Better Tools for Proving Convergence for MM Algorithms?

Eric Chi

Department of Statistics, North Carolina State University

March 21, 2019
A function $g(\theta | \tilde{\theta})$ is said to majorize the function $f(\theta)$ at $\tilde{\theta}$ if

1. $g(\tilde{\theta} | \tilde{\theta}) = f(\tilde{\theta})$ and
2. $g(\theta | \tilde{\theta}) \geq f(\theta)$ for all $\theta$.

- MM = Majorization-Minimization (or Minorization-Maximization)
- The EM algorithm is an example of the latter
MM Algorithm in Action

![Graph showing the function f(x) with x-axis labeled as very bad, optimal, and less bad, and y-axis labeled as larger and smaller. The graph illustrates the convergence of MM algorithms.]
MM Algorithm in Action

The graph illustrates the function $f(x)$ and the convergence of an MM algorithm.

- **f(x)**: The function $f(x)$ decreases as $x$ moves from **very bad** to the **optimal** point and then increases towards **less bad**.
- **X-axis**: Represents the variable $x$.
- **Y-axis**: Represents the value of $f(x)$.

The graph shows that $f(x)$ is **larger** for **very bad** values of $x$, becomes **optimal** at a specific point, and is **smaller** for **less bad** values of $x$. The orange dot indicates a point of interest in the convergence process.
MM Algorithm in Action

Convergence for MM Algorithms

March 21, 2019 3 / 22
MM Algorithm in Action

Convergence for MM Algorithms

March 21, 2019
MM Algorithm in Action

\[ f(x) \]

- larger
- smaller

\[ x \]

- very bad
- optimal
- less bad

Eric Chi (NCSU)
Strategies for Constructing Majorizations

Lipschitz differentiability
- Gradient descent, proximal gradient descent

Jensen’s inequality
- EM algorithms
- Nonnegative matrix/tensor factorizations

Supporting hyperplane property of convex functions
- SCAD, MCP

Cauchy-Schwartz inequality
- Multidimensional scaling

Majorizing zero
- Matrix completion
**Proposition (Simple Zangwill / Meyer):** Let $f$ be a continuous function on a domain $\mathcal{D}$, and let $\psi$ be a continuous iterative map from $\mathcal{D}$ into $\mathcal{D}$ such that $f(\psi(x)) < f(x)$ for all $x \in \mathcal{D}$ with $\psi(x) \neq x$. Suppose there is an $x_0$ such that the set $\mathcal{L}_f(x_0) \equiv \{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of $\psi$.
2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \to 0$.

**Clean up workflow:**

- Show that Fix-$\psi$ are optima / stationary points of $f$
- Show that there is only one optimum / finitely many stationary points of $f$
Case Study 1: The Biclustering Problem

Task

Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times p}$, find subgroups of rows & columns that go together.

- **Text mining**: similar documents share a small set of highly correlated words.
- **Collaborative filtering**: likeminded customers share similar preferences for a subset of products
- **Cancer genomics**: subtypes of cancerous tumors share similar molecular profiles over a subset of genes
Simple Solution: Cluster Dendrogram

Hierarchical Clustering

Tissue Sample

Genes
Simple Solution: Cluster Dendrogram

The Good
- Easy to interpret
- Fast computation - greedy algorithm

The Bad: Non-convex optimization problem
- Local Minimizers
- Instability (initialization, tuning parameters, or data)

The Ugly: How to choose number of biclusters?
Solution: Convex Relaxation

Solve combinatorially hard problem with a convex surrogate.
- All local minima are global minima
- Algorithms converge to global minimizer regardless of initialization

Solve a convex optimization problem to go from $A$ to $B$
Apply fusion penalties to columns and rows

\[
\min_{U} F_{\gamma}(U) = \frac{1}{2} \| X - U \|_F^2 + \gamma J(U)
\]

\[
J(U) = \sum_{i < j} w_{ij} \| U_{i} - U_{j} \|_2 + \sum_{k < l} \tilde{w}_{kl} \| U_{k} - U_{l} \|_2
\]

Leads to “checkerboard” biclustering - every (i,j)th entry is assigned to one bicluster

- ith and jth columns belong to same column cluster iff \( U_{i} - U_{j} = 0 \)
- kth and lth rows belong to same row cluster iff \( U_{k} - U_{l} = 0 \)

Need to make sure row and column penalties are on the same scale
Biclustering path

minimize $\frac{1}{2} \| \mathbf{X} - \mathbf{U} \|^2_F + \gamma \left[ \sum_{i<j} w_{ij} \| \mathbf{U}_i - \mathbf{U}_j \|_2 + \sum_{k<l} \tilde{w}_{kl} \| \mathbf{U}_k - \mathbf{U}_l \|_2 \right]$
Biclustering path

\[
\text{minimize} \quad \frac{1}{2} \|X - U\|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \|U_i - U_j\|_2 + \sum_{k<l} \tilde{w}_{kl} \|U_k - U_l\|_2 \right]
\]
The Solution Path

minimize \( \frac{1}{2} \| X - U \|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \| U \cdot i - U \cdot j \|_2 + \sum_{k<l} \tilde{w}_{kl} \| U \cdot k - U \cdot l \|_2 \right] \)
Biclustering path

\[
\text{minimize} \quad \frac{1}{2} \|X - U\|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \|U_{.i} - U_{.j}\|_2 + \sum_{k<l} \tilde{w}_{kl} \|U_{.k} - U_{.l}\|_2 \right]
\]
minimize \( \frac{1}{2} \| \mathbf{X} - \mathbf{U} \|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \| \mathbf{U}_{i} - \mathbf{U}_{j} \|_2 \right] + \sum_{k<l} \tilde{w}_{kl} \| \mathbf{U}_{k} - \mathbf{U}_{l} \|_2 \)
Biclustering path

$$\minimize \frac{1}{2} \| \mathbf{X} - \mathbf{U} \|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \| \mathbf{U}_i - \mathbf{U}_j \|_2 + \sum_{k<l} \tilde{w}_{kl} \| \mathbf{U}_k - \mathbf{U}_l \|_2 \right]$$
The Solution Path

minimize \( \frac{1}{2} \| X - U \|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \| U_i - U_j \|_2 + \sum_{k<l} \tilde{w}_{kl} \| U_k - U_l \|_2 \right] \)
Biclustering path

\[
\text{minimize } \frac{1}{2} \|X - U\|_F^2 + \gamma \left[ \sum_{i < j} w_{ij} \|U_{i} - U_{j}\|_2^2 + \sum_{k > l} \tilde{w}_{kl} \|U_{k} - U_{l}\|_2^2 \right]
\]
Biclustering path

\[ \text{minimize } \frac{1}{2} \| X - U \|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \| U_i - U_j \|_2 + \sum_{k<l} \tilde{w}_{kl} \| U_{k} - U_{l} \|_2 \right] \]
Model Selection

Validation Scheme (Wold, 1978)
- Randomly select a hold out set ($\sim 10\%$ of all entries)

$$\Theta \subset \{1, \ldots, p\} \times \{1, \ldots, n\}$$

$$= 0$$
Model Selection

- Solve the matrix completion problem for a sequence of candidate \( \gamma_m \)

\[
U_{\gamma_m}^* = \arg \min_U \frac{1}{2} \| \mathcal{P}_{\Theta^c}(X) - \mathcal{P}_{\Theta^c}(U) \|_F^2 + \gamma_m J(U)
\]

- Pick \( \gamma_m \) that minimizes prediction error on \( \Theta \)
Matrix Completion with COBRA-POD

COBRA with Partially Observed Data
- Fill in missing entries using last iterate
- Use COBRA to get next iterate
- Repeat

Algorithm 1 COBRA-POD

1: Initialize $U^{(0)}$.
2: repeat
3: $M \leftarrow P_{\Theta c}(X) + P_{\Theta}(U^{(n)})$
4: $U^{(n+1)} \leftarrow \text{COBRA}(M)$
5: until convergence

- This is a majorization-minimization (MM) algorithm
- Same majorization used in Soft-Impute: Mazumder et al. (2010)
Convergence of COBRA-POD

Proposition (Simple Zangwill / Meyer): Let $f$ be a continuous function on a domain $D$, and let $\psi$ be a continuous iterative map from $D$ into $D$ such that $f(\psi(x)) < f(x)$ for all $x \in D$ with $\psi(x) \neq x$. Suppose there is an $x_0$ such that the set $L_f(x_0) \equiv \{x \in D : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of $\psi$.
2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \to 0$.

$$f(U) = \frac{1}{2}\|\mathcal{P}_{\Theta c}(X) - \mathcal{P}_{\Theta c}(U)\|_F^2 + \gamma_m J(U)$$

Clean up workflow:
- Fix-$\psi$ are optima of $f$ (modified objective)
- Show that there is only one optimum?
Theorem 5.14 (Bauschke & Combettes) Let $\psi : \mathcal{H} \mapsto \mathcal{H}$ be a nonexpansive operator such that $\text{Fix-}\psi$ is not empty. Let $x_0 \in \mathcal{H}$ and $(\forall n \in \mathbb{N}) x_{n+1} = \psi(x_n)$. Suppose that $\psi(x_n) - x_n \rightarrow 0$. Then $(x_n)_{n \in \mathbb{N}}$ converges to a point in $\text{Fix-}\psi$.

Proposition (Simple Zangwill / Meyer): Let $f$ be a continuous function on a domain $\mathcal{D}$, and let $\psi$ be a continuous iterative map from $\mathcal{D}$ into $\mathcal{D}$ such that $f(\psi(x)) < f(x)$ for all $x \in \mathcal{D}$ with $\psi(x) \neq x$. Suppose there is an $x_0$ such that the set $\mathcal{L}_f(x_0) \equiv \{x \in \mathcal{D} : f(x) \leq f(x_0)\}$ is compact. Define $x_{n+1} = \psi(x_n)$ for $n \in \mathbb{N}$. Then

1. the sequence of iterates $\{x_n\}$ has at least one limit point and all its limit points are fixed points of $\psi$.

2. the distance between successive iterates converges to 0, i.e., $\|x_{n+1} - x_n\| \rightarrow 0$.

upshot: If MM algorithm map is nonexpansive, MM iterates converge to fixed point.
Convergence of COBRA-POD

Algorithm 2 COBRA-POD

1: Initialize $U^{(0)}$.
2: repeat
3: $M \leftarrow P_{\Theta c}(X) + P_{\Theta}(U^{(n)})$
4: $U^{(n+1)} \leftarrow \text{COBRA}(M)$
5: until convergence

\[
\text{minimize} \quad \frac{1}{2} \|X - U\|_F^2 + \gamma \left[ \sum_{i<j} w_{ij} \|U_{.i} - U_{.j}\|_2^2 + \sum_{k<l} \tilde{w}_{kl} \|U_{.k} - U_{.l}\|_2^2 \right]
\]

- COBRA is a proximal map
- Imputation step is also non-expansive

\[
\|P_{\Theta c}(X) + P_{\Theta}(U) - P_{\Theta c}(X) - P_{\Theta}(V)\| \leq \|P_{\Theta}(U) - P_{\Theta}(V)\| \leq \|U - V\|
\]
Case Study 2: Robust Low-Rank Matrix Approximation

minimize $\sum_{ij} \rho(x_{ij} - u_{ij}) + \gamma \|U\|_*$

where

$\rho(t) = -\exp \left[-\frac{\alpha t^2}{2}\right]$,

$\rho$ is Lipschitz differentiable

$g(U | \tilde{U}) = \frac{1}{2} \|\varphi(\tilde{U}) - U\|_F^2 + \gamma \|U\|_*$

$[\varphi(\tilde{U})]_{ij} = \varphi_{ij}(\tilde{U}) = w_{ij}(\tilde{U})x_{ij} + (1 - w_{ij}(\tilde{U}))\tilde{u}_{ij}$,

where

$w_{ij}(\tilde{U}) = \exp \left[-\frac{\alpha}{2} (x_{ij} - \tilde{u}_{ij})^2\right]$. 
Algorithm 3 Robust Low-Rank

1. Initialize $U^{(0)}$.
2. repeat
3. $M \leftarrow \varphi(U^{(n)})$
4. $U^{(n+1)} \leftarrow \text{prox}_{\gamma\|\cdot\|_*}(M)$
5. until convergence

- Can apply Zangwill / Meyer in this case; fixed points are stationary points - directional derivative is non-negative in all directions.
- But can’t apply Theorem 5.12
- $\varphi$ is Lipschitz continuous with constant $1 + 2e^{-3/2} \approx 1.4462603$. 
Convergence Robust Low-Rank Approximation

\[ \| \text{prox}_{\gamma \cdot \| \cdot \|_*} [\varphi(U)] \|_U \|U - V\| \]

\[ \| \text{prox}_{\gamma \cdot \| \cdot \|_*} [\varphi(U)] \|_U \|U - V\| \]

\[ \| \text{prox}_{\gamma \cdot \| \cdot \|_*} [\varphi(U)] \|_U \|U - V\| \]
Open Questions

1. Is there a way to modify the MM algorithm map to enforce non-expansivity that won’t alter the fixed point set?

2. Are there additional regularity conditions (Lipschitz smoothness of MM algorithm map?) that could ensure convergence?