Majorization-Minimization Algorithm
Theory and Applications

Ying Sun and Daniel P. Palomar

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

ELEC 5470 - Convex Optimization
Fall 2015-16, HKUST, Hong Kong
Acknowledgment

Slides of this lecture are majorly based on the following works:


1 The Majorization-Minimization Algorithm
   • Introduction
   • Construction Techniques
   • Example Algorithms
   • Applications

2 Block Successive Majorization-Minimization
   • Introduction
   • Block Coordinate Descent
   • Block Successive Majorization-Minimization
   • Example Algorithms
   • Applications

3 Distributed Algorithm for Nonlinear Programming
   • Exact Jacobi Successive Convex Approximation
   • Extensions
Outline

1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
Problem Statement

- Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathcal{X},
\end{align*}
\]

with \( \mathcal{X} \) being a closed convex set and \( f(x) \) being continuous.

- \( f(x) \) is too complicated to manipulate.

- Idea: successively minimize an approximating function \( u(x, x^k) \)

\[
x^{k+1} = \arg\min_{x \in \mathcal{X}} u(x, x^k),
\]

hoping the sequence of minimizers \( \{x^k\} \) will converge to optimal \( x^* \).

- Question: how to construct \( u(x, x^k) \)?
Consider the following optimization problem

\[ \text{minimize} \quad f(x) \]

subject to \( x \in \mathcal{X} \),

with \( \mathcal{X} \) being a closed convex set and \( f(x) \) being continuous. \( f(x) \) is too complicated to manipulate.

Idea: successively minimize an approximating function \( u(x, x^k) \)

\[ x^{k+1} = \arg \min_{x \in \mathcal{X}} u(x, x^k), \]

hoping the sequence of minimizers \( \{x^k\} \) will converge to optimal \( x^* \).

Question: how to construct \( u(x, x^k) \)?
Consider the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathcal{X},
\end{align*}
\]

with \( \mathcal{X} \) being a closed convex set and \( f(x) \) being continuous. \( f(x) \) is too complicated to manipulate.

Idea: successively minimize an approximating function

\[
\begin{align*}
\text{minimize} & \quad u(x, x^k) \\
\text{subject to} & \quad x \in \mathcal{X},
\end{align*}
\]

hoping the sequence of minimizers \( \{x^k\} \) will converge to optimal \( x^* \).

Question: how to construct \( u(x, x^k) \)?
Terminology

- **Distance from a point to a set:**
  \[
  d(x, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|x - s\|.
  \]

- **Directional derivative:**
  \[
  f'(x; d) \triangleq \liminf_{\lambda \downarrow 0} \frac{f(x + \lambda d) - f(x)}{\lambda}.
  \]

- **Stationary point:** \( x \) is a stationary point if
  \[
  f'(x; d) \geq 0, \ \forall d \text{ such that } x + d \in \mathcal{X}.
  \]
Majorization-Minimization

- Construction rule:

\[ u(y, y) = f(y), \quad \forall y \in \mathcal{X} \]  
\[ u(x, y) \geq f(x), \quad \forall x, y \in \mathcal{X} \]  
\[ u'(x, y; d)|_{x=y} = f'(y; d), \quad \forall d \text{ with } y + d \in \mathcal{X} \]  
\[ u(x, y) \text{ is continuous in } x \text{ and } y \]

- Pictorially:

\[ f(x) \quad u(x, x^k) \]  
\[ x^k \]
Algorithm

- Majorization-Minimization (Successive Upper-Bound Minimization):

1. Find a feasible point $\mathbf{x}^0 \in \mathcal{X}$ and set $k = 0$
2. repeat
3. $\mathbf{x}^{k+1} = \arg\min_{x \in \mathcal{X}} u(x, \mathbf{x}^k)$ (global minimum)
4. $k \leftarrow k + 1$
5. until some convergence criterion is met
Under assumptions A1-A4, every limit point of the sequence \( \{x^k\} \) is a stationary point of the original problem.

If further assume that the level set \( \mathcal{X}^0 = \{x \mid f(x) \leq f(x^0)\} \) is compact, then

\[
\lim_{k \to \infty} d(x^k, \mathcal{X}^*) = 0,
\]

where \( \mathcal{X}^* \) is the set of stationary points.
Outline

1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
The performance of Majorization-Minimization algorithm depends crucially on the surrogate function $u(x, x^k)$.

Guideline: the global minimizer of $u(x, x^k)$ should be easy to find.

Suppose $f(x) = f_1(x) + \kappa(x)$, where $f_1(x)$ is some “nice” function and $\kappa(x)$ is the one needed to be approximated.
Suppose \( \kappa(t) \) is convex, then

\[
\kappa \left( \sum_i \alpha_i t_i \right) \leq \sum_i \alpha_i \kappa(t_i)
\]

with \( \alpha_i \geq 0 \) and \( \sum \alpha_i = 1 \).
For example:

\[
\kappa(w^T x) = \kappa\left(w^T (x - x^k) + w^T x^k\right)
\]

\[
= \kappa\left(\sum_i \alpha_i \left(\frac{w_i (x_i - x_i^k)}{\alpha_i} + w^T x^k\right)\right)
\]

\[
\leq \sum_i \alpha_i \kappa\left(\frac{w_i (x_i - x_i^k)}{\alpha_i} + w^T x^k\right)
\]

If further assume that \(w\) and \(x\) are positive

\(\left(\alpha_i = w_i x_i^k / w^T x^k\right)\):

\[
\kappa(w^T x) \leq \sum_i \frac{w_i x_i^k}{w^T x^k} \kappa\left(\frac{w^T x^k}{x_i^k} x_i\right)
\]

The surrogate functions are separable (parallel algorithm).
Suppose $\kappa(x)$ is concave and differentiable, then

$$\kappa(x) \leq \kappa(x^k) + \nabla \kappa(x^k) (x - x^k),$$

which is a linear upper-bound.

Suppose $\kappa(x)$ is convex and twice differentiable, then

$$\kappa(x) \leq \kappa(x^k) + \nabla \kappa(x^k)^T (x - x^k) + \frac{1}{2} (x - x^k)^T M (x - x^k)$$

if $M - \nabla^2 \kappa(x) \succeq 0$, $\forall x$. 
Construction by Inequalities

- Arithmetic-Geometric Mean Inequality:
  \[
  \left( \prod_{i=1}^{n} x_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i
  \]

- Cauchy-Schwartz Inequality:
  \[
  \|x\| \geq \frac{x^T x^k}{\|x^k\|}
  \]

- Jensen’s Inequality:
  \[
  \kappa(Ex) \leq E\kappa(x)
  \]
  with \(\kappa(\cdot)\) being convex.
EM Algorithm

- Assume the complete data set \( \{x, z\} \) consists of observed variable \( x \) and latent variable \( z \).
- Objective: estimate parameter \( \theta \in \Theta \) from \( x \).
- Maximum likelihood estimator: \( \hat{\theta} = \arg \min_{\theta \in \Theta} -\log p(x|\theta) \)
- EM (Expectation Maximization) algorithm:
  - E-step: evaluate \( p(z|x, \theta^k) \)
    - “guess” \( z \) from current estimate of \( \theta \)
  - M-step: update \( \theta \) as \( \theta^{k+1} = \arg \min_{\theta \in \Theta} u(\theta, \theta^k) \), where
    
    \[
    u(\theta, \theta^k) = -E_{z|x, \theta^k} \log p(x, z|\theta)
    \]

  update \( \theta \) from “guessed” complete data set
An MM Interpretation of EM

The objective function can be written as

\[-\log p(x|\theta)\]

\[= -\log E_{z|\theta} p(x|z, \theta)\]

\[= -\log E_{z|\theta} \left( \frac{p(z|x, \theta^k) p(x|z, \theta)}{p(z|x, \theta^k)} \right)\]

\[= -\log E_{z|x, \theta^k} \left( \frac{p(x|z, \theta)}{p(z|x, \theta^k)} p(z|\theta) \right)\]

\[\leq -E_{z|x, \theta^k} \log \left( \frac{p(x|z, \theta)}{p(z|x, \theta^k)} p(z|\theta) \right) \quad \text{(Jensen’s Inequality)}\]

\[= -E_{z|x, \theta^k} \log p(x, z|\theta) + E_{z|x, \theta^k} p(z|x, \theta^k)\]

u(\theta, \theta^k)
Proximal Minimization

- $f(x)$ is convex. Solve $\min_x f(x)$ by solving the equivalent problem

$$\min_{x \in \mathcal{X}, y \in \mathcal{X}} f(x) + \frac{1}{2c} \|x - y\|^2.$$ 

- Objective function is strongly convex in both $x$ and $y$.

- Algorithm:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2c} \|x - y^k\|^2 \right\}$$

$$y^{k+1} = x^{k+1}.$$ 

- An MM interpretation:

$$x^{k+1} = \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2c} \|x - x^k\|^2 \right\}$$
Consider the unconstrained problem

\[
\min_{x \in \mathbb{R}^n} f(x),
\]

where \( f(x) = g(x) + h(x) \) with \( g(x) \) convex and \( h(x) \) concave.

DC (Difference of Convex) Programming generates \( \{x^k\} \) by solving

\[
\nabla g(x^{k+1}) = - \nabla h(x^k).
\]

An MM interpretation:

\[
x^{k+1} = \arg \min_x \left\{ g(x) + \nabla h(x^k)^T (x - x^k) \right\}.
\]
1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
[Chi-Tan-Pal-O’Ne-Jul’07] Problem: maximize system throughput. Essentially we need to solve the following problem:

\[
\min_{\mathbf{P} \in \mathcal{P}} \frac{\sum_{j \neq i} G_{ij} P_j + n_i}{\sum_j G_{ij} P_j + n_i}
\]

Objective function is the ratio of two posynomials.

Minorize a posynomial, denoted by \( g(\mathbf{x}) = \sum_i m_i(\mathbf{x}) \), by monomial:

\[
g(\mathbf{x}) \geq \prod_i \left( \frac{m_i(\mathbf{x})}{\alpha_i} \right)^{\alpha_i}
\]

where \( \alpha_i = \frac{m_i(\mathbf{x})^k}{g(\mathbf{x})} \). (Arithmetic-Geometric Mean Inequality)

Solution: approximate the denominator posynomial \( \sum_j G_{ij} P_j + n_i \) by monomial.
Reweighted $\ell_1$-norm

- Sparsity signal recovery problem

$$\begin{align*}
&\text{minimize} \quad \|x\|_0 \\
&\text{subject to} \quad Ax = b
\end{align*}$$

- $\ell_1$-norm approximation

$$\begin{align*}
&\text{minimize} \quad \|x\|_1 \\
&\text{subject to} \quad Ax = b
\end{align*}$$

- General form

$$\begin{align*}
&\text{minimize} \quad \sum_{i=1}^{n} \phi(|x_i|) \\
&\text{subject to} \quad Ax = b
\end{align*}$$
[Can-Wak-Boy’J08] Assume $\phi(t)$ is concave nondecreasing, at $x_i^k$, $\phi(|x_i|)$ is majorized by $w_i^k|x_i|$ with $w_i^k = \phi'(t)|_{t=|x_i|}$.

At each iteration a weighted $\ell_1$-norm is solved

$$\begin{align*}
\text{minimize} & \quad \sum w_i^k |x_i| \\
\text{subject to} & \quad Ax = b
\end{align*}$$
Sparse Generalized Eigenvalue Problem

- $\ell_0$-norm regularized generalized eigenvalue problem

\[
\max_x \quad x^T Ax - \rho \|x\|_0 \\
\text{subject to} \quad x^T B x = 1.
\]

- Replace $\|x_i\|_0$ by some nicely behaved function $g_p(x_i)$
  - $|x_i|^p$, $0 < p \leq 1$
  - $\log (1 + |x_i|/p) / \log (1 + 1/p)$, $p > 0$
  - $1 - e^{-|x_i|^p}$, $p > 0$.

- Take $g_p(x_i) = |x_i|^p$ for example.
[Son-Bab-Pal’J14] Majorize $g_p(x_i)$ at $x_i^k$ by quadratic function $w_i^k x_i^2 + c_i^k$.

The surrogate function for $g_p(x_i) = |x_i|^p$ is defined as

$$u(x_i, x_i^k) = \frac{p}{2} |x_i^k|^{p-2} x_i^2 + \left(1 - \frac{p}{2}\right) |x_i^k|^p.$$

Solve at each iteration the following GEVP:

$$\begin{align*}
\text{maximize} & \quad x^T A x - \rho x^T \text{diag} (w^k) x \\
\text{subject to} & \quad x^T B x = 1
\end{align*}$$

However, as $|x_i| \to 0$, $w_i \to +\infty$...
Smooth approximation of $g_p(x)$:

$$g_p^\varepsilon(x) = \begin{cases} \frac{p}{2} \varepsilon^{p-2} x^2, & |x| \leq \varepsilon \\ |x|^p - \left(1 - \frac{p}{2}\right) \varepsilon^p, & |x| > \varepsilon \end{cases}$$

- When $|x| \leq \varepsilon$, $w$ remains to be a constant.
The Majorization-Minimization Algorithm
Block Successive Majorization-Minimization
Distributed Algorithm for Nonlinear Programming

Introduction
Construction Techniques
Example Algorithms
Applications

Ying Sun and Daniel P. Palomar
Sequence Design

- Complex unimodular sequence \( \{x_n \in \mathbb{C}\}_{n=1}^N \).
- Autocorrelation: \( r_k = \sum_{n=k+1}^{N} x_n x_{n-k}^* = r_{-k}^*, \ k = 0, \ldots, N - 1 \).
- Integrated sidelobe level (ISL):
  \[
  \text{ISL} = \sum_{k=1}^{N-1} |r_k|^2.
  \]
- Problem formulation:
  \[
  \begin{align*}
  \text{minimize} & \quad \text{ISL} \\
  \{x_n\}_{n=1}^N & \\
  \text{subject to} & \quad |x_n| = 1, \ n = 1, \ldots, N.
  \end{align*}
  \]
By Fourier transform:

$$\text{ISL} \propto \sum_{p=1}^{2N} \left[ \left| a_p^H x \right|^2 - N \right]^2$$

with $x = [x_1, \ldots, x_N]^T$, $a_p = [1, e^{j\omega_p}, \ldots, e^{j\omega_p(N-1)}]^T$ and

$$\omega_p = \frac{2\pi}{2N} (p - 1).$$

Equivalent problem:

$$\text{minimize} \quad \sum_{p=1}^{2N} \left( a_p^H x x^H a_p \right)^2$$

subject to $|x_n| = 1, \forall n.$
[Son-Bab-Pal’C14] Define $A = [a_1, \ldots, a_{2N}]$, 

$$p^k = \begin{bmatrix} |a_1^H x^k|^2, & \ldots, & |a_{2N}^H x^k|^2 \end{bmatrix}^T, \quad \tilde{A} = A \left( \text{diag} \left( p^k \right) - p_{\max}^k \mathbf{I} \right) A^H.$$ 

Quadratic surrogate function: 

$$p_{\max}^k x^H \tilde{A} \tilde{A}^H x \overset{\text{const.}}{\longrightarrow} + 2 \text{Re} \left( x^H \left( \tilde{A} - 2 N^2 x^k (x^k)^H \right) x^k \right)$$ 

Equivalent to 

$$\begin{align*}
\text{minimize} & \quad \|x - y\|_2 \\
\text{subject to} & \quad |x_n| = 1, \quad \forall n
\end{align*}$$ 

with 

$$y = - \left( \tilde{A} - 2 N^2 x^k (x^k)^H \right) x^k$$ 

Closed-form solution: $x_n = e^{j \arg(y_n)}$. 

Ying Sun and Daniel P. Palomar 

MM Algorithm
Covariance Estimation

- \( x_i \sim \text{elliptical}(0, \Sigma) \)
- Fitting normalized sample \( s_i = \frac{x_i}{\|x_i\|_2} \) to Angular Central Gaussian distribution

\[
f (s_i) \propto \det (\Sigma)^{-1/2} \left( s_i^T \Sigma^{-1} s_i \right)^{-K/2}
\]

- [Sun-Bab-Pal’J14] Shrinkage penalty

\[
h(\Sigma) = \log \det (\Sigma) + \text{Tr} \left( \Sigma^{-1} \mathbf{T} \right)
\]

- Solve the following problem:

\[
\begin{align*}
\text{minimize} & \quad \log \det (\Sigma) + \frac{K}{N} \sum \log (x_i^T \Sigma^{-1} x_i) + \alpha h(\Sigma) \\
\text{subject to} & \quad \Sigma \succeq 0
\end{align*}
\]
At $\Sigma^k$, the objective function is majorized by

$$(1 + \alpha) \log \det (\Sigma) + \frac{K}{N} \sum_{i=1}^{N} \frac{x_i^T \Sigma^{-1} x_i}{x_i^T (\Sigma^k)^{-1} x_i} + \alpha \text{Tr} \left( \Sigma^{-1} T \right)$$

- Surrogate function is convex in $\Sigma^{-1}$.
- Setting the gradient to zero leads to the weighted sample average

$$\Sigma^{k+1} = \frac{1}{1 + \alpha} \frac{K}{N} \sum x_i x_i^T \left( \Sigma^k \right)^{-1} x_i + \frac{\alpha}{1 + \alpha} T$$
Outline

1 The Majorization-Minimization Algorithm
   • Introduction
   • Construction Techniques
   • Example Algorithms
   • Applications

2 Block Successive Majorization-Minimization
   • Introduction
   • Block Coordinate Descent
   • Block Successive Majorization-Minimization
   • Example Algorithms
   • Applications

3 Distributed Algorithm for Nonlinear Programming
   • Exact Jacobi Successive Convex Approximation
   • Extensions
Consider the following problem

$$\min_{x \in \mathcal{X}} f(x)$$

Set $\mathcal{X}$ possesses Cartesian product structure $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$.

Observation: the problem

$$\min_{x_i \in \mathcal{X}_i} f(x_1^0, \ldots, x_{i-1}^0, x_i, x_{i+1}^0, \ldots, x_m^0)$$

with $x_{-i}^0$ taking some feasible value, is easy to solve.
1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
Block Coordinate Descent (BCD)

- Denote $x \triangleq (x_1, \ldots, x_m)$,
  $$f(x_0^0, \ldots, x_i^0, x_i^0, x_i^{0+1}, \ldots, x_m^0) \triangleq f(x_i, x^0)$$
- Block Coordinate Descent (nonlinear Gauss-Seidel)
  1: Initialize $x^0 \in X$ and set $k = 0$.
  2: repeat
  3: $k = k + 1$, $i = (k \mod n) + 1$
  4: $x_i^k = \arg \min_{x_i \in X_i} f(x_i, x^{k-1})$
  5: $x_i^k \leftarrow x_i^{k-1}, \forall k \neq i$
  6: until some convergence criterion is met
Convergence

- **[Ber'B99]** Assume that
  - $f(x)$ is **continuously differentiable** over the set $\mathcal{X}$.
  - $x^k_i = \arg\min_{x_i \in \mathcal{X}_i} f(x_i, x^{k-1})$ has a **unique** solution.

  Then every limit point of the sequence $\{x^k\}$ is a stationary point.

- **[Gri-Sci’J00]** Generalizations
  - globally convergent for $m = 2$.
  - $f$ is component-wise strictly quasi-convex w.r.t. $m - 2$ components.
  - $f$ is pseudo-convex.
Outline

1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
BS-MM Algorithm

- Combination of MM and BCD
- Block Successive Majorization-Minimization (BS-MM):
  1: Initialize $x^0 \in \mathcal{X}$ and set $k = 0$.
  2: repeat
  3: $k = k + 1, \quad i = (k \mod n) + 1$
  4: $\mathcal{X}^k = \arg \min_{x_i \in \mathcal{X}_i} u_i (x_i, x^{k-1})$
  5: Set $x_i^k$ to be an arbitrary element in $\mathcal{X}^k$
  6: $x_i^k \leftarrow x_i^{k-1}, \quad \forall k \neq i$
  7: until some convergence criterion is met

- Generalization of BCD
Surrogate function $u_i(\cdot, \cdot)$ satisfies the following assumptions

\begin{align*}
    u_i(y_i, y) &= f(y), \quad \forall y \in \mathcal{X}, \forall i \quad (B1) \\
    u_i(x_i, y) &\geq f(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_n), \quad \forall x_i \in \mathcal{X}_i, \forall y \in \mathcal{X}, \forall i \quad (B2) \\
    u'_i(x_i, y; d_i)\big|_{x_i=y_i} &= f'(y; d), \quad \forall d = (0, \ldots, d_i, \ldots, 0) \text{ such that } y_i + d_i \in \mathcal{X}_i, \forall i \quad (B3) \\
    u_i(x_i, y) &\text{ is continuous in } (x_i, y), \quad \forall i \quad (B4)
\end{align*}

In short, $u_i(x_i, x^k)$ majorizes $f(x)$ on the $i$th block.
Under assumptions B1-B4, for simplicity additionally assume that $f$ is continuously differentiable,

- $u_i(x_i, y)$ is quasi-convex in $x_i$, each subproblem $\min_{x_i \in X_i} u_i(x_i, x^{k-1})$ has a unique solution for any $x^{k-1} \in X$, then every limit point of $\{x^k\}$ is a stationary point.
- level set $X^0 = \{x | f(x) \leq f(x^0)\}$ is compact, each subproblem $\min_{x_i \in X_i} u_i(x_i, x^{k-1})$ has a unique solution for any $x^{k-1} \in X$ for at least $m - 1$ blocks, then $\lim_{k \to \infty} d(x^k, X^*) = 0$.

- More restrictive assumption than MM due to the cyclic update behavior.
1 The Majorization-Minimization Algorithm
   • Introduction
   • Construction Techniques
   • Example Algorithms
   • Applications

2 Block Successive Majorization-Minimization
   • Introduction
   • Block Coordinate Descent
   • Block Successive Majorization-Minimization
   • Example Algorithms
   • Applications

3 Distributed Algorithm for Nonlinear Programming
   • Exact Jacobi Successive Convex Approximation
   • Extensions
Alternating Proximal Minimization

- Consider the problem

\[
\begin{align*}
\text{minimize} & \quad f(x_1, \ldots, x_m) \\
\text{subject to} & \quad x_i \in X_i,
\end{align*}
\]

with \( f(\cdot) \) being convex in each block.

- The convergence of BCD is not easy to establish since each subproblem may have multiple solutions.

- Alternating Proximal Minimization solves

\[
\begin{align*}
\text{minimize} & \quad f(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k) + \frac{1}{2c} \| x_i - x_i^k \|^2 \\
\text{subject to} & \quad x_i \in X_i
\end{align*}
\]

- Strictly convex objective \( \rightarrow \) unique minimizer
Consider the following problem

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{m} f_i(x_i) + f_{m+1}(x_1, \ldots, x_m) \\
\text{subject to} & \quad x_i \in X_i, \ i = 1, \ldots, m
\end{align*}
\]

with \( f_i \) convex and lower semicontinuous, \( f_{m+1} \) convex and

\[
\|\nabla f_{m+1}(x) - \nabla f_{m+1}(y)\| \leq \beta_i \|x_i - y_i\|
\]

Cyclically update:

\[
x_i^{k+1} = \text{prox}_{\gamma f_i} \left( x_i^k - \gamma \nabla f_{m+1}(x^k) \right),
\]

with the proximity operator defined as

\[
\text{prox}_f(x) = \arg \min_{y \in X} f(y) + \frac{1}{2} \|x - y\|^2.
\]
BS-MM interpretation:

\[
\begin{align*}
    u_i (x_i, x^k) &= f_i(x_i) + \frac{1}{2\gamma} \| x_i - x_i^k \|^2 + \nabla x_i f_{m+1}(x^k)^T (x_i - x_i^k) \\
    &\quad + \sum_{j \neq i} f_j(x_j^k) + f_{m+1}(x^k_{-i}, x_i).
\end{align*}
\]

Check:

\[
\begin{align*}
    f_{m+1}(x^k) + \frac{1}{2\gamma} \| x_i - x_i^k \|^2 + \nabla x_i f_{m+1}(x^k)^T (x_i - x_i^k) \\
    \geq f_{m+1}(x^k) + \frac{\beta_i}{2} \| x_i - x_i^k \|^2 + \nabla x_i f_{m+1}(x^k)^T (x_i - x_i^k) \\
    \geq f_{m+1}(x^k_{-i}, x_i) \quad \text{(Descent lemma)}
\end{align*}
\]

with \( \gamma \in [\varepsilon_i, 2/\beta_i - \varepsilon_i] \) and \( \varepsilon_i \in (0, \min \{1, 1/\beta_i\}) \).
Outline

1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
Robust Estimation of Location and Scatter

- $x_i \sim \text{elliptical} (\mu, R)$
- [Sun-Bab-Pal’C14] Fitting $x_i$ to a Cauchy distribution with pdf
  \[ f(x) \propto \det(R)^{-1/2} \left(1 + (x_i - \mu)^T R^{-1} (x_i - \mu)\right)^{-(K+1)/2} \]

- Shrinkage penalty
  \[ h(t, T) = K \log \left( \text{Tr} \left( R^{-1} T \right) \right) + \log \det(R) + \log \left(1 + (t - \mu)^T R^{-1} (t - \mu)\right) \]

- Solve the following problem:
  \[
  \min_{\mu, R \succeq 0} \quad \log \det(R) + \frac{K+1}{N} \sum_{i=1}^{N} \log \left(1 + (x_i - \mu)^T R^{-1} (x_i - \mu)\right) \\
  + \alpha h(t, T)
  \]
BS-MM Algorithm update:

\[
\mu_{t+1} = \frac{(K + 1) \sum_{i=1}^{N} w_i(\mu_t, R_t) x_i + N \alpha w_t(\mu_t, R_t) t}{(K + 1) \sum_{i=1}^{N} w_i(\mu_t, R_t) + N \alpha w_t(\mu_t, R_t)}
\]

\[
R_{t+1} = \frac{K + 1}{N + N \alpha} \sum_{i=1}^{N} w_i(\mu_{t+1}, R_t) (x_i - \mu_{t+1}) (x_i - \mu_{t+1})^T
\]

\[
+ \frac{N \alpha}{N + N \alpha} w_t(\mu_{t+1}, R_t) (\mu_{t+1} - t) (\mu_{t+1} - t)^T
\]

\[
+ \frac{N \alpha K}{N + N \alpha} \frac{T}{\text{Tr}(R_t^{-1} T)}
\]

where

\[
w_i(\mu, R) = \frac{1}{1 + (x_i - \mu)^T R^{-1} (x_i - \mu)}
\]

\[
w_t(\mu, R) = \frac{1}{1 + (t - \mu)^T R^{-1} (t - \mu)}.
\]
Outline

1. The Majorization-Minimization Algorithm
   - Introduction
   - Construction Techniques
   - Example Algorithms
   - Applications

2. Block Successive Majorization-Minimization
   - Introduction
   - Block Coordinate Descent
   - Block Successive Majorization-Minimization
   - Example Algorithms
   - Applications

3. Distributed Algorithm for Nonlinear Programming
   - Exact Jacobi Successive Convex Approximation
   - Extensions
Consider the following problem:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x_i \in \mathcal{X}_i
\end{align*}
\]

where the \( \mathcal{X}_i \)'s are closed and convex sets, \( f(x) = \sum_{l=1}^{L} f_l(x_1, \ldots, x_m) \).

Conditional gradient update (Frank-Wolfe):

\[
x^{k+1} = x^k + \gamma^k d^k
\]

- direction \( d^k \triangleq \bar{x}^k - x^k \) with

\[
\bar{x}^k_i = \arg \min_{x_i \in \mathcal{X}_i} \nabla_{x_i} f(x^k)^T (x_i - x_i^k)
\]

- step-size \( \gamma^k \in (0, 1] \), chosen to guarantee convergence.
Exact Jacobi SCA Algorithm

- Idea:
  - Conditional gradient update linearize all the $f_i$'s at $x^k$.
  - Each function $f_i$ might be convex w.r.t. some block $x_i$.
  - We want to preserve the convex property of $f_i(x_i, x_{-i}^k)$.

- Solution: keep the convex $f_i(x_i, x_{-i}^k)$’s and linearize the others.
Define $\mathcal{C}_i$ as the set of indices of $l$ such that $f_l(x_i, x_{-i}^k)$ is convex.

Approximate $f(x)$ on the $i$th block at point $x^k$:

$$
\tilde{f}_i(x_i, x^k) = \sum_{l \in \mathcal{C}_i} f_l(x_i, x_{-i}^k) + \pi_i (x^k)^T (x_i - x_i^k) + \pi_i (x^k) + \tau_i (x_i - x_i^k)^T H_i (x^k) (x_i - x_i^k),
$$

with

$$
\pi_i (x^k) = \sum_{l \notin \mathcal{C}_i} \nabla x_i f_l (x^k) \quad \text{and} \quad H_i (x^k) \succ c_{H_i} I.
$$
Exact Jacobi SCA update:

\[ \hat{x}_i \left( x^k, \tau_i \right) = \arg \min_{x_i \in \mathcal{X}_i} \tilde{f}_i \left( x_i, x^k \right) \]

\[ x^{k+1} = x^k + \gamma^k \left( \hat{x} - x^k \right) \]

- Step-size rule
  - constant step-size that depends on the Lipschitz constant of \( \nabla f \)
  - diminishing step-size

- Remark: update of the blocks can be done sequentially (Gauss-Seidel SCA Algorithm)
Outline

1 The Majorization-Minimization Algorithm
   • Introduction
   • Construction Techniques
   • Example Algorithms
   • Applications

2 Block Successive Majorization-Minimization
   • Introduction
   • Block Coordinate Descent
   • Block Successive Majorization-Minimization
   • Example Algorithms
   • Applications

3 Distributed Algorithm for Nonlinear Programming
   • Exact Jacobi Successive Convex Approximation
   • Extensions
Extensions

- FLEXA
  - non-smooth objective function
  - inexact update direction
  - flexible block update choice

- HyFLEXA
## Comparison

<table>
<thead>
<tr>
<th></th>
<th>BS-MM</th>
<th>FLEXA</th>
</tr>
</thead>
<tbody>
<tr>
<td>convergence</td>
<td>stationary point</td>
<td>stationary point</td>
</tr>
<tr>
<td>objective function</td>
<td>continuous \ may not be smooth</td>
<td>continuous \ may not be smooth</td>
</tr>
<tr>
<td>constraint set</td>
<td>Cartesian</td>
<td>Cartesian &amp; convex</td>
</tr>
<tr>
<td>update rule</td>
<td>sequential</td>
<td>sequential or parallel</td>
</tr>
<tr>
<td>approx. function</td>
<td>global upper-bound \ unique minimizer \ can be non-convex</td>
<td>local approximation \ not required \ convex approx.</td>
</tr>
</tbody>
</table>
Summary

We have studied

- Majorization-Minimization algorithm
- Block Coordinate Descent algorithm
- Block Successive Majorization-Minimization algorithm

We have briefly introduced

- Distributed Successive Convex Approximation algorithm
D. R. Hunter and K. Lange.
A tutorial on MM algorithms.

M. Razaviyayn, M. Hong, and Z. Luo.
A unified convergence analysis of block successive minimization methods for nonsmooth optimization.

Decomposition by partial linearization: Parallel optimization of multi-agent systems.
ISSN 1053-587X.

Power control by geometric programming.
ISSN 1536-1276.

Enhancing sparsity by reweighted l1 minimization.
J. Song, P. Babu, and D. P. Palomar. 
Sparse generalized eigenvalue problem via smooth optimization.

—. 
Optimization methods for sequence design with low autocorrelation sidelobes. 

Y. Sun, P. Babu, and D. P. Palomar. 
Regularized Tyler’s scatter estimator: Existence, uniqueness, and algorithms. 

D. P. Bertsekas. 
*Nonlinear Programming*. 

L. Grippo and M. Sciandrone. 
On the convergence of the block nonlinear gauss–seidel method under convex constraints. 
Y. Sun, P. Babu, and D. P. Palomar.
Regularized robust estimation of mean and covariance matrix under heavy tails and outliers.
Thanks

For more information visit:

http://www.ece.ust.hk/~palomar