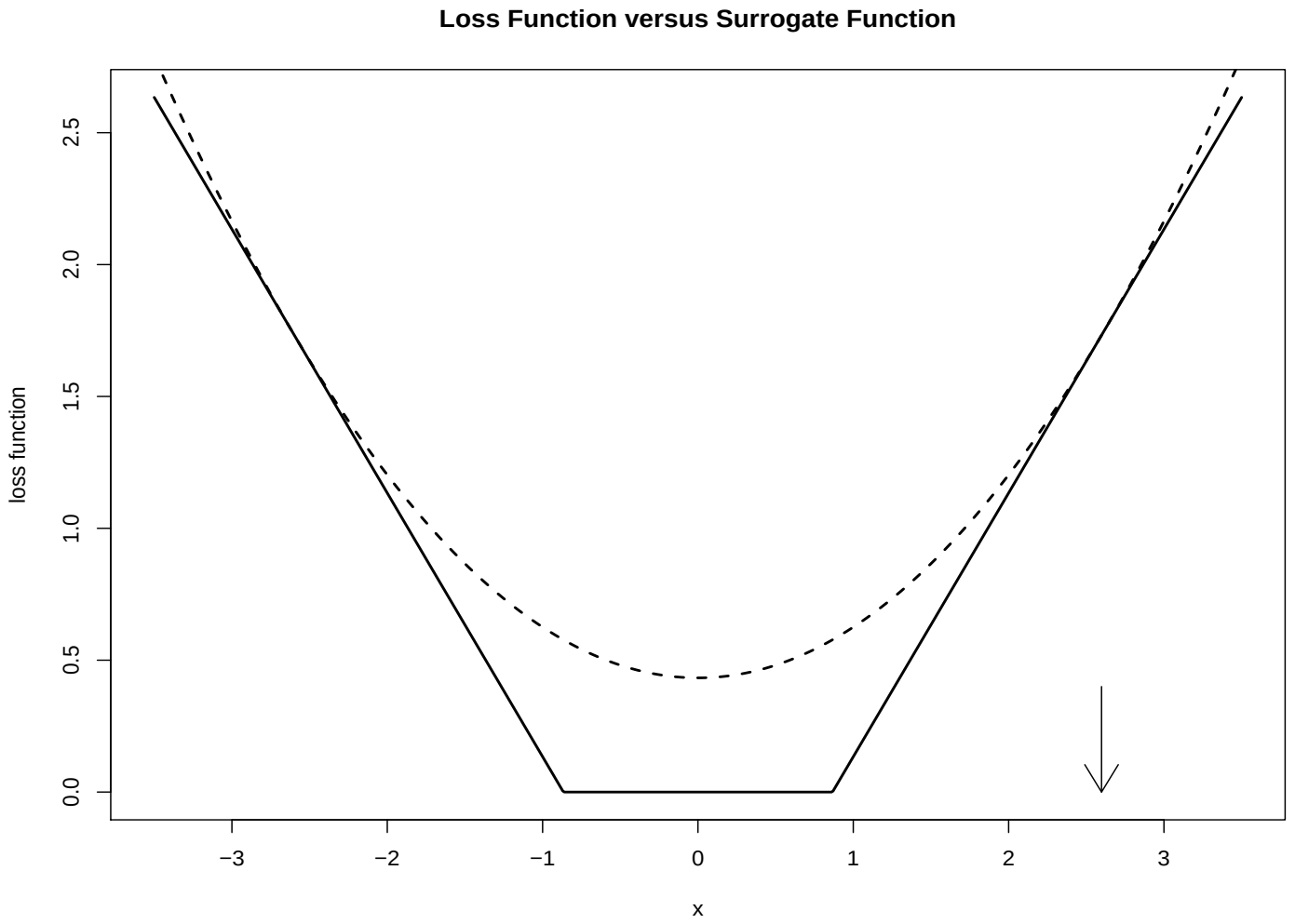
4. Unless the number of cases is much larger than the number of features, estimates of the regression coefficients tend to be poor. Hence the need for regularization.

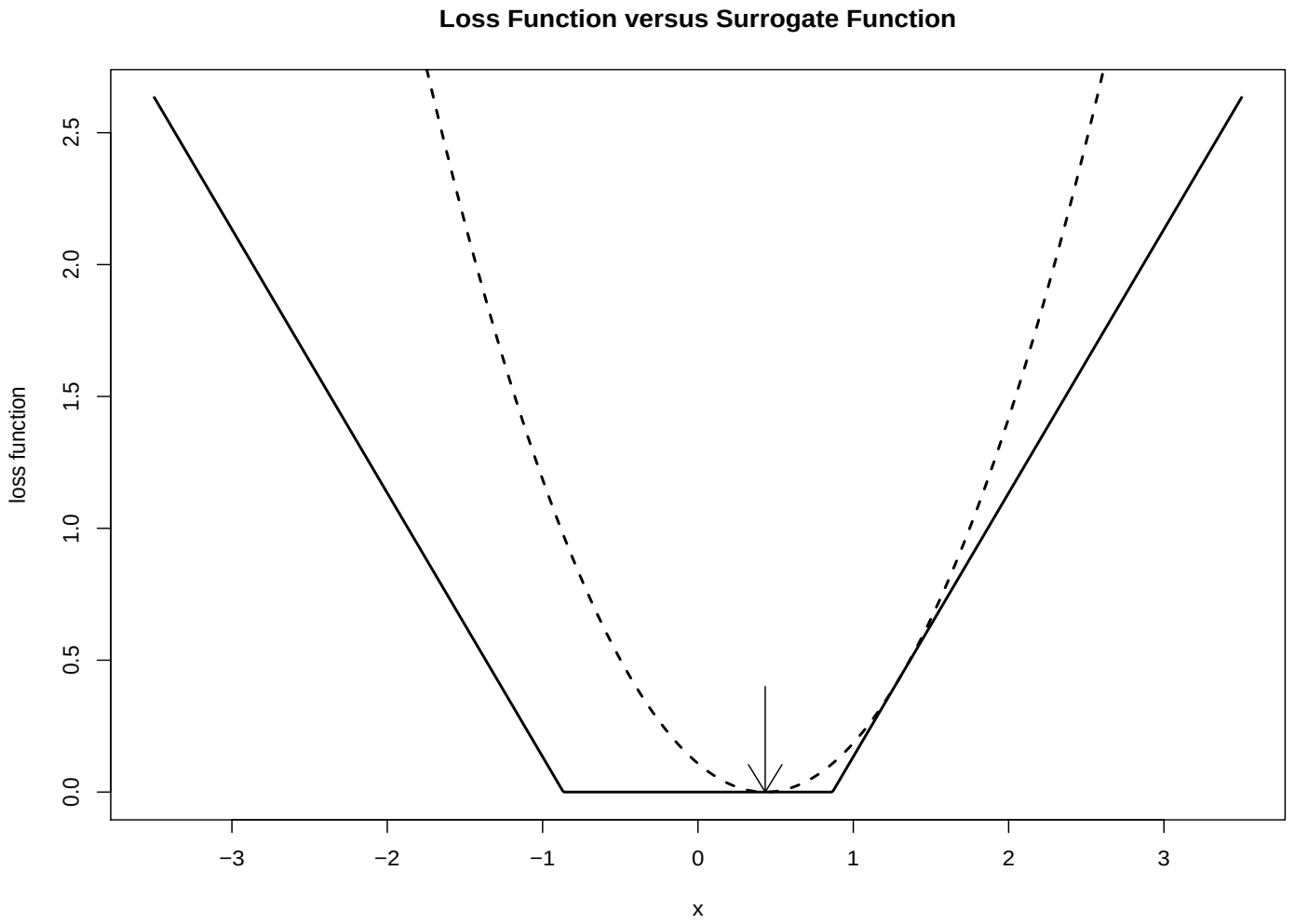
Example 5: Quadratic Majorization of $\|y\|_\epsilon$

Repeated application of the Cauchy-Schwarz inequality produces the majorizer

$$q(x \mid y) = \begin{cases} \frac{1}{2\|y\|}\|x\|^2 + \frac{1}{2}\|y\| - \epsilon & \|y\| \geq 2\epsilon \\ \frac{1}{4(\epsilon - \|y\|)}\|x - y\|^2 & \|y\| < \epsilon \\ \frac{1}{4(\epsilon - \|z\|)}\|x - z\|^2 & \epsilon < \|y\| < 2\epsilon \end{cases}$$

of $\|y\|_\epsilon$, where in the last case $z = cy$ with the positive contraction constant c chosen so that $\|z\| = 2\epsilon - \|y\|$. There is no majorization in the anomalous situation $\|y\| = \epsilon$.


Loss Function versus Surrogate Function
loss function
0.0
0.5
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3
x


Loss Function versus Surrogate Function
loss function
0.0
0.5
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3
x

Example 5: Performance of the Discriminant Method

1. On standard test problems, the machine learning method performs well compared to other methods of discriminant analysis.

2. The MM algorithm is much simpler to program then competing algorithms that solve the dual problem to risk function minimization. For $k+1$ categories, each MM update involves solving k different weighted least squares problems.

3. The MM algorithm is reasonably fast. Step doubling halves the number of iterations until convergence.

Error Rates for Examples from the UCI Data Repository

Method	Wine	Glass	Zoo	Lymphography
TDA	0	0.2970	0	0.0541
LDA	0.0112	0.4065	NA	0.0878
QDA	0.0169	NA	NA	NA
KNN	0.0506	0.2991	0.0792	0.1351
OVR	0.0169	0.3458	0.0099	0.0541
MSVM	0.0169	0.3645	NA	NA
AltMSVM	0.0169	0.3170	NA	NA
CART	0.1404	0.1449	0.1683	0.1351
Random forest	0.0674	0.1589	0.0297	0.0135

1. For the wine and glass data, the error rates are average misclassification rates based on 10-fold cross validation.

2. The error rates for the zoo and lymphography data refer to training errors.

Example 6: Convex Programming

1. In convex programming we minimize a smooth function $f(\theta)$ subject to concave inequality constraints $v_j(\theta) \geq 0$.

2. In view of the supporting hyperplane property,

$$g(\theta \mid \theta^n) \;=\; f(\theta) + \omega \sum_j [v_j(\theta^n) \ln \frac{v_j(\theta^n)}{v_j(\theta)} + \nabla v_j(\theta^n)^t(\theta - \theta^n)].$$

   majorizes $f(\theta)$. Here ω is any positive constant.

3. The presence of the term $\ln v_j(\theta)$ in $g(\theta \mid \theta^n)$ works as a barrier to prevent the event $v_j(\theta^{(m+1)}) \leq 0$ from occurring.

4. The multiplier $v_j(\theta^n)$ of $\ln v_j(\theta)$ gradually adapts and allows $v_j(\theta^{n+1})$ to tend to 0 if it is inclined to do so.

Example 6: Implementation of the MM Algorithm

1. The MM minimization algorithm must start with a point in the interior of the feasible region. All iterates stay within the interior region. Inequality constraints are implicit.

2. The minimization step of the MM algorithm can be carried out approximately by Newton's method.

3. When there are linear equality constraints $A\theta = b$ in addition to the inequality constraints $v_j(\theta) \geq 0$, these should be enforced during the minimization of $g(\theta \mid \theta^n)$. Majorization disposes of the inequality constraints.

Example 6: Derivation of the Majorizer

1. Since $-v_j(\theta)$ is convex,
$$-v_j(\theta) + v_j(\theta^n) \;\geq\; -\nabla v_j(\theta^n)^t(\theta - \theta^n).$$

2. Because $-\ln y + \ln x \geq -(y-x)x^{-1}$,
$$v_j(\theta^n)\left[-\ln v_j(\theta) + \ln v_j(\theta^n)\right] \;\geq\; v_j(\theta^n) - v_j(\theta).$$

3. Adding the last two inequalities, we see that
$$v_j(\theta^n)\left[-\ln v_j(\theta) + \ln v_j(\theta^n)\right] + \nabla v_j(\theta^n)^t(\theta - \theta^n) \;\geq\; 0,$$
with equality when $\theta = \theta^n$.

4. Summing the last inequality over j and multiplying by the positive constant ω gives the majorizer.

Example 6: Geometric Programming

Consider minimizing $f(x) = (x_1x_2x_3)^{-1} + x_2x_3$ subject to the constraint $v(x) = 4 - 2x_1x_3 - x_1x_2 \geq 0$ for positive values of the x_i. Under the change of variables $x_i = e^{\theta_i}$, the program is convex, and the MM algorithm with $\omega = 1$ gives:

Iteration $\mathbf{n}$	$\mathbf{f(x^n)}$	$\mathbf{x_1^n}$	$\mathbf{x_2^n}$	$\mathbf{x_3^n}$
1	2.0000	1.0000	1.0000	1.0000
2	1.6478	1.5732	1.0157	0.6065
3	1.5817	1.7916	0.9952	0.5340
4	1.5506	1.8713	1.0011	0.5164
5	1.5324	1.9163	1.0035	0.5090
10	1.5040	1.9894	1.0011	0.5008
15	1.5005	1.9986	1.0002	0.5001
20	1.5001	1.9998	1.0000	0.5000
25	1.5000	2.0000	1.0000	0.5000

Local Convergence of an MM Algorithm

1. The MM map $M(\theta)$ gives the unique point $M(\theta^n) = \theta^{n+1}$ that optimizes $g(\theta \mid \theta^n)$.

2. At the optimal point $\widehat{\theta}$, one can show that $M(\theta)$ has differential

$$dM(\widehat{\theta}) \;=\; -d^2 g(\widehat{\theta} \mid \widehat{\theta})^{-1} \left[d^2 f(\widehat{\theta}) - d^2 g(\widehat{\theta} \mid \widehat{\theta}) \right]$$

   in the absence of constraints.

3. The linear rate of convergence depends on the largest eigenvalue of the differential. All eigenvalues lie on [0,1).

4. In a practical sense, the rate of convergence depends on how well the surrogate function $g(\theta \mid \theta^n)$ approximates $f(\theta)$.

Global Convergence of an MM Algorithm

1. If an objective function is strictly convex or concave, then the MM algorithm will converge to the unique optimal point, assuming it exists.

2. If convexity or concavity fail, but all stationary points are isolated, then the MM algorithm will converge to one of them.

3. This point can be a local optimum, or in unusual circumstances, even a saddle point.

4. There exists methods for accelerating the convergence of MM and EM algorithms.

Remaining Challenges

1. Devise new MM algorithms, particularly for high dimensional problems.

2. Quantify the local rate of convergence of the MM algorithm in the presence of inequality constraints. When does an MM algorithm converge at a sublinear rate?

3. Estimate the computational complexity of the convex programming and other MM algorithms.

4. Devise better ways of accelerating MM and EM algorithms.

References

1. de Leeuw J, Heiser WJ (1977), Convergence of correction matrix algorithms for multidimensional scaling. *Geometric Representations of Relational Data* (editors Lingoes JC, Roskam E , Borg I), pp. 735–752, Mathesis Press

2. Hunter DR, Lange K (2004) A tutorial on MM algorithms. *Amer Statistician* 58:30–37

3. Lange, K (2004) *Optimization*. Springer-Verlag, New York

4. Lange K, Hunter DR, Yang I (2000) Optimization transfer using surrogate objective functions (with discussion). *J Comput Graphical Stat* 9:1–59